I752593

(19)中華民國智慧財產局

(12)發明說明書公告本 **(11)證書號數：TW I752593 B**

(45)公告日： 中華民國 111 (2022) 年 01 月 11 日

(21)申請案號：109127986 (22)申請日： 中華民國 109 (2020) 年 08 月 17 日

(51)Int. Cl.： *C12Q1/6869 (2018.01)* *C12N15/11 (2006.01)*
G16B20/00 (2019.01) *G16B30/00 (2019.01)*

(30)優先權：2019/08/16 美國 62/887,987
2020/02/05 美國 62/970,586
2020/03/19 美國 62/991,891
2020/05/04 美國 63/019,790
2020/07/13 美國 63/051,210

(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 (HK)；鄭 淑恒 CHENG, SUK HANG (HK)；彭文磊 PENG, WENLEI (CN)；謝 安儀 TSE, ON YEE (HK)

(74)代理人：陳長文

(56)參考文獻：

WO 2010/068289A2

Flusberg et al., "Direct detection of DNA methylation during single-molecule, real-time sequencing" Nat Methods. 2010 June ; 7 (6): 461-465

審查人員：吳姵諄

申請專利範圍項數：19 項 圖式數：133 共 331 頁

(54)名稱

核酸鹼基修飾的測定

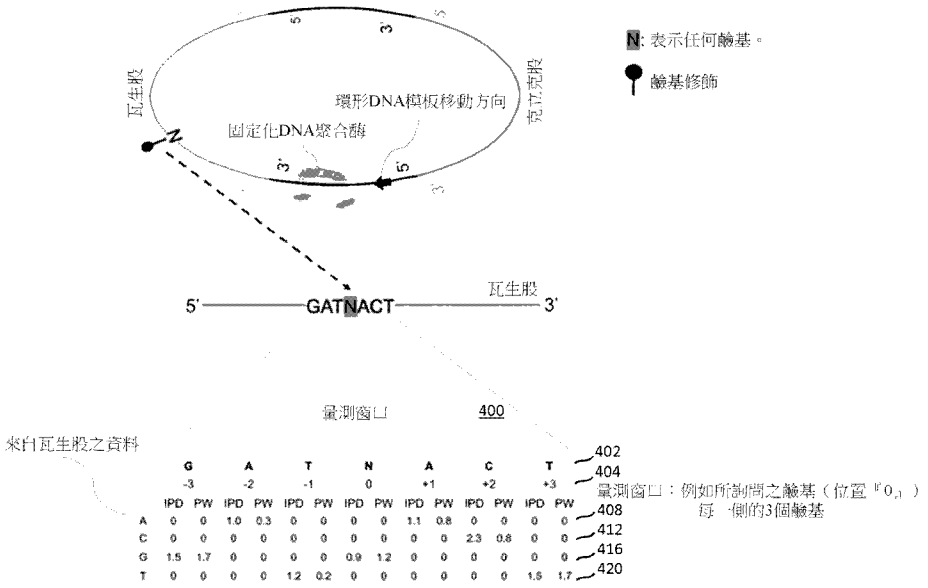
(57)摘要

本文描述使用鹼基修飾之測定分析核酸分子及獲取核酸分子分析資料的系統及方法。鹼基修飾可包括甲基化。測定鹼基修飾之方法可包括使用自定序獲得之特徵。此等特徵可包括來自定序鹼基之光信號的脈衝寬度、鹼基之脈衝間持續時間及鹼基之標識。可訓練機器學習模型以使用此等特徵檢測鹼基修飾。單倍型之間的相對修飾或甲基化程度可指示病症。修飾或甲基化狀態亦可用於檢測嵌合分子。

Systems and methods for using determination of base modification in analyzing nucleic acid molecules and acquiring data for analysis of nucleic acid molecules are described herein. Base modifications may include methylations. Methods to determine base modifications may include using features derived from sequencing. These features may include the pulse width of an optical signal from sequencing bases, the interpulse duration of bases, and the identity of the bases. Machine learning models can be trained to detect the base modifications using these features. The relative modification or methylation levels between

haplotypes may indicate a disorder. Modification or methylation statuses may also be used to detect chimeric molecules.

指定代表圖：



	G		A		T		N		A		C		T	
	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	0	0	1.0	0.3	0	0	0	0	1.1	0.8	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	2.3	0.8	0	0
G	1.5	1.7	0	0	0	0	0.9	1.2	0	0	0	0	0	0
T	0	0	0	0	1.2	0.2	0	0	0	0	0	0	1.5	1.7

【圖4】

符號簡單說明：

400:量測窗口

402:矩陣之第一列

404:矩陣之第二列

408:列

412:列

416:列

420:列

公告本

I752593

【發明摘要】

【中文發明名稱】

核酸鹼基修飾的測定

【英文發明名稱】

DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

【中文】

本文描述使用鹼基修飾之測定分析核酸分子及獲取核酸分子分析資料的系統及方法。鹼基修飾可包括甲基化。測定鹼基修飾之方法可包括使用自定序獲得之特徵。此等特徵可包括來自定序鹼基之光信號的脈衝寬度、鹼基之脈衝間持續時間及鹼基之標識。可訓練機器學習模型以使用此等特徵檢測鹼基修飾。單倍型之間的相對修飾或甲基化程度可指示病症。修飾或甲基化狀態亦可用於檢測嵌合分子。

【英文】

Systems and methods for using determination of base modification in analyzing nucleic acid molecules and acquiring data for analysis of nucleic acid molecules are described herein. Base modifications may include methylations. Methods to determine base modifications may include using features derived from sequencing. These features may include the pulse width of an optical signal from sequencing bases, the interpulse duration of bases, and the identity of the bases. Machine learning models can be trained to detect the base modifications using these features. The relative modification or methylation levels between

haplotypes may indicate a disorder. Modification or methylation statuses may also be used to detect chimeric molecules.

【指定代表圖】

圖4

【代表圖之符號簡單說明】

400: 量測窗口

402: 矩陣之第一列

404: 矩陣之第二列

408: 列

412: 列

416: 列

420: 列

【發明說明書】

【中文發明名稱】

核酸鹼基修飾的測定

【英文發明名稱】

DETERMINATION OF BASE MODIFICATIONS OF NUCLEIC ACIDS

【技術領域】

相關申請案之交叉引用

【0001】 本申請案主張 2020 年 7 月 13 日申請之名稱為「核酸鹼基修飾的測定」之美國臨時申請案第 63/051,210 號；2020 年 5 月 4 日申請之名稱為「核酸鹼基修飾的測定」之美國臨時申請案第 63/019,790 號；2020 年 3 月 19 日申請之名稱為「核酸鹼基修飾的測定」之美國臨時申請案第 62/991,891 號；2020 年 2 月 5 日申請之名稱為「核酸鹼基修飾的測定」之美國臨時申請案第 62/970,586 號；及 2019 年 8 月 16 日申請之名稱為「核酸鹼基修飾的測定」之美國臨時申請案第 62/887,987 號的優先權，該等臨時申請案之全部內容均以引用的方式併入本文中，用於所有目的。

【先前技術】

【0002】 核酸中鹼基修飾的存在在包括病毒、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）、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）。

【0003】 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的方面具有顯著作用。

【0004】 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 甲基化的畸變有關，包括但不限於致癌過程、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。

【0005】 準確量測 DNA 分子上之甲基化體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定序（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，將未甲基化胞嘧啶（以及 C）轉化為尿嘧啶。相反，甲基化胞嘧啶保持不變。隨後藉由 DNA 定序分析經亞硫酸氫鹽修飾之 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概況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。

【0006】 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 的降解，亞硫酸氫鹽定序將無法對長 DNA 分子進行定序。因此，需要在不事先進行化學處理（例如亞硫酸氫鹽轉化）及核酸擴增（例如使用 PCR）的情況下測定核酸鹼基的修飾。

【發明內容】

【0007】 吾等已開發一種新方法，在一個實施例中，該方法允許在沒有模板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。儘管此類模板DNA預處理對於鹼基修飾的測定並非必需的，但在所示的實例中，某些預處理（例如用限制酶消化）可能有助於增強本發明之態樣（例如允許富集CpG位點進行分析）。本揭示案中存在之實施例可用於檢測不同類型的鹼基修飾，例如，包括但不限於4mC、5hmC、5fC及5caC、1mA、3mA、7mA、3mC、2mG、6mG、7mG、3mT及4mT等。此類實施例可利用由定序獲得之特徵，諸如動力學特徵，其受各種鹼基修飾，以及確定甲基化狀態之目標位置周圍窗口中核苷酸之標識的影響。

【0008】 本發明之實施例可用於但不限於單分子定序。一種類型的單分子定序為單分子即時定序，其中即時監測單個DNA分子之定序進度。一種類型的單分子即時定序係由Pacific Biosciences使用其單分子即

時（SMRT）系統商業化之定序。方法可使用定序鹼基信號之脈衝寬度、鹼基之脈衝間持續時間（IPD）及鹼基之標識，以便檢測鹼基或相鄰鹼基中之修飾。另一種單分子系統係基於奈米孔定序之系統。奈米孔定序系統之一個實例係由Oxford Nanopore Technologies商業化之系統。

【0009】 吾等開發之方法可充當檢測生物樣本中鹼基修飾之工具，以評定樣本中之甲基化概況，用於各種目的，包括但不限於研究及診斷目的。檢測到的甲基化概況可用於不同的分析。甲基化概況可用於檢測DNA之來源（例如母體或胎兒、組織、細菌或自癌症患者血液中富集之腫瘤細胞獲得的DNA）。檢測組織中之異常甲基化概況有助於鑑別個體之發育障礙、鑑別及預測腫瘤或惡性腫瘤。

【0010】 本發明之實施例可包括分析生物體之單倍型的相對甲基化程度。兩個單倍型之間甲基化程度的不平衡可用於確定病症之分類。較高的不平衡性可表明存在病症或更嚴重的病症。該病症可包括癌症。

【0011】 單個分子中之甲基化模式可鑑別嵌合體及雜合DNA。嵌合及雜合分子可包括來自兩個不同基因、染色體、胞器（例如粒線體、細胞核、葉綠體）、生物體（哺乳動物、細菌、病毒等）及/或物種之序列。檢測嵌合或雜合DNA分子之接合點可允許檢測各種病症或疾病，包括癌症、產前或先天性病症之基因融合。

【0012】 可參考以下詳細描述及隨附圖式來獲得對本發明實施例之性質及優勢的較佳理解。

【圖式簡單說明】

【0013】 **圖 1** 展示根據本發明實施例之攜帶鹼基修飾之分子的 SMRT 定

序。

【0014】 **圖 2** 展示根據本發明實施例之攜帶甲基化及未甲基化 CpG 位點之分子的 SMRT 定序。

【0015】 **圖 3** 展示根據本發明實施例之脈衝間持續時間及脈衝寬度。

【0016】 **圖 4** 展示根據本發明實施例之用於檢測鹼基修飾之 DNA 之瓦生股的量測窗口的實例。

【0017】 **圖 5** 展示根據本發明實施例之用於檢測鹼基修飾之 DNA 之克立克股的量測窗口的實例。

【0018】 **圖 6** 展示根據本發明實施例之藉由組合來自 DNA 之瓦生股及其互補克立克股之資料來檢測任何鹼基修飾的量測窗口的實例。

【0019】 **圖 7** 展示根據本發明實施例之藉由組合來自 DNA 之瓦生股及其附近區域之克立克股之資料來檢測任何鹼基修飾的量測窗口的實例。

【0020】 **圖 8** 展示根據本發明實施例之用於確定 CpG 位點處之甲基化狀態之瓦生股、克立克股及兩股的量測窗口的實例。

【0021】 **圖 9** 展示根據本發明實施例之構築用於對鹼基修飾進行分類之分析、計算、數學或統計模型的一般程序。

【0022】 **圖 10** 展示根據本發明實施例之對鹼基修飾進行分類的一般程序。

【0023】 **圖 11** 展示根據本發明實施例之使用具有已知的瓦生股甲基化狀態之樣本構築用於對 CpG 位點處之甲基化狀態進行分類之分析、計算、數學或統計模型的一般程序。

【0024】 **圖 12** 展示根據本發明實施例之對未知樣本之瓦生股甲基化狀態進行分類的一般程序。

【0025】 **圖 13** 展示根據本發明實施例之使用具有已知的克立克股甲基化狀態之樣本構築用於對 CpG 位點處之甲基化狀態進行分類之分析、計算、數學或統計模型的一般程序。

【0026】 **圖 14** 展示根據本發明實施例之對未知樣本之克立克股甲基化狀態進行分類的一般程序。

【0027】 **圖 15** 展示根據本發明實施例之使用來自瓦生股及克立克股之具有已知甲基化狀態之樣本構築用於對 CpG 位點處之甲基化狀態進行分類之統計模型的一般程序。

【0028】 **圖 16** 展示根據本發明實施例之對來自瓦生股及克立克股之未知樣本的甲基化狀態進行分類的一般程序。

【0029】 **圖 17A** 及 **17B** 展示根據本發明實施例之用於確定甲基化之訓練資料集及測試資料集的效能。

【0030】 **圖 18A** 及 **18B** 展示根據本發明實施例之用於確定甲基化之訓練資料集及測試資料集的效能。

【0031】 **圖 19A** 及 **19B** 展示根據本發明實施例之用於確定甲基化之不同定序深度的訓練資料集及測試資料集的效能。

【0032】 **圖 20A** 及 **20B** 展示根據本發明實施例之用於確定甲基化之不同股的訓練資料集及測試資料集的效能。

【0033】 **圖 21A** 及 **21B** 展示根據本發明實施例之用於確定甲基化之不同量測窗口的訓練資料集及測試資料集的效能。

【0034】 **圖 22A** 及 **22B** 展示根據本發明實施例之僅使用下游鹼基確定甲基化之不同量測窗口的訓練資料集及測試資料集的效能。

【0035】 **圖 23A** 及 **23B** 展示根據本發明實施例之僅使用上游鹼基確定甲

基化之不同量測窗口的訓練資料集及測試資料集的效能。

【0036】 **圖 24** 展示根據本發明實施例之在訓練資料集中使用與下游及上游鹼基相關之動力學模式，使用不對稱側翼大小進行甲基化分析的效能。

【0037】 **圖 25** 展示根據本發明實施例之在測試資料集中使用與下游及上游鹼基相關之動力學模式，使用不對稱側翼大小進行甲基化分析的效能。

【0038】 **圖 26** 展示根據本發明實施例之關於 CpG 位點處之甲基化狀態分類之特徵的相對重要性。

【0039】 **圖 27** 展示根據本發明實施例之基於基元之 IPD 分析在不使用脈衝寬度信號之情況下進行甲基化檢測的效能。

【0040】 **圖 28** 為根據本發明實施例之使用進行甲基化分析之胞嘧啶上游 2-nt 及下游 6-nt 之主成分分析技術的圖。

【0041】 **圖 29** 為根據本發明實施例之使用主成分分析之方法及使用卷積神經網路之方法之間的效能比較圖。

【0042】 **圖 30A** 及 **30B** 展示根據本發明實施例之僅使用上游鹼基確定甲基化之不同分析、計算、數學或統計模型的訓練資料集及測試資料集的效能。

【0043】 **圖 31A** 展示根據本發明實施例之藉由全基因體擴增生成具有未甲基化腺嘌呤之分子的一種方法的實例。

【0044】 **圖 31B** 展示根據本發明實施例之藉由全基因體擴增生成具有甲基化腺嘌呤之分子的一種方法的實例。

【0045】 **圖 32A** 及 **32B** 展示根據本發明實施例之在未甲基化與甲基化資料集之間瓦生股之模板 DNA 中經定序 A 鹼基的脈衝間持續時間（IPD）值。

【0046】 **圖 32C** 展示根據本發明實施例之用於確定瓦生股中之甲基化的接收者操作特徵曲線。

【0047】 **圖 33A** 及 **33B** 展示根據本發明實施例之在未甲基化與甲基化資料集之間克立克股之模板 DNA 中經定序 A 鹼基的脈衝間持續時間（IPD）值。

【0048】 **圖 33C** 展示根據本發明實施例之用於確定克立克股中之甲基化的接收者操作特徵曲線。

【0049】 **圖 34** 展示根據本發明實施例之瓦生股的 6mA 測定。

【0050】 **圖 35** 展示根據本發明實施例之克立克股的 6mA 測定。

【0051】 **圖 36A** 及**圖 36B** 展示根據本發明實施例之使用基於量測窗口之卷積神經網路模型在 uA 與 mA 資料集之間所確定的瓦生股之經定序 A 鹼基的甲基化概率。

【0052】 **圖 37** 展示根據本發明實施例之使用基於量測窗口之 CNN 模型對瓦生股之經定序 A 鹼基進行 6mA 檢測的 ROC 曲線。

【0053】 **圖 38** 展示根據本發明實施例之基於 IPD 度量之 6mA 檢測及基於量測窗口之 6mA 檢測之間的效能比較。

【0054】 **圖 39A** 及 **39B** 展示根據本發明實施例之使用基於量測窗口之 CNN 模型在 uA 及 mA 資料集之間所確定的克立克股之彼等經定序 A 鹼基的甲基化概率。

【0055】 **圖 40** 展示根據本發明實施例之使用基於量測窗口之 CNN 模型對克立克股之經定序 A 鹼基進行 6mA 檢測的效能。

【0056】 **圖 41** 展示根據本發明實施例之包括瓦生股及克立克股之分子中 A 鹼基之甲基化狀態的實例。

【0057】 **圖 42** 展示根據本發明實施例之藉由選擇性使用 mA 資料集中 IPD 值大於其第 10 個百分位數的 A 鹼基進行增強訓練的實例。

【0058】 **圖 43** 為根據本發明實施例之 mA 資料集中未甲基化腺嘌呤之百

分比相對於各孔中子讀段之數目的圖。

【0059】 **圖 44** 展示根據本發明實施例之測試資料集中雙股 DNA 分子之瓦生股及克立克股之間的甲基腺嘌呤模式。

【0060】 **圖 45** 為顯示根據本發明實施例之訓練及測試資料集中完全未甲基化分子、半甲基化分子、完全甲基化分子及具有交錯甲基腺嘌呤模式之分子的百分比的表格。

【0061】 **圖 46** 展示根據本發明實施例之關於腺嘌呤位點之完全未甲基化分子、半甲基化分子、完全甲基化分子及具有交錯甲基腺嘌呤模式之分子的代表性分子實例。

【0062】 **圖 47** 展示根據本發明實施例之具有 CpG 島（以黃色陰影表示）之長讀段（6,265 bp）的實例。

【0063】 **圖 48** 為顯示根據本發明實施例之藉由 Pacific Biosciences SMRT 定序進行定序且與印記區域重疊之 9 種 DNA 分子的表格。

【0064】 **圖 49** 展示根據本發明實施例之基因體印記的實例。

【0065】 **圖 50** 展示根據本發明實施例之確定印記區域中之甲基化模式的實例。

【0066】 **圖 51** 展示根據本發明實施例之新方法與習知亞硫酸氫鹽定序之間推導的甲基化程度的比較。

【0067】 **圖 52A** 及 **52B** 展示根據本發明實施例之血漿 DNA 之甲基化檢測的效能。(A)預測之甲基化概率與藉由亞硫酸氫鹽定序量化之甲基化程度範圍之間的關係。(B)根據本揭示案中存在之實施例藉由 Pacific Biosciences（PacBio）定序確定之甲基化程度（y 軸）與藉由亞硫酸氫鹽定序量化之甲基化程度（x 軸）之間的相關性，解析度為 10-Mb。

【0068】 **圖 53** 展示根據本發明實施例之 Pacific Biosciences SMRT 定序與 BS-seq 之間 Y 染色體之基因體呈現（GR）的相關性。

【0069】 **圖 54** 展示根據本發明實施例之使用 CpG 塊之基於 CpG 塊之甲基化檢測的實例，每個 CpG 塊具有一系列 CpG 位點。5mC：甲基化；C：未甲基化。

【0070】 **圖 55A** 及 **55B** 展示根據本發明實施例之使用基於 CpG 塊之方法對人類 DNA 分子進行甲基化判讀的訓練及測試。(A)在訓練資料集中之效能。(B)在獨立測試資料集中之效能。

【0071】 **圖 56A** 及 **56B** 展示根據本發明實施例之腫瘤組織中之複本數變化。

【0072】 **圖 57A** 及 **57B** 展示根據本發明實施例之腫瘤組織中之複本數變化。

【0073】 **圖 58** 展示使用根據本發明實施例推導之甲基化程度自孕婦血漿繪製的血漿 DNA 組織圖譜的示意圖。

【0074】 **圖 59** 展示根據本發明實施例之胎盤對推導之母體血漿 DNA 的貢獻與藉由 Y 染色體讀數推導之胎兒 DNA 分數之間的相關性。

【0075】 **圖 60** 展示根據本發明實施例之彙總不同人類組織 DNA 樣本之定序資料的表格。

【0076】 **圖 61** 展示根據本發明實施例之分析甲基化模式之各種方式的圖示。

【0077】 **圖 62A** 及 **62B** 展示根據本發明實施例之藉由亞硫酸氫鹽定序及單分子即時定序量化之全基因體水準之甲基化密度的比較。

【0078】 **圖 63A**、**63B** 及 **63C** 展示根據本發明實施例之藉由亞硫酸氫鹽

定序及單分子即時定序量化之總體甲基化程度的不同相關性。

【0079】 **圖 64A** 及 **64B** 展示根據本發明實施例之肝細胞癌（HCC）細胞株及來自健康對照個體之白血球層樣本在 1-Mnt 解析度下之甲基化模式，其中甲基化程度藉由亞硫酸氫鹽定序及單分子即時定序來確定。

【0080】 **圖 65A** 及 **65B** 展示根據本發明實施例之藉由亞硫酸氫鹽定序及單分子即時定序確定之 HCC 細胞株（HepG2）及來自健康對照個體之白血球層樣本在 1-Mnt 解析度下之甲基化程度的散點圖。

【0081】 **圖 66A** 及 **66B** 展示根據本發明實施例之藉由亞硫酸氫鹽定序及單分子即時定序確定之 HCC 細胞株（HepG2）及來自健康對照個體之白血球層樣本在 100-knt 解析度下之甲基化程度的散點圖。

【0082】 **圖 67A** 及 **67B** 展示根據本發明實施例之 HCC 腫瘤組織及相鄰正常組織在 1-Mnt 解析度下之甲基化模式，其中甲基化程度藉由亞硫酸氫鹽定序及單分子即時定序來確定。

【0083】 **圖 68A** 及 **68B** 展示根據本發明實施例之藉由亞硫酸氫鹽定序及單分子即時定序確定之 HCC 腫瘤組織及相鄰正常組織在 1-Mnt 解析度下之甲基化程度的散點圖。

【0084】 **圖 69A** 及 **69B** 展示根據本發明實施例之藉由亞硫酸氫鹽定序及單分子即時定序確定之 HCC 腫瘤組織及相鄰正常組織在 100-knt 解析度下之甲基化程度的散點圖。

【0085】 **圖 70A** 及 **70B** 展示根據本發明實施例之 HCC 腫瘤組織及相鄰正常組織在 1-Mnt 解析度下之甲基化模式，其中甲基化程度藉由亞硫酸氫鹽定序及單分子即時定序來確定。

【0086】 **圖 71A** 及 **71B** 展示根據本發明實施例之藉由亞硫酸氫鹽定序及

單分子即時定序確定之 HCC 腫瘤組織及相鄰正常組織在 1-Mnt 解析度下之甲基化程度的散點圖。

【0087】 **圖 72A** 及 **72B** 展示根據本發明實施例之藉由亞硫酸氫鹽定序及單分子即時定序確定之 HCC 腫瘤組織及相鄰正常組織在 100-knt 解析度下之甲基化程度的散點圖。

【0088】 **圖 73** 展示根據本發明實施例之腫瘤抑制基因 *CDKN2A* 附近甲基化之異常模式的實例。

【0089】 **圖 74A** 及 **74B** 展示根據本發明實施例之藉由單分子即時定序檢測之差異性甲基化區域。

【0090】 **圖 75** 展示根據本發明實施例之使用單分子即時定序在 HCC 組織與相鄰非腫瘤組織之間的 B 型肝炎病毒 DNA 的甲基化模式。

【0091】 **圖 76A** 展示根據本發明實施例之使用亞硫酸氫鹽定序之來自患有肝硬化但無 HCC 之患者之肝臟組織中 B 型肝炎病毒 DNA 的甲基化程度。

【0092】 **圖 76B** 展示根據本發明實施例之使用亞硫酸氫鹽定序之 HCC 組織中 B 型肝炎病毒 DNA 的甲基化程度。

【0093】 **圖 77** 展示根據本發明實施例之甲基化單倍型分析。

【0094】 **圖 78** 展示根據本發明實施例之自一致序列確定之定序分子的大小分佈。

【0095】 **圖 79A**、**79B**、**79C** 及 **79D** 展示根據本發明實施例之印記區域中之對偶基因甲基化模式的實例。

【0096】 **圖 80A**、**80B**、**80C** 及 **80D** 展示根據本發明實施例之非印記區域中之對偶基因甲基化模式的實例。

【0097】 **圖 81** 展示根據本發明實施例之對偶基因特異性片段之甲基化程

度的表格。

【0098】 **圖 82** 展示根據本發明實施例之使用甲基化概況確定妊娠中血漿 DNA 之胎盤來源的實例。

【0099】 **圖 83** 展示根據本發明實施例之胎兒特異性 DNA 甲基化分析。

【0100】 **圖 84A**、**84B** 及 **84C** 展示根據本發明實施例之 SMRT-seq 之不同試劑套組之不同量測窗口大小的效能。

【0101】 **圖 85A**、**85B** 及 **85C** 展示根據本發明實施例之 SMRT-seq 之不同試劑套組之不同量測窗口大小的效能。

【0102】 **圖 86A**、**86B** 及 **86C** 展示根據本發明實施例之藉由亞硫酸氫鹽定序及 SMRT-seq（Sequel II Sequencing Kit 2.0）量化之總體甲基化程度的相關性。

【0103】 **圖 87A** 及 **87B** 展示根據本發明實施例之各種腫瘤組織與成對的相鄰非腫瘤組織之間的總體甲基化程度的比較。

【0104】 **圖 88** 展示根據本發明實施例之使用自環形一致序列（CCS）確定之序列上下文確定甲基化狀態。

【0105】 **圖 89** 展示根據本發明實施例之使用自 CCS 確定之序列上下文檢測甲基化 CpG 位點的 ROC 曲線。

【0106】 **圖 90** 展示根據本發明實施例之在沒有 CCS 資訊且沒有事先與參考基因體進行排比的情況下檢測甲基化 CpG 位點的 ROC 曲線。

【0107】 **圖 91** 展示根據本發明實施例之製備用於單分子即時定序之分子的實例。

【0108】 **圖 92** 展示根據本發明實施例之 CRISPR/Cas9 系統的圖示。

【0109】 **圖 93** 展示根據本發明實施例之用於引入跨越所關注之末端封閉

分子之兩個切口的 Cas9 複合物的實例。

【0110】 **圖 94** 展示根據本發明實施例之藉由亞硫酸氫鹽定序及單分子即時定序確定之 Alu 區的甲基化分佈。

【0111】 **圖 95** 展示根據本發明實施例之藉由使用單分子即時定序結果之模型確定的 Alu 區的甲基化程度分佈。

【0112】 **圖 96** 展示根據本發明實施例之組織及組織中 Alu 區之甲基化程度的表格。

【0113】 **圖 97** 展示根據本發明實施例之使用與 Alu 重複序列相關之甲基化信號對不同癌症類型的聚類分析。

【0114】 **圖 98A** 及 **98B** 展示根據本發明實施例之在涉及全基因體擴增及 M.SsssI 處理之測試資料集中，讀段深度對總體甲基化程度量化的影響。

【0115】 **圖 99** 展示根據本發明實施例之在使用不同子讀段深度閾值的情況下，藉由 SMRT-seq（Sequel II Sequencing Kit 2.0）及 BS-seq 確定之總體甲基化程度之間的比較。

【0116】 **圖 100** 為顯示根據本發明實施例之藉由 SMRT-seq（Sequel II Sequencing Kit 2.0）及 BS-seq 之兩次量測之間的子讀段深度對甲基化程度相關性的影響的表格。

【0117】 **圖 101** 展示根據本發明實施例之由 Sequel II Sequencing Kit 2.0 生成的資料中相對於片段大小的子讀段深度分佈。

【0118】 **圖 102** 展示根據本發明實施例之檢測核酸分子中核苷酸之修飾的方法。

【0119】 **圖 103** 展示根據本發明實施例之檢測核酸分子中核苷酸之修飾的方法。

【0120】 **圖 104** 展示根據本發明實施例之基於單倍型之相對甲基化不平衡分析。

【0121】 **圖 105A** 及 **105B** 為根據本發明實施例之病例 TBR3033 之單倍型區塊的表格，其顯示與相鄰非腫瘤組織 DNA 相比，腫瘤 DNA 中 Hap I 與 Hap II 之間的差異性甲基化程度。

【0122】 **圖 106** 為根據本發明實施例之病例 TBR3032 之單倍型區塊的表格，其顯示與相鄰正常組織 DNA 相比，腫瘤 DNA 中 Hap I 與 Hap II 之間的差異性甲基化程度。

【0123】 **圖 107A** 為根據本發明實施例之基於由 Sequel II Sequencing Kit 2.0 生成之資料，彙總顯示腫瘤與相鄰非腫瘤組織之間的兩個單倍型之間的甲基化不平衡的單倍型區塊的數量的表格。

【0124】 **圖 107B** 為根據本發明實施例之基於由 Sequel II Sequencing Kit 2.0 生成之資料，彙總顯示在不同腫瘤階段之腫瘤組織中的兩個單倍型之間的甲基化不平衡的單倍型區塊的數量的表格。

【0125】 **圖 108** 展示根據本發明實施例之基於單倍型之相對甲基化不平衡分析。

【0126】 **圖 109** 展示根據本發明實施例之對具有第一單倍型及第二單倍型之生物體的病症進行分類的方法。

【0127】 **圖 110** 展示根據本發明實施例創建人類-小鼠雜合片段，其中人類部分經甲基化，而小鼠部分未甲基化。

【0128】 **圖 111** 展示根據本發明實施例創建人類-小鼠雜合片段，其中人類部分未甲基化，而小鼠部分經甲基化。

【0129】 **圖 112** 展示根據本發明實施例，在連接後 DNA 混合物（樣本

MIX01）中 DNA 分子之長度分佈。

【0130】 **圖 113** 展示根據本發明實施例將第一 DNA（A）及第二 DNA（B）接合在一起的接合區。

【0131】 **圖 114** 展示根據本發明實施例之 DNA 混合物的甲基化分析。

【0132】 **圖 115** 展示根據本發明實施例之樣本 MIX01 中 CpG 位點的甲基化概率的盒狀圖。

【0133】 **圖 116** 展示根據本發明實施例，在樣本 MIX02 交叉連接後 DNA 混合物中 DNA 分子的長度分佈。

【0134】 **圖 117** 展示根據本發明實施例之樣本 MIX02 中 CpG 位點的甲基化概率的盒狀圖。

【0135】 **圖 118** 為根據本發明實施例比較藉由亞硫酸氫鹽定序及 Pacific Biosciences 定序確定之 MIX01 的甲基化的表格。

【0136】 **圖 119** 為根據本發明實施例比較藉由亞硫酸氫鹽定序及 Pacific Biosciences 定序確定之 MIX02 的甲基化的表格。

【0137】 **圖 120A** 及 **120B** 展示根據本發明實施例之 MIX01 及 MIX02 之僅人類及僅小鼠 DNA 的 5-Mb 面元中的甲基化程度。

【0138】 **圖 121A** 及 **121B** 展示根據本發明實施例之 MIX01 及 MIX02 之人類-小鼠雜合 DNA 片段之人類部分及小鼠部分的 5-Mb 面元中的甲基化程度。

【0139】 **圖 122A** 及 **122B** 為顯示根據本發明實施例之單個人類-小鼠雜合分子之甲基化狀態的代表圖。

【0140】 **圖 123** 展示根據本發明實施例之檢測生物樣本中之嵌合分子的方法。

【0141】 **圖 124** 展示根據本發明實施例之量測系統。

【0142】 **圖 125** 展示可與根據本發明實施例之系統及方法一起使用的實例電腦系統的方塊圖。

【0143】 **圖 126** 展示根據本發明實施例之使用 DNA 末端修復及 A 加尾之基於 *Msp*I 之靶向單分子即時定序。

【0144】 **圖 127A** 及 **127B** 展示根據本發明實施例之 *Msp*I 消化片段的大小分佈。

【0145】 **圖 128** 展示根據本發明實施例之某些選定大小範圍的 DNA 分子數的表格。

【0146】 **圖 129** 為根據本發明實施例，在限制酶消化後，CpG 島內 CpG 位點之覆蓋率百分比與 DNA 片段大小的圖。

【0147】 **圖 130** 展示根據本發明實施例之不使用 DNA 末端修復及 A 加尾之基於 *Msp*I 之靶向單分子即時定序。

【0148】 **圖 131** 展示根據本發明實施例之基於 *Msp*I 之靶向單分子即時定序，其中轉接子自連接的概率降低。

【0149】 **圖 132** 為根據本發明實施例之藉由基於 *Msp*I 之靶向單分子即時定序確定之胎盤及白血球 DNA 樣本藉由來測定之間的總體甲基化程度的圖。

【0150】 **圖 133** 展示根據本發明實施例之使用基於 *Msp*I 之靶向單分子即時定序確定之 DNA 甲基化概況對胎盤及白血球層樣本的聚類分析。

術語

【0151】「*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*」可對應於用於確定組織特異性甲基化程度之組織。來自不同個體之同一組織類型之多個樣本可用於確定該組織類型之組織特異性甲基化程度。

【0152】「*生物樣本*」係指取自人類個體之任何樣本。生物樣本可為組織生檢、細針抽吸物或血細胞。樣本亦可為例如孕婦的血漿或血清或尿液。糞便樣本亦可使用。在各種實施例中，已富集游離 DNA 之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。離心方案可包括例如 3,000 g×10 分鐘獲得流體部分，及例如 30,000 g 再離心另外 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（例如使用孔徑為 5 μm 或更小的過濾器）。

【0153】「*序列讀段*」係指自核酸分子之任何部分或全部定序之核苷酸串。舉例而言，序列讀段可為自核酸片段定序之短核苷酸串（例如 20-150）、位於核酸片段之一端或兩端的短核苷酸串或存在於生物樣本中之整個核酸片段之定序。序列讀段可以多種方式獲得，例如使用定序技術或使用探針，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。

【0154】「*子讀段*」為由環化 DNA 模板之一股中的所有鹼基生成的序列，其已藉由 DNA 聚合酶複製在一連續股中。舉例而言，子讀段對應於環化 DNA 模板 DNA 之一股。在此實例中，在環化後，一個雙股 DNA 分子將具有兩個子讀段：每個定序通道一個。在一些實施例中，生成的序列可包括一股中所有鹼基之子集，例如因為存在定序錯誤。

【0155】「*位點*」（亦稱為「*基因體位點*」）對應於單個位點，其可為單

個鹼基位置或一組相關的鹼基位置，例如 CpG 位點或較大的一組相關的鹼基位置。「基因座」可對應於包括多個位點之區域。一個基因座可僅包括一個位點，此將使得基因座在該情形下等效於一個位點。

【0156】「*甲基化狀態*」係指給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為不確定的。

【0157】 各基因體位點（例如 CpG 位點）之「*甲基化指數*」可指在該位點顯示甲基化之 DNA 片段（例如，自序列讀段或探針所確定）相比於覆蓋該位點之讀段總數之比例。「讀段」可對應於獲自 DNA 片段之資訊（例如位點處之甲基化狀態）。可使用優先與一或多個位點處具有特定甲基化狀態之 DNA 片段雜交的試劑（例如引子或探針）獲得讀段。通常，此類試劑在用根據 DNA 分子之甲基化狀態而有差異地修飾或有差異地識別 DNA 分子之方法處理後施用，該方法例如亞硫酸氫鹽轉化，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，或識別甲基胞嘧啶及羥甲基胞嘧啶之單分子定序技術（例如單分子即時定序及奈米孔定序（例如來自 Oxford Nanopore Technologies））

【0158】 區域之「*甲基化密度*」可指區域內顯示甲基化之位點處的讀段數目除以覆蓋該區域中之位點的讀段總數。位點可具有特異性特徵，例如為 CpG 位點。因此，區域之「CpG 甲基化密度」可指顯示 CpG 甲基化之讀段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）之讀段總數。舉例而言，人類基因體中每 100-kb 面元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）的總數占相對於 100-kb 區域進行定位之序列讀段所覆蓋的所有 CpG 位點的比例來確定。亦可以針對其他面元尺寸執行此分析，例如 500 bp、5 kb、10 kb、50-kb 或 1-Mb 等。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）。當區域僅包括 CpG 位點時，該 CpG 位點之甲基化指數與該區域之甲基化密度相同。「甲基化胞嘧啶之比例」可指區域中相較於所分析的胞嘧啶殘基總數（亦即包括 CpG 情形之外的胞嘧啶）的顯示為經甲基化（例如在亞硫酸氫鹽轉化後未轉化）之胞嘧啶位點「C's」數目。甲基化指數、甲基化密度、一或多個位點處甲基化之分子計數及一或多個位點處甲基化之分子（例如胞嘧啶）的比例為「*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的其他方法來查詢 DNA 分子之甲基化狀態，包括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的單分子定序（例如奈米孔定序（Schreiber 等人《國家科學院院刊（Proc Natl Acad Sci）》 2013; 110: 18910-18915）及藉由單分子即時定序（例如來自 Pacific Biosciences）（Flusberg 等人《自然-方法（Nat Methods）》 2010; 7: 461-465））。

【0159】「*甲基化體*」提供基因體中複數個位點或基因座之 DNA 甲基化量的量度。甲基化體可對應於基因體之全部、基因體之很大一部分或基因體之相對較小部分。

【0160】「*妊娠血漿甲基化體*」為自妊娠動物（例如人類）之血漿或血清確定的甲基化體。妊娠血漿甲基化體為游離甲基化體之實例，因為血漿及血清包括游離 DNA。妊娠血漿甲基化體亦為混合甲基化體之實例，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，包括但不限於紅血球系（亦即紅血球）、骨髓系（例如嗜中性白血球及其前體）及巨核細胞系之細胞。在妊娠期，血漿甲基化體可含有來自胎兒及母親之甲基化體資訊。「細胞甲基化體」對應於自患者之細胞（例如血球）確定之甲基化體。血細胞之甲基化體稱為血球甲基化體（或血液甲基化體）。

【0161】「*甲基化概況*」包括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括但不限於 CpG 位點之甲基化指數、區域中 CpG 位點之甲基化密度（簡稱 MD）、CpG 位點在相連區域上之分佈、含有一個以上 CpG 位點之區域內每個單獨 CpG 位點的甲基化模式或程度及非 CpG 甲基化。在一個實施例中，甲基化概況可包括一種以上類型之鹼基（例如胞嘧啶或腺嘌呤）之甲基化或非甲基化的模式。基因體相當大一部分之甲基化概況可視為等同於甲基化體。哺乳動物基因體中之「DNA 甲基化」通常係指在 CpG 二核苷酸中胞嘧啶殘基（亦即 5-甲基胞嘧啶）之 5'碳上添加甲基。DNA 甲基化可發生在其他情形下之胞嘧啶中，例如 CHG 及 CHH，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。

【0162】「*甲基化模式*」係指甲基化及非甲基化鹼基之順序。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甲基化鹼基之順序。舉例而言，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」表示未甲基化位點且「M」表示甲基化位點。當吾人將此概念擴展至包括但不限於甲基化之鹼基修飾時，吾人將使用術語「*修飾模式*」，其係指經修飾及未修飾之鹼基的順序。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經修飾之鹼基的順序。舉例而言，三個連續的潛在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」表示未修飾之位點且「M」表示經修飾之位點。不基於甲基化之鹼基修飾的一個實例為氧化變化，諸如在 8-側氧基-鳥嘌呤中。

【0163】 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，例如分子內甲基化鹼基或核苷酸之數目除以分子內可甲基化鹼基或核苷酸之總數。高甲基化分子為單分子甲基化程度等於或高於臨限值的分子，該臨限值可根據不同的應用而定義。該臨限值可為 5%、10%、20%、30%、40%、50%、60%、70%、80%、90%或 95%。低甲基化分子為單分子甲基化程度等於或低於臨限值的分子，該臨限值可根據不同的應用而定義，且可隨不同的應用而變化。該臨限值可為 5%、10%、20%、30%、40%、50%、60%、70%、80%、90%或 95%。

【0164】 術語「高甲基化」及「低甲基化」亦可指 DNA 分子群之甲基化程度，如藉由此等分子之多分子甲基化程度所量測。高甲基化分子群為多分子甲基化程度等於或高於臨限值的分子群，該臨限值可根據不同的應用而定義，且可隨不同的應用而變化。該臨限值可為 5%、10%、20%、30%、40%、50%、60%、70%、80%、90%或 95%。低甲基化分子群為多分子甲基化程度等於或低於臨限值的分子群，該臨限值可根據不同的應用而定義。該臨限值可為 5%、10%、20%、30%、40%、50%、60%、70%、80%、90%及 95%。在一個實施例中，分子群可與一或多個選定的基因體區域進行排比。在一個實施例中，選定的基因體區域可與諸如癌症、遺傳病症、印記病症、代謝病症或神經病症之疾病相關。選定的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。

【0165】 術語「*定序深度*」係指基因座由與基因座進行排比之序列讀段覆蓋之次數。基因座可與核苷酸一樣小，或與染色體臂一樣大，或與整個基因

體一樣大。定序深度可表示為 50×、100×等，其中「×」係指基因座由序列讀段覆蓋之次數。定序深度亦可應用於多個基因座或全基因體，在此情況下，×可指基因座或單倍體基因體或全基因體分別定序之平均次數。超深度定序可指定序深度為至少 100×。

【0166】 如本文所用之術語「*分類*」係指與樣本之特定性質相關之任何數字或其他字符。舉例而言，符號「+」（或詞語「正」）可表示樣本被分類為具有缺失或擴增。分類可為二元的（例如正或負）或具有更多的分類等級（例如自 1 至 10 或自 0 至 1 的標度）。

【0167】 術語「*閾值*」及「*臨限值*」係指操作中所用之預定數字。舉例而言，閾值尺寸可指一種尺寸，大於此尺寸則排除片段。臨限值可為一種值，高於或低於此值，則特定分類適用。此等術語中之任一者可用於此等情形中之任一者。閾值或臨限值可為代表特定分類或在兩種或更多種分類之間進行區別的「參考值」或自參考值導出。如本領域技術人員將理解的，可以各種方式確定此參考值。例如，可針對具有不同已知分類的兩個不同群組的個體確定度量，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，參考值可基於對樣本之統計分析或模擬而確定。

【0168】 術語「*癌症等級*」可指癌症是否存在（亦即存在或不存在）、癌症分期、腫瘤尺寸、是否存在轉移、身體之總腫瘤負荷、癌症對治療之反應及/或癌症嚴重程度之其他量度（例如癌症復發）。癌症等級可為數字或其他標誌，諸如符號、字母及顏色。等級可為零。癌症等級亦可包括惡化前或癌變前病況（狀態）。可以各種方式使用癌症等級。舉例而言，篩查可檢查先前未知患癌之某人是否存在癌症。評定可調查已經診斷患有癌症之某人以監測癌症隨

時間推移之進展，研究療法有效性或確定預後。在一個實施例中，預後可用患者死於癌症之機率或特定期限或時間之後癌症進展之機率或癌症轉移之機率或程度表示。檢測可能意謂「篩檢」，或可能意謂檢查具有癌症提示性特徵（例如症狀或其他陽性測試）的某人是否患有癌症。

【0169】「*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，其中等級可如上文針對癌症所描述。病理之另一實例為移植器官之排斥。其他實例病理可包括基因印記病症、自體免疫攻擊（例如損害腎臟之狼瘡性腎炎或多發性硬化症）、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。

【0170】「*妊娠相關病症*」包括以母體及/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。此等病症包括但不限於先兆子癇、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 症候群、全身性紅斑狼瘡及母親之其他免疫性疾病。

【0171】 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，即使該 DNA 片段可為單股的且不包括鹼基配對。在單股 DNA 之上下文中，「bp」可解釋為提供核苷酸長度。

【0172】 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 的長度。另外，「nt」可用於表示相對位置，諸如所分析之基因座的上游或下游。在涉及技術概念化、資料呈現、處理及分析之一些上下文中，「nt」及「bp」可互換使用。

【0173】 術語「*序列上下文*」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。

【0174】 術語「*動力學特徵*」可指源自定序，包括源自單分子即時定序之特徵。此類特徵可用於鹼基修飾分析。例示性動力學特徵包括上游及下游序列上下文、股資訊、脈衝間持續時間、脈衝寬度及脈衝強度。在單分子即時定序中，吾人連續監測聚合酶活性對 DNA 模板之影響。因此，由此類定序生成之量測結果可視為動力學特徵，例如核苷酸序列。

【0175】 術語「*機器學習模型*」可包括基於使用樣本資料（例如訓練資料）對測試資料進行預測的模型，且因此可包括監督學習。機器學習模型通常使用電腦或處理器開發。機器學習模型可包括統計模型。

【0176】 術語「*資料分析框架*」可包括可將資料作為輸入且隨後輸出預測結果的算法及/或模型。「資料分析框架」之實例包括統計模型、數學模型、機器學習模型、其他人工智慧模型及其組合。

【0177】 術語「*即時定序*」可指涉及在定序所涉及之反應進展期間進行資料收集或監測的技術。舉例而言，即時定序可涉及光學監測或拍攝 DNA 聚合酶併入新鹼基。

【0178】 術語「*約*」或「*大約*」可意謂在如本領域中一般熟習此項技術者所測定之特定值的可接受誤差範圍內，此將部分取決於如何量測或測定該值，亦即，量測系統之極限。舉例而言，根據本領域中之實踐，「約」可意謂在 1 個或大於 1 個標準差之範圍內。或者，「約」可意謂既定值之至多 20%、至多 10%、至多 5%或至多 1%之範圍。或者，尤其關於生物系統或方法，術語「約」或「大約」可意謂在值之一定數量級內、在 5 倍內且更佳在 2 倍內。當特

定值描述於本申請案與申請專利範圍中時，除非另行說明，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」可以指±10%。術語「約」可指±5%。

【實施方式】

【0179】 實現無亞硫酸氫鹽測定鹼基修飾（包括甲基化鹼基）為不同研究工作的主題，但沒有一個經證明為商業上可行的。最近，已公開一種無亞硫酸氫鹽檢測 5mC 及 5hmC 之方法（Y. Liu 等人, 2019），其使用溫和的條件進行 5mC 及 5hmC 鹼基轉化。此方法涉及多個步驟的酶促及化學反應，包括十-十一易位（TET）氧化、吡啶硼烷還原及 PCR。轉化反應各步驟之效率以及 PCR 偏差均會對 5mC 分析之最終準確性產生不利影響。舉例而言，據報導 5mC 轉化率為約 96%，假陰性率為約 3%。此表現將有可能限制吾人檢測基因體中甲基化之某些細微變化的能力。另一方面，酶促轉化將無法在整個基因體中表現同樣出色。舉例而言，5hmC 之轉化率比 5mC 之轉化率低 8.2%，非 CpG 之轉化率比 CpG 情形之轉化率低 11.4%（Y. Liu 等人, 2019）。因此，理想的情況為開發用於量測天然 DNA 分子之鹼基修飾的方法，該方法無需任何事先轉化（化學或酶促或其組合）步驟，甚至無需擴增步驟。

【0180】 存在許多概念驗證研究（Q. Liu 等人, 2019；Ni 等人, 2019），其中藉由長讀段奈米孔定序方法（例如使用由 Oxford Nanopore Technologies 開發之系統）產生的電信號使吾人能夠使用深度學習方法檢測甲基化狀態。除 Oxford Nanopore 之外，存在其他單分子定序方法可用於長讀段。一個實例為單分子即時定序。單分子即時定序之一個實例為將 Pacific Biosciences SMRT 系統商

業化。由於單分子即時定序（例如 Pacific Biosciences SMRT 系統）之原理與非光學型奈米孔系統（例如 Oxford Nanopore Technologies）之原理不同，因此針對此類非光學型奈米孔系統開發的鹼基修飾檢測方法不能用於單分子即時定序。舉例而言，非光學奈米孔系統並非設計用於捕捉基於固定化 DNA 聚合酶之 DNA 合成（由單分子即時定序，諸如 Pacific Biosciences SMRT 系統採用）所產生的螢光信號的模式。作為另一個實例，在 Oxford Nanopore 定序平台中，每個量測的電事件均與 k 聚體（例如 5 聚體）相關（Q. Liu 等人, 2019）。然而，在 Pacific Biosciences SMRT 定序平台中，每個螢光事件通常與單個併入的鹼基相關。此外，單個 DNA 分子將在 Pacific Biosciences SMRT 定序中多次定序，包括瓦生股及克立克股。相反，對於 Oxford Nanopore 長讀段定序方法，對瓦生股及克立克股各進行一次序列讀出。

【0181】 據報導，聚合酶動力學將受大腸桿菌序列中甲基化狀態的影響（Flusberg 等人, 2010）。以往的研究表明，當與 6mA、4mC、5hmC 及 8-側氧基-鳥嘌呤之檢測相比時，使用單分子即時定序之聚合酶動力學來推導單分子中特定 CpG 之甲基化狀態（5mC 與 C）更具挑戰性。原因在於甲基小且朝向大溝，不參與鹼基配對，導致 5mC 引起的動力學非常微妙的中斷（Clark 等人, 2013）。因此，缺乏在單分子水準上確定胞嘧啶甲基化狀態的方法。

【0182】 Suzuki 等人開發一種算法（Suzuki 等人, 2016），試圖結合相鄰 CpG 位點在脈衝間持續時間（IPD）比率，以提高鑑別彼等位點之甲基化狀態的可信度。然而，此算法僅允許吾人預測基因體區域為完全甲基化或完全未甲基化，但缺乏確定中間甲基化模式的能力。

【0183】 關於單分子即時定序，目前的方法僅獨立使用一或兩個參數，由於 5-甲基胞嘧啶與胞嘧啶之間的量測差異，在檢測 5mC 時獲得的準確性非常

有限。舉例而言，Flusberg 等人證明 IPD 在包括 N6-甲基腺苷、5-甲基胞嘧啶及 5-羥甲基胞嘧啶之鹼基修飾中被改變。然而，未發現定序動力學之脈衝寬度（PW）具有顯著影響。因此，在他們用於預測鹼基修飾之方法中，以檢測 N6-甲基腺苷為例，僅使用 IPD 而未使用 PW。

【0184】 在同一小組的後續出版物中（Clark 等人, 2012；Clark 等人 2013），在檢測 5-甲基胞嘧啶之算法中併入 IPD 而非 PW。在 Clark 等人 2012 中，在不轉化為 5-甲基胞嘧啶的情況下，5-甲基胞嘧啶之檢測率僅在 1.9%至 4.3%之範圍內。此外，在 Clark 等人 2013 中，作者進一步重申 5-甲基胞嘧啶動力學特徵的微妙性。為了克服檢測 5-甲基胞嘧啶之靈敏度低，Clark 等人進一步開發一種方法，該方法使用十-十一易位（Tet）蛋白將 5-甲基胞嘧啶轉化為 5-羧甲基胞嘧啶，以提高 5-甲基胞嘧啶之靈敏度（Clark 等人 2013），因為 5-羧基胞嘧啶引起之 IPD 改變比 5-甲基胞嘧啶引起之 IPD 改變多得多。

【0185】 在 Blow 等人最近的報告中，使用 Flusberg 等人先前描述之基於 IPD 比率之方法來檢測 217 種細菌及 13 種古細菌物種的鹼基修飾，每個生物體的讀取覆蓋率為 130 倍（Blow 等人, 2016）。在他們鑑別之所有鹼基修飾中，僅 5%涉及 5-甲基胞嘧啶。他們將 5-甲基胞嘧啶之此低檢測率歸因於單分子即時定序檢測 5-甲基胞嘧啶之靈敏度低。在大多數細菌中，DNA 甲基轉移酶（MT 酶）靶向一組序列基元進行甲基化（例如在大腸桿菌中 Dam 之 5'-GmATC-3'或 Dcm 之 5'-CmCWGG-3'），在基因體中幾乎所有此等基元處，此等基元位點僅一小部分保持未甲基化（Beaulaurier 等人 2019）。此外，使用基於 IPD 之方法對經 Tet 蛋白處理或未經 Tet 蛋白處理之 5'-CCWGG-3'基元中第二個 C 之甲基化狀態進行分類，得出的 5-甲基胞嘧啶之檢測率分別為 95.2%及 1.9%（Clark 等人 2013）。總體而言，沒有事先進行鹼基轉化（例如使用 Tet 蛋白）之 IPD 方法遺漏大多數

5-甲基胞嘧啶。

【0186】 在上述研究中（Clark 等人, 2012；Clark 等人, 2013；Blow 等人, 2016），使用基於 IPD 之算法，而不考慮候選鹼基修飾所在的序列上下文。其他小組已嘗試考慮核苷酸之序列上下文來檢測鹼基修飾。舉例而言，Feng 等人使用階層式模型來分析 IPD，以在各別序列上下文中檢測 4-甲基胞嘧啶及 6-甲基腺苷（Feng 等人 2013）。然而，在他們的方法中，他們僅考慮所關注鹼基處之 IPD 及與該鹼基相鄰的序列上下文，但沒有使用與所關注鹼基相連的所有鄰近鹼基的 IPD 資訊。另外，算法中未考慮 PW，且他們沒有提供關於 5-甲基胞嘧啶檢測之任何資料。

【0187】 在另一項研究中，Schadt 等人開發一種稱為條件隨機場之統計方法，用於分析所關注鹼基及鄰近鹼基的 IPD 資訊，以確定所關注鹼基是否為 5-甲基胞嘧啶（Schadt 等人, 2012）。在此項工作中，他們亦藉由將此等鹼基輸入方程式中來考慮此等鹼基之間的 IPD 相互作用。然而，他們沒有在方程式中輸入核苷酸序列，亦即 A、T、G 或 C。當他們應用該方法確定 M.Sau3AI 質體之甲基化狀態時，即使在質體序列之 800 倍序列覆蓋率下，ROC 曲線下面積仍接近 0.5。此外，在他們的方法中，他們在分析中沒有考慮 PW。

【0188】 在 Beckman 等人之另一項研究中，他們比較目標細菌基因體與例如經由全基因體擴增獲得之完全未甲基化基因體之間在基因體中共享相同 4-nt 或 6-nt 基元之所有序列的 IPD（Beckman 等人 2014）。此類分析之目的僅在於鑑別會更頻繁地受鹼基修飾影響之基元。在研究中，其僅考慮潛在修飾鹼基之 IPD，而沒有考慮鄰近鹼基之 IPD 或 PW。他們的方法不能提供關於單個核苷酸甲基化狀態之資訊。

【0189】 總而言之，此等先前僅利用 IPD 或結合鄰近核苷酸中之序列資

訊來對資料進行分組的嘗試均無法以有意義或實際的準確性確定 5-甲基胞嘧啶之鹼基修飾。在 Gouil 等人的最新綜述中，作者推斷由於信雜比低，使用單分子即時定序檢測單分子中之 5-甲基胞嘧啶為不準確的（Gouil 等人, 2019）。在此等先前的研究中，使用動力學特徵進行全基因體甲基化體分析是否可行尚不明確，尤其是對於複雜的基因體，諸如人類基因體、癌症基因體或胎兒基因體。

【0190】 與先前的研究相反，本揭示案中描述之一些實施例係基於量測及利用量測窗口內每個鹼基的 IPD、PW 及序列上下文。吾等推理，若吾等可使用多個度量之組合，例如同時利用包括上游及下游序列上下文、股資訊、IPD、脈衝寬度以及脈衝強度之特徵，則吾等可能能夠實現單鹼基解析度下鹼基修飾之精確量測（例如 mC 檢測）。序列上下文係指一段 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目標的鹼基。在一個實施例中，該段 DNA 可靠近進行鹼基修飾分析之鹼基。在另一個實施例中，該段 DNA 可遠離進行鹼基修飾分析之鹼基。該段 DNA 可為進行鹼基修飾分析之鹼基的上游及/或下游。

【0191】 在一個實施例中，用於鹼基修飾分析之上游及下游序列上下文、股資訊、IPD、脈衝寬度以及脈衝強度之特徵稱為動力學特徵。

【0192】 本揭示案中存在之實施例可用於自但不限於細胞株、生物體樣本（例如實體器官、實體組織、經由內視鏡檢獲得之樣本、血液、或孕婦之血漿或血清或尿液、絨毛膜絨毛生檢等）、自環境獲得之樣本（例如細菌、細胞污染物）、食物（例如肉類）獲得之 DNA。在一些實施例中，本揭示案中存在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）富集基因體之一部分的

步驟之後應用。儘管本發明不需要酶促或化學轉化來其作用，但在某些實施例中，可包括此類轉化步驟以進一步增強本發明之效能。

【0193】 本揭示案之實施例允許提高檢測鹼基修飾或量測修飾程度之準確性或實用性或便利性。可直接檢測修飾。實施例可避免酶促或化學轉化，其可能無法保留所有修飾資訊以供檢測。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 之兩股可一起定序，從而使一股之序列與其互補序列配對至另一股。相比之下，PCR 擴增會分開雙股 DNA 之兩股，因此此類序列配對為困難的。

【0194】 在有或沒有酶促或化學轉化之情況下確定的甲基化概況可用於分析生物樣本。在一個實施例中，甲基化概況可用於檢測細胞 DNA 之來源（例如母體或胎兒、組織、病毒或腫瘤）。檢測組織中之異常甲基化概況有助於鑑別個體之發育障礙以及鑑別及預測腫瘤或惡性腫瘤。單倍型之間甲基化程度的不平衡可用於檢測病症，包括癌症。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）。

【0195】 甲基化分析可藉由增強訓練來改進，其可包括縮小訓練集中使用的資料。可針對特定區域進行分析。在實施例中，此類靶向可涉及一種酶，該酶單獨或與其他試劑組合可基於其序列切割 DNA 序列或基因體。在一些實施例中，該酶為識別及切割特定 DNA 序列之限制酶。在其他實施例中，可組合使用一種以上具有不同識別序列之限制酶。在一些實施例中，限制酶可基於識別序列之甲基化狀態切割或不切割。在一些實施例中，該酶為 CRISPR/Cas 家族中之一種。舉例而言，可使用 CRISPR/Cas9 系統或其他基於引導 RNA（亦即短 RNA

序列，其與互補的目標 DNA 序列結合且在過程中引導酶作用於目標基因體位置）之系統來靶向所關注之基因體區域。在一些情況下，無需與參考基因體進行排比即可進行甲基化分析。

使用單分子即時定序之甲基化檢測

【0196】 本揭示案之實施例允許直接檢測鹼基修飾，而無需酶促或化學轉化。經由單分子即時定序獲得之動力學特徵（例如序列上下文、IPD 及 PW）可用機器學習進行分析，以開發模型來檢測修飾或不存在修飾。修飾程度可用於確定 DNA 分子之來源或病症之存在或程度。

【0197】 使用 Pacific Biosciences SMRT 定序作為單分子即時定序之實例進行說明，將 DNA 聚合酶分子置於充當零模波導（ZMW）之孔的底部。ZMW 為一種奈米光子器件，用於將光限制於小的觀察體積中，該觀察體積可為直徑極小的孔洞且不允許光在用於檢測之波長範圍內傳播，使得僅固定化聚合酶併入之染料標記核苷酸的光信號的發射可針對低且恆定的背景信號進行檢測（Eid 等人, 2009）。DNA 聚合酶催化經螢光標記之核苷酸併入互補核酸股中。

【0198】 **圖 1** 展示藉由單分子環形一致性定序對攜帶鹼基修飾之分子進行定序的實例。分子 102、104 及 106 攜帶鹼基修飾。DNA 分子（例如分子 106）可與髮夾轉接子連接以形成連接分子 108。連接分子 108 可隨後形成環化分子 110。環化分子可與固定化 DNA 聚合酶結合，且可啟動 DNA 合成。亦可對不攜帶鹼基修飾之分子進行定序。

【0199】 **圖 2** 展示藉由單分子即時定序對攜帶甲基化及/或未甲基化之 CpG 位點之分子進行定序的實例。DNA 分子首先與髮夾轉接子連接以形成環化分子，該等環化分子將與固定化 DNA 聚合酶結合且啟動 DNA 合成。在圖 2 中，DNA 分子 202 與髮夾轉接子連接以形成連接分子 204。連接分子 204 隨後形成

環化分子 206。亦可對無 CpG 位點之分子進行定序。環化分子 206 包括未甲基化之 CpG 位點 208，其仍可進行定序。

【0200】 一旦啟動 DNA 合成，經螢光染料標記之核苷酸將基於環形 DNA 模板藉由固定化聚合酶併入新合成之股中，從而導致光信號的發射。因為 DNA 模板已經環化，所以整個環形 DNA 模板將多次經過聚合酶（亦即 DNA 模板中之一個核苷酸將被多次定序）。由該過程產生的序列稱為子讀段，其中環化 DNA 模板中之所有鹼基全部通過 DNA 聚合酶。ZMW 中之一個分子將產生多個子讀段，因為聚合酶可圍繞整個環形 DNA 模板繼續多次。在一個實施例中，子讀段可僅含有環形 DNA 模板之序列、鹼基修飾或其他分子資訊之子集，因為在一個實施例中，存在定序錯誤。

【0201】 如**圖 3** 所示，所得螢光脈衝之到達時間及持續時間將允許吾人量測聚合酶動力學。脈衝間持續時間（IPD）為兩個發射脈衝之間的時間段長度的度量，每個發射脈衝將暗示新生股中併入之經螢光標記之核苷酸（圖 3）。如圖 3 所示，脈衝寬度（PW）為反映聚合酶動力學之另一度量，其與鹼基判讀有關之脈衝的持續時間相關聯。PW 可為在信號峰高度之 0%處的脈衝持續時間（亦即併入之經染料標記之核苷酸的螢光強度）。在一個實施例中，PW 可由例如但不限於信號峰高度之 5%、10%、20%、30%、40%、50%、60%、70%、80%或 90%處的脈衝持續時間定義。在一些實施例中，PW 可為峰下面積除以信號峰高度。

【0202】 已證明此類聚合酶動力學諸如 IPD 受合成及微生物序列（例如大腸桿菌）中諸如 N6-甲基腺嘌呤（6mA）、5-甲基胞嘧啶（5mC）及 5-羥甲基胞嘧啶（5hmC）之鹼基修飾影響（Flusberg 等人, 2010）。Flusberg 等人 2010 並未使用序列上下文及 IPD 作為獨立的輸入來檢測修飾，從而導致模型缺乏實際

有意義的檢測準確性。Flusberg 等人僅使用序列上下文來確認 GATC 中出現 6mA。Flusberg 等人未提及有關使用序列上下文與 IPD 結合作為輸入來檢測甲基化狀態。

【0203】 由於互補股中 5-甲基胞嘧啶之新鹼基併入所引起的弱中斷使得在僅使用 IPD 信號時，即使用於相對簡單的微生物基因體，甲基化判讀仍極具挑戰性，據報導甲基化基元 CmCWGG 之檢測率僅在 1.9%至 4.3%之範圍內（Clark 等人, 2013）。舉例而言，Pacific Biosciences 所提供之分析套裝軟體（SMRT Link v6.0.0）無法進行 5mC 分析。此外，先前版本 SMRT Link v5.1.0 要求吾人在甲基化分析之前使用 Tet1 酶將 5mC 轉化為 5-羧基胞嘧啶（5caC），因為與 5caC 相關之 IPD 信號將會增強（Clark 等人, 2013）。因此，毫不奇怪，沒有研究表明使用單分子即時定序以全基因體方式分析人類基因體之天然 DNA 的可行性。

量測窗口模式及機器學習模型

【0204】 需要在不進行酶促或化學轉化修飾及/或鹼基之情況下檢測鹼基中之修飾的技術。如本文所述，目標鹼基中之修飾可使用自單分子即時定序獲得之目標鹼基周圍鹼基的動力學特徵資料來檢測。動力學特徵可包括脈衝間持續時間、脈衝寬度及序列上下文。此等動力學特徵可針對目標鹼基上游及下游一定數量的核苷酸的量測窗口獲得。此等特徵（例如在量測窗口中之特定位置）可用於訓練機器學習模型。作為樣本製備之一個實例，DNA 分子之兩股可由髮夾轉接子連接，從而形成環形 DNA 分子。環形 DNA 分子允許獲得瓦生股及克立克股中任一者或兩者之動力學特徵。可基於量測窗口中之動力學特徵開發資料分析框架。此資料分析框架可隨後用於檢測修飾，包括甲基化。該部分描述檢測修飾之各種技術。

使用單股

【0205】 如**圖 4** 所示，舉例而言，吾等自 Pacific Biosciences SMRT 定序獲得瓦生股之子讀段，以分析一個特定鹼基關於鹼基修飾之狀態。在圖 4 中，進行鹼基修飾分析之鹼基每一側的 3 個鹼基定義為量測窗口 400。在一個實施例中，此 7 個鹼基（亦即 3 個核苷酸（nt）上游及下游序列及一個用於鹼基修飾分析之核苷酸）的序列上下文、IPD 及 PW 編譯為 2 維（亦即 2-D）矩陣作為量測窗口。在所示實例中，量測窗口 400 係針對瓦生股之一個子讀段。其他變型描述於本文中。

【0206】 矩陣之第一列 402 指示所研究之序列。在矩陣之第二列 404 中，位置 0 表示用於鹼基修飾分析之鹼基。相對位置-1、-2 及-3 分別指示進行鹼基修飾分析之鹼基上游的位置 1-nt、2-nt 及 3-nt。相對位置+1、+2 及+3 分別指示進行鹼基修飾分析之鹼基下游的位置 1-nt、2-nt 及 3-nt。每個位置包括 2 行，其含有相應的 IPD 及 PW 值。下面 4 列（列 408、412、416 及 420）分別對應於股（例如瓦生股）中 4 種類型之核苷酸（A、C、G 及 T）。矩陣中 IPD 及 PW 值之存在取決於在特定位置對哪種對應的核苷酸類型進行定序。如圖 4 所示，在相對位置 0 處，IPD 及 PW 值顯示在瓦生股中指示『G』之列中，表明在該位置之序列結果中判讀鳥嘌呤。行中不對應於定序鹼基之其他網格將編碼為『0』。舉例而言，對應於 2-D 數位矩陣（圖 4）之序列資訊對於瓦生股將為 5'-GATGACT-3'。

【0207】 如**圖 5** 中描繪之一個實施例所示，量測窗口可應用於克立克股之資料。吾等自單分子即時定序獲得克立克股之子讀段，以分析一個特定鹼基關於鹼基修飾之狀態。在圖 5 中，進行鹼基修飾分析之鹼基每一側的 3 個鹼基及進行鹼基修飾分析之鹼基將定義為量測窗口。在一個實施例中，此 7 個鹼基

（亦即 3 個核苷酸（nt）上游及下游序列及一個用於鹼基修飾分析之核苷酸）的序列上下文、IPD 及 PW 編譯為 2 維（亦即 2-D）矩陣作為量測窗口。矩陣之第一列指示所研究之序列。在矩陣之第二列中，位置 0 表示用於鹼基修飾分析之鹼基。相對位置-1、-2 及-3 分別指示進行鹼基修飾分析之鹼基上游的位置 1-nt、2-nt 及 3-nt。相對位置+1、+2 及+3 分別指示進行鹼基修飾分析之鹼基下游的位置 1-nt、2-nt 及 3-nt。每個位置包括 2 行，其含有相應的 IPD 及 PW 值。下面的 4 列對應於此股（例如克立克股）中 4 種類型之核苷酸（A、C、G 及 T）。矩陣中 IPD 及 PW 值之存在取決於在特定位置對哪種對應的核苷酸類型進行定序。如圖 5 所示，在相對位置 0 處，IPD 和 PW 值顯示在克立克股中指示『T』之列中，表明在該位置之序列結果中判讀胸腺嘧啶。行中不對應於定序鹼基之其他網格將編碼為『0』。舉例而言，對應於 2-D 數位矩陣（圖 5）之序列資訊對於克立克股將為 5'-ACTTAGC-3'。

使用瓦生股及克立克股

【0208】 **圖 6** 展示一個實施例，其中量測窗口可以可結合來自瓦生股及其互補克立克股之資料的方式實現。如圖 6 所示，吾等自單分子即時定序獲得瓦生股及克立克股之子讀段，以分析一個特定鹼基之修飾。在一個實施例中，來自環形 DNA 模板之克立克股的量測窗口與來自瓦生股之量測窗口互補，對其進行鹼基修飾分析。在圖 6 中，瓦生股中進行鹼基修飾分析之第一鹼基每一側的 3 個鹼基及第一鹼基將定義為第一量測窗口。克立克股中第二鹼基每一側的 3 個鹼基及第二鹼基將定義為第二量測窗口。第二鹼基與第一鹼基互補。在一個實施例中，來自瓦生股及克立克股之此 7 個鹼基（亦即 3 個核苷酸（nt）上游及下游序列及一個用於鹼基修飾分析之核苷酸）的序列上下文、IPD、PW 編譯為 2 維（亦即 2-D）矩陣。來自瓦生股及克立克股之此等量測窗口分別視為

第一及第二量測窗口。

【0209】 瓦生股及克立克股矩陣之第一列指示所研究之序列。在瓦生股矩陣之第二列中，位置 0 表示用於鹼基修飾分析之第一鹼基。克立克股矩陣之第二列中所示之位置 0 表示與第一鹼基互補之第二鹼基。相對位置-1、-2 及-3 分別指示第一及第二鹼基上游之位置 1-nt、2-nt 及 3-nt。相對位置+1、+2 及+3 分別指示第一及第二鹼基下游之位置 1-nt、2-nt 及 3-nt。自瓦生股及克立克股得出之每個位置將對應於含有相應 IPD 及 PW 值之 2 行。瓦生股及克立克股矩陣中之下面 4 列分別對應於特定股（例如克立克股）中 4 種類型之核苷酸（A、C、G 及 T）。矩陣中 IPD 及 PW 值之存在取決於在特定位置對哪種對應的核苷酸類型進行定序。

【0210】 如圖 6 所示，在相對位置 0 處，IPD 及 PW 值顯示在瓦生股中指示『A』之列及克立克股中指示『T』之列，表明在瓦生股及克立克股之該位置的序列結果中分別判讀腺嘌呤及胸腺嘧啶。行中不對應於定序鹼基之其他網格將編碼為『0』。舉例而言，對應於瓦生股之 2-D 數位矩陣（圖 6）的序列資訊將為 5'-ATAAGTT-3'。對應於克立克股之 2-D 數位矩陣（圖 6）的序列資訊將為 5'-AACTTAT-3'。

【0211】 如此實例中所示，來自瓦生股及克立克股之資料可組合形成新矩陣，該矩陣亦可視為量測窗口。此新矩陣可作為用於訓練機器學習模型之單個樣本使用。因此，新矩陣中之所有值均可視為單獨的特徵，儘管在 2D 矩陣中之特定位置可能會產生影響，例如在使用卷積神經網路（CNN）時。不同股在各個位置處的序列上下文可經由矩陣中之非零項目來傳達。

【0212】 **圖 7** 展示量測窗口可以來自瓦生股及克立克股之資料並非彼此完全互補之位置的方式來實現。如圖 7 所示，第一量測窗口為 5'-ATAAGTT-3'；

且第二量測窗口為 5'-GTAACGC-3'。在一些實施例中，瓦生股及克立克股可彼此移位，以使得位置不互補。

【0213】 **圖 8** 展示量測窗口可用於分析 CpG 位點之甲基化狀態。位置 0 對應於 CpG 位點之胞嘧啶，且因此在兩股之間存在一個位置的位移，使得 C 在兩股之 0 位置處。因此，來自瓦生股及克立克股之量測窗口中包括之序列僅一部分彼此互補。在其他實施例中，來自瓦生股及克立克股之量測窗口中的所有序列可彼此互補。在其他實施例中，來自瓦生股及克立克股之量測窗口中的序列均不彼此互補。

【0214】 在一個實施例中，對於量測窗口，圍繞進行鹼基修飾分析之鹼基的 DNA 序列段長度可為不對稱的。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 可包括但不限於 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 可包括但不限於 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。

訓練模型及檢測修飾

【0215】 **圖 9** 展示關於如何使用量測窗口確定任何鹼基修飾之一般程序。對已知未修飾及經修飾之 DNA 樣本進行單分子即時定序。經修飾之 DNA（例如經修飾之分子 902）意謂鹼基（例如鹼基 904）在該位點具有修飾（例如

甲基化）。未修飾之 DNA（例如未修飾之分子 906）意謂鹼基（例如鹼基 908）在該位點不具有修飾。兩組 DNA 均可人工創建或處理以形成經修飾/未修飾之 DNA。

【0216】 在階段 910，樣本可隨後進行單分子即時定序。作為 SMRT 定序之一部分，環形分子可藉由反覆通過固定化 DNA 聚合酶而多次定序。每次獲得之序列資訊將視為子讀段。因此，一個環形 DNA 模板將產生多個子讀段。可使用例如但不限於 BLASR（Mark J Chaisson 等人，《BMC 生物資訊學（BMC Bioinformatics）》. 2012; 13: 238）將定序子讀段與參考基因體進行排比。在各種其他實施例中，BLAST（Altschul SF 等人，《分子生物學雜誌（J Mol Biol）》. 1990;215(3):403-410）、BLAT（Kent WJ，《基因體研究（Genome Res.）》2002;12(4):656-664）、BWA（Li H 等人，《生物資訊學（Bioinformatics.）》2010;26(5):589-595）、NGMLR（Sedlazeck FJ 等人，《自然-方法》2018;15(6):461-468）、LAST（Kielbasa SM 等人，《基因體研究》 2011;21(3):487-493）及 Minimap2（Li H，《生物資訊學》2018;34(18):3094-3100）可用於將子讀段與參考基因體進行排比。排比可允許來自多個子讀段之資料組合（例如平均），因為可鑑別每個子讀段中相同位置的資料。

【0217】 在階段 912，自排比結果獲得進行鹼基修飾分析之鹼基周圍的 IPD、PW 及序列上下文。在階段 914，將 IPD、PW 及序列上下文記錄在特定結構中，例如但不限於如圖 9 所示的 2-D 矩陣。

【0218】 在階段 916，使用含有參考動力學模式衍生之具有已知鹼基修飾之分子的許多 2-D 矩陣來訓練分析、計算、數學或統計模型。在階段 918，開發由訓練產生之統計模型。為了簡單起見，圖 9 僅展示由訓練開發之統計模型，但可開發任何模型或資料分析框架。例示性資料分析框架包括機器學習模

型、統計模型及數學模型。統計模型可包括但不限於線性回歸、邏輯回歸、深度循環神經網路（例如長短期記憶，LSTM）、貝葉斯分類器（Bayes classifier）、隱式馬爾可夫模型（hidden Markov model，HMM）、線性判別分析（LDA）、k 均值聚類、具有雜訊之基於密度之空間聚類應用（DBSCAN）、隨機森林算法及支持向量機（SVM）。進行鹼基修飾分析之鹼基周圍的 DNA 段可為該鹼基上游 X-nt 及下游 Y-nt，亦即「量測窗口」。

【0219】 由於已知正確的輸出（亦即修飾狀態），因此資料結構可用於訓練過程。舉例而言，瓦生及/或克立克股之對應於鹼基上游及下游 3-nt 的 IPD、PW 及序列上下文可用於構築 2-D 矩陣，以用於訓練對鹼基修飾進行分類之統計模型。以此方式，訓練可提供一種模型，該模型可對具有先前已知狀態之核酸之位置處的鹼基修飾進行分類。

【0220】 **圖 10** 展示關於自攜帶已知鹼基修飾狀態之 DNA 樣本習得的統計模型如何可檢測鹼基修飾的一般程序。對具有未知鹼基修飾狀態之樣本進行 SMRT 定序。使用例如上述技術將定序子讀段與參考基因體進行排比。另外或替代地，子讀段可彼此進行排比。其他實施例可僅使用一個子讀段或獨立地對其進行分析，從而不進行排比。

【0221】 對於進行鹼基修飾分析之鹼基，吾人將使用與訓練步驟（圖 9）中所用相當的量測窗口在排比結果中獲得瓦生及/或克立克股之 IPD、PW 及序列上下文，並與該鹼基相關聯。在另一個實施例中，訓練與測試程序之間的量測窗口將為不同的。舉例而言，訓練與測試程序之間的量測窗口大小可能有所不同。彼等 IPD、PW 及序列上下文將轉換為 2-D 矩陣。測試樣本之此類 2-D 矩陣將與參考動力學特徵進行比較，以確定鹼基修飾。舉例而言，測試樣本之 2-D 矩陣可經由自訓練樣本習得之統計模型與參考動力學特徵進行比較，從而

可確定測試樣本中核酸分子之位點的鹼基修飾。統計模型可包括但不限於線性回歸、邏輯回歸、深度循環神經網路（例如長短期記憶，LSTM）、貝葉斯分類器、隱式馬爾可夫模型（HMM）、線性判別分析（LDA）、k 均值聚類、具有雜訊之基於密度之空間聚類應用（DBSCAN）、隨機森林算法及支持向量機（SVM）。

【0222】 **圖 11** 展示關於如何使該方法對 CpG 位點處之甲基化狀態進行分類的一般程序。對已知在 CpG 位點未甲基化及甲基化之 DNA 樣本進行單分子即時定序。將定序子讀段與參考基因體進行排比。使用瓦生股資料。

【0223】 自排比結果獲得進行甲基化分析之 CpG 位點之胞嘧啶周圍的 IPD、PW 及序列上下文，且記錄在特定結構中，例如但不限於如圖 11 所示的 2-D 矩陣。將含有參考動力學模式衍生之具有已知甲基化狀態之分子的許多 2-D 矩陣用於訓練統計模型。受詢問鹼基周圍之一段 DNA 可為該鹼基上游 X-nt 及下游 Y-nt，亦即「量測窗口」。X 可包括但不限於 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 可包括但不限於 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，瓦生股之對應於鹼基上游及下游 3-nt 之 IPD、PW 及序列上下文可用於構築 2-D 矩陣，該矩陣用於訓練統計模型以對鹼基修飾進行分類。

【0224】 **圖 12** 展示對未知樣本之甲基化狀態進行分類的一般程序。對具有未知甲基化狀態之樣本進行單分子即時定序。將定序子讀段與參考基因體進行排比。

【0225】 對於排比結果中CG位點之胞嘧啶，吾人將使用在訓練步驟（圖11）中應用、與修飾受詢問之鹼基相關聯之相當的量測窗口獲得瓦生股之IPD、PW 及序列上下文。彼等 IPD、PW 及序列上下文可轉換為 2-D 矩陣。測試樣本之此類 2-D 矩陣將與圖 11 中所示之參考動力學模式進行比較，以確定甲基化狀態。X11

【0226】 **圖 13** 及**圖 14** 展示來自克立克股之動力學特徵可用於如上詳述之訓練及測試程序，類似於瓦生股之程序。統計模型可為相同或不同的模型。當模型不同時，其可用於獲得獨立的分類，該等分類可進行比較，例如若分類一致，則鑑別為修飾狀態。若分類不一致，則可鑑別為未分類狀態。當模型相同時，可將資料組合成單個資料結構，例如圖 6 中之矩陣。

【0227】 **圖 15** 及**圖 16** 展示來自瓦生股及克立克股之動力學特徵可用於如上詳述之訓練及測試程序。對已知在 CpG 位點未甲基化及甲基化之 DNA 樣本進行單分子即時定序。將定序子讀段與參考基因體進行排比，儘管子讀段相互進行排比為可能的，如本文所述之其他方法可實現。

【0228】 對於排比結果中之子讀段，獲得進行甲基化分析之 CpG 位點之胞嘧啶周圍的 IPD、PW 及序列上下文。由於 DNA 分子經由使用兩個髮夾轉接子環化（例如遵循 SMRTBell 模板製備方案），因此環形分子可定序一次以上，從而生成一個分子之多個子讀段。子讀段可用於生成環形一致性定序（CCS）讀段。一般而言，對於本文所述之所有方法，一個 ZMW 可生成多個子讀段，但僅對應於一個 CCS 讀段。

【0229】 在一些實施例中，完全未甲基化資料集可藉由對人類 DNA 片段進行 PCR 創建。舉例而言，完全甲基化資料集可經由 CpG 甲基轉移酶 M.SssI 處理之人類 DNA 片段產生，其中假定所有 CpG 位點均經甲基化。在其他實例中，可使用另一種 CpG 甲基轉移酶，諸如 M.MpeI。在其他實施例中，可使用具有已知甲基化狀態之合成序列或預先存在的具有不同甲基化程度之 DNA 樣本，或藉由限制酶切割甲基化及未甲基化之 DNA 分子，隨後進行連接（其將產生一定比例的嵌合甲基化/未甲基化 DNA 分子）產生之雜合甲基化狀態來訓練甲基化預測模型或分類器。

【0230】 包括序列上下文、IPD 及脈衝寬度（PW）之動力學模式的轉換可為 2-D 矩陣，其包含瓦生股及克立克股之特徵，用於分析 CG 位點處之甲基化狀態，如圖 15 所示。此方法使吾等能夠準確地捕捉由甲基化胞嘧啶以及其附近的序列上下文引起之細微動力學變化。與本文所述之各種方法中之任一者一樣，對於子讀段中存在之每個 CpG，（例如 CpG 位點之胞嘧啶上游及下游的 3 個鹼基）之量測窗口可用於後續分析，從而致使總共 7 個核苷酸（包括 CpG 位點之胞嘧啶）被一起分析。可計算該 7 個核苷酸中每個鹼基的 IPD 及 PW。為了捕捉歸因於動力學變化之序列上下文，可將 IPD 及 PW 信號編譯成特定的鹼基判讀、相對定序位置及股資訊，如圖 15 所示。為了簡單起見，此類資料結構稱為動力學之 2-D 數位矩陣。

【0231】 此類 2-D 數位矩陣類似於「2-D 數位影像」。舉例而言，2-D 數位矩陣之第一列含有進行甲基化分析之 CpG 基因座之胞嘧啶周圍的相對位置，亦即該胞嘧啶位點上游及下游 3-nt。位置 0 表示待測定甲基化之胞嘧啶位點。相對位置-1 及-2 指示所討論之胞嘧啶上游的 1-nt 及 2-nt。相對位置+1 及+2 指示將使用之胞嘧啶下游的 1-nt 及 2-nt。每個位置將對應於 2 行，其含有相應的

IPD 及 PW 值。每一列對應於瓦生股及克立克股中 4 種類型之核苷酸（A、C、G 及 T）。矩陣中 IPD 及 PW 值之填充取決於在定序結果（亦即子讀段）中在特定位置預設哪種對應的核苷酸類型。

【0232】 如圖 15 所示，在相對位置 0，IPD 及 PW 值顯示在瓦生股之『C』列中，表明在該位置判讀胞嘧啶。行中不對應於定序鹼基之其他網格將編碼為『0』。舉例而言，對應於 2-D 數位矩陣（圖 15）之序列資訊對於瓦生股及克立克股分別為 5'-ATACGTT-3'及 5'-TAACGTA-3'。在此情形下，瓦生股及克立克股中 CpG 位點之胞嘧啶側翼的上游及下游序列將為不同的。由於 CpG 位點處之甲基化在瓦生股及克立克股之間為對稱的（Lister 等人, 2009），因此在一個較佳實施例中，將兩股中之動力學用於訓練甲基化預測模型。在另一個實施例中，瓦生股及克立克股可分別用於訓練甲基化預測模型。

【0233】 考慮到單分子即時定序之高資料吞吐量，在一個實施例中，深度學習算法（例如卷積神經網路（CNN））（LeCun 等人, 1989）可能適合於區分甲基化 CpG 與未甲基化 CpG。另外或替代地，亦可使用其他算法，例如但不限於線性回歸、邏輯回歸、深度循環神經網路（例如長短期記憶，LSTM）、貝葉斯分類器、隱式馬爾可夫模型（HMM）、線性判別分析（LDA）、k 均值聚類、具有雜訊之基於密度之空間聚類應用（DBSCAN）、隨機森林算法及支持向量機（SVM）等。訓練可單獨或在組合的新矩陣中使用瓦生股及克立克股，如圖 6-8 中所述。

【0234】 動力學模式之另一種轉換可為 N 維矩陣。N 可為例如 1、3、4、5、6 及 7。舉例而言，3-D 矩陣將為根據所分析之 DNA 段之串聯 CG 位點的數目分層的 2-D 矩陣之堆疊，其中第 3 維度將為該 DNA 段之串聯 CG 位點的數目。在一些實施例中，脈衝強度或脈衝幅度（例如藉由脈衝之峰值高度或脈衝信號

下的面積來量測）亦可併入矩陣中。可將脈衝強度（脈衝峰值幅度之度量，圖3）添加至在原 2-D 矩陣之基礎上與 PW 和 IPD 值相關聯之行相鄰的額外行，或添加至第 3 維度以形成 3-D 矩陣。

【0235】 作為其他實例，可將 8（列）×21（行）之 2D 矩陣轉換為包含 168 個元素之 1-D 矩陣（亦即向量）。吾等可掃描此 1-D 矩陣，例如以進行 CNN 或其他模型化。作為另一個實例，方法可將 8×21 2-D 矩陣拆分為多個較小矩陣，例如兩個 4×21 2-D 矩陣。將此兩個較小的矩陣以垂直方向放在一起，得到一個 3-D 矩陣（亦即 x=21，y=4，z=2）。方法可掃描第 1 個 2-D 矩陣，且隨後掃描第 2 個 2-D 矩陣，以形成用於機器學習之資料呈現。可進一步拆分資料以形成更高維度的矩陣。另外，可將二級結構資訊添加至資料結構中，例如在 2-D 矩陣基礎上之額外矩陣（1-D 矩陣）。此類額外矩陣可編碼量測窗口內之每個鹼基是否參與二級結構（例如莖-環結構），例如將參與「莖」之鹼基編碼為 0 且將參與「環」之鹼基編碼為 1。

【0236】 在一個實施例中，單個 DNA 分子內 CpG 位點之甲基化狀態可基於統計模型表示為甲基化之概率，而非給出「甲基化」或「未甲基化」之定性結果。概率為 1 表示，基於統計模型，CpG 位點可視為甲基化的。概率為 0 表示，基於統計模型，CpG 位點可視為未甲基化的。在後續的下游分析中，可基於概率用閾值對特定 CpG 位點分類為「甲基化」抑或「未甲基化」進行分類。閾值之可能值包括 5%、10%、15%、20%、25%、30%、35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%、90%或 95%。CpG 位點甲基化之預測概率大於預定義之閾值可分類為「甲基化」，而 CpG 位點甲基化之概率不大於預定義之閾值可分類為「未甲基化」。所需閾值將使用例如接收者操作特徵（ROC）曲線分析自訓練資料集獲得。

【0237】 圖 16 展示對來自瓦生股及克立克股之未知樣本的甲基化狀態進行分類的一般程序。對具有未知甲基化狀態之樣本進行單分子即時定序。定序子讀段可與參考基因體進行排比或與其他方法一樣相互進行排比，以確定給定位置之一致性值（例如平均值、中位數、眾數或其他統計值）。如圖所示，兩股之量測值可組合成單個 2D 矩陣。

【0238】 對於排比結果中 CG 位點之胞嘧啶，吾人將使用與訓練步驟（圖 16）中所應用相當的量測窗口（CpG 位點之胞嘧啶上游及下游的 3-nt）獲得瓦生股之 IPD、PW 及序列上下文，且與修飾受詢問之鹼基相關聯，儘管可使用不同大小的窗口。測試樣本之此類 2-D 矩陣可與圖 16 中所示之參考動力學模式進行比較，以確定甲基化狀態。

用於甲基化檢測之例示性模型訓練

【0239】 為了測試所提出方法的可行性及有效性，吾等藉由 M.SssI 處理（甲基化文庫）及 PCR 擴增（未甲基化文庫）製備胎盤 DNA 文庫，隨後進行單分子即時定序。吾等獲得甲基化及未甲基化文庫之 44,799,736 及 43,580,452 個子讀段，分別對應於 421,614 及 446,285 個環形一致序列（CCS）。因此，每個分子在甲基化及未甲基化文庫中定序之中位數為 34 及 32 次。資料集係由 Pacific Biosciences Sequel Sequencing Kit 3.0 製備之 DNA 產生。此套組係為使用原始的 Pacific Biosciences Sequel 定序儀而開發使用的。為了區分 Sequel 與其後續的 Sequel II，吾等在本文中將原始 Sequel 稱為 Sequel I。因此 Sequel Sequencing Kit 3.0 在本文中將稱為 Sequel I Sequencing Kit 3.0。為 Sequel II 定序儀設計之定序套組包括 Sequel II Sequencing Kit 1.0 及 Sequel II Sequencing Kit 2.0，其亦描述於本揭示案中。

【0240】 吾等使用自甲基化及未甲基化文庫生成之 50%之定序分子來訓

練統計模型（且使用剩餘的50%進行驗證），在此情況下，該模型為卷積神經網路（CNN）模型。舉例而言，CNN 模型可具有一或多個卷積層（例如 1D 或 2D 層）。卷積層可使用一或多個不同的濾波器，每個濾波器使用內核，該內核對特定矩陣元素局部（例如在鄰近或周圍）之矩陣值進行運算，從而為特定矩陣元素提供新的值。一種實現方式使用兩個 1D 卷積層（每個層具有 100 個內核大小為 4 的濾波器）。濾波器可單獨應用，且隨後組合（例如在加權平均中）。所得矩陣可小於輸入矩陣。

【0241】 卷積層之後可為 ReLU（整流線性單元）層，其後可為丟棄率為 0.5 之丟棄層。ReLU 為激活函數之實例，其可對各個值進行運算，從而自卷積層產生新的矩陣（影像）。亦可使用其他激活函數（例如 sigmoid、softmax 等）。可使用此類層中之一或多者。丟棄層可在 ReLU 層上或在最大池化層上使用，且充當防止過度擬合之正則化。丟棄層可在訓練過程中使用，以在作為訓練之一部分執行的最佳化程序（例如減少成本/損失函數）之不同迭代期間忽略不同（例如隨機）的值。

【0242】 在 ReLU 層之後可使用最大池化層（例如，池大小為 2）。最大池化層之作用可類似於卷積層，但不是取輸入與內核之間的點積，而是取輸入與內核重疊的區域的最大值。可使用其他卷積層。舉例而言，來自池化層之資料可輸入至另兩個 1D 卷積層（例如，每個卷積層包含 128 個內核大小為 2 的濾波器，隨後為 ReLU 層），進一步使用丟棄率為 0.5 之丟棄層。使用池大小為 2 的最大池化層。最後，可使用全連接層（例如，具有 10 個神經元，隨後為 ReLU 層）。具有一個神經元之輸出層之後可為 sigmoid 層，從而產生甲基化之概率。可調整層、濾波器及內核大小之各種設置。在此訓練資料集中，吾等使用來自甲基化及未甲基化文庫之 468,596 及 432,761 個 CpG 位點。

訓練及測試資料集之結果

【0243】 **圖 17A** 展示訓練資料集中每個單個 DNA 分子中每個 CpG 位點甲基化之概率甲基化文庫中甲基化之概率遠高於未甲基化文庫。對於甲基化概率之閾值為 0.5，正確預測 94.7%之未甲基化 CpG 位點為未甲基化的，且正確預測 84.7%之甲基化 CpG 為甲基化的。

【0244】 **圖 17B** 展示測試資料集之效能。吾等使用由訓練資料集訓練之模型來預測來自甲基化及未甲基化文庫之獨立測試資料集中 469,729 及 432,024 個 CpG 位點之甲基化狀態。對於甲基化概率之閾值為 0.5，正確預測 94.0%之未甲基化 CpG 位點為未甲基化的，且正確預測 84.1%之甲基化 CpG 為甲基化的。此等結果表明，使用新穎的動力學轉換結合序列上下文可實現 DNA（例如來自人類個體）中甲基化狀態之測定。

【0245】 吾等藉由在模型中包括特徵之子集，評估每個特徵（序列上下文、IPD 及 PW）在預測 CpG 甲基化狀態方面的能力。在訓練資料集中，具有（i）僅序列上下文、（ii）僅 IPD 及（iii）僅 PW 之模型分別給出 0.5、0.74 及 0.86 之曲線下面積（AUC）值。同時結合 IPD 及序列上下文提高效能，AUC 為 0.86。對序列上下文（「Seq」）、IPD 及 PW 之組合分析顯著提高效能，AUC 為 0.94（**圖 18A**）。獨立測試資料集之效能與訓練資料集相當（**圖 18B**）。

【0246】 吾等將 CpG 位點之子讀段深度定義為覆蓋其及其周圍 10 bp 之子讀段的平均數。如**圖 19A** 及**圖 19B** 所示，CpG 位點之子讀段深度愈高，吾等實現之甲基化檢測的準確性愈高。舉例而言，如測試資料集（**圖 19B**）中所示，若每個 CpG 位點之深度為至少 10，則預測甲基化狀態之 AUC 將為 0.93。然而，若每個 CpG 位點之子讀段深度為至少 300，則預測甲基化狀態之 AUC 將為 0.98。另一方面，即使深度為 1，吾等仍可達到 0.9 之 AUC，表明吾等方法可

在使用低定序深度之情況下實現甲基化預測。

【0247】 為了測試股資訊對甲基化分析效能之影響，根據本揭示案中存在之實施例，分別使用源自瓦生股及克立克股之序列上下文、IPD 及 PW 進行訓練。**圖 20A** 及**圖 20B** 顯示，使用單一股，亦即瓦生或克立克股進行訓練及測試為可行的，因為在訓練及測試資料集中 AUC 可達到高達 0.91 及 0.87。使用包括瓦生股及克立克股之兩股（例如，如圖 6-8 中所述）將產生最佳效能（AUC：在訓練及測試資料集中分別為 0.94 及 0.90），表明股資訊將對實現最佳效能至關重要。

【0248】 吾等進一步測試 CpG 位點上游及下游核苷酸之不同數目，以研究此參數如何影響根據本揭示案中開發之本揭示案中存在之實施例的效能。**圖 21A** 及**圖 21B** 顯示，在 CpG 之情形下，胞嘧啶上游及下游之核苷酸數目會影響甲基化預測之準確性。舉例而言，出於說明目的，考慮但不限於所分析之胞嘧啶上游及下游的 2 個核苷酸（nt）、3 nt、4 nt、6 nt、8 nt、10 nt、15 nt 及 20 nt，在訓練及測試資料集中使用所詢問之胞嘧啶上游及下游 2 nt 之方法的 AUC 將僅為 0.50，而在訓練及測試資料集中使用所詢問之胞嘧啶上游及下游 15 nt 之方法的 AUC 將增加至 0.95 及 0.92。此等結果表明，改變所分析之胞嘧啶側翼之上游及下游區域的長度將允許找出最佳效能。在一個實施例中，如圖 21B 所示，吾人將使用胞嘧啶上游及下游之 3 nt 來確定甲基化狀態，其可達到 0.89 之 AUC。

【0249】 在一個實施例中，吾人可使用所詢問之胞嘧啶側翼的不對稱序列來根據本揭示案中存在之實施例進行分析。舉例而言，可使用胞嘧啶上游 2 nt 與下游 1 nt、3 nt、4 nt、5 nt、6 nt、7 nt、8 nt、9 nt、10 nt、11 nt、12 nt、13 nt、14 nt、15 nt、16 nt、17 nt、18 nt、19 nt、20 nt、25 nt、30 nt、35 nt 及

40 nt 之組合；可使用胞嘧啶上游 3 nt 與下游 1 nt、2 nt、4 nt、5 nt、6 nt、7 nt、8 nt、9 nt、10 nt、11 nt、12 nt、13 nt、14 nt、15 nt、16 nt、17 nt、18 nt、19 nt、20 nt、25 nt、30 nt、35 nt 及 40 nt 之組合；可使用胞嘧啶上游 4 nt 與 1 nt、2 nt、3 nt、5 nt、6 nt、7 nt、8 nt、9 nt、10 nt、11 nt、12 nt、13 nt、14 nt、15 nt、16 nt、17 nt、18 nt、19 nt、20 nt、25 nt、30 nt、35 nt 及 40 nt 之組合。作為另一個實例，可使用胞嘧啶下游 2 nt 與上游 1 nt、3 nt、4 nt、5 nt、6 nt、7 nt、8 nt、9 nt、10 nt、11 nt、12 nt、13 nt、14 nt、15 nt、16 nt、17 nt、18 nt、19 nt、20 nt、25 nt、30 nt、35 nt 及 40 nt 之組合；可使用胞嘧啶下游 3 nt 與上游 1 nt、2 nt、4 nt、5 nt、6 nt、7 nt、8 nt、9 nt、10 nt、11 nt、12 nt、13 nt、14 nt、15 nt、16 nt、17 nt、18 nt、19 nt、20 nt、25 nt、30 nt、35 nt 及 40 nt 之組合；可使用胞嘧啶下游 4 nt 與上游 1 nt、2 nt、3 nt、5 nt、6 nt、7 nt、8 nt、9 nt、10 nt、11 nt、12 nt、13 nt、14 nt、15 nt、16 nt、17 nt、18 nt、19 nt、20 nt、25 nt、30 nt、35 nt 及 40 nt 之組合。藉由利用與胞嘧啶上游 n-nt 及下游 m-nt 相關聯之 IPD、PW、股資訊及序列上下文，可在某些實施例中提供提高的確定甲基化狀態之準確性。此類不同的量測窗口可應用於其他類型的鹼基修飾分析，諸如 5hmC、6mA、4mC 及 oxoG，或本文所揭示之任何修飾。此類不同的量測窗口可包括 DNA 二級結構分析，諸如 G-四聯體及莖環結構。此類實例闡述於上文。此類二級結構資訊亦可作為矩陣中之另一行添加。

【0250】 **圖 22A** 及**圖 22B** 顯示，使用僅與至少 3 個鹼基之下游鹼基相關聯的動力學模式來確定甲基化狀態為可行的。根據本揭示案中存在之實施例，在使用與胞嘧啶及其下游 3、4、6、8 及 10 個鹼基相關聯之特徵的情況下，在訓練資料集中確定甲基化狀態之 AUC 分別為 0.91、0.92、0.94、0.94 及 0.94；在測試資料集中 AUC 分別為 0.87、0.88、0.90、0.90 及 0.90。

【0251】 然而，**圖 23A** 及**圖 23B** 顯示，若吾人僅使用與上游鹼基相關聯之特徵，則分類能力似乎會降低其區分甲基化狀態之能力。在訓練資料集及測試資料集中，2 至 10 個上游鹼基之 AUC 均為 0.50。

【0252】 **圖 24** 及**圖 25** 顯示，上游及下游鹼基之不同組合將使吾人在確定甲基化狀態時達到最佳分類能力。舉例而言，與胞嘧啶上游 8 個鹼基及下游 8 個鹼基相關聯之特徵將在此資料集中實現最佳效能，在訓練及測試資料集中 AUC 分別為 0.94 及 0.91。

【0253】 **圖 26** 展示特徵在 CpG 位點處之甲基化狀態分類方面的相對重要性。括號中之『W』及『C』表示股資訊，『W』表示瓦生股且『C』表示克立克股。使用隨機森林確定包括序列上下文、IPD 及 PW 之每個特徵的重要性。隨機森林樹分析顯示，IPD 及 PW 之特徵重要性在受詢問之胞嘧啶的下游達到峰值，表明對分類能力之主要貢獻為受詢問之胞嘧啶下游的 IPD 及 PW。

【0254】 隨機森林由多個決策樹構成。在決策樹之構築過程中，使用基尼不純度（Gini impurity）來確定決策節點應採用哪種決策邏輯。對最終分類結果影響較大之重要特徵很可能出現在距離決策樹根較近的節點中，而對最終分類結果影響較小之不重要特徵很可能出現在距根較遠的節點中。因此，可藉由計算相對於隨機森林中所有決策樹之根的平均距離來估計特徵重要性。

【0255】 在一些實施例中，可進一步使用瓦生股及克立克股之間 CpG 位點處甲基化判讀之一致性來提高特異性。舉例而言，可要求將顯示甲基化之兩股稱為甲基化狀態，且將顯示未甲基化之兩股稱為未甲基化狀態。由於已知 CpG 位點處之甲基化通常為對稱的，因此自各股進行確認可提高特異性。

【0256】 在各種實施例中，來自全分子之整體動力學特徵可用於確定甲基化狀態。舉例而言，在單分子即時定序期間，全分子中之甲基化將影響全分

子之動力學。藉由將整個模板 DNA 分子之定序動力學（包括 IPD、PW、片段大小、股資訊及序列上下文）模型化，可提高關於分子是否甲基化之分類準確性。舉例而言，量測窗口可為整個模板分子。IPD、PW 或其他動力學特徵之統計值（例如，平均值、中位數、眾數、百分位數等）可用於確定全分子之甲基化。

其他分析技術之侷限性

【0257】 據報導，基於 IPD 對特定序列基元中特定 C 之甲基化檢測非常低，例如靈敏度僅為 1.9%（Clark 等人, 2013）。吾等亦試圖藉由將不同的序列基元與 IPD 組合來重新此類分析，而不使用 PW 度量，且僅使用 IPD 之閾值而非如本文所述之資料結構。舉例而言，提取所詢問之 CpG 側翼的上游及下游 3-nt。該 CpG 之 IPD 根據以該 CpG 為中心之 6-nt 側翼序列（亦即上游及下游分別 3 nt）的上下文而分層為不同的組（6 個位置，4096 組）。使用 ROC 研究同一序列基元內甲基化及未甲基化 CpG 之間的 IPD。舉例而言，比較未甲基化之「AATCGGAC」基元及甲基化之「AAT^{m}CGGAC」基元中 CpG 之 IPD，顯示 AUC 為 0.48。因此，使用特定序列組中之閾值相對於使用不同的實施例而言表現不佳。

【0258】 **圖 27** 展示上述基於基元之 IPD 分析（Beckmann 等人《BMC 生物資訊學》2014）在不使用脈衝寬度信號之情況下進行甲基化檢測的效能。垂直條形圖表示所研究之 CpG 位點側翼的不同 k 聚體基元的平均 AUC（亦即所詢問之 CpG 位點周圍的鹼基數）。圖 27 顯示，在不同的 k 聚體基元（例如所討論之 CpG 位點周圍的 2 聚體、3 聚體、4 聚體、6 聚體、8 聚體、10 聚體、15 聚體、20 聚體）中，甲基化及未甲基化胞嘧啶之間基於 IPD 之鑑別力的平均 AUC 被發現小於 60%。此等結果表明，在給定的基元上下文中考慮候選核苷酸之 IPD

而不考慮鄰近核苷酸之 IPD（Flusberg 等人, 2010）將不如本文揭示之用於測定 CpG 甲基化之方法。

【0259】 吾等亦測試 Flusberg 等人研究（Flusberg 等人, 2010）中存在之方法。吾等分析總共 5,948,348 個 DNA 區段，其為進行甲基化分析之胞嘧啶上游 2-nt 及下游 6-nt。存在 2,828,848 個甲基化區段及 3,119,500 個未甲基化區段。如**圖 28** 所示，發現使用 IPD 及 PW 自主成分分析中推導出之信號在具有甲基化胞嘧啶（mC）及未甲基化胞嘧啶（C）之片段之間基本上重疊，表明 Flusberg 等人所述之方法缺乏實際有意義的準確性。此等結果表明，如 Flusberg 等人之研究（Flusberg 等人, 2010）中所用之主成分分析將鹼基及鄰近鹼基處之 PW 及 IPD 值線性組合，無法可靠或有意義地區分 5-甲基胞嘧啶及未甲基化胞嘧啶。

【0260】 **圖 29** 顯示，在 Flusberg 等人之研究（Flusberg 等人, 2010）中使用涉及 IPD 及 PW 之兩個主成分的基於主成分分析之方法的 AUC（AUC：0.55）準確度遠低於吾等揭示內容中所示的涉及 IPD 及 PW 以及序列上下文之基於卷積神經網路之方法（AUC：0.94）。

其他數學/統計模型

【0261】 在另一個實施例中，其他數學/統計模型，例如包括但不限於隨機森林及邏輯回歸，可藉由適應上述開發之特徵來訓練。至於 CNN 模型，訓練及測試資料集係由經 M.SssI 處理（甲基化）及 PCR 擴增（未甲基化）之 DNA 構築，其用於訓練隨機森林（Breiman, 2001）。在此隨機森林分析中，吾等用 6 個特徵描述每個核苷酸：IPD、PW 及編碼鹼基標識之 4 組分二進制向量。在此類二進制向量中，A、C、G 及 T 分別用[1,0,0,0]、[0,1,0,0]、[0,0,1,0]及[0,0,0,1]編碼。對於每個所分析之 CpG 位點，吾等將其上游及下游 10 nt 之資訊併入兩股

中，形成 252 維（252 -D）向量，每個特徵代表一個維度。上述具有 252-D 向量之訓練資料集用於訓練隨機森林模型以及邏輯回歸模型。經訓練之模型用於預測獨立測試資料集中之甲基化狀態。隨機森林由 100 個決策樹構成。在樹的構築過程中，使用自助樣本。在拆分每個決策樹之節點時，採用基尼不純度來確定最佳拆分，且在每個拆分中最多考慮 15 個特徵。另外，要求決策樹之每片葉子含有至少 60 個樣本。

【0262】 **圖 30A** 及**圖 30B** 展示使用隨機森林及邏輯回歸進行甲基化預測之方法的效能圖 30A 展示 CNN、隨機森林及邏輯回歸之訓練資料集中的 AUC 值。圖 30B 展示 CNN、隨機森林及邏輯回歸之測試資料集中的 AUC 值。使用隨機森林之方法在訓練及測試資料集中的 AUC 分別達到 0.93 及 0.86。

【0263】 用相同的 252-D 向量描述之訓練資料集用於訓練邏輯回歸模型。經訓練之模型用於預測獨立測試資料集中之甲基化狀態。將具有 L2 正則化之邏輯回歸模型（Ng 及 Y., 2004）與訓練資料集擬合。如圖 30A 及圖 30B 所示，使用邏輯回歸之方法在訓練及測試資料集中的 AUC 將分別達到 0.87 及 0.83。

【0264】 因此，此等結果表明，使用吾等在本揭示案中開發之特徵及分析方案，除 CNN 以外之某些模型（例如但不限於隨機森林及邏輯回歸）可用於甲基化分析。此等結果亦表明，根據本揭示案中之實施例實施的 CNN 在測試資料集（圖 30B）中之 AUC 為 0.90，優於隨機森林（AUC：0.86）及邏輯回歸（AUC：0.83）。

測定核酸之 6mA 修飾

【0265】 除甲基化之 CpG 之外，本文所述之方法亦可檢測其他 DNA 鹼基修飾。舉例而言，可檢測甲基化腺嘌呤，包括呈 6mA 之形式。

使用動力學特徵及定序上下文進行 6mA 檢測

【0266】 為了評估所揭示之用於測定核酸鹼基修飾之實施例的效能及實用性，吾等進一步分析 N6-腺嘌呤甲基化（6mA）。在一個實施例中，經由全基因體擴增，用未甲基化腺嘌呤（uA）、未甲基化胞嘧啶（C）、未甲基化鳥嘌呤（G）及未甲基化胸腺嘧啶（T）擴增大約 1 ng 人類 DNA（例如自胎盤組織提取），以獲得 100 ng DNA 產物。

【0267】 **圖 31A** 展示藉由全基因體擴增生成具有未甲基化腺嘌呤之分子之一種方法的實例。在圖中，「uA」表示未甲基化腺嘌呤，「mA」表示甲基化腺嘌呤。使用抗外切核酸酶之經硫代磷酸酯修飾之無規六聚體作為引子進行全基因體擴增，該等引子在基因體上隨機結合，從而允許聚合酶（例如 Φ29 DNA 聚合酶）擴增 DNA（例如，藉由等溫線性擴增）。在階段 3102，使雙股 DNA 變性。在階段 3106，當許多無規六聚體（例如 3110）與變性的模板 DNA（亦即單股 DNA）黏接時，引發擴增反應。如 3114 所示，當股 3118 之六聚體介導之 DNA 合成沿 5'至 3'方向進行且到達下一個六聚體介導之 DNA 合成位點時，聚合酶置換新合成之 DNA 股（3122）且繼續股延伸。經置換之股成為單股 DNA 模板，供無規六聚體再次結合，且可能啟動新的 DNA 合成。在等溫過程中重複之六聚體黏接及股置換將導致擴增之 DNA 產物的高產率。本文所述之此擴增可能屬於多重置換擴增（MDA）技術。

【0268】 將擴增的 DNA 產物進一步片段化為例如但不限於大小為 100 bp、200 bp、300 bp、400 bp、500 bp、600 bp、700 bp、800 bp、900 bp、1 kb、5 kb、10 kb、20 kb、30 kb、40 kb、50 kb、60 kb、70 kb、80 kb、90 kb、100 kb 或其他所需大小範圍之片段。片段化方法可包括酶解、霧化、流體動力剪切及音波處理等。因此，原始鹼基修飾諸如 6mA 可藉由用未甲基化之 A（uA）進行全基因體擴增而幾乎消除。圖 31A 展示 DNA 產物之可能片段

（3126、3130 及 3134），其中兩股均具有未甲基化之 A。對此類全基因體擴增之無 mA 的 DNA 產物進行單分子即時定序，以產生 uA 資料集。

【0269】 **圖 31B** 展示藉由全基因體擴增生成具有甲基化腺嘌呤之分子之一種方法的實例。在圖中，「uA」表示未甲基化腺嘌呤，「mA」表示甲基化腺嘌呤。經由全基因體擴增，用 6mA 及未甲基化之 C、G 及 T 擴增大約 1 ng 人類 DNA，以獲得 10 ng DNA 產物。甲基化腺嘌呤可經由一系列化學反應產生（J D Engel 等人《生物化學雜誌（J Biol Chem.）》 1978;253:927-34）。如圖 31B 所示，使用抗外切核酸酶之經硫代磷酸酯修飾之無規六聚體作為引子進行全基因體擴增，該等引子在基因體上隨機結合，從而允許聚合酶（例如 Φ29 DNA 聚合酶）擴增 DNA（例如藉由等溫線性擴增），類似於圖 31A。抗外切核酸酶之經硫代磷酸酯修飾之無規六聚體對校對 DNA 聚合酶之 3'→5'外切核酸酶活性具有抗性。因此，在擴增期間，將保護無規六聚體免於降解。

【0270】 當許多無規六聚體與變性的模板 DNA（亦即單股 DNA）黏接時，引發擴增反應。當六聚體介導之 DNA 合成沿 5'至 3'方向進行且到達下一個六聚體介導之 DNA 合成位點時，聚合酶置換新合成之 DNA 股且繼續股延伸。經置換之股成為單股 DNA 模板，供無規六聚體再次結合，且啟動新的 DNA 合成。在等溫過程中重複之六聚體黏接及股置換將導致擴增之 DNA 產物的高產率。

【0271】 將擴增的 DNA 產物進一步片段化為例如但不限於大小為 100 bp、200 bp、300 bp、400 bp、500 bp、600 bp、700 bp、800 bp、900 bp、1 kb、5 kb、10 kb、20 kb、30 kb、40 kb、50 kb、60 kb、70 kb、80 kb、90 kb、100 kb 或其他長度組合之片段。如圖 31B 所示，擴增的 DNA 產物將在每股之腺嘌呤位點上包括不同形式之甲基化模式。舉例而言，雙股分子之兩股可相對於

腺嘌呤甲基化（分子 I），當兩股在全基因體擴增期間自 DNA 合成衍生而來時，將產生該分子。

【0272】 作為另一個實例，雙股分子之一股可含有遍及腺嘌呤位點之交錯甲基化模式（分子 II）。交錯甲基化模式定義為包括 DNA 股中存在之甲基化及未甲基化鹼基之混合物的模式。在以下實例中，吾等使用交錯的腺嘌呤甲基化模式，該模式包括 DNA 股中存在之甲基化及未甲基化腺嘌呤之混合物。此類型之雙股分子（分子 II）將可能生成，因為含有未甲基化腺嘌呤之未甲基化六聚體與 DNA 股結合且啟動 DNA 延伸。將對此類擴增的 DNA 產物進行定序，該產物含有具未甲基化腺嘌呤之六聚體。或者，此類型之雙股分子（分子 II）將由來自含有未甲基化腺嘌呤之原始模板 DNA 的片段化 DNA 啟動，因為此類片段化 DNA 可作為引子結合至 DNA 股。將對此類擴增的 DNA 產物進行定序，該產物含有股中具有未甲基化腺嘌呤之原始 DNA 的一部分。由於未甲基化之六聚體引子僅為所得 DNA 股之一小部分，因此大多數片段仍將含有 6mA。

【0273】 作為另一個實例，雙股 DNA 分子之一股可在腺嘌呤位點上甲基化，但另一股可為未甲基化的（分子 III）。當提供無甲基化腺嘌呤之原始 DNA 股作為模板 DNA 分子以產生具有甲基化腺嘌呤之新股時，可生成此類型之雙股分子。

【0274】 兩股可為未甲基化的（分子 IV）。此類型之雙股分子可由於無甲基化腺嘌呤之兩個原始 DNA 股重新黏接而產生。

【0275】 片段化方法可包括酶解、霧化、流體動力剪切及音波處理等。此類全基因體擴增之 DNA 產物可主要以 A 位點甲基化。對此具有 mA 之 DNA 進行單分子即時定序，以生成 mA 資料集。

【0276】 對於 uA 資料集，吾等使用單分子即時定序對 262,608 個長度中

位數為 964 bp 之分子進行定序。中位子讀段深度為 103 x。在子讀段中，48%可使用 BWA 排比器與人類參考基因體進行排比（Li H 等人《生物資訊學》2009;25:1754-60）。舉例而言，吾人可採用 Sequel II System（Pacific Biosciences）進行單分子即時定序。使用 SMRTbell Express Template Prep Kit 2.0（Pacific Biosciences）對片段化之 DNA 分子進行單分子即時（SMRT）定序模板構築。用 SMRT Link v8.0 軟體（Pacific Biosciences）計算定序引子黏接及聚合酶結合條件。簡言之，使定序引子 v2 與定序模板黏接，且隨後使用 Sequel II Binding and Internal Control Kit 2.0（Pacific Biosciences）使聚合酶與模板結合。在 Sequel II SMRT Cell 8M 上進行定序。用 Sequel II Sequencing Kit 2.0（Pacific Biosciences）在 Sequel II 系統上收集定序影片 30 小時。

【0277】 對於 mA 資料集，吾等使用單分子即時定序對 804,469 個長度中位數為 826 bp 之分子進行定序。中位子讀段深度為 34 x。在子讀段中，27%可使用 BWA 排比器與人類參考基因體進行排比（Li H 等人《生物資訊學》2009;25:1754-60）。

【0278】 在一個實施例中，以股特異性方式分析包括但不限於 IPD 及 PW 之動力學特徵。對於源自瓦生股之定序結果，使用自 uA 資料集隨機選取之 644,318 個無甲基化之 A 位點及自 mA 資料集隨機選取之 718,586 個有甲基化之 A 位點構成訓練資料集。此類訓練資料集用於建立分類模型及/或臨限值，以區分甲基化及未甲基化腺嘌呤。測試資料集由 639,702 個無甲基化之 A 位點及 723,320 個有甲基化之 A 位點構成。此類測試資料集用於驗證自訓練資料集推導出之模型/臨限值的效能。

【0279】 吾等分析源自瓦生股之定序結果。**圖 32A** 展示 uA 及 mA 資料集之訓練資料集的脈衝間持續時間（IPD）值。對於訓練資料集，在 mA 資料集中

觀察到的經定序 A 位點之 IPD 值（中位數：1.09；範圍：0-9.52）高於 uA 資料集（中位數：0.20；範圍：0-9.52）（*P* 值＜0.0001；Mann Whitney U 檢驗）。

【0280】 **圖 32B** 展示 uA 及 mA 資料集之測試資料集的 IPD。當吾等研究測試資料集中經定序 A 位點之 IPD 值時，吾等觀察到 mA 資料集中之 IPD 值高於 uA 資料集（中位數 1.10 對 0.19；*P* 值＜0.0001；Mann Whitney U 檢驗）。

【0281】 **圖 32C** 展示使用 IPD 閾值之接收者操作特徵（ROC）曲線下面積。真陽性率在 y 軸上，假陽性率在 x 軸上。使用相應的 IPD 值區分模板 DNA 分子中具有及不具有甲基化之經定序 A 鹼基時，訓練及測試資料集兩者之接收者操作特徵曲線下面積（AUC）為 0.86。

【0282】 除了來自瓦生股之結果外，吾等分析源自克立克股之定序結果。**圖 33A** 展示 uA 及 mA 資料集之訓練資料集的 IPD 值。對於訓練資料集，在 mA 資料集中觀察到的經定序 A 位點之 IPD 值（中位數：1.10；範圍：0-9.52）高於 uA 資料集（中位數：0.19；範圍：0-9.52）（*P* 值＜0.0001；Mann Whitney U 檢驗）。

【0283】 **圖 34B** 展示 uA 及 mA 資料集之測試資料集的 IPD 值。與 uA 資料集相比，在測試資料集之 mA 資料集中亦觀察到經定序 A 位點之 IPD 值較高（中位數 1.10 對 0.19；*P* 值＜0.0001；Mann Whitney U 檢驗）。

【0284】 **圖 33C** 展示 ROC 曲線下面積。真陽性率在 y 軸上，假陽性率在 x 軸上。使用相應的 IPD 值區分模板 DNA 分子中具有及不具有甲基化之經定序 A 鹼基時，訓練及測試資料集之 ROC 曲線下面積（AUC）值分別為 0.86 及 0.87。

【0285】 **圖 34** 展示根據本發明之實施例，使用量測窗口對瓦生股進行 6mA 測定的圖示。此類量測窗口可包括動力學特徵，諸如 IPD 及 PW 及附近的序列上下文。6mA 之測定可與甲基化 CpG 之測定類似地進行。

【0286】 **圖 35** 展示根據本發明之實施例，使用量測窗口對克立克股進行 6mA 測定的圖示。此類量測窗口可包括動力學特徵，諸如 IPD 及 PW 及附近的序列上下文。

【0287】 舉例而言，將自所詢問之模板 DNA 中經定序 A 鹼基每一側的 10 個鹼基用於構築量測窗口。將包括 IPD、PW 及序列上下文之特徵值用於根據本文所揭示之方法使用卷積神經網路（CNN）訓練模型。在其他實施例中，統計模型可包括但不限於線性回歸、邏輯回歸、深度循環神經網路（例如長短期記憶，LSTM）、貝葉斯分類器、隱式馬爾可夫模型（HMM）、線性判別分析（LDA）、k 均值聚類、具有雜訊之基於密度之空間聚類應用（DBSCAN）、隨機森林算法及支持向量機（SVM）等。

【0288】 **圖 36A** 及**圖 36B** 展示使用基於量測窗口之 CNN 模型所確定的 uA 及 mA 資料集之間瓦生股之經定序 A 鹼基的甲基化概率。圖 36A 展示自訓練資料集習得的 CNN 模型。舉例而言，CNN 模型使用兩個 1D 卷積層（每個卷積層具有 64 個內核大小為 4 的濾波器，隨後為 ReLU（整流線性單元）層），隨後為丟棄率為 0.5 之丟棄層。使用池大小為 2 的最大池化層。隨後流入兩個 1D 卷積層（每個卷積層具有 128 個內核大小為 2 的濾波器，隨後為 ReLU 層），進一步使用丟棄率為 0.5 之丟棄層。使用池大小為 2 的最大池化層。最後，具有 10 個神經元之全連接層，隨後為 ReLU 層，具有一個神經元之輸出層，隨後為 sigmoid 層，從而得出甲基化概率。層、濾波器、內核大小之其他設置可進行調整，例如，如本文針對其他甲基化（例如 CpG）所述。在此關於瓦生股定序結果之訓練資料集中，吾等使用來自未甲基化及甲基化文庫之 644,318 及 718,586 個 A 鹼基。

【0289】 基於 CNN 模型，對於瓦生股相關資料，與 uA 資料集中存在之

彼等 A 鹼基相比，mA 資料庫之模板 DNA 分子中經定序之 A 鹼基在訓練及測試資料集中均引起高得多的甲基化概率（*P* 值＜0.0001；Mann Whitney U 檢驗）。對於訓練資料集，uA 資料集中 A 位點甲基化之中位概率為 0.13（四分位數範圍，IQR：0.09-0.15），而 mA 資料集中該值為 1.000（IQR：0.998-1.000）。

【0290】 圖 36A 展示針對測試資料集所確定之甲基化概率。對於測試資料集，uA 資料集中 A 位點甲基化之中位概率為 0.13（IQR：0.10-0.15），而 mA 資料集中該值為 1.000（IQR：0.997-1.000）。圖 36A 及 36B 顯示，可訓練基於量測窗口之 CNN 模型以檢測測試資料集中之甲基化。

【0291】 **圖 37** 為使用基於量測窗口之 CNN 模型對瓦生股之經定序 A 鹼基進行 6mA 檢測的 ROC 曲線。真陽性率在 y 軸上，假陽性率在 x 軸上。該圖顯示，對於由瓦生股定序結果組成之訓練及測試資料集，使用 CNN 模型區分具有及不具有甲基化之經定序 A 位點的 AUC 值分別為 0.94 及 0.93。其表明，使用本文揭示內容利用瓦生股之資料確定 A 位點之甲基化狀態為可行的。若吾等使用確定的甲基化概率 0.5 作為閾值，則 6mA 檢測可達到 99.3%之特異性及 82.6%之靈敏度。圖 37 顯示，可使用基於量測窗口之 CNN 模型以高特異性及靈敏度來檢測 6mA。該模型之準確性可與僅使用 IPD 度量之技術進行比較。

【0292】 **圖 38** 展示基於 IPD 度量之 6mA 檢測與基於量測窗口之 6mA 檢測之間的效能比較。靈敏度標繪在 y 軸上，特異性標繪在 x 軸上。圖 38 顯示，根據本文揭示內容使用基於量測窗口之 6mA 分類的效能（AUC：0.94）優於僅使用 IPD 度量之習知方法（AUC：0.87）（P 值＜0.0001；DeLong 檢驗）。基於量測窗口之 CNN 模型優於基於 IPD 度量之檢測。

【0293】 **圖 39A** 及 **39B** 展示使用基於量測窗口之 CNN 模型確定之在 uA 及 mA 資料集之間的克立克股之彼等經定序 A 鹼基的甲基化概率。圖 39A 展示

訓練資料集，圖 39B 展示測試資料集。兩幅圖均在 y 軸上標繪甲基化概率。圖 39A 及 39B 顯示，基於 CNN 模型，對於克立克股相關資料，mA 資料庫之模板 DNA 分子中之經定序 A 鹼基在訓練及測試資料集中產生的甲基化概率比 uA 資料庫中存在之彼等 A 鹼基高得多（P 值＜0.0001；Mann-Whitney U 檢驗）。

【0294】 **圖 40** 展示使用基於量測窗口之 CNN 模型對克立克股之經定序 A 鹼基進行 6mA 檢測的效能。真陽性率在 y 軸上。假陽性率在 x 軸上。圖 40 顯示，對於由克立克股定序結果組成之訓練及測試資料集，使用 CNN 模型區分具有及不具有甲基化之經定序 A 位點的 AUC 值分別為 0.95 及 0.94。亦證明使用本文所揭示之 CNN 方法的效能（AUC：0.94）優於僅使用 IPD 度量的效能（0.87）（*P* 值＜0.0001）。結果表明，使用本文揭示內容利用瓦生股之資料確定 A 位點之甲基化狀態為可行的。若吾等使用確定的甲基化概率 0.5 作為閾值，則 6mA 檢測可達到 99.3%之特異性及 83.0%之靈敏度。圖 40 顯示，可使用基於量測窗口之 CNN 模型以高特異性及靈敏度來檢測 6mA。

【0295】 **圖 41** 展示包括瓦生股及克立克股之分子中 A 鹼基之甲基化狀態的實例。白點表示未甲基化之腺嘌呤。黑點表示甲基化之腺嘌呤。帶點的水平線表示雙股 DNA 分子之股。分子 1 顯示，瓦生股及克立克股均確定為 A 鹼基未甲基化。分子 2 顯示，瓦生股幾乎全部未甲基化，而克立克股幾乎全部甲基化。分子 3 顯示，瓦生股及克立克股均確定為 A 鹼基幾乎全部甲基化。

使用選擇性資料集進行增強訓練

【0296】 如圖 36A、36B、39A 及 39B 所示，在 mA 資料集中，模板 DNA 分子中經定序之 A 鹼基的甲基化概率存在雙峰分佈。換言之，mA 資料集中存在一些具有 uA 信號之分子。此點藉由 mA 資料集中存在完全未甲基化之分子及半甲基化之分子進一步證明（圖 41）。一個可能的原因可為 DNA 模板中具有 uA

之分子在全基因體擴增後之 mA 資料集中仍會占相當大的比例，因為具有 6mA 之分子會導致全基因體擴增步驟期間擴增 DNA 的效率降低。此解釋得到以下事實的支持：在相同的擴增條件下，用 6mA 擴增之 1 ng 基因體 DNA 僅會產生 10 ng DNA 產物，而用未甲基化之 A 擴增的 1 ng 基因體 DNA 將產生 100 ng DNA 產物。因此，對於 mA 資料集，腺嘌呤通常未甲基化（例如 0.051%）之原始模板 DNA 分子（Xiao CL 等人，《分子細胞學（Mol Cell.）》2018;71:306-318）將占總腺嘌呤之大約 10%。

【0297】 在一個實施例中，當吾人試圖訓練用於區分 mA 與 uA 之 CNN 模型時，吾人將選擇性地使用 mA 資料集中彼等具有相對較高 IPD 值的 A 鹼基，以減少 uA 資料對訓練 mA 檢測模型之影響。僅可使用 IPD 值高於某一閾值的 A 鹼基。閾值可對應於百分位數。在一個實施例中，吾人將使用 mA 資料集中 IPD 值大於第 10 百分位數處之值的彼等 A 鹼基。在一些實施例中，吾人將使用 IPD 值大於第 1、5、15、20、30、40、50、60、70、80、90 或 95 百分位數處之值的彼等 A。百分位數可基於參考樣本或多個參考樣本中所有核酸分子之資料。

【0298】 **圖 42** 展示藉由選擇性地使用 mA 資料集中 IPD 值大於其第 10 百分位數之 A 鹼基進行增強訓練的效能。圖 42 展示 y 軸上之真陽性率及 x 軸上之假陽性率。該圖顯示，藉由使用 mA 資料集中 IPD 值大於第 10 百分位數之 A 鹼基訓練 CNN 模型，區分 mA 鹼基與 uA 鹼基之 AUC 將增加至 0.98，優於訓練前不根據 IPD 值進行選擇之資料訓練的模型（AUC：0.94）。其表明使用 IPD 值選擇 mA 位點來創建訓練資料集將有助於提高鑑別力。

【0299】 為了進一步證實 mA 資料集中存在具有 uA 鹼基之分子，吾等假設 mA 資料集中 uA 之百分比會富集在彼等具有較多子讀段之孔中，因為與無 6mA 之分子相比，分子中存在之 6mA 會減緩聚合酶生成新股時的延伸。

【0300】 **圖 43** 展示 mA 資料集中未甲基化腺嘌呤之百分比與各孔中子讀段數目的圖式。y 軸顯示 mA 資料集中 uA 之百分比。x 軸顯示各孔中之子讀段數目。使用增強的模型重新分析測試資料集，該模型藉由在移除 IPD 值低於第 10 百分位數之 A 位點後使用 mA 位點進行訓練。隨著每孔子讀段數目的增加，包括每個定序孔 1 至 10 個子讀段至每孔 10 至 20 個子讀段、至每孔 40 至 50 個子讀段、每孔 60 至 70 個子讀段及 70 個以上，觀察到 uA 逐漸增加（亦即自 14.6 上升至 55.05%）。因此，具有大量子讀段之孔往往具有低 mA。A 之甲基化可延緩定序反應之進展。因此，具有高子讀段深度之定序孔更可能對 A 未甲基化。可利用此行為，使用與分子相關聯之子讀段數目的閾值檢測未甲基化之分子，例如大於 70 個子讀段可鑑別為多數未甲基化。

【0301】 **圖 44** 顯示測試資料集中雙股 DNA 分子之瓦生股及克立克股之間的甲基腺嘌呤模式。A 之甲基化為不對稱的，且因此兩股之間的行為不同。大多數分子由於併入 mA 而甲基化，仍有一些殘餘的未甲基化之 A。y 軸顯示克立克股之甲基腺嘌呤水準。x 軸顯示瓦生股之甲基腺嘌呤水準。每個點代表一個雙股分子。使用由所選 mA 位點訓練的增強模型，雙股分子可根據各股之甲基化程度分為以下不同的組：

【0302】 對於雙股 DNA 分子，瓦生股及克立克股之甲基腺嘌呤水準均大於 0.8。此類雙股分子定義為關於腺嘌呤位點之完全甲基化分子（圖 44，A 區）。一股之甲基腺嘌呤水準定義為確定為甲基化之 A 位點在該股之總 A 位點中的百分比。

【0303】 對於雙股 DNA 分子，一股之甲基腺嘌呤水準大於 0.8，而另一股小於 0.2。此類分子定義為關於腺嘌呤位點之半甲基化分子（圖 44，區域 B1 及 B2）。

【0304】 對於雙股 DNA 分子，瓦生股及克立克股之甲基腺嘌呤水準均小於 0.2。此類雙股分子定義為關於腺嘌呤位點之完全未甲基化分子（圖 44，C 區）。

【0305】 對於雙股 DNA 分子，瓦生股及克立克股之甲基腺嘌呤水準不屬於 a、b 及 c 組。此類雙股分子定義為具有關於腺嘌呤位點之交錯甲基化模式的分子（圖 44，D 區）。交錯甲基化模式定義為存在於 DNA 股中之甲基化及未甲基化腺嘌呤的混合物。

【0306】 在一些其他實施例中，用於定義未甲基化股之甲基腺嘌呤水準的閾值可為但不限於小於 0.01、0.05、0.1、0.2、0.3、0.4 及 0.5。用於定義甲基化股之甲基腺嘌呤水準之閾值將為但不限於大於 0.5、0.6、0.7、0.8、0.9、0.95 及 0.99。

【0307】 **圖 45** 為顯示訓練及測試資料集中完全未甲基化分子、半甲基化分子、完全甲基化分子及具有交錯甲基腺嘌呤模式之分子之百分比的表格。測試資料集中之分子可分類為關於腺嘌呤位點之完全未甲基化分子（7.0%）、半甲基化分子（9.8%）、完全甲基化分子（79.4%）及具有交錯甲基腺嘌呤模式之分子（3.7%）。此等結果與訓練資料集中所示之結果相當，在該資料集中存在關於腺嘌呤位點之完全未甲基化分子（7.0%）、半甲基化分子（10.0%）、完全甲基化分子（79.4%）及具有交錯甲基腺嘌呤模式之分子（3.6%）。

【0308】 **圖 46** 展示關於腺嘌呤位點之完全未甲基化分子、半甲基化分子、完全甲基化分子及具有交錯甲基腺嘌呤模式之分子的代表性分子實例。白點表示未甲基化之腺嘌呤。黑點表示甲基化之腺嘌呤。帶點的水平線表示雙股 DNA 分子之股。

【0309】 在實施例中，吾人可藉由增加用於訓練 CNN 模型之 6mA 鹼基

的純度來改良區分甲基化及未甲基化腺嘌呤的效能。為此，吾人可增加 DNA 擴增反應之持續時間，使得增加的新產生的 DNA 產物可稀釋由原始 DNA 模板貢獻之未甲基化腺嘌呤的影響。在其他實施例中，吾人可在用 6mA 進行 DNA 擴增期間併入生物素化鹼基。用 6mA 新產生的 DNA 產物可使用抗生蛋白鏈菌素包被之磁珠拉下且富集。

6 mA 甲基化概況之用途

【0310】 DNA 6mA 修飾存在於細菌、古菌、原生生物及真菌之基因體中（Didier W 等人《自然綜述微生物（Nat Rev Micorbiol.）》2009;4:183-192）。亦據報導，6mA 存在於人類基因體中，佔總腺嘌呤之 0.051%（Xiao CL 等人《分子細胞學》2018;71:306-318）。考慮到 6mA 在人類基因體中之含量低，在一個實施例中，吾人可藉由在全基因體擴增步驟中調整 6mA 在 dNTP 混合物（N 代表未修飾之 A、C、G 及 T）中之比率來創建訓練資料集。舉例而言，吾人可使用的 6mA 與 dNTP 之比率為 1:10、1:100、1:1000、1:10000、1:100000 或 1:1000000。在另一個實施例中，腺嘌呤 DNA 甲基轉移酶 M. EcoGII 可用於創建 6mA 訓練資料集。

【0311】 胃癌及肝癌組織中 6mA 之量較低，且此 6mA 下調與腫瘤發生之增加相關（Xiao CL 等人《分子細胞學》2018;71:306-318）。另一方面，據報導膠質母細胞瘤中存在較高水準之 6mA（Xie 等人《細胞（Cell）》2018;175:1228-1243）。因此，如本文所揭示之用於 6mA 之方法將可用於研究癌症基因體學（Xiao CL 等人《分子細胞學》2018;71:306-318；Xie 等人《細胞》2018;175:1228-1243）。另外，發現 6mA 在哺乳動物粒線體 DNA 中更為普遍及豐富，顯示與低氧相關聯（Hao Z 等人《分子細胞學》2020; doi:10.1016/j.molcel.2020.02.018）。因此，本揭示案中用於 6mA 檢測之方法將可

用於研究在不同臨床條件諸如妊娠、癌症及自體免疫疾病下之粒線體應激反應。

結果與應用

檢測甲基化

【0312】 對於不同的生物樣本及基因體區域，使用上述方法檢測 CpG 位點之甲基化。舉例而言，使用單分子即時定序對孕婦血漿中之游離 DNA 進行甲基化測定，相對於使用亞硫酸氫鹽定序進行的甲基化測定進行驗證。甲基化結果可用於不同的應用，包括確定複本數及診斷病症。下述方法不限於 CpG 位點，且亦可應用於本文所述之任何修飾。

胎盤組織中長 DNA 分子之甲基化檢測

【0313】 單分子即時定序可對長度為千鹼基之 DNA 分子進行定序（Nattestad 等人, 2018）。使用本文所述之發明對 CpG 位點之甲基化狀態的解密將允許吾人藉由協同使用單分子即時定序之長讀段資訊推斷甲基化狀態之單倍型資訊。為了證明推斷長讀段甲基化狀態以及其單倍型資訊的可行性，吾等對胎盤組織 DNA 進行定序，得到 478,739 個分子，該等分子由 28,913,838 個子讀段覆蓋。存在 7 個大小大於 5 kb 之分子。每個分子平均由 3 個子讀段覆蓋。

【0314】 **圖 47** 展示沿著大小為 6,265 bp 之長 DNA 分子（亦即單倍型區塊）之甲基化狀態，該分子在 ZMW 孔號為 m54276_180626_162240/40763503 之 ZMW 中定序且相對於人類基因體中 chr1:113246546-113252811 的基因體位置進行定位。『-』代表非 CpG 核苷酸；『U』代表 CpG 位點之未甲基化狀態；且『M』代表 CpG 位點之甲基化狀態。以黃色突出顯示之區域 4710 指示 CpG 島區域，該區域已知一般為未甲基化的（圖 47）。該 CpG 島中之大多數 CpG 位點經推導為未甲基化的（96%）。相比之下，CpG 島外之 75%的 CpG 位點經推導為未

甲基化的。此等結果表明，CpG 島外（例如 CpG 島岸/島架）之甲基化程度高於 CpG 島之甲基化程度。在該 CpG 島外的區域中，以單倍型排列之甲基化及未甲基化狀態的混合物將表明甲基化模式之變化。此類觀察結果大體上與當前的理解一致（Zhang 等人, 2015；Feinberg 及 Irizarry, 2010）。因此，本揭示案使吾人能夠沿著長分子判讀不同的甲基化狀態，包括甲基化及未甲基化狀態，其意味著甲基化狀態之單倍型資訊可為階段性的。單倍型資訊係指一段連續的 DNA 上 CpG 位點之甲基化狀態的連接。

【0315】 在一個實施例中，吾等可使用此本文中之方法來分析沿著單倍型之甲基化狀態，以檢測及分析印記區域。對印記區域進行表觀遺傳調控，以親源方式引起甲基化狀態。舉例而言，一個重要的印記區域位於人類染色體 11p15.5 上，且含有印記基因 *IGF2*、H19 及 *CDKN1C*（$P57^{kip2}$），其為胎兒生長之強調節因子（Brioude 等人, 《自然綜述內分泌學（Nat Rev Endocrinol.）》2018;14:229-249）。印記區域之遺傳及表觀遺傳畸變將與疾病相關聯。貝克威思-威德曼症候群（BWS）為一種過度生長症候群，患者在兒童早期常常表現為巨舌畸形、腹壁缺損、偏側發育過度、腹腔器官增大及胚胎腫瘤之風險增加。BWS 被認為由 11p15.5 區域內之遺傳或表觀遺傳缺陷引起（Brioude 等人, 《自然綜述內分泌學》2018;14:229-249）。位於 *H19* 與 *IGF2* 之間的一個稱為 ICR1（印記控制區 1）之區域在父本對偶基因上有差異地甲基化。ICR1 指導 *IGF2* 之親源特異性表現。因此，ICR1 之遺傳及表觀遺傳畸變將導致 *IGF2* 之異常表現，其為導致 BWS 之可能原因之一。因此，沿著印記區域檢測甲基化狀態將具有臨床意義。

【0316】 吾等自公共資料庫下載 92 個印記基因之資料，該公共資料庫展出當前報告之印記基因（http://www.geneimprint.org/）。此等印記基因上游及

下游 5-kb 之區域用於進一步分析。在此等區域中，160 個 CpG 島與此等印記基因相關聯。吾等自胎盤樣本獲得 324,248 個環形一致序列。移除質量低且與 CpG 島重疊區域短（例如小於該相關 CpG 島之長度的 50%）之環形一致序列後，吾等獲得與 9 個 CpG 島重疊之 9 個環形一致序列，其對應於 8 個印記基因。

【0317】 **圖 48** 為顯示 9 個 DNA 分子之表格，該等分子藉由單分子即時定序來定序且與印記區域重疊，該等印記區域包括 H19、WT1-AS、WT1、DLK1、MEG3、ATP10A、LRRTM1 及 MAGI2。第 6 行含有與涉及印記區域之 CpG 島重疊的 DNA 段。『U』代表 CpG 上下文之未甲基化胞嘧啶；『M』代表 CpG 上下文之甲基化胞嘧啶。『*』代表定序結果未覆蓋之 CpG 位點；『-』代表非 CpG 位點之核苷酸；若分子與單核苷酸多形現象（SNP）重疊，則在括號中註明基因型。第 7 行指示整個分子之甲基化狀態。若根據本揭示案中存在之實施例顯示大部分 CpG 位點（例如大於 50%）經甲基化，則可將分子稱為甲基化的；否則將其稱為未甲基化的。

【0318】 在 9 個 DNA 分子中，5 個 DNA 分子（55.6%）稱為甲基化的，其沒有顯著偏離 50%之 DNA 分子將甲基化之預期。如圖 48 之表格的第 6 行中所示，大部分 CpG 位點顯示為以一致的方式甲基化或未甲基化，亦即作為甲基化單倍型。一個實施例為，若根據本揭示案中存在之實施例顯示大部分 CpG 位點（例如大於 50%）經甲基化，則將分子稱為甲基化的，否則將其稱為未甲基化的。可使用其他閾值來確定分子是否甲基化，例如但不限於分子中至少 10%、20%、30%、40%、50%、60%、70%、80%、90%及 100%之 CpG 位點經分析被認為經甲基化。

【0319】 在另一個實施例中，吾等可使用同時包含至少一個 SNP 及至少一個 CpG 位點分析之分子來確定區域是否可能與印記區域相關聯，或已知的印

記基因是否可能為異常的（例如印記喪失）。出於說明之目的，**圖 49** 展示來自印記區域之第一分子攜帶對偶基因『A』；而來自印記區域之第二分子攜帶對偶基因『G』。假設印記區域為父本印記的，來自母本單倍型之第一分子為完全未甲基化的；而來自父本單倍型之第二分子為完全甲基化的。在一個實施例中，此類假設將提供甲基化狀態之實況，從而允許根據本揭示案中存在之實施例測試鹼基修飾檢測之效能。

【0320】 圖 49 展示測定印記區域中甲基化模式之實例。提取生物樣本中之 DNA 且與髮夾轉接子連接以形成環形 DNA 分子。關於彼等環形 DNA 分子之序列資訊及鹼基修飾（例如 CpG 位點之甲基化狀態）為未知的。對彼等環形 DNA 分子進行單分子即時定序。在將子讀段相對於參考基因體進行定位之後，確定源自彼等環形 DNA 分子之每個子讀段中鹼基的 IPD、PW 及序列上下文。另外，已確定彼等分子之基因型。與 CG 位點相關聯之量測窗口中的 IPD、PW 及序列上下文將與根據本揭示案中存在之實施例的參考動力學模式進行比較，以確定每個 CpG 之甲基化狀態。若具有不同對偶基因之兩個分子以一個完全未甲基化且另一個完全甲基化之方式顯示不同的甲基化模式，則與此兩個分子相關聯之基因體區域將為印記區域。在一個實施例中，若此類基因體區域恰好為已知的印記區域，例如，如圖 49 所示，則此兩個分子之甲基化模式與正常情形下之預期甲基化模式（亦即實況）一致。其可表明根據本揭示案中存在之實施例之甲基化狀態分類方法的準確性。在一個實施例中，根據本揭示案中存在之實施例所量測之甲基化模式與預期的甲基化模式之間的推導將表明印記之畸變，例如印記之喪失。

【0321】 **圖 50** 展示測定印記區域中甲基化模式之實例。在一個實施例中，可經由分析該區域在某一譜系樹上之甲基化模式來進一步確定印記模式。

舉例而言，可進行跨父本、母本基因體及後代之甲基化模式及對偶基因資訊的分析。此類譜系樹可進一步包括父本或母本祖父、父本或母本祖母的基因體或其他相關基因體。在另一個實施例中，此類分析可擴展至特定人群中的家庭三人組（母親、父親及孩子）資料集，例如根據本文中存在之實施例獲得每個個體的甲基化及基因型資訊。

【0322】 如分類後所示，可確定基因型（對偶基因在盒中）及甲基化狀態。對於每一個分子，可提供每個位點之甲基化模式（例如，全部甲基化或全部未甲基化），以鑑別分子遺傳自哪個親本。或者，可確定甲基化密度，且一或多個閾值可對分子為高甲基化（例如，>80%或其他%且來自一個親本）抑或低甲基化（例如，<20%或其他%且來自另一個親本）進行分類。

cfDNA 分子之甲基化檢測

【0323】 作為另一個實例，游離 DNA（cfDNA）甲基化亦已愈來愈多地視為非侵入性產前檢測之重要分子信號。舉例而言，吾等已證明，來自攜帶組織特異性甲基化區域之 cfDNA 分子可用於確定孕婦血漿中不同組織諸如嗜中性白血球、T 細胞、B 細胞、肝臟、胎盤之貢獻比例（Sun 等人, 2015）。亦已證明使用孕婦血漿 DNA 甲基化檢測第 21 對染色體三體症之可行性（Lun 等人, 2013）。母本血漿中之 cfDNA 分子經片段化以使得中位數大小為 166 bp，比大小大約 500 bp 之人工片段化的大腸桿菌 DNA 短得多。據報導，cfDNA 為非隨機片段化的，例如，血漿 DNA 之末端基元與組織起源諸如來自胎盤相關聯。游離 DNA 之此類特性使其序列上下文與人工片段化之大腸桿菌 DNA 極為不同。因此，通常對於游離 DNA 分子而言，尚不清楚此類聚合酶動力學是否能夠定量地推導甲基化程度。本專利申請案中之揭示內容將適用於但不限於孕婦血漿中之游離 DNA 甲基化分析，例如藉由使用自上述組織 DNA 分子訓練之甲基化預測

模型。

【0324】 使用單分子即時定序，對懷有男胎之孕婦的六個血漿 DNA 樣本進行定序，中位數為 30,738,399 個子讀段（範圍：1,431,215-105,835,846），對應於中位數 111,834 CCS（範圍：61,010-503,582）。每個血漿 DNA 之定序中位數為 262 次（範圍：173-320）。資料集由 Sequel I Sequencing Kit 3.0 製備之 DNA 生成。

【0325】 為了評估 cfDNA 分子之甲基化檢測，吾等使用亞硫酸氫鹽定序（Jiang 等人, 2014）分析上述 6 個孕婦血漿 DNA 樣本之甲基化。吾等獲得 6600 萬個配對末端讀段之中位數（5800-8200 萬個配對末端讀段）。發現總甲基化中位數為 69.6%（67.1%-72.0%）。

【0326】 **圖 51** 展示藉由新方法及習知亞硫酸氫鹽定序推導之甲基化程度的比較。y 軸為根據本專利申請案中存在之實施例預測的甲基化程度。x 軸為藉由亞硫酸氫鹽定序推導之甲基化程度。對單分子即時定序生成之血漿 DNA 結果進行中位數 314,675 個 CpG 位點（範圍：144,546-1,382,568）的分析。預測經甲基化之 CpG 位點的中位比例為 64.7%（範圍：60.8-68.5%），其似乎與亞硫酸氫鹽定序推導之結果相當。如圖 51 所示，採用本發明甲基化預測方法之單分子即時定序及亞硫酸氫鹽定序推導之總體甲基化程度之間存在良好的相關性（r：0.96，p 值=0.0023）。

【0327】 由於亞硫酸氫鹽定序之深度較淺，其可能不適合推導人類基因體中每個 CpG 之甲基化程度（亦即經定序之 CpG 的甲基化分數）。取而代之的是，吾等藉由彙總覆蓋基因體區域之 CpG 位點的讀取信號來計算一些具有多個 CpG 位點之區域的甲基化程度，其中任何兩個連續的 CpG 位點在 50 nt 內且 CpG 位點之數目為至少 10。在一個區域之 CpG 位點上經定序之胞嘧啶及胸腺嘧啶之

和中，經定序之胞嘧啶的百分比指示該區域之甲基化程度。根據區域甲基化程度將區域分為不同的組。隨著甲基化程度增加，自先前訓練資料集（亦即組織DNA）習得之模型所預測的甲基化概率相應地升高（**圖 52A**）。此等結果進一步表明使用單分子即時定序預測孕婦 cfDNA 分子甲基化狀態的可行性及有效性。**圖 52B** 顯示，根據本揭示案中存在之實施例，使用單分子即時定序估計之 10-Mb 基因體窗口的甲基化程度藉由亞硫酸氫鹽定序之甲基化程度很好地校正（r＝0.74；p 值＜0.0001）。

【0328】 **圖 53** 顯示，藉由單分子即時定序量測之孕婦母本血漿中 Y 染色體之基因體呈現（GR）與藉由 BS-seq 量測之基因體呈現很好地相關（r=0.97；P 值=0.007）。此等結果表明，單分子即時定序亦能夠準確定量源自非造血組織諸如胎盤之 DNA 分子，該等組織貢獻的 DNA 一般占少數。換言之，本揭示案證明在定序之前不進行任何鹼基轉化及擴增之情況下，同時分析天然分子之複本數畸變及甲基化狀態的可行性。

基於 CpG 塊之方法

【0329】 一些實施例可對許多基因體區域進行甲基化分析，該等基因體區域具有多個 CpG 位點，例如但不限於 2、3、4、5、10、20、30、40、50、100 個 CpG 位點等。此類基因體區域之大小可為例如但不限於 50、100、200、300 及 500 nt 等。此區域中 CpG 位點之間的距離可為例如但不限於 10、20、30、40、50、100、200、300 nt 等。在一個實施例中，吾等可合併 50 nt 內之任何兩個連續 CpG 位點，形成 CpG 塊，使得此塊中之 CpG 位點數超過 10 個。在此類基於塊之方法中，可將多個區域合併成一個窗口，該窗口表示為單個矩陣，有效地將該等區域一起處理。

【0330】 舉例而言，如**圖 54** 所示，將與 CpG 塊相關聯之所有子讀段的

動力學用於甲基化分析。將該塊中每個 CpG 側翼的上游及下游 10 nt 的投射 IPD 概況相對於 CpG 位點進行人工排比，以計算平均 IPD 概況（**圖 54**）。「投射」一詞意謂吾等已將子讀段動力學信號與所討論之每個相應的 CpG 位點進行排比。CpG 塊之平均 IPD 概況用於訓練模型（例如使用人工神經網路，簡稱 ANN）以鑑別每塊之甲基化狀態。ANN 分析包括一個輸入層、兩個隱藏層及一個輸出層。每個 CpG 塊之特徵為 21 個 IPD 值之特徵向量，其將輸入 ANN 中。第一隱藏層包括 10 個以 ReLu 作為激活函數之神經元。第二隱藏層包括 5 個以 ReLu 作為激活函數之神經元。最後，輸出層包括 1 個以 Sigmoid 作為激活函數之神經元，其將輸出甲基化概率。顯示甲基化概率>0.5 之 CpG 位點被視為甲基化，否則視為未甲基化。平均 IPD 概況可用於分析整個分子之甲基化狀態。若高於臨限值（例如 0、1、2、3 等）之一定數量的位點經甲基化，或若分子具有一定的甲基化密度，則可認為整個分子經甲基化。

【0331】 在未甲基化及甲基化文庫中存在 9,678 及 9,020 個 CpG 塊，每一塊含有至少 10 個 CpG 位點。彼等 CpG 塊覆蓋未甲基化及甲基化文庫之 176,048 及 162,943 個 CpG 位點。如**圖 55A** 及**圖 55B** 所示，在訓練資料集及測試資料集中，吾等在預測甲基化狀態方面可達到大於 90%之總體準確度。然而，此類依賴於 CpG 塊之實施例將大大減少能夠評定之 CpG 的數量。根據定義，對最少數量 CpG 位點之要求會將甲基化分析限制於一些特定的基因體區域（例如優先分析 CpG 島）。

確定起源或疾病

【0332】 甲基化概況可用於檢測組織來源或確定病症之分類。甲基化概況分析可與其他臨床資料結合使用，包括影像學、習知血液檢查及其他醫學診斷資訊。甲基化概況可使用本文所述之任何方法確定。

確定複本數畸變

【0333】 此部分表明，SMRT 對於測定複本數為準確的，且因此可同時分析甲基化概況及複本數概況。

【0334】 已顯示，複本數畸變可藉由對腫瘤組織進行定序來揭露（Chan (2013)）。此處，吾等表明，癌症相關之複本數畸變可藉由使用單分子即時定序對腫瘤組織進行定序來鑑別。舉例而言，對於病例 TBR3033，吾等分別獲得腫瘤 DNA 及其配對的相鄰非腫瘤肝組織 DNA 的 589,435 及 1,495,225 個一致序列（用於構築每個一致序列之子讀段的最低要求為 5）。資料集由 Sequel II Sequencing Kit 1.0 製備之 DNA 生成。在一個實施例中，基因體經由電腦分為 2-Mb 窗口。計算相對於每個窗口進行定位之一致序列的百分比，從而得到 2-Mb 解析度之基因體呈現（GR）。GR 可由一個位置之許多讀段來確定，該等讀段由整個基因體之總序列讀段標準化。

【0335】 **圖 56A** 展示使用單分子即時定序之腫瘤與其配對的相鄰非腫瘤組織 DNA 之間的 GR 比率。腫瘤 DNA 與配對的相鄰正常組織 DNA 之間的複本數比率顯示在 y 軸上，包括第 1 至 22 號染色體之每個 2-Mb 窗口的基因體面元指數顯示在 x 軸上。對於此圖，將 GR 比率高於所有 2-Mb 窗口之第 95 百分位數的區域分類為具有複本數增加，而將 GR 比率低於所有 2-Mb 窗口之第 5 百分位數的區域分類為具有複本數損失。吾等觀察到第 13 號染色體之複本數損失，而第 20 號染色體之複本數增加。此類增加及損失均為正確的結果。

【0336】 **圖 56B** 展示使用亞硫酸氫鹽定序之腫瘤與其配對的相鄰非腫瘤組織之間的 GR 比率。腫瘤 DNA 與配對的相鄰正常組織 DNA 之間的複本數比率顯示在 y 軸上，包括第 1 至 22 號染色體之每個 2-Mb 窗口的基因體面元指數顯示在 x 軸上。圖 56A 中藉由單分子即時定序鑑別之複本數變化在圖 56B 中匹配

之亞硫酸氫鹽定序結果中得到驗證。

【0337】 對於病例 TBR3032，吾等分別獲得腫瘤 DNA 及其配對的相鄰非腫瘤組織 DNA 的 413,982 及 2,396,054 個一致序列（用於構築每個一致序列之子讀段的最低要求為 5）。在一個實施例中，基因體經由電腦分為 2-Mb 窗口。計算相對於每個窗口進行定位之一致序列的百分比，亦即 2-Mb 基因體呈現（GR）。

【0338】 **圖 57A** 展示使用單分子即時定序之腫瘤與其配對的相鄰非腫瘤組織 DNA 之間的 GR 比率。腫瘤 DNA 與配對的相鄰正常組織 DNA 之間的複本數比率顯示在 y 軸上，包括第 1 至 22 號染色體之每個 2-Mb 窗口的基因體面元指數顯示在 x 軸上。對於此圖，將 GR 比率高於所有 2-Mb 窗口之第 95 百分位數的區域分類為具有複本數增加，而將 GR 比率低於所有 2-Mb 窗口之第 5 百分位數的區域分類為具有複本數損失。吾等觀察到第 4、6、11、13、16 及 17 號染色體之複本數損失，而第 5 及 7 號染色體之複本數增加。

【0339】 **圖 57B** 展示使用亞硫酸氫鹽定序之腫瘤與其配對的相鄰非腫瘤組織之間的 GR 比率。腫瘤 DNA 與配對的相鄰正常組織 DNA 之間的複本數比率顯示在 y 軸上，包括第 1 至 22 號染色體之每個 2-Mb 窗口的基因體面元指數顯示在 x 軸上。圖 57A 中藉由單分子即時定序鑑別之複本數變化在圖 57B 中匹配之亞硫酸氫鹽定序結果中得到驗證。

【0340】 因此，可同時分析甲基化概況及複本數概況。在此例證中，由於腫瘤組織之腫瘤純度一般未必總是 100%，因此擴增區域會相對增加腫瘤 DNA 貢獻，而缺失區域將相對降低腫瘤 DNA 貢獻。由於腫瘤基因體之特徵在於全局性低甲基化，因此與缺失區域相比，擴增區域將進一步降低甲基化程度。作為例證，對於病例 TBR3033，使用本發明所量測之第 22 號染色體（複本數增加）

的甲基化程度為 48.2%，低於第 3 號染色體（複本數損失）的甲基化程度（甲基化程度：54.0%）。對於病例 TBR3032，使用本發明所量測之染色體 5p 臂（複本數增加）的甲基化程度為 46.5%，低於染色體 5q 臂（複本數損失）的甲基化程度（甲基化程度：54.9%）。

孕婦血漿 DNA 組織圖譜

【0341】 如**圖 58** 所示，吾等推理，甲基化分析之準確性將使吾等能夠將孕婦之血漿 DNA 甲基化概況與不同參考組織（例如肝臟、嗜中性白血球、淋巴細胞、胎盤、T 細胞、B 細胞、心臟、大腦等）之甲基化概況進行比較。因此，孕婦血漿 DNA 池中來自不同細胞類型之 DNA 貢獻可使用以下程序來推導。將根據本揭示案中存在之實施例確定的 DNA 混合物（例如血漿 DNA）之 CpG 甲基化程度記錄在向量（X）中，且將檢索到的不同組織之參考甲基化程度記錄在矩陣（M）中，該矩陣可藉由但不限於亞硫酸氫鹽定序來定量。不同組織對 DNA 混合物之貢獻比例（p）可藉由但不限於二次規劃來求解。此處，吾等使用數學方程式來說明不同器官對所分析之 DNA 混合物的貢獻比例的推論。DNA 混合物中不同位點之甲基化密度與不同組織中相應位點之甲基化密度的數學關係可表示為：

$$\bar{X}_i = \sum_k (p_k \times M_{ik})，$$

其中$\bar{\boldsymbol{X}}_{\boldsymbol{i}}$代表 DNA 混合物中 CpG 位點 i 之甲基化密度；$\boldsymbol{p}_{\boldsymbol{k}}$代表細胞類型 k 對 DNA 混合物之貢獻比例；$\boldsymbol{M}_{\boldsymbol{ik}}$代表細胞類型 k 中 CpG 位點 i 之甲基化密度。當位點數目等於或大於器官數目時，可確定單個$\boldsymbol{p}_{\boldsymbol{k}}$之值。為了提高資訊量，棄去在所有參考組織類型中顯示甲基化程度變異性小的 CpG 位點。在一個實施例中，吾等使用一組特定的 CpG 位點來進行分析。舉例而言，彼等 CpG 位點之特徵為不同組織中甲基化程度的變異係數（CV）大於 30%，且組織間最大及最小甲基

化程度之間的差異大於 25%。在一些其他實施例中，亦可使用 5%、10%、20%、30%、40%、50%、60%、80%、90%、100%、110%、200%、300%等 之 CV；且可使用組織間最大及最小甲基化程度之間的差異大於 5%、10%、15%、20%、25%、30%、40%、50%、60%、70%、80%、90%、100%等。

【0342】 可在算法中包括額外準則以提高準確性。舉例而言，所有細胞類型之總貢獻將被限制為 100%，亦即

$\sum_k p_k = 100\%$ 。

此外，所有器官之貢獻均必須為非負的：

$p_k \geq 0, \forall k$

【0343】 由於生物學變異，觀察到的總體甲基化模式可能與自組織甲基化推導之甲基化模式不完全相同。在此類情況下，需要進行數學分析，以確定各個組織最可能的貢獻比例。在此方面，DNA 中觀察到的甲基化模式與自組織推導之甲基化模式之間的差異由 W 表示：

$$W = \bar{X}_i - \sum_k (p_k \times M_{ik})$$

【0344】 每個p_k之最有可能之值可藉由將 W 降至最低來確定，W 為觀察到的甲基化模式與推導的甲基化模式之間的差異。此方程式可使用數學算法來求解，例如藉由但不限於使用二次規劃、線性/非線性回歸、期望最大化（EM）算法、最大似然算法、最大後驗估計及最小平方法。

【0345】 如**圖 59** 所示，吾等觀察到，使用圖 58 中存在之血漿 DNA 組織作圖方法，懷有男胎之孕婦的母本血漿中胎盤 DNA 貢獻與 Y 染色體讀段估計之胎兒 DNA 分數有很好的相關性。此結果表明使用動力學追蹤孕婦血漿 DNA 之來源組織的可行性。

區域甲基化程度量化

【0346】 此部分描述用於確定所選基因體區域之代表性甲基化程度的技術，其可使用相對較低水準之定序來完成。可使用甲基化位點之數量及甲基化位點之總數來確定每股或每分子或每個區域的甲基化程度。亦分析各種組織之甲基化程度。

【0347】 吾等對 11 個人類組織 DNA 樣本進行定序，每個樣本之中位數為 3070 萬個子讀段（範圍：910 萬-8860 萬），可與人類參考基因體（hg19）進行排比。每個樣本之子讀段由中位數 380 萬個 Pacific Biosciences 單分子即時（SMRT）定序孔（範圍：110 萬-1150 萬）產生，每個孔含有至少一個可與人類參考基因體進行排比之子讀段。平均而言，SMRT 孔中每個分子平均定序 9.9 次（範圍：6.5-13.4 次）。人類組織 DNA 樣本包括 1 個妊娠個體之母本白血球層樣本、1 個胎盤樣本、2 個肝細胞癌（HCC）腫瘤組織、2 個與 2 個先前提及之 HCC 組織配對的相鄰非腫瘤組織、4 個健康對照個體之白血球層樣本（M1 及 M2 來自男性個體；F1 及 F2 來自女性個體）、1 個 HCC 細胞株（HepG2）。定序資料彙總之詳情顯示於**圖 60** 中。

【0348】 圖 60 展示第一行中之不同組織組及第二行中之樣本名稱。「總子讀段」指示自 SMRT 孔產生之序列總數，包括來自瓦生股及克立克股之序列。「經定位之子讀段」列出可與人類參考基因體進行排比之子讀段的數量。「子讀段可定位性」係指可與人類參考基因體進行排比之子讀段的比例。「每個 SMRT 孔之平均子讀段深度」指示由每個 SMRT 孔產生之子讀段的平均數量。「SMRT 孔之數量」係指產生可檢測子讀段之 SMRT 孔的數量。「可定位孔」指示含有至少一個可排比子讀段之孔的數量。「可定位孔率（%）」為含有至少一個可排比子讀段之孔的百分比。

甲基化程度及模式分析技術

【0349】 在一個實施例中，吾人可量測單個核酸股（例如 DNA 或 RNA）之甲基化密度，其定義為股內甲基化鹼基數除以股內可甲基化鹼基總數。此量測亦稱為「單股甲基化程度」。此單股量測在本揭示案之上下文中特別可行，因為單分子即時定序平台可自雙股 DNA 分子之兩股中之每一者獲得定序資訊。此舉藉由在製備定序文庫時使用髮夾轉接子，使雙股 DNA 分子之瓦生股及克立克股以環形形式連接且一起定序來促進。實際上，此結構亦可使同一雙股 DNA 分子之配對瓦生股及克立克股在同一反應中定序，從而可單獨確定且直接比較任何雙股 DNA 分子之瓦生股及克立克股上相應互補位點的甲基化狀態（例如圖 20A 及 20B）。

【0350】 此等基於股之甲基化分析無法用其他技術輕易實現。因為在不使用如本申請案所揭示之直接甲基化分析方法的情況下，吾人將需要應用另一種手段來區分甲基化鹼基與未甲基化鹼基，例如藉由亞硫酸氫鹽轉化。亞硫酸氫鹽轉化需要用亞硫酸氫鈉處理 DNA，以便可將甲基化胞嘧啶及未甲基化胞嘧啶分別區分為胞嘧啶及胸腺嘧啶。在許多亞硫酸氫鹽轉化方案之變性條件下，雙股 DNA 分子之兩股相互解離。在許多定序應用中，使用例如 Illumina 平台，亞硫酸氫鹽轉化之 DNA 隨後藉由聚合酶鏈反應（PCR）進行擴增，聚合酶鏈反應涉及將雙股 DNA 解離成單股。

【0351】 藉由 Illumina 定序，吾人可在亞硫酸氫鹽轉化之前使用甲基化轉接子製備無 PCR 定序文庫。即使使用此策略，雙股 DNA 分子之每個 DNA 股將被隨機選擇在流動槽中進行橋式擴增。由於定序之無規性，來自同一 DNA 分子之每股不太可能在同一反應中定序。即使在同一運行中分析來自同一基因座之一個以上序列讀段，亦不存在簡單的手段來確定兩個讀段來自一個雙股 DNA 分子之配對瓦生股及克立克股中之每一者，抑或來自兩個不同的雙股 DNA 分

子。此類考慮為重要的，因為在本發明之某些實施例中，雙股 DNA 分子之兩股可表現出不同的甲基化模式。當量測多個核酸股（例如 DNA 或 RNA）之單股甲基化密度時，吾人亦可基於**圖 61** 中關於「所關注之基因體區域的甲基化程度」的概念及方程式來確定「多股甲基化程度」。

【0352】 圖 61 展示分析甲基化模式之各種方式。將具有未知序列及甲基化資訊之雙股 DNA 分子（X）用轉接子連接，在一個實例中形成髮夾環結構。因此，在此實例中，包括瓦生 X(a)股及克立克 X(b)股之 DNA 分子的兩個單股以環形形式以物理方式配對在一起。瓦生股及克立克股中之位點的甲基化狀態可使用本揭示案中所描述之方法（例如，使用來自定序儀之動力學信號、電子信號、電磁信號、光信號或其他類型之物理信號）獲得。環化 DNA 分子中之瓦生股及克立克股可在同一反應中進行詢問。定序後，修剪掉轉接子序列。

【0353】 可藉由分析確定不同的甲基化程度。在圖 61 之(I)中，可分析僅單股分子諸如 X(a)或 X(b)的甲基化模式。此分析可稱為單股甲基化模式分析。該分析可包括但不限於確定位點之甲基化狀態或甲基化模式。在圖 61 中，單股分子 X(a)顯示甲基化模式 5'-UMMUU-3'，其中「U」表示未甲基化之位點且「M」表示甲基化之位點，而互補的單股分子 X(b)顯示甲基化模式 3'-UMUUU-5'。因此，X(b)具有與 X(a)不同的甲基化模式。X(a)及 X(b)之相應的單股甲基化程度分別為 40%及 20%。

【0354】 相比之下，如(II)所示，吾人可在單個雙股 DNA 分子水準上分析甲基化模式（亦即考慮瓦生股及克立克股之甲基化模式。此分析可稱為單分子、雙股 DNA 甲基化模式分析。此範例分子 X 之單分子、雙股 DNA 甲基化程度為 30%。此分析之一種變化形式為將來自瓦生股及克立克股之動力學信號組合起來分析修飾。特別地，由於 CpG 位點上之甲基化一般為對稱的，因此在確

定位點之甲基化狀態之前，可將來自瓦生股及克立克股之動力學信號組合用於位點。在一些情形中，使用自分子之瓦生股及克立克股組合之動力學信號確定鹼基修飾的效能將優於獨立使用單股動力學信號的效能。舉例而言，如圖 20B 所示，與獨立使用單股（AUC：0.85）相比，組合使用來自包括瓦生股及克立克股之兩股的動力學信號將在測試資料集中產生較大 AUC（0.90）。

【0355】 在圖 61 之(III)中，確定所關注之基因體區域的甲基化程度，其中攜帶不同分子大小及不同數目之可甲基化位點（例如 CpG 位點）的不同 DNA 分子可對所關注之基因體區域作出貢獻。此分析可稱為多股甲基化程度分析。術語「多股」可指多個單股 DNA 分子，或多個雙股 DNA 分子，或其任何組合。在此實例中，存在三個雙股 DNA 分子覆蓋所關注之基因體區域：分子「X」、「Y」及「Z」，每個分子具有「a」及「b」股。此區域之相應甲基化程度為 9/28，亦即 32%。待分析之基因體區域的大小可具有 1 nt、10 nt、20 nt、30 nt、40 nt、50 nt、100 nt、1 knt（千核苷酸，亦即一千個核苷酸）、2 knt、3 knt、4 knt、5 knt、10 knt、20 knt、30 knt、40 knt、50 knt、100 knt、200 knt、300 knt、400 knt、500 knt、1 Mnt（兆核苷酸，亦即 100 萬個核苷酸）、2 Mnt、3 Mnt、4 Mnt、5 Mnt、10 Mnt、20 Mnt、30 Mnt、40 Mnt、50 Mnt、100 Mnt 或 200 Mnt。基因體區域可為染色體臂或整個基因體。

【0356】 在確定分子中各位點之甲基化狀態後，亦可確定甲基化模式。舉例而言，在一種情境中，在單個雙股 DNA 分子上存在三個連續的 CpG 位點，瓦生股及克立克股各自的甲基化模式可揭露為三個位點之甲基化（M）、非甲基化（N）及甲基化（M）。例如瓦生股之此模式 MNM 可稱為瓦生股此區域之「甲基化單倍型」。由於 DNA 甲基化維持活性之存在，雙股 DNA 分子之瓦生股及克立克股的甲基化模式可為彼此互補的。舉例而言，若瓦生股上之 CpG 位點

經甲基化，則克立克股上之互補 CpG 位點亦可經甲基化。類似地，瓦生股上之非甲基化 CpG 位點可與克立克股上之非甲基化 CpG 位點互補。

【0357】 在一個實施例中，吾人可量測單個 DNA 分子之甲基化程度，其定義為分子內甲基化鹼基或核苷酸之數目除以該分子內可甲基化鹼基或核苷酸之總數。此量測亦稱為「單分子甲基化程度」。此單分子量測在本揭示案之上下文中可能特別有用，因為單分子即時定序平台可能具有長讀取長度。當量測多個 DNA 分子之單分子甲基化程度時，吾人亦可基於圖 61 中之概念及方程式確定「多分子甲基化程度」。舉例而言，「多分子甲基化程度」可為單分子甲基化程度之平均值或中位數。

【0358】 在一些實施例中，可對 DNA 分子上之一或多個遺傳多形現象（例如單核苷酸多形現象（SNP））以及分子上某一位點之甲基化狀態進行分析，從而揭露該分子之遺傳及表觀遺傳資訊。此類分析將揭示所分析之 DNA 分子的「分階段甲基化單倍型」。分階段甲基化單倍型分析可用於例如母本血漿（含有攜帶母本及胎兒遺傳及表觀遺傳特徵之游離 DNA 分子的混合物）中之基因體印記及游離核酸的研究中。

甲基化結果比較

【0359】 圖 60 之表格中組織之全基因體水準的甲基化密度係使用亞硫酸氫鹽定序及使用如本揭示案中所述之單分子即時定序來確定。**圖 62A** 在 y 軸上顯示藉由亞硫酸氫鹽定序量化之甲基化密度，在 x 軸上顯示組織類型。**圖 62B** 在 y 軸上顯示藉由如本揭示案所述之單分子即時定序量化的甲基化密度，在 x 軸上顯示組織類型。

【0360】 圖 62A 展示使用亞硫酸氫鹽定序（亦即樣本經亞硫酸氫鹽轉化且隨後進行 Illumina 定序）之不同組織的甲基化密度（Lister 等人《自然》

2009;462:315-322），包括 HepG2、HCC 腫瘤組織、與 HCC 腫瘤相鄰的匹配正常肝組織（亦即相鄰的正常組織）、胎盤組織及白血球層樣本。HepG2 呈現最低的甲基化程度，甲基化程度為 40.4%。白血球層樣本呈現最高的甲基化程度，甲基化程度為 76.5%。發現 HCC 腫瘤組織之平均甲基化密度（51.2%）低於匹配的相鄰正常組織之平均甲基化密度（71.0%）。此與 HCC 腫瘤與相鄰正常組織相比在全基因體水準上處於低甲基化的預期一致（Ross 等人《表觀基因體學（Epigenomics）》2010;2:245-69）。資料集由 Sequel II Sequencing Kit 1.0 製備之 DNA 生成。

【0361】 使用單分子即時定序及根據本揭示案之方法對相同組織的部分進行甲基化分析。結果顯示於圖 62B 中。使用本揭示案之單分子即時定序方法進行甲基化分析能夠展示 HepG2 細胞株之甲基化程度最低，其次為所分析之 HCC 腫瘤組織，隨後為胎盤組織。相鄰的非腫瘤肝組織樣本之甲基化程度高於包括 HCC 及胎盤組織之其他組織，其中白血球層之甲基化程度最高。

【0362】 **圖 63A、63B** 及 **63C** 顯示根據本文所述之方法藉由亞硫酸氫鹽定序及單分子即時定序量化的總體甲基化程度的相關性。圖 63A 在 x 軸上顯示藉由亞硫酸氫鹽定序量化之甲基化程度，在 y 軸上顯示使用本文所述之方法藉由單分子即時定序量化之甲基化程度。黑色實線為擬合的回歸線。虛線為兩個量測值相等的地方。

【0363】 根據本文所揭示之發明，亞硫酸氫鹽定序與單分子即時定序之間的甲基化程度存在非常高的相關性（r = 0.99；*P* 值＜0.0001）。此等資料表明，使用特此揭示之單分子即時定序方法進行甲基化分析為確定組織間甲基化程度的有效手段，且能夠比較此等組織間的甲基化狀態及概況。對於甲基化程度之兩個量測值，吾等注意到圖 63A 中回歸線之斜率偏離 1。此等結果表明，

與習知大規模平行亞硫酸氫鹽定序相比，在使用根據本揭示案之單分子即時定序測定甲基化程度時，兩個量測值之間可能存在偏差（在某些情況下，此偏差可稱為偏倚）。

【0364】 在一個實施例中，吾等可使用線性或 LOESS（局部加權平滑）回歸來量化偏差。舉例而言，若吾等將大規模平行亞硫酸氫鹽定序（Illumina）視為參考，則根據本揭示案之單分子即時定序確定的結果可使用回歸係數進行轉換，從而調和不同平台之間的讀取結果。在圖 63A 中，線性回歸公式為 Y=aX+b，其中「*Y*」表示根據本揭示案之單分子即時定序確定的甲基化程度；「*X*」表示亞硫酸氫鹽定序確定的甲基化程度；「*a*」表示回歸線之斜率（例如 a=0.62）；「*b*」表示 y 軸之截距（例如 b=17.72）。在此情況下，由單分子即時定序確定之調和甲基化值將由*(Y-b)/a* 計算。在另一個實施例中，吾人可使用兩個量測值之間的偏差（***ΔM***）與兩個量測值的相應平均值（$\bar{M}$）的關係，其由以下公式(1)及(2)定義：

$$\Delta M = S - \textit{基於亞硫酸氫鹽之甲基化}, \quad (1)$$

$$\bar{M} = \frac{S + \textit{基於亞硫酸氫鹽之甲基化}}{2}, \quad (2)$$

其中「***S***」表示根據本發明之單分子即時定序確定的甲基化程度，「***基於亞硫酸氫鹽之甲基化***」表示亞硫酸氫鹽定序確定的甲基化程度。

【0365】 圖 63B 展示ΔM與$\bar{M}$之間的關係。兩個量測值之平均值（$\bar{M}$）標繪在 x 軸上，兩個量測值之間的偏差（***ΔM***）標繪在 y 軸上。虛線代表一條水平過零的線，在該線上資料點表明兩個量測值之間沒有差異。此等結果表明，偏差根據平均值而變化。兩個量測值之平均值愈高，偏差之幅度就愈大。ΔM值之中位數為-8.5%（範圍：-12.6%至+2.5%），表明方法之間存在差異。

【0366】 圖 63C 在 x 軸上顯示兩個量測值之平均值（$\bar{M}$），在 y 軸上顯示

相對偏差（*RD*）。相對偏差由以下公式定義：

$$RD = \frac{\Delta M}{M} \times 100\%, \quad (3)。$$

虛線代表一條水平過零的線，在該線上資料點表明兩個量測值之間沒有差異。此等結果表明，偏差根據平均值而變化。兩個量測值之平均值愈大，相對推導之幅度就愈大。*RD* 值之中位數為-12.5%（範圍：-18.1%至+6.0%）。

【0367】 據報導，習知全基因體亞硫酸氫鹽定序（Illumina）引入顯著的偏向性序列輸出且高估全局甲基化，在特定基因體區域，不同方法之間的甲基化程度量化存在顯著變化（Olova 等人《基因體生物學（Genome Biol.）》2018;19:33）。本文所揭示之方法可在沒有亞硫酸氫鹽轉化之情況下進行，亞硫酸氫鹽轉化會急劇降解 DNA，且可在沒有 PCR 擴增之情況下進行，PCR 擴增可能會使過程複雜化，或可能會在確定甲基化程度時引入額外的誤差。

【0368】 **圖 64A** 及 **64B** 展示在 1-Mb 解析度下之甲基化模式。圖 64A 展示 HCC 細胞株（HepG2）之甲基化模式。圖 64B 展示來自健康對照個體之白血球層樣本的甲基化模式。染色體表意文字（每個圖中之最外環）以順時針方向自短臂末端至長臂末端排列。自外部之第二環（亦描述為中間環）顯示亞硫酸氫鹽定序確定的甲基化程度。最內環顯示根據本揭示案之單分子即時定序確定的甲基化程度。甲基化程度分類為 5 個等級，亦即 0-20%（淺綠色）、20-40%（綠色）、40-60%（藍色）、60-80%（淺紅色）及 80-100%（紅色）。如圖 64A 及 64B 所示，亞硫酸氫鹽定序（中間軌道）與根據本揭示案之單分子即時定序（最內側軌道）之間在 1-Mb 解析度下的甲基化概況為一致的。顯示母本白血球層樣本之甲基化程度高於 HCC 細胞株（HepG2）。

【0369】 **圖 65A** 及 **65B** 展示在 1-Mb 解析度下量測之甲基化程度的散點圖。圖 65A 展示 HCC 細胞株（HepG2）之甲基化程度。圖 65B 展示來自健康對

照個體之白血球層樣本的甲基化程度。對於圖 65A 及圖 65B，藉由亞硫酸氫鹽定序量化之甲基化程度在 x 軸上，而藉由根據本揭示案之單分子即時定序量測的甲基化程度在 y 軸上。實線為擬合的回歸線。虛線為兩種量測技術相等的地方。對於 HCC 細胞株，藉由單分子即時定序以 1-Mb 解析度測定之甲基化程度與藉由亞硫酸氫鹽定序量測之甲基化程度有良好的相關性（r=0.99；P＜0.0001）（圖 65A）。亦觀察到來自白血球層樣本之資料的相關性（r=0.87，P＜0.0001）（圖 65B）。

【0370】 **圖 66A** 及 **66B** 展示以 100-kb 解析度量測之甲基化程度的散點圖。圖 66A 展示 HCC 細胞株（HepG2）之甲基化程度。圖 66B 展示來自健康對照個體之白血球層樣本的甲基化程度。對於圖 66A 及圖 66，藉由亞硫酸氫鹽定序量化之甲基化程度在 x 軸上，而藉由根據本揭示案之單分子即時定序量測的甲基化程度在 y 軸上。實線為擬合的回歸線。虛線為兩種量測技術相等的地方。當分析之解析度增加至每 100-kb（或 100-knt）窗口時，亦觀察到在 1-Mb（或 1-Mnt）解析度下兩種方法在甲基化定量量測之間的高度相關性。所有此等資料表明，本揭示案之單分子即時方法為量化基因體區域內甲基化程度或甲基化密度的有效工具，其解析度不同，例如在 1-Mb（或 1-Mnt）下或在 100-kb（或 100-knt）下。資料亦表明，本發明為評定區域間或樣本間甲基化概況或甲基化模式的有效工具。

【0371】 **圖 67A** 及 **67B** 展示在 1-Mb 解析度下之甲基化模式。圖 67A 展示 HCC 腫瘤組織（TBR3033T）之甲基化模式。圖 67B 展示相鄰正常組織（TBR3033N）之甲基化模式。染色體表意文字（每個圖中之最外環）以順時針方向自短臂末端至長臂末端排列。自外部之第二環（亦描述為中間環）顯示亞硫酸氫鹽定序確定的甲基化程度。最內環顯示根據本揭示案之單分子即時定序

確定的甲基化程度。甲基化程度分類為 5 個等級，亦即 0-20%（淺綠色）、20-40%（綠色）、40-60%（藍色）、60-80%（淺紅色）及 80-100%（紅色）。如圖 67A 所示，吾等可檢測 HCC 腫瘤組織 DNA（TBR3033T）之低甲基化，其可與圖 67B 中之相鄰正常肝組織 DNA（TBR3033N）區分開。藉由亞硫酸氫鹽定序（中間軌道）及根據本揭示案之單分子即時定序（最內側軌道）確定的甲基化程度及模式為一致的。顯示相鄰正常組織 DNA 的甲基化程度高於 HCC 腫瘤組織 DNA 的甲基化程度。

【0372】 **圖 68A** 及 **68B** 展示以 1-Mb 解析度量測之甲基化程度的散點圖。圖 68A 展示 HCC 腫瘤組織（TBR3033T）之甲基化程度。圖 68B 展示相鄰正常組織之甲基化程度。對於圖 68A 及圖 68B，藉由亞硫酸氫鹽定序量化之甲基化程度在 x 軸上，而藉由根據本揭示案之單分子即時定序量測的甲基化程度在 y 軸上。實線為擬合的回歸線。虛線為兩種量測技術相等的地方。對於 HCC 腫瘤組織 DNA，藉由單分子即時定序以 1-Mb 解析度量測之甲基化程度與藉由亞硫酸氫鹽定序測定之甲基化程度有良好的相關性（r=0.96；*P* 值＜0.0001）（圖 68A）。來自相鄰正常肝組織樣本之資料亦為相關的（r=0.83，*P* 值＜0.0001）（圖 68B）。

【0373】 **圖 69A** 及 **69B** 展示以 100-kb 解析度量測之甲基化程度的散點圖。圖 69A 展示 HCC 腫瘤組織（TBR3033T）之甲基化程度。圖 69B 展示相鄰正常組織（TBR3033N）之甲基化程度。對於圖 69A 及圖 69B，藉由亞硫酸氫鹽定序量化之甲基化程度在 x 軸上，而藉由根據本揭示案之單分子即時定序量測的甲基化程度在 y 軸上。實線為擬合的回歸線。虛線為兩種量測技術相等的地方。當以更高的解析度（例如以 100-kb 窗口）進行甲基化程度之量測時，亦觀察到兩種方法在 1-Mb 解析度下甲基化定量資料之如此高程度的相關性。

【0374】 **圖 70A** 及 **70B** 展示其他腫瘤組織及正常組織在 1-Mb 解析度下之甲基化模式。圖 70A 展示 HCC 腫瘤組織（TBR3032T）之甲基化模式。圖 70B 展示相鄰正常組織（TBR3032N）之甲基化模式。染色體表意文字（每個圖中之最外環）以順時針方向自短臂末端至長臂末端排列。自外部之第二環（亦描述為中間環）顯示亞硫酸氫鹽定序確定的甲基化程度。最內環顯示根據本揭示案之單分子即時定序確定的甲基化程度。甲基化程度分類為 5 個等級，亦即 0-20%（淺綠色）、20-40%（綠色）、40-60%（藍色）、60-80%（淺紅色）及 80-100%（紅色）。如圖 70A 所示，吾等可檢測 HCC 腫瘤組織 DNA（TBR3032T）之低甲基化，其可與圖 70B 中之相鄰正常肝組織 DNA（TBR3032N）區分開。藉由亞硫酸氫鹽定序（中間軌道）及使用本發明之單分子即時定序（最內側軌道）確定的甲基化程度及模式為一致的。顯示相鄰正常組織 DNA 的甲基化程度高於 HCC 腫瘤組織 DNA 的甲基化程度。

【0375】 **圖 71A** 及 **71B** 展示以 1-Mb 解析度量測之甲基化程度的散點圖。圖 71A 展示 HCC 腫瘤組織（TBR3032T）之甲基化程度。圖 71B 展示相鄰正常組織之甲基化程度。對於圖 71A 及圖 71B，藉由亞硫酸氫鹽定序量化之甲基化程度在 x 軸上，而藉由根據本揭示案之單分子即時定序量測的甲基化程度在 y 軸上。實線為擬合的回歸線。虛線為兩種量測技術相等的地方。對於 HCC 腫瘤組織 DNA，藉由單分子即時定序以 1-Mb 解析度量測之甲基化程度與藉由亞硫酸氫鹽定序測定之甲基化程度有良好的相關性（r=0.98；P＜0.0001）（圖 71A）。來自相鄰正常肝組織樣本之資料亦為相關的（r=0.87，P＜0.0001）（圖 71B）。

【0376】 **圖 72A** 及 **72B** 展示以 100-kb 解析度量測之甲基化程度的散點圖。圖 72A 展示 HCC 腫瘤組織（TBR3032T）之甲基化程度。圖 72B 展示相鄰正常組織（TBR3032N）之甲基化程度。對於圖 72A 及圖 72B，藉由亞硫酸氫鹽定

序量化之甲基化程度在 x 軸上，而藉由根據本揭示案之單分子即時定序量測的甲基化程度在 y 軸上。實線為擬合的回歸線。虛線為兩種量測技術相等的地方。當以更高的解析度（例如以 100-kb 窗口）進行甲基化程度之量測時，亦觀察到兩種方法在 1-Mb 解析度下甲基化定量資料之如此高程度的相關性。

腫瘤與相鄰正常組織之間的甲基化差異區域

【0377】 在癌症基因體之區域中經常發現甲基化體畸變。此類畸變之一個實例為所選基因體區域之低甲基化及高甲基化（Cadieux 等人《癌症研究（Cancer Res.）》2006;66:8469-76；Graff 等人《癌症研究》1995;55:5195-9；Costello 等人《自然遺傳學（Nat Genet.）》2000;24:132-8）。另一個實例為所選基因體區域中甲基化及未甲基化鹼基的異常模式。此部分表明，測定甲基化之技術可用於進行定量分析及分析腫瘤之診斷。

【0378】 圖 73 展示腫瘤抑制基因 *CDKN2A* 附近之甲基化異常模式的實例。用藍色突出顯示並加下劃線的座標表示 CpG 島。黑色填充點表示甲基化之位點。未填充之點表示未甲基化之位點。每條帶點的水平線右側括號內的數字表示片段的大小、單分子甲基化密度及 CpG 位點的數量。舉例而言，(3.3 kb, MD:17.9%, CG:39)意謂此片段之大小為 3.3 kb，此片段之甲基化程度為 17.9%且 CpG 位點之數量為 39。MD 代表甲基化密度。

【0379】 如圖 73 所示，*CDKN2A*（細胞週期素依賴性激酶抑制劑 2A）基因編碼包括 INK4A（p16）及 ARF（p14）之兩種蛋白質，充當腫瘤抑制劑。在與腫瘤組織相鄰的非腫瘤組織中，有兩個分子（分子 7301 及分子 7302）覆蓋 *CDKN2A* 基因之重疊區域。分子 7301 及分子 7302 之單個雙股 DNA 分子的甲基化程度分別顯示為 17.9%及 7.6%。相反，發現腫瘤組織中存在之分子 7303 之單個雙股 DNA 分子的甲基化程度為 93.9%，遠高於配對的相鄰非腫瘤組織中存在

之分子的甲基化程度。另一方面，吾人亦可使用與腫瘤組織相鄰的非腫瘤組織中存在之分子 7301 及 7302 來計算多股甲基化程度。結果，多股甲基化程度為 9.7%，低於腫瘤組織的甲基化程度（93.9%）。不同的甲基化程度表明，吾人可使用單個雙股分子甲基化程度及/或多股甲基化程度來檢測或監測疾病，諸如癌症。

【0380】 **圖 74A** 及 **74B** 展示根據本發明實施例之藉由單分子即時定序檢測之差異性甲基化區域。圖 74A 展示癌症基因體中之低甲基化。圖 74B 展示癌症基因體中之高甲基化。x 軸表示 CpG 位點之座標。用藍色突出顯示並加下劃線的座標表示 CpG 島。黑色填充點表示甲基化之位點。未填充之點表示未甲基化之位點。每條帶點的水平線右側括號內的數字表示片段的大小、片段級甲基化密度及 CpG 位點的數量。舉例而言，(3.1 kb, MD:88.9%, CG:180)意謂此片段之大小為 3.1 kb，此片段之甲基化密度為 88.9%且 CpG 位點之數量為 180。

【0381】 圖 74A 展示接近 *GNAS* 基因之區域，與相鄰的正常肝組織相比，HCC 腫瘤組織中呈現更多的低甲基化片段。圖 74B 展示接近 *ESR1* 基因之區域，其在 HCC 組織中呈現高甲基化片段，但與相應區域進行排比的來自配對相鄰非腫瘤組織的 DNA 片段反而顯示出低甲基化。如圖 74B 所示，當癌症樣本與非癌症樣本進行比較時，個別 DNA 分子之甲基化概況或甲基化單倍型足以揭示彼等基因體區域，亦即 *GNAS* 及 *ESR1* 之異常甲基化狀態。

【0382】 此等資料表明，特此揭示之單分子即時定序甲基化分析可確定個別 DNA 片段上每個 CpG 位點之甲基化狀態（無論甲基化或未甲基化）。單分子即時定序之讀取長度比 Illumina 定序之讀取長度長得多（大約千鹼基長），Illumina 定序每次讀取通常可跨越 100-300 nt 長度（De Maio 等人《微生物基因體學（Micob Genom.）》2019;5(9)）。將單分子即時定序之長讀取長度特性與吾

等特此揭示之甲基化分析方法相結合，吾人可容易地確定沿著任何單個 DNA 分子存在的多個 CpG 位點的甲基化單倍型。甲基化概況係指自基因體之一個座標至一段連續 DNA（例如在同一染色體上，或在細菌質體內，或在病毒基因體之單個 DNA 段內）內之另一個座標的 CpG 位點的甲基化狀態。

【0383】 由於單分子即時定序無需事先進行擴增即可單獨分析每個 DNA 分子，因此對任何單個 DNA 分子確定之甲基化概況實際上為甲基化單倍型，意味著同一 DNA 分子自一端至另一端之 CpG 位點的甲基化狀態。若自同一基因體區域對一或多個分子進行定序，則可使用如圖 61 所示之相同公式，自多個 DNA 片段之資料彙總基因體區域內所有經定序 CpG 位點中每個 CpG 位點的甲基化%（亦即甲基化程度或甲基化密度）。對於所有經定序 CpG 位點，可報告每個 CpG 位點之甲基化%，從而提供經定序基因體區域之甲基化概況。或者，可自經定序基因體區域內之所有讀段及所有位點彙總資料，提供該區域之一個甲基化%值，亦即以與圖 64 至 72 所示之 1-Mb 或 1-kb 區域之甲基化程度計算方式相同的方式。

病毒 DNA 甲基化分析

【0384】 此部分表明，本揭示案之甲基化技術可用於準確測定病毒 DNA 之甲基化程度。

【0385】 **圖 75** 展示使用單分子即時定序之兩對 HCC 組織樣本與相鄰非腫瘤組織樣本之間的 B 型肝炎病毒 DNA 的甲基化模式。每個箭頭代表 HBV 基因體中之一個基因註釋。帶有『P』、『S』、『X』及『C』之箭頭表示關於 HBV 基因體之基因註釋：分別編碼聚合酶、表面抗原、X 蛋白及核心蛋白。吾等鑑別出一個大小為 1,183 bp 之片段（分子 I），其來源於相鄰的非腫瘤組織，跨度為 2,278 至 3,141 的 HBV 基因體，以虛線矩形突出顯示，顯示甲基化程度為 12%。

吾等亦鑑別出三個片段（分子 II、III 及 IV），分別為 3,215 bp、2,961 bp 及 3,105 bp，均來源於腫瘤組織。其中，HCC 腫瘤中之兩個片段（分子 III 及 IV）與非腫瘤組織中分子 I 所跨越的 HBV 基因體區域重疊。與虛線矩形中突出顯示之 HBV 區域（HBV 基因體位置：2,278 -3,141）的低甲基化程度（12%）相比，HCC 組織中彼等片段（分子 III 及 IV）之甲基化程度更高（亦即 24%及 30%）。此等結果表明，使用單分子即時定序之方法為可行的，可確定病毒基因體中之甲基化模式，且能夠鑑別 HCC 與及 HCC 組織之間 HBV 的差異甲基化區域（DMR）。因此，根據本揭示案，使用單分子即時定序確定整個病毒基因體的甲基化狀態將提供一種使用組織活檢體研究臨床相關性的新工具。

【0386】 此 DMR 區域恰好與基因 P、C 及 S 重疊。據報導，與具有 HBV 感染但無癌症之肝組織相比，此區域在 HCC 組織中亦被證明為高甲基化的（Jain 等人《科學報告（Sci Rep.）》 2015;5:10478；Fernandez 等人《基因體研究》 2009;19:438-51）。

【0387】 吾等將四名患有肝硬化但無 HCC 之患者之肝組織的亞硫酸氫鹽定序結果進行彙總，獲得 1,156 個 HBV 片段用於甲基化分析。**圖 76A** 顯示患有肝硬化但無 HCC 之患者的肝組織中 B 型肝炎病毒 DNA 的甲基化程度。另外，吾等將 15 名患者之 HCC 腫瘤組織的亞硫酸氫鹽定序結果進行彙總，獲得 736 個 HBV 片段用於甲基化分析。**圖 76B** 展示 HCC 腫瘤組織中 B 型肝炎病毒 DNA 的甲基化程度。如圖 76A 及圖 76B 所示，吾等亦藉由大規模平行亞硫酸氫鹽定序觀察到 HBV 之 DMR 區域（HBV 基因體位置：1,982 - 2,435）在 HCC 組織中之甲基化程度高於肝硬化肝組織。此等結果表明，確定病毒基因體甲基化狀態之方法為有效的。

變體相關之甲基化分析

【0388】 不同的對偶基因可能與不同的甲基化概況相關聯。舉例而言，印記基因可具有一個對偶基因之甲基化程度高於另一個對偶基因。此部分表明，甲基化概況可用於區分某些基因體區域之對偶基因。

【0389】 一個含有單個 DNA 模板之單分子即時定序孔將產生許多子讀段。子讀段包括動力學特徵[例如脈衝間持續時間（IPD）及脈衝寬度（PW）]及核苷酸組成。在一個實施例中，來自一個單分子即時定序孔之子讀段可用於產生一致序列（亦稱為環形一致序列，CCS），其可顯著減少定序錯誤（例如錯配、插入或缺失）。本文描述 CCS 之其他細節。在一個實施例中，可使用與人類參考基因體進行排比之彼等子讀段構築一致序列。在另一個實施例中，一致序列可藉由將子讀段相對於同一單分子即時定序孔中之最長子讀段進行定位來構築。

【0390】 **圖 77** 說明分階段甲基化單倍型分析之原理。填充的棒棒糖代表分類為甲基化之 CpG 位點。未填充的棒棒糖代表分類為未甲基化之 CpG 位點。

【0391】 如圖 77 中之一個實施例所示，將子讀段與人類參考基因體進行排比。將來自一個單分子即時定序孔之經排比子讀段進行摺疊以形成一致序列。一致序列一般可使用存在於每個排比位置之子讀段中的最頻繁的核苷酸來確定。因此，核苷酸變體，包括但不限於單核苷酸變體、插入及缺失，可自一致序列中鑑別出。根據本揭示案，可使用由核苷酸變體標記之同一分子中的平均 IPD 及 PW 來確定甲基化模式。因此，吾等可進一步確定變體相關之甲基化模式。同一分子中之甲基化狀態可視為甲基化單倍型。甲基化單倍型可能不容易直接由兩個或更多個短 DNA 分子構築，因為可能沒有分子標記可區分兩個或更多個片段化之短 DNA 分子是來源於原始單個分子還是由兩個或更多個不同的原始分子貢獻。合成長讀段技術（諸如 10X Genomics 開發之連鎖讀段定序）提

供一種可能性，亦即將單個長 DNA 分子分佈至一個分區（諸如液滴）中，且用相同的分子條形碼序列標記源自該長 DNA 分子之短 DNA 分子。然而，此條形碼步驟涉及不會保留原始甲基化狀態之 PCR 擴增。

【0392】 此外，若吾人試圖使用亞硫酸氫鹽處理長 DNA 分子，則亞硫酸氫鹽處理前的第一步涉及在破壞性條件下之 DNA 變性，將雙股 DNA 變為單股 DNA，因為亞硫酸氫鹽僅可在某些化學條件下作用於單股 DNA 分子。此 DNA 變性步驟將使長的 DNA 分子降解為短的片段，導致原始甲基化單倍型資訊的丟失。基於亞硫酸氫鹽之甲基化分析的第二個缺點將使雙股 DNA 在亞硫酸氫鹽轉化步驟中變性為單股 DNA，亦即瓦生股及克立克股。對於一個分子，有 50%的幾率對瓦生股進行定序，50%的幾率對克立克股進行定序。在數以百萬計的瓦生股及克立克股中，同時對一個分子之瓦生股及克立克股進行定序的幾率極低。即使假設一個分子之瓦生股及克立克股均已經定序，但仍無法明確判定此類瓦生股及克立克股是來源於原始的單個片段還是由兩個或更多個不同的原始片段貢獻。Liu 等人最近介紹一種不含亞硫酸氫鹽之定序方法，用於檢測甲基化胞嘧啶及羥甲基胞嘧啶（Liu 等人《自然生物技術（Nat Biotechnol.）》2019;37:424-429），該方法在溫和的條件下使用基於十-十一易位（TET）酶之轉化，導致 DNA 的降解較少。然而，其涉及酶促反應之兩個連續步驟。酶促反應之任一步驟的轉化率低均會顯著影響總體轉化率。另外，即使對於此不含亞硫酸氫鹽之檢測甲基化胞嘧啶的定序方法，定序結果中仍存在區分分子之瓦生股及克立克股的困難。

【0393】 相反，在本發明之實施例中，分子之瓦生股及克立克股經由鐘形轉接子共價連接以形成環形 DNA 分子。因此，分子之瓦生股及克立克股均在同一反應孔中定序，且可確定各股之甲基化狀態。

【0394】 本發明實施例之一個優點為能夠確定長的連續 DNA 分子（例如長度為千鹼基或千核苷酸）之甲基化及遺傳（亦即序列）資訊。使用短讀段定序技術產生此類資訊較為困難。對於短讀段定序技術，吾人必須使用遺傳或表觀遺傳特徵的支架將多個短讀段之定序資訊結合起來，從而可推斷出一長段的甲基化及遺傳資訊。然而，由於此類遺傳或表觀遺傳錨點之間的距離，此舉在許多情形中可證明具有挑戰性。舉例而言，平均每 1 kb 就有一個 SNP，而目前的短讀段定序技術通常可對每個讀段進行至多 300 nt 的定序，即使在成對末端的格式中亦產生 600 nt。

【0395】 在一個實施例中，變體相關之甲基化單倍型分析可用於研究印記基因之甲基化模式。印記區域以親源方式經受表觀遺傳調控（例如 CpG 甲基化）。舉例而言，對圖 60 之表格中之一個白血球層 DNA 樣本（M2）進行定序，獲得約 1.52 億個子讀段。對於此樣本，53%之單分子即時定序孔產生至少一個可與人類參考基因體進行排比的子讀段。每個 SMRT 孔之平均子讀段深度為 7.7 倍。吾等總共獲得約 300 萬個一致序列。約 91%之參考基因體由一致序列覆蓋至少一次。對於覆蓋區域，定序深度為 7.9 倍。資料集由 Sequel II Sequencing Kit 1.0 製備之 DNA 生成。

【0396】 **圖 78** 展示根據一致序列確定之定序分子的大小分佈，中位數大小為 6,289 bp（範圍：66-198,109 bp）。片段大小（bp）顯示在 x 軸上，與片段大小相關的頻率（%）顯示在 y 軸上。

【0397】 **圖 79A**、**79B**、**79C** 及 **79D** 展示印記區域中對偶基因甲基化模式的實例。x 軸表示 CpG 位點之座標。用藍色突出顯示並加下劃線的座標表示 CpG 島。黑色填充點表示甲基化之 CpG 位點。未填充之點表示未甲基化之 CpG 位點。每個水平系列之填充及未填充之點（亦即 CpG 位點）當中嵌入之字母表示

SNP 位點之對偶基因。每個水平系列點右側括號內的數字表示片段的大小、片段級甲基化密度及 CpG 位點的數量。舉例而言，(10.0 kb, MD:79.1%, CG:139)表明相應片段之大小為 10.0 kb，片段之甲基化密度為 79.1%且 CpG 位點之數量為 139。虛線矩形勾勒出每個基因內甲基化差異最大的區域。

【0398】 **圖 79A** 展示 11 個定序片段，中位數大小為 11.2 kb（範圍：1.3-25 kb），源自 *SNURF* 基因。*SNURF* 基因為母本印記的，意味著個體自母親繼承之基因複本經甲基化且為轉錄沉默的。如圖 79A 所示，在虛線矩形中，***C*** 對偶基因相關片段為高度甲基化的，而 ***T*** 對偶基因相關片段為高度未甲基化的。高度甲基化可表明超過 70%、80%、90%、95%或 99%之位點經甲基化。對偶基因特異性甲基化模式可在其他印記基因中觀察到，包括 *PLAGL1*（**圖 79B**）、*NAP1L5*（**圖 79C**）及 *ZIM2*（**圖 79D**）。圖 79B 顯示，對於 *PLAGL1*，***T*** 對偶基因相關片段為高度未甲基化的，而 ***C*** 對偶基因相關片段為高度甲基化的。**圖 79C** 顯示，對於 *NAP1L5*，***C*** 對偶基因相關片段為高度未甲基化的，而 ***T*** 對偶基因相關片段為高度甲基化的。**圖 79D** 顯示，對於 *ZIM2*，***C*** 對偶基因相關片段為高度未甲基化的，而 ***T*** 對偶基因相關片段為高度甲基化的。

【0399】 **圖 80A**、**80B**、**80C** 及 **80D** 展示非印記區域中對偶基因甲基化模式的實例。x 軸表示 CpG 位點之座標。用藍色突出顯示並加下劃線的座標表示 CpG 島。黑色填充點表示甲基化之 CpG 位點。未填充之點表示未甲基化之 CpG 位點。每個水平系列之填充及未填充之點（亦即 CpG 位點）當中嵌入之字母表示單核苷酸多形現象（SNP）位點之對偶基因。每個水平系列點右側括號內的數字表示片段的大小、片段級甲基化密度及 CpG 位點的數量。虛線矩形表示隨機選擇的區域，用於計算括號中報告的甲基化密度。與圖 79A-79D 中之結果相反，在非印記基因中不存在此類可觀察的對偶基因甲基化模式。圖 80A 顯示在

chr7 區域中沒有不同的對偶基因甲基化模式。圖 80B 顯示在 chr12 區域中沒有不同的對偶基因甲基化模式。圖 80C 顯示在 chr1 區域中沒有不同的對偶基因甲基化模式。圖 80D 顯示在另一個 chr1 區域中沒有不同的對偶基因甲基化模式。

【0400】 **圖 81** 展示對偶基因特異性片段之甲基化程度的表格。第一行列出「印記基因」及「隨機選擇的區域」的類別。第二行列出特定基因。第三行列出基因中 SNP 之第一個對偶基因。第四行列出基因中 SNP 之第二個對偶基因。第五行顯示與第一個對偶基因相連之片段的甲基化程度。第六行顯示與第二個對偶基因相連之片段的甲基化程度。對於彼等印記基因，與對偶基因 2 相連之片段的甲基化程度（平均值：88.6%；範圍 84.6-91.1%）遠高於彼等與對偶基因 1 相連之片段（平均值：12.2%；範圍 7.6-15.7%）（*P* 值=0.03），表明存在對偶基因特異性甲基化。相比之下，彼等隨機選擇的區域之間的甲基化程度沒有顯著變化（*P* 值=1），表明不存在對偶基因特異性甲基化。

妊娠期游離 DNA 分析

【0401】 在此例證中，證明特此揭示之方法適用於分析自懷有至少一個胎兒之婦女獲得之血漿或血清中的游離核酸。在妊娠期間，在母體循環中發現來自胎盤細胞之游離 DNA 及游離 RNA 分子。此類胎盤來源之游離核酸分子亦稱為母體血漿中之游離胎兒核酸或循環游離胎兒核酸。游離胎兒核酸存在於母體血漿中之母體游離核酸之背景中。舉例而言，循環游離胎兒 DNA 分子作為少量物種存在於母體血漿及血清中之游離母體 DNA 背景中。

【0402】 為了區分母體血漿或血清中之游離胎兒 DNA 與游離母體 DNA，眾所周知，吾人可使用遺傳或表觀遺傳手段或結合使用。在遺傳學上，胎兒基因體可藉由父本遺傳的胎兒特異性 SNP 對偶基因、父本遺傳的突變或重生突變而與母本基因體不同。表觀遺傳學上，與母體血細胞之甲基化體相比，胎盤甲

基化體一般為低甲基化的（Lun 等人《臨床化學（Clin Chem.）》2013;59:1583-94）。由於胎盤為游離胎兒 DNA 之主要貢獻者，而母體血細胞為母體循環（血漿或血清）中游離母體 DNA 之主要貢獻者，因此與血漿或血清中游離母體 DNA 相比，游離胎兒 DNA 分子一般為低甲基化的。存在特定的基因體基因座，其中與母體血細胞相比，胎盤為高甲基化的。舉例而言，*RASSF1A* 之啟動子及外顯子 1 區域在胎盤中的甲基化程度高於母體血細胞（Chiu 等人《美國病理學雜誌（Am J Pathol.）》2007;170:941-950）。因此，與來自同一基因座之循環游離母體 DNA 相比，來源於此 *RASSF1A* 基因座之循環游離胎兒 DNA 將為高甲基化的。

【0403】 在實施例中，可基於兩個循環核酸池之間的差異性甲基化狀態，將游離胎兒 DNA 與游離母體 DNA 分子區分開。舉例而言，發現沿著游離 DNA 分子之 CpG 位點大部分為未甲基化的，此分子可能來自胎兒。若發現沿著游離 DNA 分子之 CpG 位點大部分甲基化，則此分子可能來自母親。本領域中熟習此項技術者已知有數種方法來確定此類分子是否確實來自胎兒或母親。一種方法為將定序分子之甲基化模式與胎盤或母體血細胞中相應基因座之已知甲基化概況進行比較。

【0404】 **圖 82** 展示使用甲基化概況確定妊娠中血漿 DNA 之胎盤來源的實例。用藍色突出顯示並加下劃線的座標表示 CpG 島。黑色填充點表示甲基化之位點。未填充之點表示未甲基化之位點。每條帶點的水平線附近括號內的數字表示片段的大小、單分子甲基化密度及 CpG 位點的數量。

【0405】 如圖 82 所示，若母體血漿游離 DNA 分子與 *RASSF1A* 之啟動子區域（已知在胎盤組織中特異性甲基化之區域）進行排比且使用本發明方法生成之定序資料為高甲基化的，則此分子可能來源於胎兒或胎盤。相反，顯示低甲基化之分子可能來源於母體背景 DNA（主要為造血來源）。

【0406】 **圖 83** 展示用於胎兒特異性甲基化分析之方法。該方法包括利用含有胎兒特異性 SNP 對偶基因或胎兒特異性突變（例如父本遺傳的或自然界中重生的）的定序分子。當鑑別出此類胎兒特異性遺傳特徵時，存在於同一游離 DNA 分子上之鹼基的甲基化狀態反映游離胎兒 DNA 或胎盤甲基化體的甲基化概況。當血漿游離 DNA 定序揭露母本基因體中不存在之對偶基因或突變時（例如藉由分析母本基因體 DNA），或藉由分析父本 DNA 或已知在家族中傳播的 DNA（例如藉由分析先證者之 DNA），可發現胎兒特異性遺傳特徵。

【0407】 胎兒特異性 DNA 分子之甲基化可藉由分析彼等攜帶與母本基因體中同型接合對偶基因不同的對偶基因的 DNA 片段來確定。可預期胎兒 DNA 分子之甲基化低於母體 DNA 分子之甲基化。

【0408】 舉例而言，對一名孕婦之白血球層 DNA 及其匹配的胎盤 DNA 進行定序，分別獲得 59x 及 58x 單倍體基因體覆蓋率。吾等鑑別出總共 822,409 個資訊性 SNP，其中母親為同型接合子且胎兒為異型接合子。吾等經由單分子即時定序，在母體血漿（M13160）中發現 2,652 個胎兒特異性片段及 24,837 個共享片段（亦即攜帶共享對偶基因之片段；主要來源與母體）。胎兒 DNA 分數為 19.3%。根據本揭示案，推導出彼等胎兒特異性片段及共享片段之甲基化概況。結果發現，胎兒特異性片段之甲基化程度為 57.4%，而共享片段之甲基化程度為 69.9%。此發現與目前孕婦血漿中胎兒 DNA 之甲基化程度低於母體 DNA 的認識一致（Lun 等人，《臨床化學》2013;59:1583-94）。

【0409】 甲基化模式可用於診斷或監測目的。舉例而言，母體血漿樣本之甲基化概況已用於確定胎齡（https://www.ncbi.nlm.nih.gov/pubmed/27979959）。一種應用為作為品質控制步驟。另一種潛在應用為監測妊娠之「生物」與「時間」年齡。此應用可用於

早產之檢測或風險評估。其他實施例可用於分析母體血液中之胎兒細胞。在其他實施例中，此類胎兒細胞可藉由基於抗體之方法或藉由使用細胞標誌物（例如，在細胞表面上或在細胞質中）之選擇性染色來鑑別，或藉由流動式細胞測量術或顯微操縱或顯微解剖或物理方法（例如，經由腔室、表面或容器之差異性流速）來富集。

使用不同試劑進行甲基化檢測

【0410】 此部分表明，甲基化技術不限於特定的試劑系統。

【0411】 使用不同的試劑系統進行甲基化分析，以確認技術可應用。舉例而言，使用 Sequel II System（Pacific Biosciences）進行 SMRT-seq，以進行單分子即時定序。使用 SMRTbell Express Template Prep Kit 2.0（Pacific Biosciences）對剪切的 DNA 分子進行單分子即時（SMRT）定序模板構築。用 SMRT Link v8.0 軟體（Pacific Biosciences）計算定序引子黏接及聚合酶結合條件。簡言之，使定序引子 v2 與定序模板黏接，且隨後使用 Sequel II Binding and Internal Control Kit 2.0（Pacific Biosciences）使聚合酶與模板結合。在 Sequel II SMRT Cell 8M 上進行定序。用 Sequel II Sequencing Kit 2.0（Pacific Biosciences）在 Sequel II 系統上收集定序影片 30 小時。在其他實施例中，其他化學試劑及反應緩衝液將用於 SMRT-seq。在一個實施例中，聚合酶將根據其甲基化狀態具有沿著 DNA 模板股併入核苷酸之不同動力學特徵（Huber 等人《核酸研究（Nucleic Acids Res.）》2016;44:9881-9890）。在本揭示案中，除非另外指出，否則使用定序引子 v1 生成結果。

【0412】 為了證明在本文所述之揭示內容中本發明在使用不同試劑之情況下的用途，吾等分析基於不同定序套組產生之 SMRT-seq 資料，該等套組包括但不限於 Sequel I Sequencing Kit 3.0、RS II、Sequel II Sequencing Kit 1.0 及 Sequel

II Sequencing Kit 2.0。RS II 包括每個 SMRT 細胞 150,000 ZMW。Sequel 每個 SMRT 細胞使用 1,000,000 ZMW。Sequel II 每個 SMRT 細胞使用 800 萬 ZMW，且具有兩個定序套組（1.0 及 2.0）。此分析涉及兩個資料集。第一資料集係基於全基因體擴增後之 DNA 製備的，代表未甲基化狀態。第二類型資料集係基於 M.SsssI 甲基轉移酶處理後之 DNA 製備的，代表甲基化狀態。此等資料使用 Sequel 定序儀中之 Sequel Sequencing Kit 3.0；Sequel II 定序儀中之 Sequel II Sequencing Kit 1.0 及 Sequel II Sequencing Kit 2.0 生成。因此，吾等獲得具有不同試劑（例如聚合酶）產生之動力學概況的三個資料集。將每個資料集分成一個訓練資料集及一個測試資料集，用於評估使用根據本揭示案之 CNN 模型的效能。

量測窗口

【0413】 **圖 84A**、**84B** 及 **84C** 展示 SMRT-seq 之不同試劑套組的不同量測窗口大小在包含全基因體擴增資料（未甲基化之 CpG 位點）及 M.SsssI 處理之資料（甲基化之 CpG 位點）之訓練資料集中的效能。真陽性率標繪在 y 軸上，假陽性率標繪在 x 軸上。圖 84A 展示基於 Sequel Sequencing Kit 3.0 生成之 SMRT-seq 資料。圖 84B 展示基於 Sequel II sequencing Kit 1.0 生成之 SMRT-seq 資料。圖 84C 展示基於 Sequel II Sequencing Kit 2.0 生成之 SMRT-seq 資料。在圖中，『-』表示所分析之 CpG 胞嘧啶位點的上游信號。『+』表示所分析之 CpG 胞嘧啶位點的下游信號。舉例而言，『-6 nt』表示所分析之 CpG 胞嘧啶位點的 6 nt 上游信號。『+6 nt』表示所分析之 CpG 胞嘧啶位點的 6 nt 下游信號。『±6 nt』表示包括所分析之 CpG 胞嘧啶位點的 6 nt 上游信號及 6 nt 下游信號（亦即 CpG 胞嘧啶位點側翼總共 12 nt 序列）。

【0414】 對於基於 Sequel Sequencing Kit 3.0 之訓練資料集，如圖 84A 所示，使用包含所分析之 CpG 胞嘧啶上的信號及該胞嘧啶位點之 6 nt 上游信號

（由-6 nt 表示）（例如 IPD、PW、相對位置及序列組成）的量測窗口，AUC 值為 0.50，表明沒有區分甲基化之 CpG 胞嘧啶與未甲基化之 CpG 胞嘧啶的鑑別力。然而，對於基於 Sequel II Sequencing Kit 1.0 及 2.0 之訓練資料集，相應的 AUC 值為 0.62（圖 84B）及 0.75（圖 84C）。此等資料證明，SMRT-seq 中使用的不同試劑具有不同的固有動力學概況。此等資料表明，本文所揭示之方法易於適應不同試劑的使用。此外，隨著試劑的進一步發展，例如使用不同的聚合酶及其他化學試劑，可潛在地提高檢測鹼基修飾之準確性。

【0415】 作為另一個實例，對於基於 Sequel Sequencing Kit 3.0 之訓練資料集，如圖 84A 所示，使用包含 CpG 胞嘧啶位點之 10 bp 上游信號（由-10 nt 表示）的量測窗口，AUC 值為 0.50，表明沒有區分甲基化之 CpG 胞嘧啶與未甲基化之 CpG 胞嘧啶的鑑別力。然而，對於基於 Sequel II Sequencing Kit 1.0 及 2.0 之訓練資料集，相應的 AUC 值為 0.66（圖 84B）及 0.79（圖 84C），其表明與包含 6 nt 上游信號之量測窗口相比有所改進。此等資料證實，用於 SMRT-seq 之不同試劑具有不同的固有動力學概況。此等資料表明，本文所揭示之方法易於適應不同試劑的使用。

【0416】 與具有上游信號之量測窗口相比，具有下游信號之量測窗口可導致分類效能之更大改進。舉例而言，對於基於 Sequel Sequencing Kit 3.0 之訓練資料集，如圖 84A 所示，使用包含 CpG 胞嘧啶位點之 6 nt 下游信號（+6 nt）的量測窗口，AUC 值為 0.94，遠大於使用 6 nt 上游信號的 AUC 值（AUC：0.5）。對於基於 Sequel II Sequencing Kit 1.0 及 2.0 之訓練資料集，相應的 AUC 值分別為 0.95（圖 84B）及 0.92（圖 84C），表明與包含上游 6 nt 之量測窗口相比有所改進。此等資料表明，與序列上下文相關聯之動力學特徵將改良使用但不限於 CNN 模型之分類能力。此等資料亦表明，經由調整量測窗口，本文中之揭示內

容將適用於由不同試劑及定序條件（例如不同聚合酶、其他化學試劑、其濃度及定序反應參數（例如持續時間））產生的資料集。使用包括 CpG 胞嘧啶位點之 10 nt 下游信號的量測窗口進行分析將得出類似的結論（圖 84A、84B 及 84C）。

【0417】 在另一個實施例中，吾人可使用包含所分析之胞嘧啶上的信號以及該胞嘧啶之上游及下游信號的量測窗口。舉例而言，如圖 84A、84B 及 84C 所示，使用包含 6 nt 上游信號及 6 nt 下游信號（由±6 nt 表示）之量測窗口，發現基於 Sequel Sequencing Kit 3.0、Sequel II Sequencing Kit 1.0 及 2.0 之訓練資料集的 AUC 值分別為 0.94、0.95 及 0.92。使用包含 10 nt 上游信號及 10 nt 下游信號（由±10 nt 表示）之量測窗口，發現基於 Sequel Sequencing Kit 3.0、Sequel II Sequencing Kit 1.0 及 2.0 之訓練資料集的 AUC 值分別為 0.94、0.95 及 0.94。此等資料表明，本文中之揭示內容將廣泛適用於由不同試劑及定序反應參數產生之資料集。

【0418】 **圖 85A、85B** 及 **85C** 顯示，當應用自訓練資料集訓練之 CNN 模型時，自具有不同定序套組之不同量測窗口的測試資料集獲得結果。真陽性率標繪在 y 軸上，假陽性率標繪在 x 軸上。圖例中之標註相當於圖 84A、84B 及 84C 中使用的標註。圖 85A 展示基於 Sequel Sequencing Kit 3.0 生成之 SMRT-seq 資料。圖 85B 展示基於 Sequel II sequencing Kit 1.0 生成之 SMRT-seq 資料。圖 85C 展示基於 Sequel II Sequencing Kit 2.0 生成之 SMRT-seq。所有在訓練資料集中得出的結論均可在沒有參與訓練過程之此等獨立測試資料集中得到驗證。另外，在三個獨立的測試資料集中，對涉及 Sequel II Sequencing Kit 1.0 及 2.0 之兩個資料集（2/3）之分析表明，使用包括 10 nt 上游和下游信號（由±10 nt 表示）之量測窗口優於其他資料集。

與亞硫酸氫鹽定序之比較

【0419】 **圖 86A、86B** 及 **86C** 展示藉由亞硫酸氫鹽定序及 SMRT-seq（Sequel II Sequencing Kit 2.0）量化之總體甲基化程度的相關性。圖 86A 在 y 軸上顯示藉由 SMRT-seq 量化之百分比形式的甲基化程度。圖 86B 在 x 軸上顯示藉由亞硫酸氫鹽定序量化之百分比形式的甲基化程度。黑線為擬合的回歸線。虛線為兩個量測值相等的對角線。圖 86B 展示布蘭德-奧特曼圖（Bland-Altman plot）。x 軸表示根據本揭示案之 SMRT-seq 及亞硫酸氫鹽定序量化之甲基化程度的平均值。y 軸表示根據本揭示案之 SMRT-seq 與亞硫酸氫鹽定序之間甲基化程度的差異（亦即 Pacific Biosciences 甲基化-基於亞硫酸氫鹽之甲基化）。虛線對應一條水平過零的線，在該線上兩個量測值之間沒有差異。偏離虛線之資料點表明量測值之間存在偏差。圖 86C 展示相對於藉由亞硫酸氫鹽定序量化之值的百分比變化。x 軸表示根據本揭示案之 SMRT-seq 及亞硫酸氫鹽定序量化之甲基化程度的平均值。y 軸表示兩個量測值之間的甲基化程度差異相對於甲基化程度平均值的百分比。虛線對應一條水平過零的線，在該線上兩個量測值之間沒有差異。偏離虛線之資料點表明量測值之間存在偏差。

【0420】 對於圖 86A，線性回歸公式為 Y=aX+b，其中「*Y*」代表根據本揭示案之 SMRT-seq 確定的甲基化程度；「*X*」代表藉由亞硫酸氫鹽定序確定之甲基化程度；「*a*」代表回歸線之斜率（例如 a=1.45）；「*b*」代表 y 軸上之截距（例如 b= -20.98）。在此情況下，SMRT-seq 確定之甲基化值將由*(Y-b)/a* 計算。該圖顯示，對於 Sequel II Sequencing Kit 2.0，與 Sequel II Sequencing Kit 1.0 一樣，藉由 SMRT-seq 確定之甲基化程度可轉換為藉由亞硫酸氫鹽定序確定之甲基化程度，反之亦然。

【0421】 圖 86B 為顯示根據本揭示案之 SMRT-seq 與亞硫酸氫鹽定序之間

甲基化定量的偏差的布蘭德-奧爾特曼圖，其中 x 軸表示根據本揭示案之 SMRT-seq 及亞硫酸氫鹽定序量化之甲基化程度的平均值，y 軸表示根據本揭示案之 SMRT-seq 及亞硫酸氫鹽定序量化之甲基化程度的差異。兩個量測值之間的中位數差異為-6.85%（範圍：-10.1-1.7%）。藉由本揭示案量化之甲基化程度相對於藉由亞硫酸氫鹽定序之值的中位數百分比變化為-9.96%（範圍：-14,76 - 3.21%）。差異視平均值而變化。兩個量測值的平均值愈高，偏差愈大。

【0422】 圖 86C 顯示與圖 86B 相同的資料，但甲基化程度的差異除以兩個甲基化程度的平均值。圖 86C 亦顯示，兩個量測值的平均值愈高，偏差愈大。

【0423】 誤差可能與亞硫酸氫鹽定序有關，而與 SMRT-seq 的方法無關。據報導，習知全基因體亞硫酸氫鹽定序（Illumina）引入明顯偏向的序列輸出且高估全局甲基化，不同方法在特定基因體區域的甲基化程度量化存在很大差異（Olova 等人《基因體生物學》2018;19:33）。本文所揭示之實施例具有許多例示性優點，由此其可在沒有會使 DNA 急劇降解的亞硫酸氫鹽轉化的情況下進行，且可在沒有 PCR 擴增的情況下進行。

組織起源

【0424】 吾等根據本揭示案中之實施例，使用單分子即時定序（SMRT-seq，Pacific Biosciences）對各種癌症類型進行甲基化分析。用於 SMRT-seq 之癌症類型包括但不限於結腸直腸癌（n=3）、食道癌（n=2）、乳癌（n=2）、腎細胞癌（n=2）、肺癌（n=2）、卵巢癌（n=2）、前列腺癌（n=2）、胃癌（n=2）及胰臟癌（n=1）。其匹配的相鄰非腫瘤組織亦納入 SMRT-seq。資料基由 Sequel II Sequencing Kit 2.0 製備之 DNA 生成。

【0425】 **圖 87A** 及 **87B** 展示各種腫瘤組織與配對的相鄰非腫瘤組織之間

總體甲基化程度的比較。y 軸上為百分比形式的甲基化程度。在圖 87A 中，甲基化程度藉由 SMRT-seq 量化。在圖 87B 中，甲基化程度藉由亞硫酸氫鹽定序量化。組織的類型（亦即腫瘤組織或相鄰的非腫瘤組織）在 x 軸上。不同的符號代表不同的起源組織。

【0426】 圖 87A 顯示，包括乳癌、結腸直腸癌、食道癌、肝癌、肺癌、卵巢癌、胰臟癌、腎細胞癌及胃癌之腫瘤組織的總體甲基化程度分別顯著低於相應的非腫瘤組織（*P* 值=0.006，配對樣本 Wilcoxon 符號秩檢驗），包括乳房、結腸、食道、肝臟、肺、卵巢、胰臟、前列腺、腎臟及胃。腫瘤與配對的非腫瘤組織之間甲基化程度的中位數差異為-2.7%（IQR：-6.4 ~ - 0.8%）。

【0427】 圖 84B 證實腫瘤組織中較低的甲基化程度。因此，此等結果表明，根據本揭示案之 SMRT-seq 可準確地測定各種癌症類型及組織的甲基化模式，意味著本揭示案在組織活檢的基礎上，在癌症的早期檢測、預後、診斷及治療方面有廣泛的應用。各種腫瘤類型之甲基化程度降低的程度不同，可能表明甲基化模式與癌症類型相關，從而可確定癌症的起源組織。

增強檢測及其他技術

【0428】 在一些實施例中，鹼基修飾（例如甲基化）之分析可使用以下一或多個參數來進行：序列上下文、IPD 及 PW。IPD 及 PW 可自定序反應中確定，而無需與參考基因體進行排比。單分子即時定序方法之態樣可進一步提高確定序列上下文、IPD 及 PW 之準確性。一個態樣為環形一致性定序之效能，其中可多次量測定序模板之特定部分，因此允許基於經由此等多次讀出之值的平均值或分佈來量測序列上下文、IPD 及 PW。在某些實施例中，在沒有排比過程的情況下對鹼基修飾之分析可提高計算效率，減少周轉時間且可降低分析成本。儘管可在沒有排比過程的情況下執行實施例，但在其他實施例中，可使用

排比過程且可為較佳的，例如，若排比過程用於確定檢測到的鹼基修飾之臨床或生物學含義（例如，若腫瘤抑制因子為高甲基化的）；或若排比過程用於選擇對應於某些所關注之基因體區域之定序資料的子集，以進行進一步分析。對於需要來自所選基因體區域之資料的實施例，此等實施例可能需要使用一或多種能夠在基因體中所關注區域中裂解的酶或基於酶之方法論，例如限制酶或CRISPR-Cas9 系統來靶向此類區域。CRISPR-Cas9 系統可能優於基於 PCR 之方法，因為 PCR 擴增通常不保存有關 DNA 鹼基修飾之資訊。可分析此類所選（生物資訊學[例如經由排比]或經由諸如 CRISPR-Cas9 之方法）區域的甲基化程度，得到關於組織起源、胎兒病症、妊娠病症及癌症之資訊。

使用子讀段進行甲基化分析，無需與參考基因體進行排比

【0429】 在實施例中，可使用包含子讀段之動力學特徵及序列上下文的量測窗口進行甲基化分析，而無需與參考基因體進行排比。如圖 88 所示，源自零模式波導（ZMW）之子讀段用於構築一致序列 8802（亦稱為環形一致序列，CCS）。計算 CCS 中每個位置之平均動力學值，包括但不限於 PW 及 IPD 值。基於 CpG 位點之上游及下游序列，自 CCS 確定該 CpG 位點周圍的序列上下文。因此，將構築如本揭示案中所定義之量測窗口進行訓練，其中該量測窗口包括根據相對於 CCS 具有動力學特徵之子讀段的 PW、IPD 值及序列上下文。此程序避免子讀段與參考基因體之排比。

【0430】 為了測試圖 88 所示的原理，吾等使用源自全基因體擴增之 DNA 的 601,942 個未甲基化之 CpG 位點及源自 CpG 甲基轉移酶（例如 M.SssI）處理之 DNA 的 163,527 個甲基化之 CpG 位點，形成訓練資料集。吾等使用源自全基因體擴增之 DNA 的 546,393 個未甲基化之 CpG 位點及源自 CpG 甲基轉移酶（例如 M.SssI）處理之 DNA 的 193,641 個甲基化之 CpG 位點，形成測試資料集。資

料集由 Sequel II Sequencing Kit 2.0 製備之 DNA 生成。

【0431】 如**圖 89** 所示，在一個實施例中，使用與子讀段及 CCS 相關聯之動力學特徵及序列上下文訓練用於確定甲基化之卷積神經網路（CNN）模型，吾人可實現在測試及訓練資料集中區分甲基化之 CpG 位點與未甲基化之 CpG 位點的 AUC 值分別為 0.94 及 0.95。在其他實施例中，可使用其他神經網路模型、深度學習算法、人工智慧及/或機器學習算法。

【0432】 若吾等為甲基化概率設置 0.2 之閾值，則吾等可在檢測甲基化之 CpG 位點時獲得 82.4%的靈敏度及 91.7%的特異性。此等結果說明，吾人可使用具有動力學特徵之子讀段區分甲基化及未甲基化之 CpG 位點，而無需事先與參考基因體進行排比。

【0433】 在另一個實施例中，為了確定 CpG 位點之甲基化狀態，吾人亦可使用直接來自子讀段之動力學特徵以及序列上下文，而無需 CCS 資訊及事先與參考基因體進行排比。吾等使用動力學特徵，包括跨越子讀段中存在之 CpG 上游 20-nt 及下游 20-nt 位置的 PW 及 IPD 值，來訓練確定甲基化狀態之 CNN 模型。如圖 90 所示，根據本揭示案中之實施例，在訓練及測試資料集中，使用與子讀段相關之動力學特徵檢測甲基化之 CpG 位點之 ROC 曲線的 AUC 分別為 0.70 及 0.69。此等資料表明，使用本揭示案中之實施例來使用與子讀段相關聯之動力學特徵來推斷 DNA 分子之甲基化模式為可行的，但無需事先排比及構築一致序列。然而，此實施例中確定甲基化之效能不如組合利用如本揭示案中所述之排比資訊或一致序列的實施例。吾等會設想，在生成子讀段及動力學值方面增強的精度將改進使用子讀段及其相關動力學特徵確定鹼基修飾的效能。

使用靶向的單分子即時定序對缺失區域進行甲基化分析

【0434】 本文所述之方法亦可應用於分析一或多個所選基因體區域。在

一個實施例中，所關注之區域可首先藉由雜交方法進行富集，該方法允許來自所關注之區域的 DNA 分子與具有互補序列之合成寡核苷酸雜交。對於使用本文所述之方法分析鹼基修飾，目標 DNA 分子不能在進行定序之前藉由 PCR 擴增，因為原始 DNA 分子中之鹼基修飾資訊不會轉移至 PCR 產物。已開發數種方法來富集此等目標區域，而無需進行 PCR 擴增。

【0435】 在另一個實施例中，可經由使用 CRISPR-Cas9 系統來富集目標區域（Stevens 等人《公共科學圖書館·綜合（PLOS One）》 2019;14(4):e0215441；Watson 等人《實驗室研究（Lab Invest）》 2020;100:135-146）。在一個實施例中，首先將 DNA 樣本中 DNA 分子之末端去磷酸化，以使其不易直接連接至定序轉接子。隨後，所關注之區域由 Cas9 蛋白與引導 RNA（crRNA）引導，以產生雙股切口。隨後將雙股切口兩側側翼之所關注區域連接至所選定序平台指定的定序轉接子。在另一個實施例中，可用外切核酸酶處理 DNA，以使 Cas9 蛋白未結合之 DNA 分子降解（Stevens 等人《公共科學圖書館·綜合》 2019;14(4):e0215441）。由於此等方法不涉及 PCR 擴增，因此可對具有鹼基修飾之原始 DNA 分子進行定序，且確定鹼基修飾。在一個實施例中，此方法可用於靶向大量共享同源序列之區域，例如長散佈核元件（LINE）重複序列。在一個實例中，此類分析可用於分析母體血漿中之循環游離 DNA，以檢測胎兒非整倍體（Kinde 等人《公共科學圖書館·綜合》2012;7(7):e41162。

【0436】 如**圖 91** 所示，可藉由使用 CRISPR（成簇規律間隔短回文重複序列）/Cas9（CRISPR 相關蛋白 9）系統來實現靶向的單分子即時定序。對攜帶 5' 磷醯基（亦即 5'-P）及 3'羥基（亦即 3'-OH）之 DNA 片段（例如分子 9102）進行末端封閉處理，由此移除 5'-P 且將 3'-OH 與雙脫氧核苷酸（亦即 ddNTP）連接。因此，末端已經修飾之所得分子（例如分子 9104）無法與轉接子連接，用

於隨後的 DNA 文庫製備。然而，末端封閉之分子受到 CRISPR/Cas9 系統介導之目標特異性裂解，將 5'-P 及 3'-OH 末端引入所關注之分子。攜帶 5'-P 及 3'-OH 末端之此類新裂解之 DNA 分子（例如分子 9106）獲得與髮夾轉接子連接之能力，以形成環形分子 9108。用外切核酸酶 III 及 VII 消化未連接之轉接子、線性 DNA 及僅進行一次裂解之分子。結果，與兩個髮夾轉接子連接之分子被富集，且進行單分子即時定序。此等目標分子適合於根據本揭示案中存在之實施例進行鹼基修飾分析（亦即靶向的單分子即時定序）。

【0437】 如**圖 92** 所示，CRISPR/Cas9 系統中之 Cas9 蛋白與引導 RNA（亦即 gRNA）相互作用，引導 RNA 包括 CRISPR RNA（crRNA，負責 DNA 靶向）及反式激活 crRNA（tracrRNA，負責與 Cas9 形成複合物）（Pickar-Oliver 等人《自然分子細胞生物學綜述（Nat Rev Mol Cell biol.）》2019;20:490-507）。彎曲的形狀代表 Cas9 蛋白，其為一種使用 CRISPR 序列作為引導來識別及切割與 CRISPR 序列之一部分互補之 DNA 的特定股的酶。crRNA 與 tracrRNA 黏接。在一個實施例中，合成的單個 RNA 序列含有 crRNA 及 tracrRNA 序列，稱為單引導 RNA（sgRNA）。crRNA 中之一個區段命名為間隔序列，將經由與目標區域之互補鹼基配對引導 Cas9 蛋白識別且切割雙股 DNA（dsDNA）之特定股。在一個實施例中，間隔序列與靶向 dsDNA 之間的互補性中不涉及錯配。在另一個實施例中，間隔序列與靶向 dsDNA 之間的互補鹼基配對將允許錯配。舉例而言，錯配之數量為但不限於 1、2、3、4、5、6、7、8 等。在一個實施例中，CRISPR 序列將為可程式化的，視不同的 CRISPR/Cas 複合物設計之切割效率、特異性、靈敏度及多工能力而定。

【0438】 如**圖 93** 所示，吾等設計一對 CRISPR/Cas9 複合物，靶向跨越人類基因體中 Alu 元件的兩個切口。『XXX』表示 Cas9 核酸酶切割位點側翼的三個

核苷酸。『YYY』表示與『XXX』互補的三個相應核苷酸。5'-NGG 表示前間隔序列鄰近基元（PAM）序列。在其他 CRISPR/Cas 系統中，PAM 序列可為不同的，且 Cas 核酸酶切割位點側翼的序列可為不同的。在該圖中，Alu 區域之大小為 223 bp。在人類基因體中，存在 1,175,329 個 Alu 區域，每個區域均含有此類 Alu 元件之同源物。此 Alu 元件中存在 5 個 CpG 位點的中位數（範圍：0-34）。舉例而言，此設計含有 36-nt crRNA，其含有 20-nt 間隔序列。詳細的 gRNA 序列資訊如下所示：

【0439】 用於引入第一切口之第一 CRISPR/Cas9 複合物：（所有序列自 5' 至 3'）

crRNA：GCCUGUAAUCCCAGCACUUUGUUUUAGAGCUAUGCU

tracrRNA：

AGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUU

【0440】 用於引入第二切口之第二 CRISPR/Cas9 複合物：

crRNA：AGGGUCUCGCUCUGUCGCCCGUUUUAGAGCUAUGCU

tracrRNA：

AGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUU

【0441】 將 crRNA 分子與 tracrRNA（例如 67-nt）黏接以形成 gRNA 之骨架。具有經設計之 gRNA 的 Cas9 核酸酶可以一定水準之特異性裂解具有靶向切割位點之末端封閉分子的兩股。人類基因體中存在 116,184 個所關注之 Alu 區域，該等區域應該被所設計之 CRISPR/Cas9 複合物切割。因此，Cas9 複合物靶向切割後之彼等 Alu 區域可與髮夾轉接子連接。與髮夾轉接子連接之彼等分子

可藉由單分子即時定序進行定序。可有針對性地確定彼等 Alu 區域之甲基化模式。在一個實施例中，來自兩個 Cas9 複合物之間隔序列可與雙股 DNA 受質之同一股（例如瓦生股或克立克股）鹼基配對。在一個實施例中，來自兩個 Cas9 複合物之 gRNA 中的間隔序列可與雙股 DNA 受質的不同股鹼基配對。舉例而言，Cas9 複合物中之一個間隔序列與雙股 DNA 受質之瓦生股互補，而 Cas9 複合物中之另一間隔序列與雙股 DNA 受質之克立克股互補，反之亦然。

【0442】 在一個實施例中，與髮夾轉接子連接之 DNA 分子為環形，其將對外切核酸酶消化具有抗性。因此，吾人可用外切核酸酶（例如外切核酸酶 III 及 VII）處理轉接子連接之 DNA 產物，以移除線性 DNA（例如脫靶的 DNA 分子）。此使用外切核酸酶之步驟可進一步富集靶向的分子。待定序之靶向的分子的大小取決於由一或多個 Cas9 核酸酶引入之兩個切割位點之間的跨度大小，例如，包括但不限於 10 bp、20 bp、30 bp、40 bp、50 bp、100 bp、200 bp、300 bp、400 bp、500 bp、1000 bp、2000 bp、3000 bp、4000 bp、5000 bp、10 kb、20 kb、30 kb、40 kb、50 kb、100 kb、200 kb、300 kb、500 kb 及 1 Mb。

【0443】 舉例而言，使用具有靶向 Alu 區域之 gRNA 的 Cas9，吾等使用單分子即時定序對人類肝細胞癌（HCC）腫瘤組織樣本中之 187,010 個分子進行定序。其中，113,491 個分子攜帶靶向切口（亦即目標裂解率為約 60.7%之分子）。資料集由 Sequel II Sequencing Kit 2.0 製備之 DNA 生成。換言之，在此實例中由 Cas9 複合物引入所關注分子中之切割位點的特異性為 60.7%。在其他實施例中，由 Cas9 或其他 Cas 複合物引入所關注分子中之切割位點的特異性將為變化的，包括但不限於 1%、5%、10%、20%、30%、40%、50%、60%、70%、80%、90%及 100%。自 CCS 及未與參考基因體進行排比之子讀段獲得的 IPD、PW 值及序列上下文用於確定 Alu 序列中 CpG 位點之甲基化狀態。

【0444】 如**圖 94** 所示，吾等觀察到藉由亞硫酸氫鹽定序及根據本揭示案之單分子即時定序確定之甲基化程度之間的類似甲基化分佈。圖 94 展示亞硫酸氫鹽定序及單分子即時定序（Pacific Biosciences）之甲基化密度（以百分比計）的直方圖。y 軸表示樣本中具有 x 軸上所示之特定甲基化密度之分子的比例。此結果表明，使用 Cas9 介導之靶向單分子即時定序來確定甲基化模式為可行的。此結果亦表明，吾人可使用子讀段相關之動力學特徵（包括 PW 及 IPD 值）來確定甲基化，而無需與參考基因體進行排比。如圖 94 所示，吾等觀察到大量 Alu 區域顯示低甲基化，其與癌症基因體將在 Alu 重複區域中去甲基化之先前知識一致（Rodriguez 等人《核酸研究》2008; 36:770-784）。

【0445】 **圖 95** 在 y 軸上顯示根據本揭示案之單分子即時定序確定之甲基化程度的分佈，在 x 軸上顯示亞硫酸氫鹽定序確定之甲基化密度。如圖 95 所示，根據亞硫酸氫鹽定序之結果，將 Alu 區域之甲基化程度分為 5 個類別，亦即 0 - 20%、20 - 40%、40 - 60%、60 - 80%及 80 - 100%。吾等模型使用量測窗口進一步確定同一組 Alu 區域之甲基化程度，該等量測窗口包括每一類 Alu 區域之動力學特徵及序列上下文（y 軸）。吾等模型所確定之甲基化程度的分佈按照各分組類別之甲基化程度的升序逐漸增加。同樣，此等結果表明，使用 Cas9 介導之靶向單分子即時定序來確定甲基化模式為可行的。吾人可使用子讀段相關之動力學特徵（包括 PW 及 IPD 值）來確定甲基化，而無需與參考基因體進行排比。

【0446】 在另一個實施例中，吾人可使用其他類型的 CRISPR/Cas 系統，例如但不限於 Cas12a、Cas3 及其他直系同源物（例如金黃色葡萄球菌 Cas9）或經工程改造之 Cas 蛋白（增強型胺基酸球菌屬 Cas12a）來進行靶向單分子即時定序。

【0447】 吾人可使用無核酸酶活性之去活化的 Cas9（dCas9）來富集靶向的分子，而無需裂解。舉例而言，靶向的 DNA 分子由包含生物素化之 dCas9 及目標序列特異性 gRNA 之複合物結合。此類靶向的 DNA 分子可能不會被 dCas9 切割，因為 dCas9 為核酸酶缺陷的。經由使用抗生蛋白鏈菌素包覆之磁珠，可富集靶向的 DNA 分子。

【0448】 在一個實施例中，吾人可使用外切核酸酶來消化與 Cas 蛋白一起培育後的 DNA 混合物。外切核酸酶可降解 Cas 蛋白未結合之 DNA 分子，而外切核酸酶可不降解或可大大降低降解 Cas 蛋白結合之 DNA 分子的效率。因此，在最終的定序結果中，有關 Cas 蛋白結合之目標分子的資訊可能會進一步豐富。

【0449】 **圖 96** 展示組織及組織中 Alu 區域之甲基化程度的表格。許多組織顯示甲基化程度在 85-92%之範圍內，包括在 88%至 92%之範圍內。HCC 腫瘤組織及胎盤組織顯示甲基化程度低於 80%。如圖 96 中所見，HCC 腫瘤在吾等設計所針對之 Alu 區域中顯示出頻繁的低甲基化。因此，本揭示案中存在之 Alu 區域的甲基化測定可用於使用自腫瘤活檢體或其他組織或細胞提取之 DNA 在腫瘤進展或治療期間檢測、分期及監測癌症。

【0450】 胎盤組織在整個 Alu 區域之低甲基化可用於使用孕婦血漿 DNA 進行非侵入性產前檢測。舉例而言，較高程度的低甲基化可表明孕婦之胎兒 DNA 分數較高。在另一個實例中，若女性懷有染色體非整倍體之胎兒，則藉由此方法檢測到的源自受影響染色體之 Alu 片段的數量可能與懷有整倍體胎兒之女性在數量上有所不同（亦即增加或減少）。因此，若胎兒患有第 21 對染色體三體症，則當與懷有整倍體胎兒之女性相比時，藉由此方法檢測到的源自第 21 號染色體之 Alu 片段的數量可能會增加。另一方面，若胎兒具有單體染色體，

則當與懷有整倍體胎兒之女性相比時，藉由此方法檢測到的源自該染色體之 Alu 片段的數量可能會減少。與未受影響之染色體相比，測定血漿中受影響染色體（13、18 或 21）之額外低甲基化的呈現可用作區分懷有正常胎兒及異常胎兒之女性的分子指標。

Cas9 複合物靶向的 Alu 區域針對不同類型癌症的甲基化分析

【0451】 即使吾等靶向的 Alu 重複序列在不同的組織中高度甲基化，吾等假設不同的癌症類型在彼等 Alu 重複序列中具有不同的去甲基化模式。在一個實施例中，根據本文呈現之解釋內容，吾人可使用基於 Cas9 之靶向單分子即時定序來分析甲基化模式，以確定不同的癌症類型。

【0452】 **圖 97** 展示不同癌症類型之與 Alu 重複序列相關之甲基化信號的聚類分析。使用微陣列技術（Infinium HumanMethylation450 BeadChip, Illumina Inc）分析來自 TCGA 資料庫（www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga）之癌症個體之 CpG 位點的甲基化狀態。分析微陣列晶片中存在的且與 CRISPR/Cas9 複合物靶向之 Alu 區域重疊的 3,024 個 CpG 位點的甲基化狀態。有許多患者之 CpG 源自所關注之 Alu 區域。藉由微陣列量化每個 CpG 之甲基化程度（亦稱為甲基化指數，或 β 值）。吾等基於彼等 CpG 位點之甲基化程度在患者中的數量進行階層式聚類分析。因此，彼等 CpG 位點之甲基化程度模式類似的患者會聚在一起，形成一個分枝系。不同患者中甲基化模式之相似性將由聚類樹狀圖中的高度值來表示。在此實例中，高度係根據歐幾里得距離（Euclidean distance）來計算。在其他實施例中，將使用其他距離度量，包括但不限於閔可夫斯基（Minkowski）距離、切比雪夫（Chebychev）距離、馬氏（Mahalanobism）距離、曼哈頓（Manhattan）距離、餘弦距離、相關距離、斯皮爾曼（Spearman）距離、漢明

（Hamming）距離、傑卡德（Jaccard）距離等。本文所用之高度表示聚類之間距離度量的值，反映聚類之間的相關性。舉例而言，若吾人觀察到兩個聚類合併的高度為 x，則表明彼等聚類之間的距離為 x（例如所有聚類間患者之間的平均距離）。

【0453】 藉由使用 CpG 位點之甲基化狀態，在聚類分析之結果中，根據癌症類型將患者聚類為不同的不同組。癌症類型包括膀胱尿道上皮癌（BLCA）、乳房侵襲性癌（BRCA）、卵巢漿液性囊腺癌（OV）、胰臟腺癌（PAAD）、HCC、肺腺癌（LUAD）、胃腺癌（STAD）、皮膚黑素瘤（SKCM）及子宮癌肉瘤（UCS）。圖中癌症類型之後的數字表示患者。因此，聚類表明，吾等選擇之 Alu 重複序列中的甲基化信號為分類癌症類型（包括圖 97 中未顯示之癌症類型）提供資訊。在一個實施例中，吾人可基於組織活檢中之甲基化模式來區分原發性及繼發性腫瘤。

子讀段深度及大小閾值

【0454】 此部分顯示，子讀段深度及/或大小閾值可用於提高甲基化檢測之準確性及/或效率。為了測試某些子讀段深度或大小，可修改文庫製備。

【0455】 基於 Sequel II Sequencing Kit 2.0，吾等分析測試資料集中讀段深度對總體甲基化程度量化的影響，該等資料集由全基因體擴增或 M.SsssI 處理後之樣本產生。吾等研究之基因體位點由至少具有一定閾值之子讀段覆蓋，例如但不限於≥1x、10x、20x、30x、40x、50x、60x、70x、80x、90x、100x 等。

【0456】 **圖 98A** 展示在涉及全基因體擴增之測試資料集中，讀段深度對總體甲基化程度量化的影響。**圖 98B** 展示在涉及 M.SsssI 處理之測試資料集中，讀段深度對整體甲基化程度量化的影響。y 軸顯示百分比形式的總體甲基化程度。x 軸顯示子讀段深度。虛線表示總體甲基化程度的預期值。

【0457】 如圖 98A 所示，對於涉及全基因體擴增之資料集，在最初的幾個閾值，諸如但不限於 1x、10x、20x、40x、50x，總體甲基化下降，範圍介於 5.7%至 5.2%。在 50x 或更高的閾值，甲基化程度逐漸穩定在 5%左右。

【0458】 另一方面，在圖 98B 中，對於由 M.SsssI 處理後之樣本產生的資料集，在最初的幾個閾值，諸如但不限於 1x、10x、20x、40x、50x，整體甲基化增加，範圍介於 70%至 83%。在 50x 或更高的閾值，甲基化程度逐漸穩定在 83%左右。

【0459】 在一個實施例中，吾人可調整子讀段深度閾值，使得鹼基修飾分析之效能適合於不同的應用。在其他實施例中，吾人可使用較不嚴格的子讀段深度閾值來獲得更多適合下游分析的 ZMW（亦即分子數）。在另一個實施例中，吾人可將根據本揭示案之 SMRT-seq 確定的甲基化程度的讀數校準為第二量測，例如但不限於 BS-seq、數位液滴 PCR（在亞硫酸氫鹽轉化之樣本上）、甲基化特異性 PCR 或甲基化胞嘧啶結合抗體或其他蛋白質。在另一個實施例中，藉由對 5mC 保留之全基因體擴增後的 DNA 分子進行 BS-seq、數位液滴 PCR（在亞硫酸氫鹽轉化之樣本上）、甲基化特異性 PCR 或甲基-CpG 結合域（MBD）蛋白質富集之基因體定序（MBD-seq）來獲得第二量測。舉例而言，5mC 保留之全基因體擴增可由 DNA 引子酶 TthPrimPol、聚合酶 Φ29 及 DNMT1（DNA 甲基轉移酶 1）介導。

【0460】 吾等分析各種癌症類型及非腫瘤組織之不同子讀段深度的甲基化程度。根據本揭示案之 SMRT-seq 確定的甲基化程度亦與 BS-seq 定序結果進行比較。使用 Sequel II Sequencing Kit 2.0，吾等獲得 4300 萬個子讀段的中位數（四分位數範圍（IQR）：3000 萬-5200 萬），從而可生成與人類參考基因體進行排比的 460 萬個環形一致序列（CCS）的中位數（IQR：280 萬-580 萬）。在彼等

樣本中，亦對 22 個樣本進行完善的大規模平行亞硫酸氫鹽定序（BS-seq），以確定甲基化模式，為甲基化程度的比較提供第二量測。

【0461】 **圖 99** 展示在使用不同子讀段深度閾值的情況下，藉由根據本揭示案之 SMRT-seq（Sequel II Sequencing Kit 2.0）及 BS-seq 確定之總體甲基化程度之間的比較。藉由 SMRT-seq 確定之百分比形式的甲基化程度顯示在 y 軸上。藉由亞硫酸氫鹽定序確定之百分比形式的甲基化程度在 x 軸上。符號表示 1x、10x 及 30x 之不同子讀段深度。三條對角線表示不同子讀段深度的擬合線。

【0462】 圖 99 顯示，當分析由子讀段覆蓋至少一次的基因體位點時（亦即子讀段深度閾值≥1x），根據本揭示案之 SMRT-seq 確定之 CpG 位點之甲基化程度與 BS-seq 確定之 CpG 位點之甲基化程度具有良好的相關性（r = 0.8；*P* 值＜0.0001）。此等結果表明，本揭示案中存在之實施例可用於量測不同組織類型之甲基化程度，包括但不限於結腸直腸癌、結腸直腸組織、食道癌、食道組織、乳癌、非癌性乳房組織、腎細胞癌、腎組織、肺癌及肺組織。吾等亦觀察到，隨著子讀段深度閾值增加至 10x 及 30x，此兩個量測值之間的相關性分別提高至 0.87（*P* 值＜0.0001）及 0.95（*P* 值＜0.0001）。在一些實施例中，增加子讀段深度或選擇覆蓋更多子讀段之基因體區域將改良根據本揭示案之基於 SMRT-seq 之甲基化測定的效能。

【0463】 **圖 100** 為顯示子讀段深度對藉由 SMRT-seq（Sequel II Sequencing Kit 2.0）及 BS-seq 之兩次量測之間甲基化程度相關性之影響的表格。第一行展示子讀段深度閾值。第二行展示 Pearson's r，亦即相關係數。第三行展示與閾值相關聯之 CpG 位點的數量，括號內為位點數量的範圍。

【0464】 如圖 100 所示，藉由 SMRT-seq 及 BS-seq 之兩次測量之間甲基化程度的相關性根據不同的子讀段深度閾值而變化。在一個實施例中，吾人可

利用子讀段深度閾值與兩次測量之間的相關係數（例如 Pearson 相關係數）之間的關係來確定用於區分甲基化胞嘧啶與未甲基化胞嘧啶之最佳子讀段深度閾值。圖 100 顯示，在子讀段深度閾值為 30x（亦即≥30x）時，根據本揭示案之 SMRT-seq 量測的甲基化程度與 BS-seq 產生的結果具有最高的相關性（Pearson's r＝0.952）。在其他實施例中，吾人可使用但不限於 1x、10x、30x、40x、50x、60x、70x、80x、900x、100x、200x、300x、400x、500x、600x、700x、800x 等之子讀段深度閾值。

【0465】 如圖 100 所示，用於甲基化分析之 CpG 位點的數量隨著子讀段深度閾值的增加而減少。在子讀段深度閾值為 100x 的情況下，與子讀段深度閾值為 30x（Pearson's r＝0.952）相比，觀察到甲基化程度的兩次測量之間的相關性較低（Pearson's r＝0.875）。較高的子讀取閾值的較低相關性可歸因於滿足更嚴格的子讀段深度閾值之 CpG 位點的數量較少。在一個實施例中，吾人可考慮子讀段深度之要求與可用於甲基化分析之分子數量之間的權衡。舉例而言，若吾人旨在掃描全基因體之甲基化模式，則可能需要更多的分子。若吾人使用靶向 SMRT-seq 專注於特定區域，則可能需要更高的子讀段深度以獲得該區域之甲基化模式。

【0466】 **圖 101** 展示 Sequel II Sequencing Kit 2.0 生成之資料中相對於片段大小的子讀段深度分佈。子讀段深度顯示在 y 軸上，DNA 分子之長度顯示在 x 軸上。DNA 分子之長度由環形一致序列（CCS）的大小推導得出。

【0467】 由於子讀段深度可能會影響使用 SMRT-seq 資料進行甲基化測定的效能，且子讀段深度為經定序之 DNA 分子之長度的函數，因此 DNA 分子之大小對於獲得用於分析樣本中甲基化模式之最佳子讀段深度可能至關重要。如圖 101 所示，DNA 愈長，子讀段深度愈低。舉例而言，對於大小為 1 kb 之分子

群體，中位數子讀段深度為 50x。對於大小為 10 kb 之分子群體，中位數子讀段深度為 15x。

【0468】 在一個實施例中，如圖 100 所示，子讀段深度之最佳閾值可為至少 30x，其導致最高的相關係數。為了進一步提高將滿足 30x 之最佳子讀段深度閾值之分子的通量，吾人可利用子讀段深度與 DNA 模板分子長度之間的關係。舉例而言，在圖 101 中，30x 為長度為約 4 kb 之分子的中位數子讀段深度。因此，吾人可在 SMRT-seq 文庫製備之前分級分離 4-kbDNA 分子，且將定序限制於 4-kb DNA 分子。在其他實施例中，可使用用於 DNA 分子分級分離之其他大小閾值，包括但不限於 100 bp、200 bp、300 bp、400 bp、500 bp、600 bp、700 bp、800 bp、900 bp、1 kb、2 kb、3 kb、4 kb、5 kb、6 kb、7 kb、9 kb、10 kb、20 kb、30 kb、40 kb、50 kb、60 kb、70 kb、80 kb、90 kb、100 kb、500 kb、1 Mb 或大小閾值之不同組合。

基於限制酶之靶向單分子即時定序

【0469】 此部分描述使用限制酶來提高檢測修飾之實用性及/或通量及/或成本效益。用限制酶生成之 DNA 片段可用於確定樣本之來源。

使用限制酶消化 DNA 分子

【0470】 在實施例中，吾人可在單分子即時定序（例如使用 Pacific Biosciences 系統）之前，使用一或多種限制酶消化 DNA 分子。因為限制酶之識別位點的分佈會不均勻地存在於人類基因體中，所以由限制酶消化之 DNA 可能會產生傾斜的大小分佈。具有較多限制酶識別位點之基因體區域可消化成較小的片段，而具有較少限制酶識別位點之基因體區域可消化成較長的片段。在實施例中，根據大小範圍，吾人可選擇性獲得源自一或多個區域之 DNA 分子，該等區域具有類似的一或多種限制酶之切割模式。可藉由對一或多種限制酶之電

腦切割分析來確定用於大小選擇之所需大小範圍。吾人可使用電腦程式來確定參考基因體（例如人類參考基因體）中所關注之限制酶的識別位點數量。根據彼等識別位點，將此類參考基因體電腦模擬剪切成片段，從而提供所關注之基因體區域的大小資訊。

【0471】 **圖 126** 展示一種基於 *Msp*I 之靶向單分子即時定序的方法，該方法使用 DNA 末端修復及 A 加尾。在如圖 126 所示之實施例中，吾人可使用識別 5'C^CGG3'位點之 *Msp*I 消化生物體之 DNA 樣本，例如但不限於人類 DNA 樣本。對消化後具有 5' CG 突出端之 DNA 片段進行大小選擇，富集源自 CpG 島之 DNA 分子。富集 G 及 C 殘基之基因體區域（亦稱為 GC 含量）可產生較短的片段。因此，吾人可基於所關注區域之 GC 含量確定片段大小的範圍以進行選擇。本領域中熟習此項技術者可使用各種 DNA 片段大小選擇工具，包括但不限於凝膠電泳、尺寸排阻電泳、毛細管電泳、層析、質譜分析、過濾方法、基於沈澱之方法、微流體及奈米流體。對經大小分級之 DNA 分子進行 DNA 末端修復及 A 加尾，使得所需 DNA 產物可與攜帶 5' T 突出端之髮夾轉接子連接，形成環形 DNA 模板。

【0472】 在例如但不限於使用外切核酸酶（例如外切核酸酶 III 及 VII）移除未連接之轉接子、線性 DNA 及不完全環形 DNA 後，與髮夾轉接子連接之 DNA 分子可用於單分子即時定序，以確定 IPD、PW 及序列上下文，從而確定如本文所揭示之甲基化概況。藉由分析富集 CpG 之基因體區域，可藉由本揭示案之定序資料分析方法確定之甲基化概況，對自不同組織或不同疾病及/或生理條件之組織或生物樣本獲得的 DNA 進行區分及分類。

【0473】 對於圖 126 中涉及大小選擇之步驟，在實施例中，所需的大小範圍可藉由 *Msp*I 之電腦模擬切割分析來確定。吾等確定人類參考物中總共

2,286,541 個 *Msp*I 切割位點。根據彼等 *Msp*I 切割位點，將人類參考基因體電腦模擬剪切成片段。吾等獲得總共 2,286,565 個片段。每個單個片段的大小藉由該片段之核苷酸總數來確定。

【0474】 **圖 127A** 及 **127B** 展示經 *Msp*I 消化之片段的大小分佈。此等圖之 y 軸為特定大小片段之頻率（百分比）。圖 127A 之 x 軸具有 50 至 500,000 bp 之對數標度。圖 127B 之 x 軸具有 50 至 1,000 bp 之線性標度。

【0475】 如圖 127A 及 127B 所示，經 *Msp*I 消化之 DNA 分子具有傾斜的大小分佈。經 *Msp*I 消化之片段的中位數大小為 404 bp（IQR：98 - 1,411 bp）。約 53%之彼等經 *Msp*I 消化之片段小於 1 kb。大小概況中存在一系列可能由重複元件引起的尖峰。某些重複元件可能具有相似的 *Msp*I 切割位點模式，導致由 *Msp*I 消化衍生出的一組分子擁有相似的片段大小。舉例而言，頻率最高的尖峰（亦即總共 49,079 個）對應的大小為 64 bp。其中，45,894 個（94%）與 Alu 重複序列重疊。吾人可選擇大小為 64 bp 之 DNA 分子來富集源自 Alu 重複序列之 DNA 分子。該資料表明，大小選擇可用於富集根據本揭示案之下游甲基化分析所需的 DNA 分子。

【0476】 **圖 128** 展示具有某些選定大小範圍之 DNA 分子數量的表格。第一行顯示以鹼基對為單位的大小範圍。第二行顯示大小範圍內之分子相對於總片段的百分比。第三行顯示大小範圍內與 CpG 島重疊的分子數量。第四行顯示大小範圍內之分子與 CpG 島重疊的百分比。第五行顯示經定序之 CpG 位點的數量。第六行顯示落入 CpG 島內之 CpG 位點的數量。第七行顯示藉由大小選擇靶向且落入 CpG 島內之 CpG 位點的百分比。如圖 128 所示，自人類基因體進行 *Msp*I 消化產生之 DNA 分子的數量根據所討論之不同大小範圍而變化。與 CpG 島重疊之 DNA 分子的數量隨不同的大小範圍而變化。

【0477】 由於 CCGG 基元優先出現在 CpG 島中，因此選擇大小小於特定閾值之分子可使源自 CpG 島之 DNA 分子富集。舉例而言，大小範圍為 50 至 200 bp 之分子的數量為 526,543 個，占經 *Msp*I 消化之人類基因體衍生之總 DNA 片段的 23.03%。在 526,543 個 DNA 分子中，有 104,079 個（19.76%）與 CpG 島重疊。大小範圍為 600 至 800 bp 之分子的數量為 133,927 個，占經 *Msp*I 消化之人類基因體衍生之總 DNA 片段的 5.86%。在 133,927 個分子中，有 3,673 個（2.74%）分子與 CpG 島重疊。舉例而言，吾人可選擇 50 至 200 bp 之大小來富集源自 CpG 島之 DNA 片段。

【0478】 為了經由基於 *Msp*I 之靶向單分子即時定序計算與 CpG 島重疊之 CpG 位點的富集程度，吾等對藉由音波處理剪切之 DNA 進行模擬，吾等在正態分佈的基礎上模擬 ZMW 產生之 526,543 個片段，平均大小為 200 bp，標準差為 20 bp。僅有 0.88%之 DNA 分子與 CpG 島重疊。共有 71,495 個 CpG 位點與 CpG 島重疊。如圖 128 所示，選擇範圍介於 50 至 200 bp 之經 *Msp*I 消化之片段將導致 19.8%之片段與 CpG 島重疊。因此，此等資料表明，與藉由音波處理製備之 DNA 相比，藉由 *Msp*I 消化製備之 DNA 可能具有 22.5 倍的源自 CpG 島之 DNA 片段富集。此外，吾等分析經由 *Msp*I 消化在 CpG 島中富集之 CpG 位點。選擇範圍介於 50 至 200 bp 之經 *Msp*I 消化之片段可產生 885,041 個 CpG 位點與 CpG 島重疊，占該大小範圍內經定序片段之總 CpG 位點的 37.5%。與藉由音波處理製備之 DNA 相比，與 CpG 島重疊之 CpG 位點有 12.3 倍（亦即 885,041/71,495）的富集。基於圖 128 中所示之資訊，可選擇適合之大小範圍，以包括 CpG 位點之期望數量及 CpG 島內 CpG 位點之期望倍數富集。

【0479】 **圖 129** 為限制酶消化後 CpG 島內 CpG 位點之覆蓋率百分比與 DNA 片段大小的圖。y 軸顯示由具有給定大小之片段覆蓋的 CpG 島內 CpG 位點

的百分比。x 軸顯示限制酶消化後之 DNA 片段的大小範圍的上限。圖 129 展示藉由擴大大小選擇範圍覆蓋之 CpG 島內 CpG 位點的百分比。在圖 129 中，大小範圍為 50 bp 至 x 軸所示之大小。在其他實施例中，大小範圍之下限可自定義，例如但不限於 60 bp、70 bp、80 bp、90 bp、100 bp、200 bp、300 bp、400 bp 及 500 bp。隨著藉由增加大小上限來擴大大小範圍，吾等可觀察到 CpG 島內 CpG 位點之覆蓋率百分比逐漸增加且穩定在 65%。一些 CpG 位點未經覆蓋，因為其位於 50 bp 以下之 DNA 片段內，或其位於極長分子（例如>100,000 bp）內之片段內。

【0480】 在一些實施例中，可使用兩種或更多種不同的限制酶（具有不同的限制位點）來分析 DNA 樣本，以便增加 CpG 島內 CpG 位點之覆蓋率。藉由不同的酶消化 DNA 樣本可在單獨的反應中進行，因此每個反應中僅存在一種限制酶。舉例而言，可使用識別 CG^CG 位點之 *Acc*II 在 CpG 島上優先切割。在其他實施例中，可使用具有 CG 二核苷酸作為識別位點之一部分的其他限制酶。在人類基因體內，有 678,669 個 *Acc*II 切割位點。吾等使用 *Acc*II 限制性對人類參考基因體進行電腦模擬切割，獲得總共 678,693 個片段。隨後，吾等根據上文關於 *Msp*I 消化所述之方法，對此等片段進行電腦模擬大小選擇，且計算 CpG 島內 CpG 位點之覆蓋率百分比。吾等可觀察到隨著大小選擇範圍的擴大，CpG 位點之覆蓋率百分比逐漸增加。覆蓋率百分比在 50%左右趨於平穩。結合兩個酶消化實驗（亦即 *Msp*I 消化及 *Acc*II 消化）之資料，CpG 位點之覆蓋率進一步增加。經由選擇大小為 50 bp 至 400 bp 之 DNA 片段，覆蓋 CpG 島內 80%之 CpG 位點。此百分比高於單獨使用該兩種酶中之任一者進行消化實驗的相應數字。經由使用其他限制酶分析 DNA 樣本，可進一步提高覆蓋率。若將 DNA 樣本分成兩個等分試樣。一個等分試樣用 *Msp*I 消化，另一個用 *Acc*II 消化。將兩個經消

化之 DNA 樣本以等莫耳混合在一起，且使用單分子即時定序以 500 萬個 ZMW 進行定序。基於電腦模擬分析，就環形一致序列而言，CpG 島內 83%之 CpG 位點（亦即 1,734,345 個）將定序至少 4 次。

【0481】 **圖 130** 展示不使用 DNA 末端修復及 A 加尾之基於 *Msp*I 之靶向單分子即時定序。在實施例中，經消化之 DNA 分子與髮夾轉接子之間的連接可在無 DNA 末端修復及 A 加尾過程之情況下進行。吾人可直接將攜帶 5' CG 突出端之經消化之 DNA 分子與攜帶 5' CG 突出端之髮夾轉接子進行連接，形成用於單分子即時定序之環形 DNA 模板。在清除未連接之轉接子及自連接之轉接子二聚體之後，且在一些實施例中，在移除未連接之轉接子、線性 DNA 及不完全環形 DNA 之後，與髮夾轉接子連接之 DNA 分子可適用於單分子即時定序，以獲得 IPD、PW 及序列上下文。根據本揭示案，將使用 IPD、及序列上下文來確定單分子之甲基化概況。

【0482】 **圖 131** 展示轉接子自連接之概率降低的基於 *Msp*I 之靶向單分子即時定序。加下劃線的胞嘧啶鹼基表示沒有 5'磷酸基團之鹼基。在一些實施例中，為了使在轉接子連接過程中可能發生的自連接之轉接子二聚體之形成的可能性降至最低，吾人可使用去磷酸化之髮夾轉接子與彼等經 *Msp*I 消化之 DNA 分子進行轉接子連接。彼等去磷酸化之髮夾轉接子可能不會形成自連接之轉接子二聚體，因為缺乏 5'磷酸基團。連接後，對產物進行轉接子清除步驟，以純化與髮夾轉接子連接之 DNA 分子。與髮夾轉接子連接之可能攜帶缺口的 DNA 分子進一步進行磷酸化（例如 T4 多核苷酸激酶）及 DNA 連接酶（例如 T4 DNA 連接酶）之缺口密封。在實施例中，吾人可進一步進行未連接之轉接子、線性 DNA 及不完全環形 DNA 之移除。與髮夾轉接子連接之 DNA 分子適用於單分子即時定序，以獲得 IPD、PW 及序列上下文。根據本揭示案，將使用 IPD、及序

列上下文來確定單分子之甲基化概況。

【0483】 除 *Msp*I 之外，亦可使用其他限制酶，諸如 *Sma*I，具有識別位點 CCCGGG。

【0484】 在一些實施例中，可在 DNA 末端修復步驟之後進行所需的大小選擇過程。在一些實施例中，當確定髮夾轉接子對大小選擇結果之影響時，可在髮夾轉接子連接之後進行所需的大小選擇過程。在此等及其他實施例中，基於 *Msp*I 之靶向單分子即時定序中所涉及之程序步驟的順序可根據實驗情況而改變。

【0485】 在實施例中，將使用基於凝膠電泳及/或基於磁珠之方法進行大小選擇。在實施例中，限制酶可包括但不限於 *Bgl*II、*Eco*RI、*Eco*RII、*Bam*HI、*Hind*III、*Taq*I、*Not*I、*Hin*FI、*Pvu*II、*Sau*3AI、*Sma*I、*Hae*III、*Hga*I、*Hpa*II、*Alu*I、*Eco*RV、*Eco*P15I、*Kpn*I、*Pst*I、*Sac*I、*Sal*I、*Sca*I、*Spe*I、*Sph*I、*Stu*I、*Xba*I 及其組合。

用甲基化區分生物樣本類型

【0486】 此部分描述使用藉由限制酶消化生成之片段確定的甲基化概況，以便於區分不同的生物樣本。

【0487】 吾等根據本揭示案中之實施例，使用基於 *Msp*I 之單分子即時定序確定之甲基化概況評定生物樣本之間甲基化概況的差異。吾等以胎盤組織 DNA 及白血球層 DNA 樣本為例。吾等在基於 *Msp*I 之靶向單分子即時定序的基礎上，進行電腦模擬，以生成有關胎盤及白血球層 DNA 樣本之資料。該模擬係基於先前使用 Sequel II Sequencing Kit 1.0 藉由 SMRT 對胎盤組織 DNA 及白血球層 DNA 進行定序達到全基因體覆蓋生成之每個核苷酸的動力學值，包括 IPD 及 PW。隨後，吾等模擬對胎盤 DNA 及白血球層 DNA 樣本進行 *Msp*I 消化之條件，

隨後使用 50 至 200 bp 之大小範圍進行基於凝膠之大小選擇。將所選 DNA 分子與髮夾轉接子連接，形成環形 DNA 模板。對環形 DNA 模板進行單分子即時定序，以獲得有關 IPD、PW 及序列上下文之資訊。

【0488】 假設有 500,000 個 ZMW 產生 SMRT 定序子讀段，彼等子讀段遵循經 *Msp*I 消化之片段在 50 至 200 bp 大小範圍內之基因體分佈，如表 1 所示。假定胎盤及白血球層 DNA 樣本之子讀段深度為 30x。吾等分別對胎盤 DNA 樣本及白血球層 DNA 樣本重複模擬 10 次。因此，獲得藉由經 *Msp*I 消化之靶向單分子即時定序電腦模擬生成的資料集，其包含總共 10 個胎盤 DNA 樣本及 10 個白血球層 DNA 樣本。根據本揭示案，藉由 CNN 對資料集進行進一步分析，確定每個樣本之甲基化概況。吾等獲得來自 CpG 島之 9,198 個 CpG 位點的中位數（範圍：5,497 - 13,928），占總定序之 CpG 位點的 13.6%（範圍：45,304- 90,762）。每個分子中每個 CpG 位點之甲基化狀態藉由根據本揭示案之 CNN 模型來確定。

【0489】 **圖 132** 為藉由基於 *Msp*I 之靶向單分子即時定序確定之胎盤與白血球 DNA 樣本之間的總體甲基化程度的圖。y 軸為百分比形式的甲基化程度。x 軸上列出樣本類型。圖 132 顯示，與白血球層樣本（中位數：69.5%；範圍：68.9%-70.4%）相比，胎盤樣本之總體甲基化程度（中位數：57.6%；範圍：56.9%-59.1%）較低（*P* 值＜0.0001，Mann-Whitney *U* 檢驗）。此等結果表明，藉由基於 *Msp*I 之單分子即時定序確定的甲基化概況可用於基於組織樣本或生物樣本之甲基化差異對其進行區分。由於此等資料表明，胎盤 DNA 可與白血球層 DNA 因其藉由基於 *Msp*I 之單分子即時定序檢測到的甲基化差異而區分開，因此吾人可應用此方法量測母體血漿中之胎兒 DNA 分數。由於母體血漿或母體血清中之胎兒 DNA 來自胎盤，而樣本中之其餘 DNA 分子主要來源於母體白血球層細胞，因此可使用甲基化來量測胎兒 DNA 分數。在實施例中，此技術將為區分

不同組織或具有不同疾病及/或生理條件之組織或生物樣本的有用工具。

【0490】 為了使用 CpG 島之甲基化概況進行胎盤 DNA 樣本與白血球層 DNA 樣本之間的聚類分析，吾等使用分類為甲基化之 CpG 位點在該 CpG 島之總 CpG 位點中的比例來計算 CpG 島之 DNA 甲基化程度。出於說明之目的，吾等使用 CpG 島區域之甲基化程度進行聚類分析。

【0491】 **圖 133** 展示使用基於 *Msp*I 之靶向單分子即時定序確定的 DNA 甲基化概況對胎盤及白血球層樣本進行聚類分析。不同患者中 CpG 島之甲基化模式的相似性由聚類樹狀圖中的高度值來表示。在此實例中，高度係根據歐幾里得距離來計算。在一個實施例中，吾人可使用高度閾值 100 將聚類樹切割成兩組，從而可以 100%之靈敏度及特異性區分胎盤及白血球層樣本。在其他實施例中，吾人可使用其他高度閾值，包括但不限於 50、60、70、80、90、120、130、140 及 150 等。圖 133 顯示，使用根據本揭示案之基於 *Msp*I 之單分子即時定序確定之 CpG 島的甲基化概況，將 10 個胎盤 DNA 樣本及 10 個白血球層 DNA 樣本分別清晰地聚類為兩組。

訓練及檢測方法

【0492】 此部分展示訓練用於檢測鹼基修飾之機器學習模型及使用機器學習模型檢測鹼基修飾的例示性方法。

模型訓練

【0493】 **圖 102** 展示檢測核酸分子中核苷酸之修飾的例示性方法 1020。例示性方法 1020 可為訓練用於檢測修飾之模型的方法。該修飾可包括甲基化。甲基化可包括本文所述之任何甲基化。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，核苷酸可能有兩個以上的狀態（分類）。

【0494】 在方塊 1022，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 4-16 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別核酸分子中定序之核苷酸的各別窗口。與第一複數個資料結構相關聯之每個窗口可包括 4 個或更多個連續核苷酸，包括 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量之連續核苷酸。窗口可為重疊的。每個窗口可包括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。第一資料結構亦可為窗口內之每個核苷酸包括股特性之值。股特性可指示核苷酸存在於第一股或第二股。窗口可包括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，每個窗口可包括第一核酸分子之僅一股上的核苷酸。

【0495】 第一核酸分子可為環形 DNA 分子。環形 DNA 分子可藉由使用 Cas9 複合物切割雙股 DNA 分子形成經切割之雙股 DNA 分子來形成。可將髮夾轉接子連接至經切割之雙股 DNA 分子的末端。在實施例中，雙股 DNA 分子之兩端可經切割及連接。舉例而言，切割、連接及後續分析可如圖 91 所述進行。

【0496】 第一複數個第一資料結構可包括 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括至少 1,000、10,000、50,000、100,000、500,000、1,000,000、5,000,000 或更多個核酸分子。作為另一個實例，可產生至少 10,000 或 50,000 或 100,000 或 500,000 或 1,000,000 或 5,000,000 個序列讀段。

【0497】 藉由量測與核苷酸對應之信號中的脈衝，對第一核酸分子中之

每一者進行定序。該信號可為螢光信號，或其他類型的光信號（例如化學發光、光度）。該信號可由核苷酸或與核苷酸相關之標籤產生。

【0498】 修飾在每個第一核酸分子之每個窗口之目標位置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。可已知修飾存在於第一核酸分子中，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括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。

【0499】 目標位置可為各別窗口之中心。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，目標位置可在各別窗口之任何其他位置，包括第一位置或最後位置。舉例而言，若窗口跨越一股之 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。

【0500】 每個第一資料結構包括窗口內之特性的值。該等特性可為針對窗口內之每個核苷酸的。該等特性可包括核苷酸之標識。該標識可包括鹼基（例如，A、T、C 或 G）。該等特性亦可包括核苷酸相對於各別窗口內之目標位置的位置。舉例而言，位置可為相對於目標位置之核苷酸距離。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，而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。

【0501】 該等特性可包括對應於核苷酸之脈衝的寬度。脈衝之寬度可為脈衝最大值一半時的寬度。該等特性可進一步包括脈衝間持續時間（IPD），其

表示對應於核苷酸之脈衝與對應於鄰近核苷酸之脈衝之間的時間。脈衝間持續時間可為與核苷酸相關聯之脈衝的最大值及與鄰近核苷酸相關聯之脈衝的最大值之間的時間。鄰近核苷酸可為相鄰核苷酸。該等特性亦可包括對應於窗口內之每個核苷酸之脈衝的高度。該等特性可進一步包括股特性之值，其指示核苷酸存在於第一核酸分子之第一股抑或第二股上。股之指示可類似於圖 6 中所示之矩陣。

【0502】 複數個第一資料結構中之每個資料結構可排除 IPD 或寬度低於閾值的第一核酸分子。舉例而言，可僅使用 IPD 值大於第 10 百分位數（或第 1、第 5、第 15、第 20、第 30、第 40、第 50、第 60、第 70、第 80、第 90 或第 95 百分位數）的第一核酸分子。百分位數可基於一或多個參考樣本中所有核酸分子之資料。寬度之閾值亦可對應於百分位數。

【0503】 在方塊 1024，儲存複數個第一訓練樣本。每個第一訓練樣本包括第一複數個第一資料結構中之一者及指示目標位置之核苷酸之修飾的第一狀態的第一標記。

【0504】 在方塊 1026，接收第二複數個第二資料結構。方塊 1026 可為視情況選用的。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之各別核酸分子中定序之核苷酸的各別窗口。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括與第一複數個第一資料結構相同之特性的值。

【0505】 複數個第一訓練樣本可使用多重置換擴增（MDA）生成。在一些實施例中，複數個第一訓練樣本可藉由使用一組核苷酸擴增第一複數個核酸

分子來生成。該組核苷酸可包括指定比率之第一類型的甲基化（例如，6mA 或任何其他甲基化[例如 CpG]）。指定比率可包括相對於未甲基化核苷酸之 1:10、1:100、1:1000、1:10000、1:100000 或 1:1000000。複數個第二核酸分子可使用多重置換擴增由第一類型之未甲基化核苷酸生成。

【0506】 在方塊 1028，儲存複數個第二訓練樣本。方塊 1028 可為視情況選用的。每個第二訓練樣本包括第二複數個第二資料結構中之一者及指示目標位置之核苷酸之修飾的第二狀態的第二標記。

【0507】 在方塊 1029，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，藉由基於模型之輸出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使模型之參數最佳化來進行訓練。模型之輸出指定在各別窗口中目標位置之核苷酸是否具有修飾。該方法可僅包括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。

【0508】 在一些實施例中，模型之輸出可包括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。

【0509】 該模型可包括卷積神經網路（CNN）。CNN 可包括一組卷積濾波器，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。過濾器可為本文所述之任何過濾器。每層之濾波器的數量可為 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。濾波器之內核大小可為 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 可包括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 亦可包括複數個隱藏層，其包括複數個節點。複數個隱藏層中之第一層耦合至輸入層。CNN 可進一步包括輸出層，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。輸出資料結構可包括特性。

【0510】 該模型可包括監督學習模型。監督學習模型可包括不同的方法及算法，包括分析學習、人工神經網路、反向傳播、提昇（元算法）、貝氏統計（Bayesian statistics）、基於病例之推理、決策樹學習、歸納邏輯程式化、高斯過程回歸（Gaussian process regression）、遺傳程式化、資料處理之分組方法、內核估計器、學習自動機、學習分類器系統、最小訊息長度（決策樹、決策圖等）、多線子空間學習、樸素貝葉斯分類器（naive Bayes classifier）、最大熵分類器、條件隨機場、最近鄰算法、可能近似正確學習（PAC）學習、鏈波下降規則、知識獲取方法、符號機器學習算法、亞符號機器學習算法、支持向量機、最小複雜度機器（MCM）、隨機森林、分類器集成、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（一種多準則分類算法）。該模型可為線性回歸、邏輯回歸、深度循環神經網路（例如長短期記憶，LSTM）、貝葉斯分類器、隱式馬爾可夫模型（HMM）、線性判別分析（LDA）、k均值聚類、具有雜訊之基於密度之空間聚類應用（DBSCAN）、隨機森林算法、支持向量機（SVM）或本文所述之任何模型。

【0511】 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，以提供對目標位置之核苷酸的修飾進行分類之經最佳化之精度。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）來驗證模型之準確性。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。該模型可包含複數個子模型，從而提供集合模型。子模型可為較弱的模型，一旦組合就提供更精確的最終模型。

【0512】 在一些實施例中，嵌合或雜合核酸分子可用於驗證模型。複數個第一核酸分子中之至少一些各自包括對應於第一參考序列之第一部分及對應於第二參考序列之第二部分。第一參考序列可來自與第二參考序列不同的染色體、組織（例如腫瘤或非腫瘤）、生物體或物種。第一參考序列可為人類的，且第二參考序列可來自不同的動物。每個嵌合核酸分子可包括對應於第一參考序列之第一部分及對應於第二參考序列之第二部分。第一部分可具有第一甲基化模式，第二部分可具有第二甲基化模式。第一部分可用甲基化酶處理。第二部分可不用甲基化酶處理，且可對應於第二參考序列之未甲基化部分。

修飾之檢測

【0513】 **圖 103** 展示用於檢測核酸分子中核苷酸之修飾的方法 1030。該修飾可為圖 102 之方法 1020 描述的任何修飾。

【0514】 在方塊 1032，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定序之核苷酸的窗口。樣本核酸分子可藉由量測對應於核苷酸之光信號中的脈衝來定序。窗口可為圖 102 之方塊 1022 描述的任何窗口，且定序可為圖 102 之方塊 1022 描述的任何定序。輸入資料結構可包括圖 102 之方塊 1022 描述的相同特性之值。方法 1030 可包括對樣本核酸分子進行定序。

【0515】 窗口內之核苷酸可或可不與參考基因體進行排比。窗口內之核苷酸可使用環形一致序列（CCS）確定，而無需將經定序之核苷酸與參考基因體進行排比。每個窗口中之核苷酸可藉由 CCS 而非與參考基因體進行排比來鑑別。在一些實施例中，可在沒有 CCS 且沒有將經定序之核苷酸與參考基因體進

行排比的情況下確定窗口。

【0516】 窗口內之核苷酸可經富集或過濾。富集可藉由涉及 Cas9 之方法來進行。Cas9 方法可包括使用 Cas9 複合物切割雙股 DNA 分子以形成經切割之雙股 DNA 分子，且將髮夾轉接子連接至經切割之雙股 DNA 分子的末端，類似於圖 91。過濾可藉由選擇大小在大小範圍內之雙股 DNA 分子來進行。核苷酸可來自此等雙股 DNA 分子。可使用保留分子之甲基化狀態的其他方法（例如甲基結合蛋白）。

【0517】 在方塊 1034，將輸入資料結構輸入至模型中。該模型可藉由圖 102 中之方法 1020 來訓練。

【0518】 在一些實施例中，嵌合核酸分子可用於驗證模型。複數個第一核酸分子中之至少一些各自包括對應於第一參考序列之第一部分及對應於與第一參考序列不相接之第二參考序列之第二部分。第一參考序列可來自與第二參考序列不同的染色體、組織（例如腫瘤或非腫瘤）、胞器（例如粒線體、細胞核、葉綠體）、生物體（哺乳動物、病毒、細菌等）或物種。第一參考序列可為人類的，且第二參考序列可來自不同的動物。每個嵌合核酸分子可包括對應於第一參考序列之第一部分及對應於第二參考序列之第二部分。第一部分可具有第一甲基化模式，第二部分可具有第二甲基化模式。第一部分可用甲基化酶處理。第二部分可不用甲基化酶處理，且可對應於第二參考序列之未甲基化部分。

【0519】 在方塊 1036，使用模型確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。

【0520】 輸入資料結構可為複數個輸入資料結構中之一個輸入資料結構。每個輸入資料結構可對應於複數個樣本核酸分子之各別樣本核酸分子中定

序之核苷酸的各別窗口。複數個樣本核酸分子可自個體之生物樣本獲得。生物樣本可為本文所述之任何生物樣本。可針對每個輸入資料結構重複方法 1030。該方法可包括接收複數個輸入資料結構。可將複數個輸入資料結構輸入至模型中。可使用模型確定在每個輸入資料結構之各別窗口中目標位置處之核苷酸中是否存在修飾。

【0521】 複數個樣本核酸分子中之每個樣本核酸分子的大小可大於閾值大小。舉例而言，閾值大小可為 100 bp、200 bp、300 bp、400 bp、500 bp、600 bp、700 bp、800 bp、900 bp、1 kb、2 kb、3 kb、4 kb、5 kb、6 kb、7 kb、9 kb、10 kb、20 kb、30 kb、40 kb、50 kb、60 kb、70 kb、80 kb、90 kb、100 kb、500 kb 或 1 Mb。具有大小閾值可導致更高的子讀段深度，其中之任一者均可增加修飾檢測之準確性。在一些實施例中，該方法可包括在對 DNA 分子進行定序之前，針對特定的大小對 DNA 分子進行分級分離。

【0522】 複數個樣本核酸分子可與複數個基因體區域進行排比。對於複數個基因體區域中之每個基因體區域，可將許多樣本核酸分子與基因體區域進行排比。樣本核酸分子之數量可大於閾值數量。閾值數量可為子讀段深度閾值。子讀段深度閾值數可為 1x、10x、30x、40x、50x、60x、70x、80x、900x、100x、200x、300x、400x、500x、600x、700x 或 800x。可確定子讀段深度閾值數以提高或優化準確性。子讀段深度閾值數可與複數個基因體區域之數量相關。舉例而言，子讀段深度閾值數愈高，複數個基因體區域之數量愈低。

【0523】 可確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確定病症之分類。病症之分類可包括使用修飾之數量。可將修飾之數量與臨限值進行比較。替代或另外地，分類可包括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段與參考基因體進行排

比來確定。若已知與病症相關之某些位置顯示為具有修飾，則可確定病症。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，且可基於比較確定病症。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。該病症可為癌症或本文所述之任何病症（例如，妊娠相關病症、自體免疫疾病）。

【0524】 可分析統計學上顯著數量之核酸分子，以便為病症、組織起源或臨床相關之 DNA 分數提供準確的測定。在一些實施例中，分析至少 1000 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 個核酸分子或更多。作為另一個實例，可產生至少 10,000 或 50,000 或 100,000 或 500,000 或 1,000,000 或 5,000,000 個序列讀段。

【0525】 該方法可包括確定病症之分類為個體患有該病症。分類可包括使用修飾之數量及/或修飾之位點的病症等級。

【0526】 臨床相關之 DNA 分數、胎兒甲基化概況、母體甲基化概況、印記基因區域之存在或起源組織（例如，來自含有不同細胞類型混合物之樣本）可使用一或多個核苷酸處之修飾的存在來確定。臨床相關之 DNA 分數包括但不限於胎兒 DNA 分數、腫瘤 DNA 分數（例如，來自含有腫瘤細胞及非腫瘤細胞混合物之樣本）及移植 DNA 分數（例如來自含有供體細胞及受體細胞混合物之樣本）。

【0527】 該方法可進一步包括治療病症。可根據所確定之病症等級、經鑑別之修飾及/或起源來源（例如，自癌症患者之循環中分離的腫瘤細胞）提供治療。舉例而言，可用特定的藥物或化學療法靶向經鑑別之修飾。起源組織可用於指導手術或任何其他形式之治療。並且，病症等級可用於確定用任何類型之治療的積極性。

【0528】 實施例可包括在確定患者之病症等級之後治療患者之病症。治療可包括任何適合之療法、藥物、化學療法、放療或手術，包括本文提及之參考文獻中所述之任何治療。參考文獻中關於治療之資訊以引用之方式併入本文中。

單倍型分析

【0529】 在腫瘤組織樣本中發現兩個單倍型之間的甲基化概況存在差異。因此，單倍型之間的甲基化不平衡可用於確定癌症或其他病症的等級分類。單倍型之不平衡亦可用於鑑別胎兒對單倍型之遺傳。胎兒病症亦可經由分析單倍型之間的甲基化不平衡來鑑別。細胞 DNA 可用於分析單倍型之甲基化程度。

單倍型相關之甲基化分析

【0530】 單分子即時定序技術可鑑別個別 SNP。自單分子即時定序孔產生之長讀段（例如長達數千鹼基）允許藉由利用每個一致讀段中存在之單倍型資訊對基因體中之變異進行分期（Edge 等人《基因體研究》 2017;27:801-812；Wenger 等人 《自然生物技術》 2019;37:1155-1162）。如圖 77 所示，可自 CCS 與各別單倍型上之對偶基因相連之 CpG 位點的甲基化程度來分析單倍型之甲基化概況。此分期甲基化單倍型分析可用於解決關於同源染色體之兩個複本在不同的臨床相關病況（諸如癌症）中是否具有相似或不同的甲基化模式的問題。在一個實施例中，單倍型甲基化將為由分配給該單倍型之許多 DNA 片段貢獻的總甲基化程度。單倍型可為不同大小的塊，包括但不限於 50 nt、100 nt、200 nt、300 nt、400 nt、500 nt、1 knt、2 knt、3 knt、4 knt、5 knt、10 knt、20 knt、30 knt、40 knt、50 knt、100 knt、200 knt、300 knt、400 knt、500 knt、1 Mnt、2 Mnt 及 3 Mnt。

基於單倍型之相對甲基化不平衡分析

【0531】 **圖 104** 展示基於單倍型之相對甲基化不平衡分析。藉由分析單分子即時定序結果確定單倍型（亦即 Hap I 及 Hap II）。使用彼等根據圖 77 中所述之方法確定甲基化概況的單倍型相關片段，可確定與每個單倍型相關之甲基化模式。從而，可比較 Hap I 與 Hap II 之間的甲基化模式。

【0532】 為了量化 Hap I 與 Hap II 之間的甲基化差異，計算 Hap I 與 Hap II 之間甲基化程度的差異（10}ΔF）。差異ΔF 計算如下：

$$\Delta F = M_{HapI} - M_{HapII}$$

其中ΔF 代表 Hap I 與 Hap II 之間甲基化程度的差異，M_{HapI}及M_{HapII}分別代表 Hap I 及 Hap II 之甲基化程度。ΔF 為正值表明 Hap I 之 DNA 甲基化程度高於 Hap II。

基於單倍型之 HCC 腫瘤 DNA 的相對甲基化不平衡分析

【0533】 在一個實施例中，單倍型甲基化分析可用於檢測癌症基因體中之甲基化畸變。舉例而言，將分析基因體區域內之兩個單倍型之間的甲基化變化。基因體區域內之單倍型定義為單倍型區塊。單倍型區塊可視為染色體上已定相之一組對偶基因。在一些實施例中，單倍型區塊將根據支持染色體上兩個對偶基因物理連接之一組序列資訊儘可能地延長。對於病例 3033，吾等自相鄰正常組織 DNA 之定序結果中獲得 97,475 個單倍型區塊。單倍型區塊的中位數大小為 2.8 kb。25%之單倍型區塊的大小大於 8.2 kb。單倍型區塊的最大大小為 282.2 kb。資料集由 Sequel II Sequencing Kit 1.0 製備之 DNA 生成。

【0534】 出於說明之目的，吾等使用許多準則來鑑別潛在的單倍型區塊，該等單倍型區塊表現出與相鄰的非腫瘤組織 DNA 相比，腫瘤 DNA 中 Hap I 與 Hap II 之間的差異性甲基化。準則為：(1)所分析之單倍型區塊含有至少 3 個

三個 CCS 序列，其分別自三個定序孔產生；(2)相鄰非腫瘤組織 DNA 中 Hap I 與 Hap II 之間甲基化程度的絕對差異小於 5%；(3)腫瘤組織 DNA 中 Hap I 與 Hap II 之間甲基化程度的絕對差異大於 30%。吾等鑑別出 73 個符合上述準則之單倍型區塊。

【0535】 **圖 105A** 及 **105B** 為病例 TBR3033 之 73 個單倍型區塊的表格，顯示與相鄰非腫瘤組織 DNA 相比，HCC 腫瘤 DNA 中 Hap I 與 Hap II 之間的差異性甲基化程度。第一行顯示與單倍型區塊相關聯之染色體。第二行顯示染色體內單倍型區塊之起始座標。第三行顯示單倍型區塊之結束座標。第四行展示單倍型區塊之長度。第四行列出單倍型區塊 id。第五行顯示與腫瘤組織相鄰的非腫瘤組織中 Hap I 的甲基化程度。第六行顯示非腫瘤組織中 Hap II 的甲基化程度。第七行顯示腫瘤組織中 Hap I 的甲基化程度。第八行顯示腫瘤組織中 Hap II 的甲基化程度。

【0536】 與 73 個單倍型區塊顯示腫瘤組織 DNA 之單倍型之間的甲基化程度差異大於 30%相反，僅一個單倍型區塊顯示非腫瘤組織 DNA 之差異大於 30%，但腫瘤組織 DNA 之差異小於 5%。在一些實施例中，可使用另一組準則來鑑別呈現差異性甲基化之單倍型區塊。可使用其他最大及最小臨限值差異。舉例而言，最小臨限值差異可為 10%、15%、20%、25%、30%、35%、40%、45%、50%或更多。作為實例，最大臨限值差異可為 1%、5%、10%、15%、20%或 30%。此等結果表明，單倍型之間甲基化差異的變化可充當一種新的生物標誌物，用於癌症診斷、檢測、監測、預後及指導治療。

【0537】 在一些實施例中，當研究甲基化模式時，長的單倍型區塊將電腦模擬分割成較小的塊。

【0538】 對於病例 3032，吾等自相鄰非腫瘤組織 DNA 之定序結果中獲得

61,958 個單倍型區塊。單倍型區塊的中位數大小為 9.3 kb。25%之單倍型區塊的大小大於 27.6 kb。單倍型區塊的最大大小為 717.8 kb。作為說明，吾等使用上述相同的三個準則來鑑別潛在的單倍型區塊，該等單倍型區塊表現出與相鄰的正常組織 DNA 相比，腫瘤 DNA 中 Hap I 與 Hap II 之間的差異性甲基化。吾等鑑別出 20 個符合上述準則之單倍型區塊。資料集由 Sequel II Sequencing Kit 1.0 製備之 DNA 生成。

【0539】 **圖 106** 為病例 TBR3032 之 20 個單倍型區塊的表格，顯示與相鄰正常組織 DNA 相比，腫瘤 DNA 中 Hap I 與 Hap II 之間的差異性甲基化程度。第一行顯示與單倍型區塊相關聯之染色體。第二行顯示染色體內單倍型區塊之起始座標。第三行顯示單倍型區塊之結束座標。第四行展示單倍型區塊之長度。第四行列出單倍型區塊 id。第五行顯示與腫瘤組織相鄰的非腫瘤組織中 Hap I 的甲基化程度。第六行顯示非腫瘤組織中 Hap II 的甲基化程度。第七行顯示腫瘤組織中 Hap I 的甲基化程度。第八行顯示腫瘤組織中 Hap II 的甲基化程度。

【0540】 與圖 106 中 20 個單倍型區塊顯示 HCC 腫瘤組織中之差異相反，僅一個單倍型區塊顯示在非腫瘤組織中之差異大於 30%，但在腫瘤組織中之差異小於 5%。此等結果進一步表明，單倍型之間甲基化差異的變化將充當一種新的生物標誌物，用於癌症診斷、檢測、監測、預後及指導治療。對於其他實施例，可使用其他準則來鑑別呈現差異性甲基化之單倍型區塊。

基於單體型之其他腫瘤類型 DNA 的相對甲基化不平衡分析

【0541】 如上所述，單倍型之間甲基化程度的分析顯示，與配對的相鄰非腫瘤組織相比，HCC 腫瘤組織具有更多表現出甲基化不平衡的單倍型區塊。作為一個實例，腫瘤組織中顯示甲基化不平衡之單倍型區塊的準則為：(1)所分析之單倍型區塊含有至少三個 CCS 序列，其自三個定序孔產生；(2)相鄰非腫瘤

組織 DNA 或基於歷史資料之正常組織 DNA 中 Hap I 與 Hap II 之間甲基化程度的絕對差異小於 5%；(3)腫瘤組織 DNA 中 Hap I 與 Hap II 之間甲基化程度的絕對差異大於 30%。納入準則(2)係因為非腫瘤/正常組織在甲基化程度上顯示單倍型不平衡可能指示印記區域而非腫瘤區域。非腫瘤組織中顯示甲基化不平衡之單倍型區塊的準則為：(1)所分析之單倍型區塊含有至少三個 CCS 序列，其自三個定序孔產生；(2)相鄰非腫瘤組織 DNA 或基於歷史資料之正常組織 DNA 中 Hap I 與 Hap II 之間甲基化程度的絕對差異大於 30%；(3)腫瘤組織 DNA 中 Hap I 與 Hap II 之間甲基化程度的絕對差異小於 5%。

【0542】 在其他實施例中，可使用其他準則。舉例而言，為了鑑別不平衡單倍型 I 癌症基因體，非腫瘤組織中 Hap I 與 Hap II 之間的甲基化程度差異可小於 1%、5%、10%、20%、40%、50%或 60%等，而腫瘤組織中 Hap I 與 Hap II 之間的甲基化程度差異可大於 1%、5%、10%、20%、40%、50%或 60%等。為了鑑別不平衡單倍型 I 非癌症基因體，非腫瘤組織中 Hap I 與 Hap II 之間的甲基化程度差異可大於 1%、5%、10%、20%、40%、50%或 60%等，而腫瘤組織中 Hap I 與 Hap II 之間的甲基化程度差異可小於 1%、5%、10%、20%、40%、50%或 60%等。

【0543】 **圖 107A** 為基於 Sequel II Sequencing Kit 2.0 生成之資料，總結腫瘤與相鄰非腫瘤組織之間顯示兩個單倍型之間甲基化不平衡之單倍型區塊的數量的表格。第一行列出組織類型。第二行列出腫瘤組織中顯示兩個單倍型之間甲基化不平衡之單倍型區塊的數量。第三行列出配對的相鄰非腫瘤組織中顯示兩個單倍型之間甲基化不平衡之單倍型區塊的數量。列顯示腫瘤組織比配對的相鄰非腫瘤組織具有更多的顯示兩個單倍型之間甲基化不平衡的單倍型區塊。

【0544】 此分析中涉及之單倍型區塊的中位數長度為 15.7 kb（IQR：

10.3-26.1 kb）。包括肝臟之 HCC 結果，資料顯示 7 種組織類型之腫瘤組織具有更多的具有甲基化不平衡之單倍型區塊。除肝臟之外，其他組織包括結腸、乳房、腎臟、肺、前列腺及胃組織。因此，在一些實施例中，吾人可使用具有甲基化不平衡之單倍型區塊的數量檢測患者是否患有腫瘤或癌症。

【0545】 **圖 107B** 為基於 Sequel II Sequencing Kit 2.0 生成之資料，總結不同腫瘤階段之腫瘤組織中顯示兩個單倍型之間之甲基化不平衡之單倍型區塊的數量的表格。第一行顯示具有腫瘤之組織類型。第二行顯示腫瘤組織中兩個單倍型之間甲基化不平衡之單倍型區塊的數量。第三行列出使用惡性腫瘤之 TNM 分類的腫瘤分期資訊。T3 及 T3a 為大小大於 T2 之腫瘤。

【0546】 該表顯示更多的單倍型區塊，顯示乳房及腎臟之較大腫瘤的甲基化不平衡。舉例而言，對於乳腺組織，分類為腫瘤等級 T3（TNM 分期）、ER 陽性且表現出 *ERBB2* 擴增之組織的顯示甲基化不平衡之單倍型區塊（57）比分類為腫瘤等級 T2（TNM 分期）、PR（孕酮受體）/ER（雌激素受體）陽性且無 *ERBB2* 擴增之組織的單倍型區塊（18）更多。對於腎臟組織，分類為腫瘤等級 T3a 之組織之顯示甲基化不平衡的單倍型區塊（68）比分類為腫瘤等級 T2 之組織的單倍型區塊（0）更多。

【0547】 在一些實施例中，吾人可利用顯示甲基化不平衡之單倍型區塊對腫瘤進行分類，且與其臨床行為（例如進展、預後或治療反應）相關。此等資料表明，基於單倍型之甲基化不平衡程度可充當腫瘤之分類器，且可併入臨床研究或試驗或最終的臨床服務中。腫瘤之分類可包括大小及嚴重程度。

基於單倍型之母體血漿游離 DNA 的甲基化分析

【0548】 可確定父母雙方或父母一方的單倍型。單倍型分析方法可包括長讀段單分子定序、連鎖短讀段定序（例如 10x 基因體學）、長程單分子 PCR 或

群體推斷。若已知父本單倍型，則可藉由連接多個游離 DNA 分子之甲基化概況來組裝游離胎兒 DNA 甲基化體，每個游離 DNA 分子含有至少一個沿著父本單倍型存在之父本特異性 SNP 對偶基因。換言之，父本單倍型用作連接胎兒特異性讀段序列之骨架。

【0549】 **圖 108** 展示單倍型之相對甲基化不平衡的分析。若已知母本單倍型，則兩個單倍型（亦即 Hap I 及 Hap II）之間的甲基化不平衡可用於確定胎兒遺傳之母本單倍型。如圖 108 所示，使用單分子即時定序技術對來自孕婦之血漿 DNA 分子進行定序。根據本文之揭示內容可確定甲基化及對偶基因資訊。在一個實施例中，與致病基因相關之 SNP 指定為 Hap I。若胎兒遺傳 Hap I，則與攜帶 Hap II 對偶基因之片段相比，母體血漿中會存在更多攜帶 Hap I 對偶基因之片段。來源於胎兒之 DNA 片段的低甲基化會使 Hap I 之甲基化程度低於 Hap II 之甲基化程度。因此，若 Hap I 之甲基化程度顯示低於 Hap II，則胎兒遺傳母本 Hap I 之可能性較大。否則，胎兒遺傳母本 Hap II 之可能性較大。在臨床實踐中，基於單倍型之甲基化不平衡分析可用於確定未出生的胎兒是否遺傳與遺傳病症相關聯之母本單倍型，該等遺傳病症例如但不限於單基因病症，包括 X 脆折症候群、肌肉營養不良、亨廷頓氏病（Huntington disease）或 β-地中海貧血症。

例示性病症分類方法

【0550】 **圖 109** 展示對具有第一單倍型及第二單倍型之生物體的病症進行分類的例示性方法 1090。方法 1090 涉及比較兩個單倍型之間的相對甲基化程度。

【0551】 在方塊 1091，分析來自生物樣本之 DNA 分子，以鑑別其在對應於生物體之參考基因體中的位置。DNA 分子可為細胞 DNA 分子。舉例而言，可

對 DNA 分子進行定序以獲得序列讀段，且可將序列讀段相對於參考基因體進行定位（排比）。若生物體為人類，則參考基因體將為參考人類基因體，可能來自特定亞群。作為另一個實例，可用不同的探針（例如按照 PCR 或其他擴增方法）分析 DNA 分子，其中每個探針對應於基因體位置，該位置可覆蓋異型接合子及一或多個 CpG 位點，如下所述。

【0552】 此外，可分析 DNA 分子以確定 DNA 分子之各別對偶基因。舉例而言，DNA 分子之對偶基因可自定序獲得之序列讀段或自與 DNA 分子雜交之特定探針來確定，其中兩種技術均可提供序列讀段（例如，當存在雜交時，探針可視為序列讀段）。可確定 DNA 分子之一或多個位點（例如 CpG 位點）中之每一者的甲基化狀態。

【0553】 在方塊 1092，鑑別第一染色體區域之第一部分的一或多個異型接合基因座。每個異型接合基因座可包括第一單倍型中之相應第一對偶基因及第二單倍型中之相應第二對偶基因。一或多個異型接合基因座可為第一複數個異型接合基因座，其中第二複數個異型接合基因座可對應於不同的染色體區域。

【0554】 在方塊 1093，鑑別第一組複數個 DNA 分子。複數個 DNA 分子中之每一者均位於方塊 1096 之異型接合基因座中之任一者，且包括相應的第一對偶基因，因此 DNA 分子可鑑別為對應於第一單倍型。DNA 分子有可能位於一個以上的異型接合基因座，但是通常一個讀段將僅包括一個異型接合基因座。第一組 DNA 分子中之每一者亦包括 N 個基因體位點中之至少一者，其中該等基因體位點用於量測甲基化程度。N 為整數，例如大於或等於 1、2、3、4、5、10、20、50、100、200、500、1,000、2,000 或 5,000。因此，DNA 分子之讀段可表示覆蓋 1 個位點、2 個位點等。1 個基因體位點可包括存在 CpG 核苷酸之

位點。

【0555】 在方塊 1094，使用第一組複數個 DNA 分子確定第一單倍型之第一部分的第一甲基化程度。第一甲基化程度可藉由本文所述之任何方法確定。第一部分可對應於單個位點或包括許多位點。第一單倍型之第一部分可長於或等於 1 kb。舉例而言，第一單倍型之第一部分可長於或等於 1 kb、5 kb、10 kb、15 kb 或 20 kb。甲基化資料可為來自細胞 DNA 之資料。

【0556】 在一些實施例中，可針對第一單倍型之複數個部分確定複數個第一甲基化程度。每一部分之長度可大於或等於 5 kb 或本文所揭示之第一單倍型之第一部分的任何大小。

【0557】 在方塊 1095，鑑別第二組複數個 DNA 分子。複數個 DNA 分子中之每一者均位於方塊 1096 之異型接合基因座中之任一者，且包括相應的第二對偶基因，因此 DNA 分子可鑑別為對應於第二單倍型。第二組 DNA 分子中之每一者亦包括 N 個基因體位點中之至少一者，其中該等基因體位點用於量測甲基化程度。

【0558】 在方塊 1096，使用第二組複數個 DNA 分子確定第二單倍型之第一部分的第二甲基化程度。第二甲基化程度可藉由本文所述之任何方法確定。第二單倍型之第一部分可長於或等於 1 kb 或第一單倍型之第一部分的任何大小。第一單倍型之第一部分可與第二單倍型之第一部分互補。第一單倍型之第一部分及第二單倍型之第一部分可形成環形 DNA 分子。第一單倍型之第一部分的第一甲基化程度可使用來自環形 DNA 分子之資料來確定。舉例而言，環形 DNA 之分析可包括圖 1、圖 2、圖 4、圖 5、圖 6、圖 7、圖 8、圖 50 或圖 61 所述之分析。

【0559】 環形 DNA 分子可藉由使用 Cas9 複合物切割雙股 DNA 分子形成

經切割之雙股 DNA 分子來形成。可將髮夾轉接子連接至經切割之雙股 DNA 分子的末端。在實施例中，雙股 DNA 分子之兩端可經切割及連接。舉例而言，切割、連接及後續分析可如圖 91 所述進行。

【0560】 在一些實施例中，可針對第二單倍型之複數個部分確定複數個第二甲基化程度。第二單倍型之複數個部分中的每一部分可與第一單倍型之複數個部分中的一部分互補。

【0561】 在方塊 1097，使用第一甲基化程度及第二甲基化程度計算參數之值。該參數可為分離值。分離值可為兩個甲基化程度之間的差或兩個甲基化程度之比率。

【0562】 若使用第二單倍型之複數個部分，則對於第二單倍型之複數個部分中之每一部分，可使用第二單倍型之一部分的第二甲基化程度及使用第一單倍型之互補部分的第一甲基化程度計算分離值。可將分離值與閾值進行比較。

【0563】 閾值可自未患病症之組織確定。參數可為第二單倍型之分離值超過閾值之部分的數量。舉例而言，第二單倍型之分離值超過閾值之部分的數量可類似於圖 105A、圖 105B 及圖 106 中所示之具有大於 30%之差異的區域的數量。在圖 105A、圖 105B 及圖 106 中，分離值為比率，且閾值為 30%。在一些實施例中，閾值可自患有病症之組織確定。

【0564】 在另一個實例中，可對每一部分之分離值進行彙總，例如求和，其可藉由加權和或各個分離值之函數之和來完成。此類彙總可提供參數之值。

【0565】 在方塊 1098，將參數之值與參考值進行比較。參考值可使用無病症之參考組織來確定。參考值可為分離值。舉例而言，參考值可表示兩個單

倍型之甲基化程度之間應不存在顯著差異。舉例而言，參考值可為 0 之統計差異或約 1 之比率。當使用複數個部分時，參考值可為健康生物體中兩個單倍型顯示超過閾值之分離值之部分的數量。在一些實施例中，參考值可使用患有病症之參考組織來確定。

【0566】 在方塊 1099，使用參數之值與參考值之比較確定生物體中病症之分類。若參數之值超過參考值，則可確定該病症存在或更可能存在。該病症可包括癌症。癌症可為本文所述之任何癌症。病症之分類可為病症之可能性。病症之分類可包括病症之嚴重程度。舉例而言，指示具有單倍型不平衡之部分數量較大的較大參數值可指示更嚴重的癌症形式。

【0567】 雖然圖 109 所述之方法涉及病症之分類，但類似的方法可用於確定可能由單倍型之間甲基化程度的不平衡導致的任何病況或特徵。舉例而言，來自胎兒 DNA 之單倍型的甲基化程度可能低於來自母體 DNA 之單倍型的甲基化。甲基化程度可用於將核酸分類為母體或胎兒的。

【0568】 當病症為癌症時，腫瘤之不同染色體區域可能表現出甲基化之此類差異。視哪些區域受影響而定，可提供不同的治療。此外，具有表現出此類甲基化差異之不同區域的個體可具有不同的預後。

【0569】 具有足夠分離（例如，大於閾值）之染色體區域（部分）可鑑別為異常（或具有異常分離）。可將異常區域之模式（潛在地說明哪種單倍型高於另一種單倍型）與參考模式（例如，自患有癌症、可能為特定類型之癌症之個體或健康個體確定）進行比較。若兩個模式在臨限值內（例如，小於指定數量之區域/部分不同）與具有特定分類之參考模式相同，則可將個體鑑別為具有該分類之病症。此類分類可包括印記病症，例如，如本文所述。

雜合分子之單分子甲基化分析

【0570】 為了進一步評估本文所揭示之關於測定核酸之鹼基修飾之實施例的效能及效用，吾等人為地創建人類及小鼠雜合 DNA 片段，其中人類部分為甲基化的，而小鼠部分為未甲基化的，反之亦然。確定雜合或嵌合 DNA 分子之接合點可允許檢測包括癌症之各種病症或疾病的基因融合。

創建人類及小鼠雜合 DNA 片段之方法

【0571】 此部分描述創建雜合 DNA 片段，且隨後描述確定片段之甲基化概況的程序。

【0572】 在一個實施例中，人類 DNA 係經由全基因體擴增而擴增，從而將消除人類基因體中之原始甲基化特徵，因為全基因體擴增將不會保留甲基化狀態。全基因體擴增可使用抗外切核酸酶之硫代磷酸酯修飾的簡併六聚體作為引子來進行，該等引子可在基因體上隨機結合，從而使聚合酶（例如 Φ29 DNA 聚合酶）無需熱循環即可擴增 DNA。經擴增之 DNA 產物將為未甲基化的。經擴增之人類 DNA 分子用 M.SssI（一種 CpG 甲基轉移酶）進一步處理，該酶理論上將使雙股、非甲基化或半甲基化的 DNA 中 CpG 上下文處之所有胞嘧啶完全甲基化。因此，此類由 M.SssI 處理之經擴增之人類 DNA 將變成甲基化之 DNA 分子。

【0573】 相比之下，對小鼠 DNA 進行全基因體擴增，因此將產生未甲基化之小鼠 DNA 片段。

【0574】 **圖 110** 展示創建人類-小鼠雜合 DNA 片段，其中人類部分為甲基化的，而小鼠部分為未甲基化的。經填充之棒棒糖代表甲基化之 CpG 位點。未填充之棒棒糖代表未甲基化之 CpG 位點。帶對角線條紋之粗條 11010 代表甲基化之人類部分。帶垂直條紋之粗條 11020 代表未甲基化之小鼠部分。

【0575】 為了產生雜合人類-小鼠 DNA 分子，在一個實施例中，將全基因體擴增且經 M.SssI 處理之 DNA 分子進一步用 *Hind*III 及 *Nco*I 消化，產生黏性末

端以便於下游連接。在一個實施例中，將甲基化之人類 DNA 片段與未甲基化之小鼠 DNA 片段進一步以等莫耳比混合。對此類人類-小鼠 DNA 混合物進行連接過程，在一個實施例中，該過程由 DNA 連接酶在 20℃下介導 15 分鐘。如圖 110 所示，此連接反應將產生 3 種類型之所得分子，包括人類-小鼠雜合 DNA 分子（a：人類-小鼠雜合片段）；僅人類 DNA 分子（b：人類-人類連接，及 c：未連接之人類 DNA）；及僅小鼠 DNA 分子（d：小鼠-小鼠連接及 e：未連接之小鼠 DNA）。對連接後之 DNA 產物進行單分子即時定序。根據本文提供之揭示內容分析定序結果，以確定甲基化狀態。

【0576】 **圖 111** 展示創建人類-小鼠雜合 DNA 片段，其中人類部分為未甲基化的，而小鼠部分為甲基化的。經填充之棒棒糖代表甲基化之 CpG 位點。未填充之棒棒糖代表未甲基化之 CpG 位點。帶對角線條紋之粗條 11110 代表甲基化之小鼠部分。帶垂直條紋之粗條 11120 代表未甲基化之人類部分。

【0577】 對於圖 111 中之實施例，小鼠 DNA 分子係經由全基因體擴增而擴增，從而將消除小鼠基因體中之原始甲基化。經擴增之 DNA 產物將為未甲基化的。經擴增之小鼠 DNA 將用 M.SssI 進一步處理。因此，此類由 M.SssI 處理之經擴增之小鼠 DNA 將變成甲基化之 DNA 分子。相反，對人類 DNA 片段進行全基因體擴增，從而將獲得未甲基化之人類片段。在一個實施例中，將甲基化之人類片段與未甲基化之片段進一步以等莫耳比混合。對此類人類-小鼠 DNA 混合物進行由 DNA 連接酶介導之連接過程。如圖 111 所示，此連接反應將產生 3 種類型之所得分子，包括人類-小鼠雜合 DNA 分子（a：人類-小鼠雜合片段）；僅人類 DNA 分子（b：人類-人類連接，及 c：未連接之人類 DNA）；及僅小鼠 DNA 分子（d：小鼠-小鼠連接及 e：未連接之小鼠 DNA）。對連接後之 DNA 產物進行單分子即時定序。根據本文提供之揭示內容分析定序結果，以確定甲基化

狀態。

【0578】 根據圖 110 所示之實施例，吾等製備人工 DNA 混合物（命名為樣本 MIX01），其包含人類-小鼠雜合 DNA 分子、僅人類 DNA 及僅小鼠 DNA，其中人類相關之 DNA 分子為甲基化的，而小鼠 DNA 分子為未甲基化的。對於樣本 MIX01，吾等獲得 1.66 億個子讀段，其可與人類或小鼠參考基因體進行排比，或部分與人類基因體進行排比且部分與小鼠基因體進行排比。此等子讀段係自大約 500 萬個 Pacific Biosciences 單分子即時（SMRT）定序孔中產生。單分子即時定序孔中之每個分子平均定序 32 次（範圍：1-881 次）。

【0579】 為了確定雜合片段中之人類 DNA 及小鼠 DNA 部分，吾等首先藉由組合孔中所有相關子讀段之核苷酸資訊來構築一致序列。吾等總共獲得樣本 MIX01 之 3,435,657 個一致序列。資料集由 Sequel II Sequencing Kit 1.0 製備之 DNA 生成。

【0580】 將一致序列與包含人類及小鼠參考基因體之參考基因體進行排比。吾等獲得 320 萬個排比的一致序列。其中，39.6%之序列分類為僅人類 DNA 類型；26.5%之序列分類為僅小鼠 DNA 類型，且 30.2%之序列分類為人類-小鼠雜合 DNA。

【0581】 圖 112 展示連接後之 DNA 混合物（樣本 MIX01）中 DNA 分子之長度分佈。x 軸顯示 DNA 分子之長度。y 軸顯示與 DNA 分子長度相關聯之頻率。如圖 112 所示，人類-小鼠雜合 DNA 分子具有較長的長度分佈，其與該等分子為至少兩種類型分子之組合的事實一致。

【0582】 **圖 113** 展示將第一 DNA（A）及第二 DNA（B）接合在一起之接合區域。DNA（A）及 DNA（B）可用限制酶消化。在一個實施例中，為了提高使用交錯末端之連接效率，吾等使用分別識別 A^AGCTT 及 C^CATGG 位點的限制

酶*Hind*III及*Nco*I，在連接步驟之前消化人類及小鼠DNA。隨後可連接DNA（A）及DNA（B）。在698,492個具有接合區域之人類-小鼠雜合DNA分子中，吾等發現88%之人類-小鼠雜合DNA分子攜帶A^AGCTT及C^CATGG之酶識別位點，進一步表明人類及小鼠DNA片段之間已發生連接。該接合區域定義為第一DNA片段及第二DNA片段以物理方式接合在一起的區域或位點。由於接合點包括DNA（A）及 DNA（B）共同的序列，因此不能僅藉由序列來確定對應於接合點之一股的一部分為DNA（A）或DNA（B）之一部分。分析對應於接合點之一股之一部分的甲基化模式或密度可用於確定該部分係來自 DNA（A）抑或 DNA（B）。舉例而言，DNA（A）可為病毒 DNA，DNA（B）可為人類 DNA。精確接合點之測定可告知此類整合DNA是否以及如何破壞蛋白質結構。

【0583】 **圖 114** 展示 DNA 混合物之甲基化分析。帶對角線條紋之條 11410 指示在排比分析中觀察到的接合區域，該區域將在連接之前藉由限制酶處理引入。「RE位點」表示限制酶（RE）識別位點。

【0584】 如圖 114 所示，在一個實施例中，將排比之一致序列分為如下三個類別：

【0585】 (1)參照一或多個排比準則，經定序之DNA僅與人類參考基因體進行排比，而不與小鼠參考基因體進行排比。在一個實施例中，一個排比準則可定義為但不限於經定序之 DNA 之 100%、95%、90%、80%、70%、60%、50%、40%、30%或 20%的連續核苷酸可與人類參考基因體進行排比。在一個實施例中，一個排比準則將為未與人類參考基因體進行排比之經定序片段的剩餘部分不能與小鼠參考基因體進行排比。在一個實施例中，一個排比準則為經定序之DNA可與人類參考基因體中之單個區域進行排比。在一個實施例中，排比可為完美的。在其他實施例中，排比可容納核苷酸差異，包括插入、錯配及缺

失，其限制條件為此類差異小於某些臨限值，諸如但不限於排比序列長度之1%、2%、3%、4%、5%、10%、20%或30%。在另一個實施例中，排比可為參考基因體中一個以上的位置。在其他實施例中，與參考基因體中一或多個位點之排比可以概率方式（例如指示錯誤排比之幾率）陳述，且概率量測可用於後續處理。

【0586】 (2)參照一或多個排比準則，經定序之DNA僅與小鼠參考基因體進行排比，而不與人類參考基因體進行排比。在一個實施例中，一個排比準則可定義為但不限於經定序之 DNA 之 100%、95%、90%、80%、70%、60%、50%、40%、30%或 20%的連續核苷酸可與小鼠參考基因體進行排比。在一個實施例中，一個排比準則將為剩餘部分不能與人類參考基因體進行排比。在一個實施例中，一個排比準則為經定序之DNA可與小鼠參考基因體中之單個區域進行排比。在一個實施例中，排比可為完美的。在其他實施例中，排比可容納核苷酸差異，包括插入、錯配及缺失，其限制條件為此類差異小於某些臨限值，諸如但不限於排比序列長度之 1%、2%、3%、4%、5%、10%、20%或 30%。在另一個實施例中，排比可為參考基因體中一個以上的位置。在其他實施例中，與參考基因體中一或多個位點之排比可以概率方式（例如指示錯誤排比之幾率）陳述，且概率量測可用於後續處理。

【0587】 (3)經定序之DNA的一部分與人類參考基因體進行唯一排比，而另一部分與小鼠參考基因體進行唯一排比。在一個實施例中，若在連接之前使用限制酶，則在排比分析中會觀察到接合區域，對應於限制酶切割位點。在一些實施例中，由於定序及排比誤差，人類及小鼠DNA部分之間的接合區域僅可在一定區域內大致確定。在一些實施例中，若連接涉及無限制酶切割之分子（例如，若存在平端連接），則限制酶識別位點將無法在人類-小鼠雜合 DNA 片

段之接合區域中觀察到。

【0588】 脈衝間持續時間（IPD）、脈衝寬度（PW）及 CpG 位點周圍的序列上下文係自對應於一致序列之彼等子讀段獲得。從而，可根據本揭示案中存在之實施例確定每個 DNA 分子之甲基化，包括僅人類、僅小鼠及人類-小鼠雜合 DNA。

甲基化結果

【0589】 此部分描述雜合 DNA 片段之甲基化結果。甲基化密度可用於鑑別雜合 DNA 片段之不同部分的起源。

【0590】 **圖 115** 展示樣本 MIX01 中 CpG 位點之甲基化概率的盒狀圖。x 軸顯示樣本 MIX01 中存在之三種不同的分子：僅人類 DNA、僅小鼠 DNA 及人類-小鼠雜合 DNA（包括人類部分及小鼠部分）。y 軸顯示特定單個 DNA 分子之 CpG 位點的甲基化概率。此分析係以人類 DNA 甲基化程度較高而小鼠 DNA 未甲基化程度較高之方式來進行。

【0591】 如圖 115 所示，僅人類 DNA 中 CpG 位點之甲基化概率（中位數：0.66；範圍：0-1）顯著高於僅小鼠 DNA 中 CpG 位點之甲基化概率（中位數：0.06；範圍：0-1）（*P* 值＜0.0001）。此等結果與分析設計一致，其中人類 DNA 由於 CpG 甲基轉移酶 M.SssI 之處理而甲基化程度較高，而小鼠 DNA 由於在全基因體擴增期間不能保留甲基化而未甲基化程度較高。此外，人類-小鼠雜合 DNA 分子之人類 DNA 部分內的 CpG 位點與小鼠 DNA 部分內的 CpG 位點（中位數：0.06；範圍：0-1）相比，顯示甲基化概率較高（中位數：0.69；範圍：0-1）（*P* 值<0.0001）。此等資料表明，所揭示之方法可準確地確定 DNA 分子以及 DNA 分子內之區段的甲基化狀態。

【0592】 甲基化概率係指基於所使用之統計模型估計的單分子內特定

CpG 位點之概率。概率為 1 表示，基於統計模型，使用所測得參數（包括 IPD、PW 及序列上下文）之 100%的 CpG 位點將被甲基化。概率為 0 表示，根據統計模型，使用所測得參數（包括 IPD、PW 及序列上下文）之 0%的 CpG 位點將被甲基化。換言之，使用所測得參數之所有 CpG 位點將為未甲基化的。圖 115 展示甲基化概率之分佈，僅人類 DNA 及人類部分之分佈比小鼠對應物更寬。亞硫酸氫鹽定序用於量測類似樣本之甲基化，以確認甲基化不完全，結果如下所示。圖 115 展示人類與小鼠 DNA 中甲基化之間的顯著差異。

【0593】 根據圖 111 所示之實施例，吾等製備人工 DNA 混合物（命名為樣本 MIX02），其包含人類-小鼠雜合 DNA 分子、僅人類 DNA 及僅小鼠 DNA，其中人類部分為未甲基化的，而小鼠部分為甲基化的。對於樣本 MIX02，吾等獲得 1.4 億個子讀段，其可與人類或小鼠參考基因體進行排比，或部分與人類基因體進行排比且部分與小鼠基因體進行排比。此等子讀段係自大約 500 萬個 Pacific Biosciences 單分子即時（SMRT）定序孔中產生。單分子即時定序孔中之每個分子平均定序 27 次（範圍：1-1028 次）。

【0594】 吾等亦藉由組合孔中所有相關子讀段之核苷酸資訊來構築一致序列。吾等總共獲得樣本 MIX02 之 3,265,487 個一致序列。使用 BWA 將一致序列與包含人類及小鼠參考基因體之參考基因體進行排比（Li H 等人, 《生物資訊學》2010;26(5):589-595）。吾等獲得 300 萬個排比的一致序列。其中，30.5%分類為僅人類 DNA 類型；32.2%分類為僅小鼠 DNA 類型，33.8%分類為人類-小鼠雜合 DNA。資料集由 Sequel II Sequencing Kit 1.0 製備之 DNA 生成。

【0595】 **圖 116** 展示樣本 MIX02 交叉連接後 DNA 混合物中 DNA 分子之長度分佈。x 軸顯示 DNA 分子之長度。y 軸顯示與 DNA 分子長度相關聯之頻率。如圖 116 所示，人類-小鼠雜合 DNA 分子具有較長的長度分佈，其與該等分

子經由一個以上分子連接產生的事實一致。

【0596】 **圖 117** 展示樣本 MIX02 中 CpG 位點之甲基化概率的盒狀圖。根據本文所述之方法確定甲基化狀態。x 軸顯示樣本 MIX01 中存在之三種不同的分子：僅人類 DNA、僅小鼠 DNA 及人類-小鼠雜合 DNA（包括人類部分及小鼠部分）。y 軸顯示 CpG 位點之甲基化概率。此分析係以人類 DNA 未甲基化而小鼠 DNA 甲基化之方式來進行。

【0597】 如圖 117 所示，僅人類 DNA 中 CpG 位點之甲基化概率（中位數：0.06；範圍：0-1）顯著低於僅小鼠 DNA 中 CpG 位點之甲基化概率（中位數：0.93；範圍：0-1）（*P* 值<0.0001）。此等結果與分析設計一致，其中人類 DNA 由於在全基因體擴增期間不能保留甲基化而未甲基化程度較高，而小鼠 DNA 由於 CpG 甲基轉移酶 M.SssI 之處理而甲基化程度較高。此外，人類-小鼠雜合 DNA 分子之人類 DNA 部分內的 CpG 位點與小鼠 DNA 部分內的 CpG 位點（中位數：0.93；範圍：0-1）相比，顯示甲基化概率較低（中位數：0.07；範圍：0-1）（*P* 值<0.0001）。此等資料表明，所揭示之方法可準確地確定 DNA 分子以及 DNA 分子內之區段的甲基化狀態。

【0598】 根據本揭示案中之實施例，亞硫酸氫鹽定序用於量測人類-小鼠雜合片段之甲基化，其甲基化模式由單分子即時定序確定。將樣本 MIX01（人類 DNA 經甲基化而小鼠 DNA 未甲基化）及 MIX02（人類 DNA 未甲基化而小鼠 DNA 經甲基化）經由音波處理剪切，得到中位數 DNA 片段大小為 196 bp（四分位數範圍：161-268）之混合物。隨後在 MiSeq 平台（Illumina）進行雙端亞硫酸氫鹽定序（BS-Seq），讀段長度為 300 bp x2。吾等分別獲得 MIX01 及 MIX02 之 370 萬及 290 萬個定序片段，該等片段與人類或小鼠參考基因體進行排比，或部分與人類基因體進行排比且部分與小鼠基因體進行排比。對於 MIX01，41.6%

之排比片段分類為僅人類 DNA，56.6%分類為僅小鼠 DNA，1.8%分類為人類-小鼠雜合 DNA。對於 MIX02，61.8%之排比片段分類為僅人類 DNA，36.3%分類為僅小鼠 DNA，1.9%分類為人類-小鼠雜合 DNA。在 BS-Seq 中確定為人類-小鼠雜合 DNA 之定序片段的百分比（<2%）遠低於 Pacific Biosciences 定序結果中觀察到的百分比（>30%）。值得注意的是，長片段（中位數為約 2 kb）係藉由 Pacific Biosciences 定序來定序，而長片段共享成適於 MiSeq 之短片段（中位數為約 196 bp）。此類剪切過程會極大地稀釋人類-小鼠雜合片段。

【0599】 **圖 118** 展示比較藉由亞硫酸氫鹽定序及 Pacific Biosciences 定序確定之 MIX01 之甲基化的表格。該表之最左側部分展示 DNA 之類型：1)僅人類；2)僅小鼠；及 3)人類-小鼠雜合，分為人類部分及小鼠部分。該表之中間部分展示亞硫酸氫鹽定序之詳情，包括 CG 位點之數量及甲基化密度。該表之最右側部分展示 Pacific Biosciences 定序之詳情，包括 CG 位點之數量及甲基化密度。

【0600】 如圖 118 所示，在亞硫酸氫鹽定序及 Pacific Biosciences 定序結果中，對於 MIX01，僅人類 DNA 始終顯示出比僅小鼠 DNA 更高的甲基化密度。對於人類-小鼠雜合片段，在亞硫酸氫鹽定序結果中，確定人類部分及小鼠部分之甲基化程度分別為 46.8%及 2.3%。此等結果證實，如根據本揭示案之 Pacific Biosciences 定序所確定，與小鼠部分相比，人類部分之甲基化密度更高。藉由 Pacific Biosciences 定序，觀察到人類部分之甲基化密度為 57.4%，且觀察到小鼠部分之甲基化密度較低，為 12.1%。此等結果表明，根據本揭示案藉由 Pacific Biosciences 定序確定之甲基化可能為可行的。特定言之，Pacific Biosciences 定序可用於確定不同的甲基化密度，包括在 DNA 中具有比另一部分更高的甲基化密度的部分。吾等觀察到，根據本揭示案藉由 Pacific Biosciences 定序確定之甲

基化密度相對於亞硫酸氫鹽定序更高。此類估計可使用此兩種技術確定的結果之間的差異進行調整，以便比較各個技術的結果。

【0601】 **圖 119** 展示比較藉由亞硫酸氫鹽定序及 Pacific Biosciences 定序確定之 MIX02 之甲基化的表格。該表之最左側部分展示 DNA 之類型：1)僅人類；2)僅小鼠；及 3)人類-小鼠雜合，分為人類部分及小鼠部分。該表之中間部分展示亞硫酸氫鹽定序之詳情，包括 CG 位點之數量及甲基化密度。該表之最右側部分展示 Pacific Biosciences 定序之詳情，包括 CG 位點之數量及甲基化密度。

【0602】 如圖 119 所示，在亞硫酸氫鹽定序及 Pacific Biosciences 定序結果中，對於 MIX02，僅人類 DNA 始終顯示出比僅小鼠 DNA 更低的甲基化密度。對於人類-小鼠雜合片段，在亞硫酸氫鹽定序結果中，確定人類部分及小鼠部分之甲基化程度分別為 1.8%及 67.4%。此等結果進一步證實，如根據本揭示案之 Pacific Biosciences 定序所確定，與小鼠部分相比，人類部分之甲基化密度更低。藉由 Pacific Biosciences 定序，如根據本揭示案藉由 Pacific Biosciences 定序所確定，觀察到人類部分之甲基化密度為 13.1%，且觀察到小鼠部分之甲基化密度較高，為 72.2%。其亦表明，根據本揭示案藉由 Pacific Biosciences 定序確定甲基化為可行的。特定言之，Pacific Biosciences 定序可用於確定不同的甲基化密度，包括在 DNA 中具有比另一部分更低的甲基化密度的部分。吾等亦觀察到，根據本揭示案藉由 Pacific Biosciences 定序確定之甲基化密度相對於亞硫酸氫鹽定序更高。此類估計可使用此兩種技術確定的結果之間的差異進行調整，以便比較各個技術的結果。

【0603】 **圖 120A** 展示 MIX01 之僅人類及僅小鼠 DNA 在 5-Mb 面元中之甲基化程度。**圖 120B** 展示 MIX02 之僅人類及僅小鼠 DNA 在 5-Mb 面元中之甲

基化程度。在兩圖中，百分比形式之甲基化程度顯示在 y 軸上。僅人類 DNA 及僅小鼠 DNA 中之每一者的亞硫酸氫鹽定序及 Pacific Biosciences 定序顯示在 x 軸上。

【0604】 發現根據本揭示案藉由 Pacific Biosciences 定序確定之圖 120A 及圖 120B 中之結果在樣本 MIX01 及 MIX02 中跨面元系統性較高。

【0605】 **圖 121A** 展示 MIX01 之人類-小鼠雜合 DNA 片段之人類部分及小鼠部分在 5-Mb 面元中之甲基化程度。**圖 121B** 展示 MIX02 之人類-小鼠雜合 DNA 片段之人類部分及小鼠部分在 5-Mb 面元中之甲基化程度。在兩圖中，百分比形式之甲基化程度顯示在 y 軸上。人類部分 DNA 及小鼠部分 DNA 中之每一者的亞硫酸氫鹽定序及 Pacific Biosciences 定序顯示在 x 軸上。

【0606】 圖 121A 及圖 121B 均顯示，當使用 Pacific Biosciences 定序時，與亞硫酸氫鹽定序相比，甲基化程度增加。此增加與圖 120A 及圖 120B 中用僅人類 DNA 及僅小鼠 DNA 所見的 Pacific Biosciences 定序之甲基化程度的增加相似。雜合片段之亞硫酸氫鹽定序結果中存在的跨 5-Mb 面元之甲基化程度的變異性增加可能係由於用於分析之 CpG 位點數量較少。

【0607】 **圖 122A** 及 **122B** 為顯示單個人類-小鼠雜合分子中甲基化狀態的代表性圖。圖 122A 展示樣本 MIX01 中之人類-小鼠雜合片段。圖 122B 展示樣本 MIX02 中之人類-小鼠雜合片段。經填充之圓圈指示甲基化之位點，未填充之圓圈指示未甲基化之位點。根據本文所述之實施例確定此等片段中之甲基化狀態。

【0608】 如圖 122A 所示，確定來自樣本 MIX01 之雜合分子的人類部分甲基化程度更高。相反，確定小鼠 DNA 部分之甲基化程度更低。相反，圖 122B 顯示，確定來自樣本 MIX02 之雜合分子的人類部分甲基化程度更低，而確定小

鼠 DNA 部分甲基化程度更高。

【0609】 此等結果表明，本揭示案中存在之實施例允許吾人確定單個 DNA 分子中之甲基化變化，其中在分子之不同部分中甲基化模式不同。在一個實施例中，可量測基因或其他基因體區域之甲基化狀態，其中基因或基因體區域之不同部分會表現出不同的甲基化狀態（例如啟動子與基因主體）。在另一個實施例中，本文提出之方法可檢測人類-小鼠雜合片段，提供一種通用的方法來檢測相對於參考基因體含有非連續片段之 DNA 分子（亦即嵌合分子），且分析其甲基化狀態。舉例而言，吾等可使用此方法來分析但不限於基因融合、基因體重排、轉譯、倒位、重複、結構變異、病毒 DNA 整合、減數分裂重組等。

【0610】 在一些實施例中，此等雜合片段可在定序之前使用基於探針之雜交方法或 CRISPR-Cas 系統或其用於目標 DNA 富集之變異方法來富集。最近，據報導，來自藍細菌，亦即霍氏雙歧藍細菌（*Scytonema hofmanni*）之 CRISPR 相關轉座酶能夠將 DNA 區段插入所關注之靶位點附近的區域（Strecker 等人《科學》2019;365:48-53）。CRISPR 相關轉座酶可像 Tn7 介導之轉座一樣其作用。在一個實施例中，吾等可調整此 CRISPR 相關之轉座酶，以在 gRNA 的引導下，將例如用生物素標記之註釋序列插入一或多個所關注之基因體區域。吾等可使用塗覆有例如抗生蛋白鏈菌素之磁珠來捕捉註釋序列，從而根據本揭示案中之實施例同時拉下目標 DNA 序列進行定序及甲基化分析。

【0611】 在一些實施例中，片段可藉由使用限制酶富集，該等限制酶可包括本文所揭示之任何限制酶。

例示性嵌合分子檢測方法

【0612】 **圖 123** 展示檢測生物樣本中之嵌合分子的方法 1230。嵌合分子可包括來自兩個不同基因、染色體、胞器（例如粒線體、細胞核、葉綠體）、

生物體（哺乳動物、細菌、病毒等）及/或物種之序列。方法 1230 可應用於來自生物樣本之複數個 DNA 分子中之每一者。在一些實施例中，複數個 DNA 分子可為細胞 DNA。在其他實施例中，複數個 DNA 分子可為來自孕婦血漿之游離 DNA 分子。

【0613】 在方塊 1232，可對 DNA 分子進行單分子定序，以獲得提供 N 個位點中之每一者的甲基化狀態的序列讀段。N 可為 5 或更多，包括 5 至 10、10 至 15、15 至 20 或大於 20。序列讀段之甲基化狀態可形成甲基化模式。DNA 分子可為複數個 DNA 分子中之一個 DNA 分子，且可對複數個 DNA 分子進行方法 1230。甲基化模式可採取各種形式。舉例而言，模式可為 N（例如 2、3、4 等）個甲基化位點，隨後為 N 個未甲基化位點，反之亦然。此類甲基化變化可指示接合點。經甲基化之連續位點的數量可不同於未甲基化之連續位點的數量。

【0614】 在方塊 1234，甲基化模式可滑移至一或多個參考模式上，該等參考模式對應於具有來自參考人類基因體之兩部分的兩個部分的嵌合分子。參考模式可充當過濾器，以鑑別指示接合點之匹配模式。可跟蹤與參考模式匹配之位點的數量，以使得匹配位置對應於最大數量之匹配位點（亦即，甲基化狀態與參考模式匹配的數量）。參考人類基因體之兩部分可為參考人類基因體之不連續部分。參考人類基因體之兩部分可相隔超過 1kb、5kb、10kb、100kb、1 Mb、5 Mb 或 10 Mb。該兩部分可來自兩個不同的染色體臂或染色體。一或多個參考模式可包括甲基化狀態與未甲基化狀態之間的變化。

【0615】 在方塊 1236，可在甲基化模式與一或多個參考模式之第一參考模式之間鑑別匹配位置。匹配位置可鑑別序列讀段中參考人類基因體之兩部分之間的接合點。匹配位置可對應於參考模式與甲基化模式之間的重疊函數中的

最大值。重疊函數可使用多個參考模式，其中輸出可能為集合函數上的最大值（亦即，每個參考模式對輸出值有貢獻）或跨參考模式鑑別之單個最大值。

【0616】 在方塊 1238，接合點可輸出為嵌合分子中基因融合之位置。基因融合之位置可與包括癌症之各種病症或疾病之基因融合的參考位置進行比較。自其中獲得生物樣本之生物體可對病症或疾病進行治療。

【0617】 匹配位置可輸出至排比函數。基因融合之位置可經細化。細化基因融合之位置可包括將序列讀段之第一部分與參考人類基因體之第一部分進行排比。第一部分可在接合點之前。細化基因融合之位置可包括將序列讀段之第二部分與參考人類基因體之第二部分進行排比。第二部分可在接合點之後。參考人類基因體之第一部分可與人類參考基因體之第二部分相隔至少 1 kb。舉例而言，參考人類基因體之第一部分及人類參考基因體之第二部分可相隔 1.0 至 1.5 kb、1.5 至 2.0 kb、2.0 至 2.5 kb、2.5 至 3.0 kb、3 至 5 kb 或 5 kb 以上。

【0618】 多個嵌合分子之接合點可相互比較，以確認基因融合之位置。

結論

【0619】 吾等已開發一種有效的方法來預測單鹼基解析度下核酸之鹼基修飾（例如甲基化）程度。此新方法實施一種新的方案，用於同時捕捉所詢問之鹼基周圍的聚合酶動力學、序列上下文及股資訊。此類新的動力學轉換使得動力學脈衝中出現的細微中斷可經鑑別及模型化。與先前僅使用 IPD 之方法相比，本專利申請案中提出的新方法大大提高甲基化分析之解析度及準確性。此新方案可容易地擴展用於其他目的，例如檢測 5hmC（5-羥甲基胞嘧啶）、5fC（5-甲醯基胞嘧啶）、5caC（5-羧基胞嘧啶）、4mC（4-甲基胞嘧啶）、6mA（N6-甲基腺嘌呤）、8oxoG（7,8-二氫-8-側氧基鳥嘌呤）、8oxoA（7,8-二氫-8-側氧基腺嘌呤）及其他形式之鹼基修飾以及 DNA 損傷。在另一個實施例中，此新方案

（例如類似於本申請案中存在之 2-D 數位矩陣的動力學轉換）可用於使用奈米孔定序系統進行鹼基修飾分析。

【0620】 甲基化檢測之此實現方式可用於不同來源的核酸樣本，例如細胞核酸、環境採樣（例如細胞污染物）的核酸、病原體（例如細菌及真菌）的核酸及孕婦血漿中之 cfDNA。其將為基因體研究及分子診斷打開許多新的可能性，諸如非侵入性產前檢測、癌症檢測及移植監測。對於基於 cfDNA 之非侵入性產前診斷，此項新發明使得在診斷中同時使用每個分子之複本數畸變、大小、突變、片段末端及鹼基修飾成為可行，而不需要在定序前進行 PCR 及實驗轉換，從而提高靈敏度。可使用本文所述之方法檢測單倍型之間甲基化程度的不平衡。此類不平衡可表明 DNA 分子之起源（例如，自病症提取，諸如自癌症患者血液中分離的癌細胞）或病症之起源。

例示性系統

【0621】 **圖 124** 展示根據本發明之一個實施例的量測系統 12400。如圖所示之系統包括樣本固持器 12410 內之樣本 12405，諸如 DNA 分子，其中樣本 12405 可與分析法 12408 接觸，以提供物理特徵 12415 的信號。樣本固持器之一實例可為包括分析法之探針及/或引子的流槽或液滴藉以移動之管（其中液滴包括分析法）。偵測器 12420 檢測樣本之物理特徵 12415（例如，螢光強度、電壓或電流）。偵測器 12402 可按時間間隔（例如，週期性時間間隔）進行量測，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為數位形式。樣本固持器 12401 及偵測器 12402 可以形成分析法裝置，例如根據本文所述之實施例進行定序的定序裝置。資料信號 12425 自偵測器 12402 發送至邏輯系統 12403。資料信號 12425 可以儲存在本地記憶體 12435、外部記憶體 12404 或儲存裝置 12445 中。

【0622】 邏輯系統 12403 可為或可包括電腦系統、ASIC、微處理器等。其亦可包括顯示器（例如，監視器、LED 顯示器等）及使用者輸入裝置（例如，滑鼠、鍵盤、按鈕等）或與其耦接。邏輯系統 12403 及其他組件可以係獨立的或網路連接的電腦系統之一部分，或者其可以直接連接至或併入至包括偵測器 12402 及/或樣本固持器 12401 之裝置（例如定序裝置）中。邏輯系統 12403 亦可包括在處理器 12405 中執行之軟體。邏輯系統 12403 可包括電腦可讀媒體，該電腦可讀媒體儲存用於控制系統 12400 以執行本文所述之任何方法的指令。舉例而言，邏輯系統 12403 可以向包括樣本固持器 12401 之系統提供命令，從而執行定序或其他物理操作。可按特定順序執行此類物理操作，例如，以特定順序添加及移除試劑。此類物理操作可由機器人系統（例如包括機器人臂）執行，如可用於獲得樣本並執行分析法。

【0623】 本文中提及之任何電腦系統均可利用任何適合數目之子系統。此類子系統之實例展示於**圖 125** 中之電腦系統 10 中。在一些實施例中，電腦系統包括單一電腦設備，其中子系統可為電腦設備之組件。在其他實施例中，電腦系統可包括具有內部組件之多個電腦設備，其各自為一個子系統。電腦系統可包括桌上型及膝上型電腦、平板電腦、行動電話、其他移動裝置及基於雲端之系統。

【0624】 圖 125 中所示之子系統經由系統匯流排 75 互連。示出附加的子系統，諸如印表機 74、鍵盤 78、儲存裝置 79、監測器 76（例如，顯示屏幕，諸如 LED），其耦接至顯示器配接器 82，及其他子系統。耦接至輸入/輸出（I/O）控制器 71 之周邊裝置及 I/O 裝置可利用本領域中已知的任何數目的手段（諸如輸入/輸出（I/O）埠 77（例如 USB、FireWire®））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、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 之互連允許中央處理器 73 與每個子系統通信並控制來自系統記憶體 72 或儲存裝置 79（例如，固接磁碟，諸如硬碟機或光碟）之複數個指令之執行，以及子系統之間的資訊交換。系統記憶體 72 及/或儲存裝置 79 可體現為電腦可讀媒體。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。

【0625】 電腦系統可包括複數個相同的組件或子系統，例如，利用外部介面 81、利用內部介面或經由可卸除式儲存裝置連接在一起，該等可卸除式儲存裝置可自一個組件連接至另一組件或將一個組件自另一組件卸除。在一些實施例中，電腦系統、子系統或設備可經由網路進行通信。在此等情況下，可將一台電腦視為用戶端且另一台電腦視為伺服器，其中每一者可為同一電腦系統之一部分。用戶端及伺服器可各自包括多個系統、子系統或組件。

【0626】 實施例之各態樣可使用硬體電路（例如，特殊應用積體電路或場可程式化閘陣列）及/或使用具有一般可程式化處理器之電腦軟體以模組化或積體方式以控制邏輯的形式來實施。如本文所用，處理器可包括單核處理器、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，本領域中一般熟習此項技術者將知曉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。

【0627】 描述於本申請案中之任何軟體組件或功能可作為待由處理器執行的使用任何適合之電腦語言（諸如 Java、C、C++、C#、Objective-C、Swift）或腳本語言（諸如 Perl 或 Python）的軟體程式碼使用例如習知或物件導向技術

來執行。軟體程式碼可以一系列指令或命令形式儲存於電腦可讀媒體上以用於儲存及/或傳輸。適合的非暫時性電腦可讀媒體可包括隨機存取記憶體（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 DVD（數位化通用光碟）或藍光碟、快閃記憶體及其類似者。電腦可讀媒體可為此等儲存或傳輸裝置之任何組合。

【0628】 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（包括網際網路）傳輸的載波信號來編碼及傳輸。因此，電腦可讀媒體可使用以此類程式編碼之資料信號建立。可將用程式碼編碼之電腦可讀媒體與相容裝置封裝在一起，或者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駐留在單個電腦產品（例如，硬碟機、CD 或整個電腦系統）上或內部，且可存在於系統或網路內之不同電腦產品上或內部。電腦系統可包括用於向使用者提供本文所提及之任何結果的監測器、印表機、或其他適合的顯示器。

【0629】 本文所描述之任何方法可完全或部分地使用電腦系統來進行，該電腦系統包括一或多個處理器，該一或多個處理器可經組態以進行該等步驟。因此，實施例可針對於經組態以執行本文所描述之任何方法之步驟的電腦系統，潛在地用不同組件執行各別步驟或各別步驟組。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，此等步驟之各部分可與其他方法之其他步驟之各部分一起使用。另外，步驟之全部或各部分可視情況選用。此外，任何方法之任何步驟可使用用於進行此等步驟之系統的模組、單元、電路或其他構件來進行。

【0630】 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每一個

別態樣或此等個別態樣之特定組合相關的特定實施例。

【0631】 為了說明及描述之目的，已呈現本發明之實例實施例的以上描述。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。

【0632】 除非有相反的特定說明，否則「一（a/an）」或「該（the）」之敍述意欲意謂「一或多個（種）」。除非有相反的特定說明，否則「或」之使用意欲意謂「包含性的或」，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，否則對「第一」或「第二」組件之提及不將所提及之組件限制在特定位置。術語「基於」意欲意謂「至少部分地基於」。

【0633】 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用於所有目的。不承認任一者為先前技術。

參考文獻

Albert, T.J.等人(2007)藉由微陣列雜交直接選擇人類基因體基因座.《自然方法（*Nat. Methods*）》, **4**, 903-905。

Beckmann 等人(2014)在低覆蓋率及宏基因體學設置中檢測表觀遺傳基元.《BMC 生物資訊學》, 15(增刊 9): S16。

Beaulaurier, J.等人(2019) 使用現代定序技術破譯細菌表觀基因體.《自然綜述遺傳學（*Nature Reviews Genetics*）》, 20:157-172。

Blow, M.J.等人(2016) 原核生物之表觀遺傳學前景.《公共科學圖書館•遺傳學（*PLOS Genet.*）》, **12**, e1005854。

Breiman, L.(2001)隨機森林。《機器學習（*Mach. Learn.*）》, **45**, 5-32。

Chan, K.C.A.等人(2013) 藉由血漿 DNA 亞硫酸氫鹽定序對癌症相關之全基因體低甲基化及複本數畸變進行非侵入性檢測.《美國科學學院學報（*Proc. Natl. Acad. Sci. U. S. A.*）》, **110**, 18761-8。

Clark, T.A.等人(2013) 經由 Tet1 氧化在單分子即時定序中增強 5-甲基胞嘧啶檢測.《BMC 生物學（*BMC Biol.*）》, **11**, 4。

Clark, T.A.等人(2012) 使用單分子即時 DNA 定序表徵 DNA 甲基轉移酶特異性.《核酸研究》, 40:e29。

Eid, J.等人(2009) 自單一聚合酶分子進行即時 DNA 定序.《科學》**323**, 133-138。

Feinberg, A.P.及 Irizarry, R.A.(2010) 隨機表觀遺傳變異作為發展、進化適應及疾病之驅動力.《美國科學學院學報》, **107**, 1757-1764。

Feng, Z.等人(2013) 藉由對聚合酶動力學之序列上下文依賴性模型化自 SMRT 定序資料檢測 DNA 修飾.《公共科學圖書館·計算生物學（*PLoS Comput Biol.*）》, 9:e1002935。

Flusberg, B.A.等人(2010) 在單分子即時定序期間直接檢測 DNA 甲基化.《自然方法》, **7**, 461-465。

Frommer, M.等人(1992) 在個別 DNA 股中產生 5-甲基胞嘧啶殘基之陽性顯示的基因體定序方案.《美國科學學院學報》, **89**, 1827-1831。

Gai, W.等人(2018) 用於研究具有或不具有肝轉移之結腸直腸癌的血漿中肝臟及結腸特異性 DNA 甲基化標記.《臨床化學》, **64**, 1239-1249。

Gouil, Q.等人(2019) 研究 DNA 甲基化之最新技術.《生物化學短評（*Essays*

Biochem.)》63(6):639-648.

Grunau, C.(2001) 亞硫酸氫鹽基因體定序：關鍵實驗參數之系統研究.《核酸研究》, **29**, 65e – 65。

Herman, J.G.等人(1996) 甲基化特異性 PCR：CpG 島甲基化狀態之新穎 PCR 分析法.《美國科學學院學報》, **93**, 9821-9826。

Jiang, P.等人(2014) 甲基管道：用於全基因體亞硫酸氫鹽定序資料分析之整合式生物資訊學管道.《公共科學圖書館》, **9**, e100360。

LeCun, Y.等人(1989) 應用於手寫郵遞區號識別之反向傳播.《神經計算（*Neural Comput.*)》, **1**, 541-551。

Lee, E.-J.等人(2011) 藉由溶液雜合選擇及大規模平行定序進行有針對性的亞硫酸氫鹽定序.《核酸研究》, **39**, e127-e127。

Lehmann-Werman, R.等人(2016) 使用循環 DNA 之甲基化模式鑑別組織特異性細胞死亡.《美國科學學院學報》, **113**, E1826-E1834。

Lister, R.等人(2009) 鹼基解析度下之人類 DNA 甲基化體顯示廣泛的表觀基因體差異.《自然》, **462**, 315-322。

Liu, Q.等人(2019) 藉由深度循環神經網路在牛津奈米孔定序資料上檢測 DNA 鹼基修飾.《自然通訊（*Nature Commun.*)》, **10**, 2449。

Liu, Y.等人(2019) 鹼基解析度下無亞硫酸氫鹽直接檢測 5-甲基胞嘧啶及 5-羥甲基胞嘧啶.《自然生物技術》, **37**, 424-429。

Lun, F.M.F.等人(2013) 藉由母體血漿 DNA 之全基因體亞硫酸氫鹽定序進行非侵入性產前甲基化體分析.《臨床化學》**59**, 1583-1594。

Nattestad, M.等人(2018) 藉由乳癌細胞株之長讀段 DNA 及 RNA 定序揭露複雜的重排及致癌基因擴增.《基因體研究》, **28**, 1126-1135

Ng, A.Y.(2004) 特徵選擇，L_1與L_2正則化以及旋轉不變性.第二十一屆機器學習國際會議 - ICML '04.ACM Press, New York, New York, USA, 第 78 頁。

Ni, P.等人(2019) 深度信號：使用深度學習自奈米孔定序讀段檢測 DNA 甲基化狀態.《生物資訊學》,**35**, 4586-4595

Okou, D.T.等人(2007) 高通量再定序之基於微陣列之基因體選擇.《自然方法》, **4**, 907-909。

Olova, N.等人(2018) 全基因體亞硫酸氫鹽定序文庫製備策略之比較鑑別影響 DNA 甲基化資料之偏差來源.《基因體生物學》, **19**, 33。

Robertson, K.D.(2005) DNA 甲基化與人類疾病.《自然綜述遺傳學》, **6**, 597-610。

Smith, Z.D.及 Meissner, A.(2013) DNA 甲基化：在哺乳動物發育中的作用.《自然綜述遺傳學》, **14**, 204-20。

Schadt, E.E.等人(2013) 第三代 DNA 定序資料中模型化動力學速率變化以檢測對 DNA 鹼基之推定修飾.《基因體研究》, 23(1):129-41。

Sun, K.等人(2015) 藉由全基因體甲基化定序之血漿 DNA 組織定位用於非侵入性產前、癌症及移植評定.《美國科學學院學報》, **112**, E5503-E5512。

Suzuki, Y.等人(2016) AgIn：量測個別重複元件之 CpG 甲基化態勢.《生物資訊學》, **32**, 2911-2919。

Watson, C.M.等人(2019) 基於 Cas9 之富集及單分子定序對基因體重複進行

精確表徵.《實驗室研究（*Lab.Investig*）》, **100**, 135-146。

Zhang, W.等人(2015)使用甲基化標記、基因體位置及 DNA 調控元件預測全基因體 DNA 甲基化.《基因體生物學》, **16**, 14。

【符號說明】

【0634】

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: 印表機

75: 系統匯流排

76: 監測器

77: 輸入/輸出（I/O）端口

78: 鍵盤

79: 存儲裝置

81: 外部介面

82: 顯示器配接器

85: 資料收集裝置

102: 分子

104: 分子

106: 分子

108: 分子

110: 環化分子

202: DNA 分子

204: 連接分子

206: 環化分子

208: 未甲基化之 CpG 位點

400: 量測窗口

402: 矩陣之第一列

404: 矩陣之第二列

408: 列

412: 列

416: 列

420: 列

902: 經修飾之分子

904: 鹼基

906: 未修飾之分子

908: 鹼基

910: 階段

912: 階段

914: 階段

916: 階段

918: 階段

3102: 階段

3106: 階段

3110: 無規六聚體

3114: 股置換擴增

3118: 股

3122: 新合成之 DNA 股

3126: 可能片段

3130: 可能片段

3134: 可能片段

4710: 區域

7301: 非腫瘤組織中存在之分子

7302: 非腫瘤組織中存在之分子

7303: 腫瘤組織中存在之分子

9102: 分子

9104: 分子

9106: 分子

1020: 方法

1022: 方塊

1024: 方塊

1026: 方塊

1028: 方塊

1029: 方塊

1030: 方法

1032: 方塊

1034: 方塊

1036: 方塊

1090: 方法

1091: 方塊

1092: 方塊

1093: 方塊

1094: 方塊

1095: 方塊

1096: 方塊

1097: 方塊

1098: 方塊

1099: 方塊

1230: 方法

1232: 方塊

1234: 方塊

1236: 方塊

1238: 方塊

11010: 帶對角線條紋之粗條

11020: 帶垂直條紋之粗條

11110: 帶對角線條紋之粗條

11120: 帶垂直條紋之粗條

11410: 帶對角線條紋之條

12400: 量測系統

12401: 樣本固持器

12402: 偵測器

12403: 邏輯系統

12404: 外部記憶體

12405: 樣本

12408: 分析法

12415: 物理特徵

12425: 資料信號

12435: 本地記憶體

12445: 存儲裝置

【序列表】

<110> 香港中文大學(THE CHINESE UNIVERSITY OF HONG KONG)

<120> 核酸鹼基修飾的測定

<140> TW 109127986
<141> 2020-08-17

<150> US63/051,210
<151> 2020-07-13

<150> US63/019,790
<151> 2020-05-04

<150> US62/991,891
<151> 2020-03-19

<150> US62/970,586
<151> 2020-02-05

<150> US62/887,987
<151> 2019-08-16

<160> 5

<170> PatentIn version 3.5

<210> 1
<211> 36
<212> RNA
<213> 人工序列

<220>
<223> 人工序列之描述：合成寡核苷酸

<400> 1
gccuguaauc ccagcacuuu guuuuagagc uaugcu 36

<210> 2
<211> 67
<212> RNA
<213> 人工序列

<220>
<223> 人工序列之描述：合成寡核苷酸

<400> 2
agcauagcaa guuaaaauaa ggcuaguccg uuaucaacuu gaaaaagugg caccgagucg 60

gugcuuu 67

<210> 3
<211> 36
<212> RNA
<213> 人工序列

<220>
<223> 人工序列之描述：合成寡核苷酸

<400> 3
agggucucgc ucugucgccc guuuuagagc uaugcu 36

<210> 4
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 人工序列之描述：合成寡核苷酸

<400> 4
atacgtacgt 10

<210> 5
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 人工序列之描述：合成寡核苷酸

<400> 5
atacgtacgt 10

【發明申請專利範圍】

【請求項1】

一種用於檢測核酸分子中核苷酸之修飾的方法，該方法包含：

(a) 接收藉由量測對應於樣本核酸分子中定序之核苷酸之光信號中的脈衝所獲取的資料，及自該資料獲取以下特性之值：

對於每個核苷酸：

核苷酸之標識，

該樣本核酸分子中之核苷酸的位置，

對應於該核苷酸之脈衝的寬度，及

脈衝間持續時間，其表示對應於該核苷酸之脈衝與對應於鄰近核苷酸之脈衝之間的時間；

(b) 產生輸入資料結構，該輸入資料結構包含該樣本核酸分子中定序之核苷酸的窗口，其中對於窗口內之各核苷酸，該輸入資料結構包含以下特性：

該核苷酸之標識，

該核苷酸相對於該各別窗口內目標位置的位置，

對應於該核苷酸之脈衝的寬度，及

脈衝間持續時間；

(c) 將該輸入資料結構輸入至模型中，該模型如下進行訓練：

接收第一複數個第一資料結構，該第一複數個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別核酸分子中定序之核苷酸的各別窗口，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光信號中的脈衝來定序，其中該修飾

在每個第一核酸分子之每個窗口中之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含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，每個樣本包括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，使用該複數個第一訓練樣本，基於該模型之輸出匹配或不匹配該等第一標記之相應標記來使該模型之參數最佳化，其中該模型之輸出指定該各別窗口中目標位置處之核苷酸是否具有該修飾，

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

【請求項2】

如請求項1之方法，其中：

該輸入資料結構為複數個輸入資料結構中之一個輸入資料結構，

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，

該複數個樣本核酸分子係自個體之生物樣本獲得，及

每個輸入資料結構對應於該複數個樣本核酸分子之各別樣本核酸分子中定序之核苷酸的各別窗口，及

該方法進一步包含：

接收該複數個輸入資料結構，

將該複數個輸入資料結構輸入至該模型中，及

使用該模型確定在每個輸入資料結構之各別窗口中之目標位置處

之核苷酸中是否存在修飾。

【請求項3】

如請求項2之方法，其進一步包含：

確定該修飾存在於一或多個核苷酸處，及

使用在一或多個核苷酸處之該修飾的存在來確定病症之分類。

【請求項4】

如請求項3之方法，其中該病症包含癌症。

【請求項5】

如請求項4之方法，其進一步包含：

確定該病症之分類為該個體患有該病症，及

藉由化學療法、放療或手術治療該個體之該病症。

【請求項6】

如請求項3之方法，其中確定該病症之分類使用修飾之數量或該等修飾之位點。

【請求項7】

如請求項2之方法，其中該修飾為甲基化，該方法進一步包含：

確定該修飾存在於一或多個核苷酸處，及

使用在一或多個核苷酸處之該修飾的存在確定臨床相關之DNA分數、胎兒甲基化概況、母體甲基化概況、印記基因區域之存在或起源組織。

【請求項8】

如請求項2之方法，其中該複數個樣本核酸分子中之每個樣本核酸分子的大小大於閾值大小。

【請求項9】

如請求項2之方法，其中：

該複數個樣本核酸分子與複數個基因體區域進行排比，

對於該複數個基因體區域中之每個基因體區域：

許多樣本核酸分子與該基因體區域進行排比，

樣本核酸分子之數量大於閾值數量。

【請求項10】

如請求項1之方法，其進一步包含對該樣本核酸分子進行定序。

【請求項11】

如請求項1之方法，其中該模型包括機器學習模型、主成分分析、卷積神經網路或邏輯回歸。

【請求項12】

如請求項1之方法，其中：

對應於該輸入資料結構之核苷酸的窗口包含該樣本核酸分子之第一股上的核苷酸及該樣本核酸分子之第二股上的核苷酸，及

該輸入資料結構進一步包含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13】

如請求項12之方法，其中該樣本核酸分子為藉由以下形成之環形DNA分子：

使用Cas9複合物切割雙股DNA分子，形成經切割之雙股DNA分子，及

將髮夾轉接子連接至該經切割之雙股DNA分子的末端。

【請求項14】

如請求項1之方法，其中該窗口內之該等核苷酸係使用環形一致序列確定的，且無需將經定序之核苷酸與參考基因體進行排比。

【請求項15】

如請求項1之方法，其中該窗口內之每個核苷酸均經富集或過濾。

【請求項16】

如請求項15之方法，其中該窗口內之每個核苷酸藉由以下富集：

使用Cas9複合物切割雙股DNA分子，形成經切割之雙股DNA分子，且將髮夾轉接子連接至該經切割之雙股DNA分子的末端，或

藉由以下過濾：

選擇具有大小範圍內之大小的雙股DNA分子。

【請求項17】

如請求項1之方法，其中該窗口內之核苷酸無需使用環形一致序列且無需將經定序之核苷酸與參考基因體進行排比即可確定。

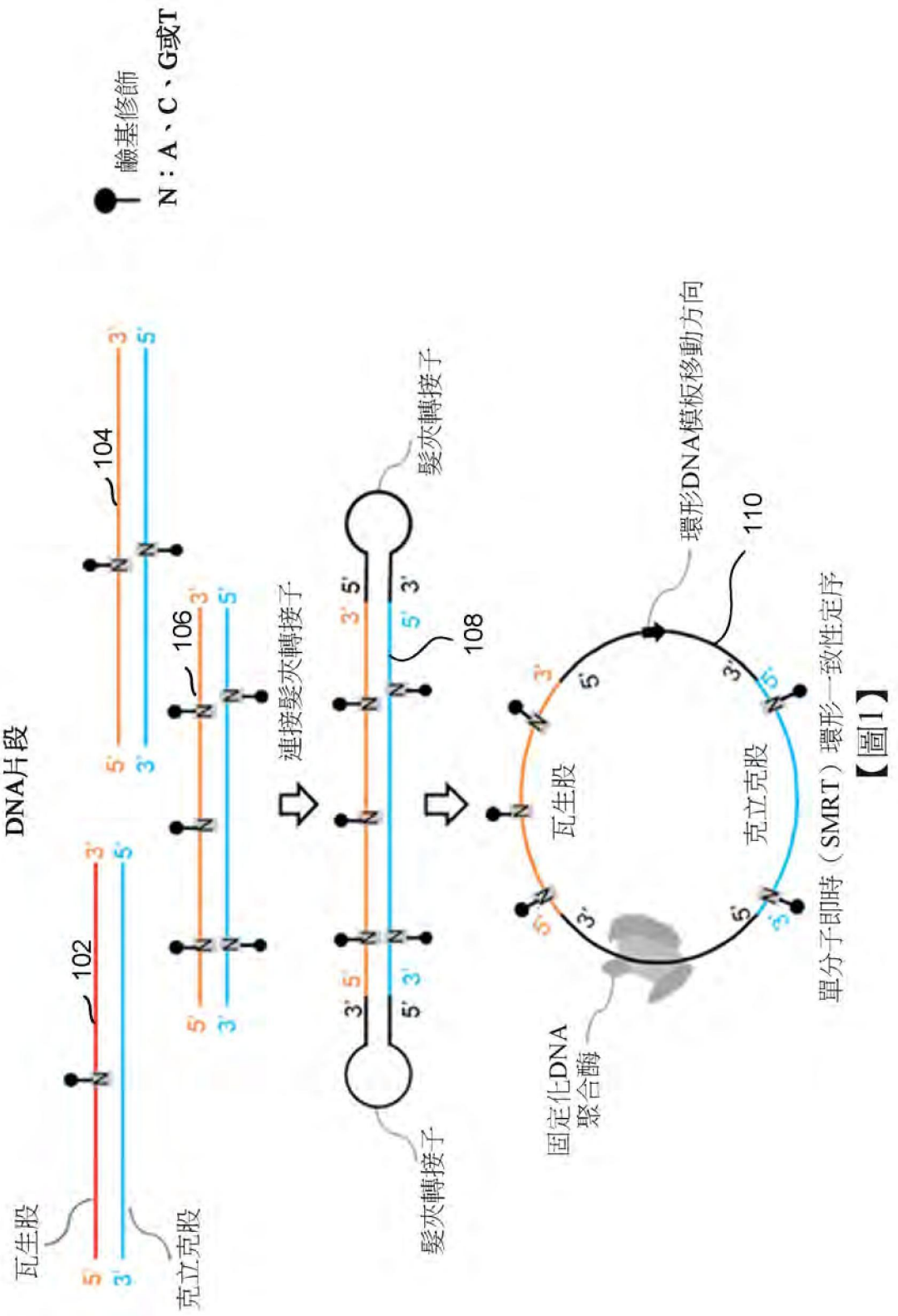
【請求項18】

如請求項1之方法，其中該光信號為來自染料標記之核苷酸的螢光信號。

【請求項19】

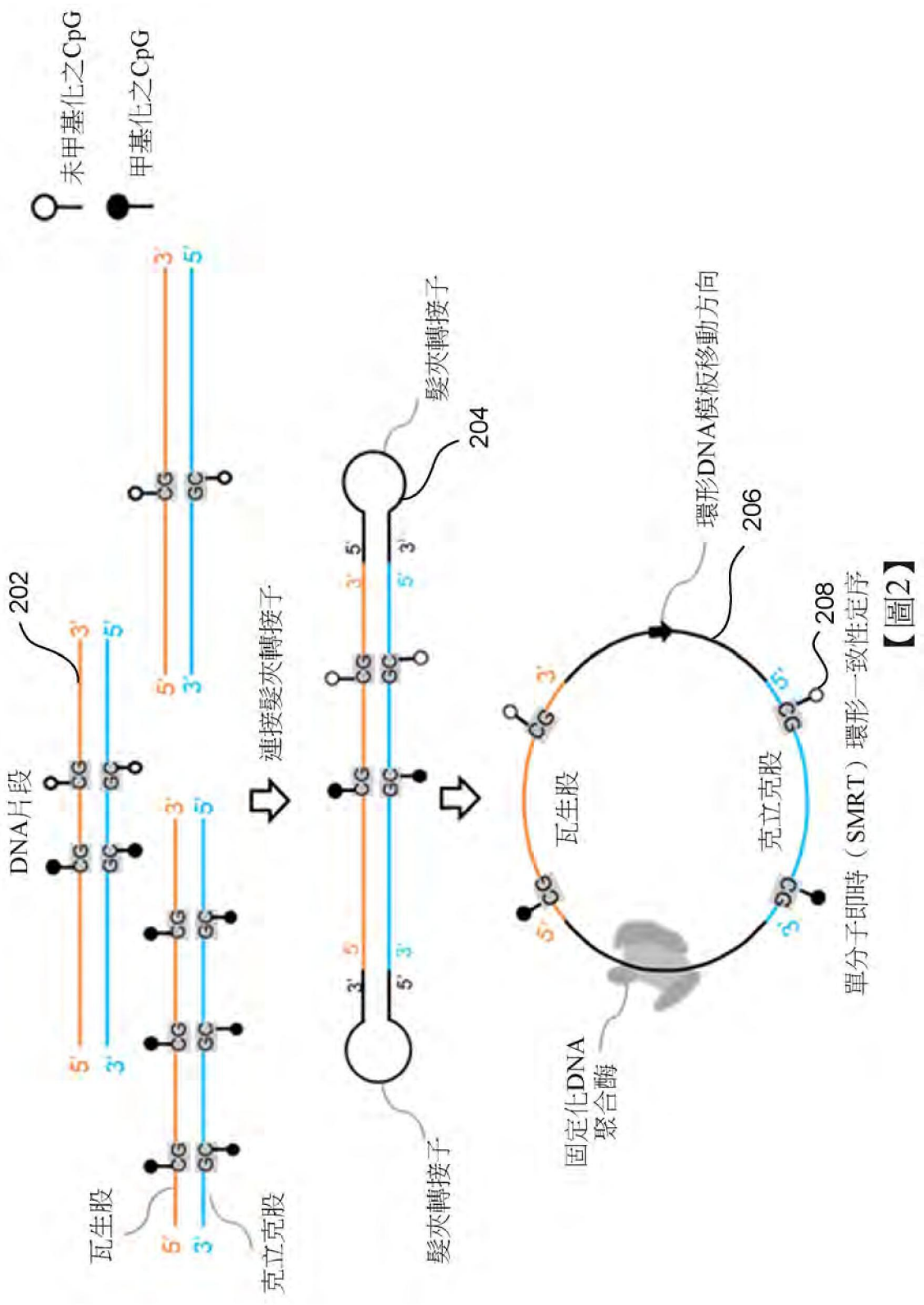
如請求項1之方法，其中與該第一複數個資料結構相關聯之每個窗口包含每個第一核酸分子之第一股上的4個連續核苷酸。

【發明圖式】



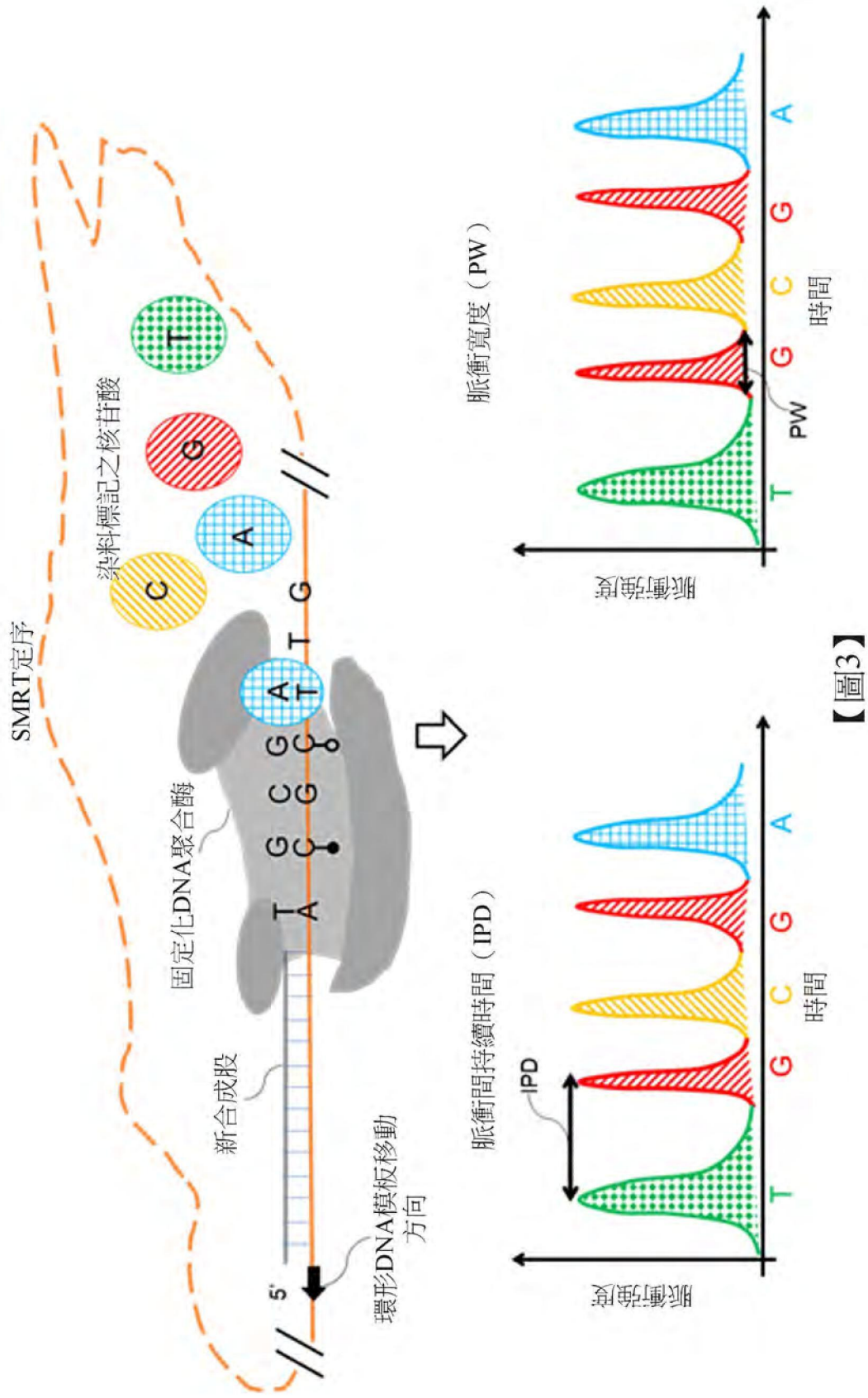
單分子即時（SMRT）環形一致性定序

【圖1】

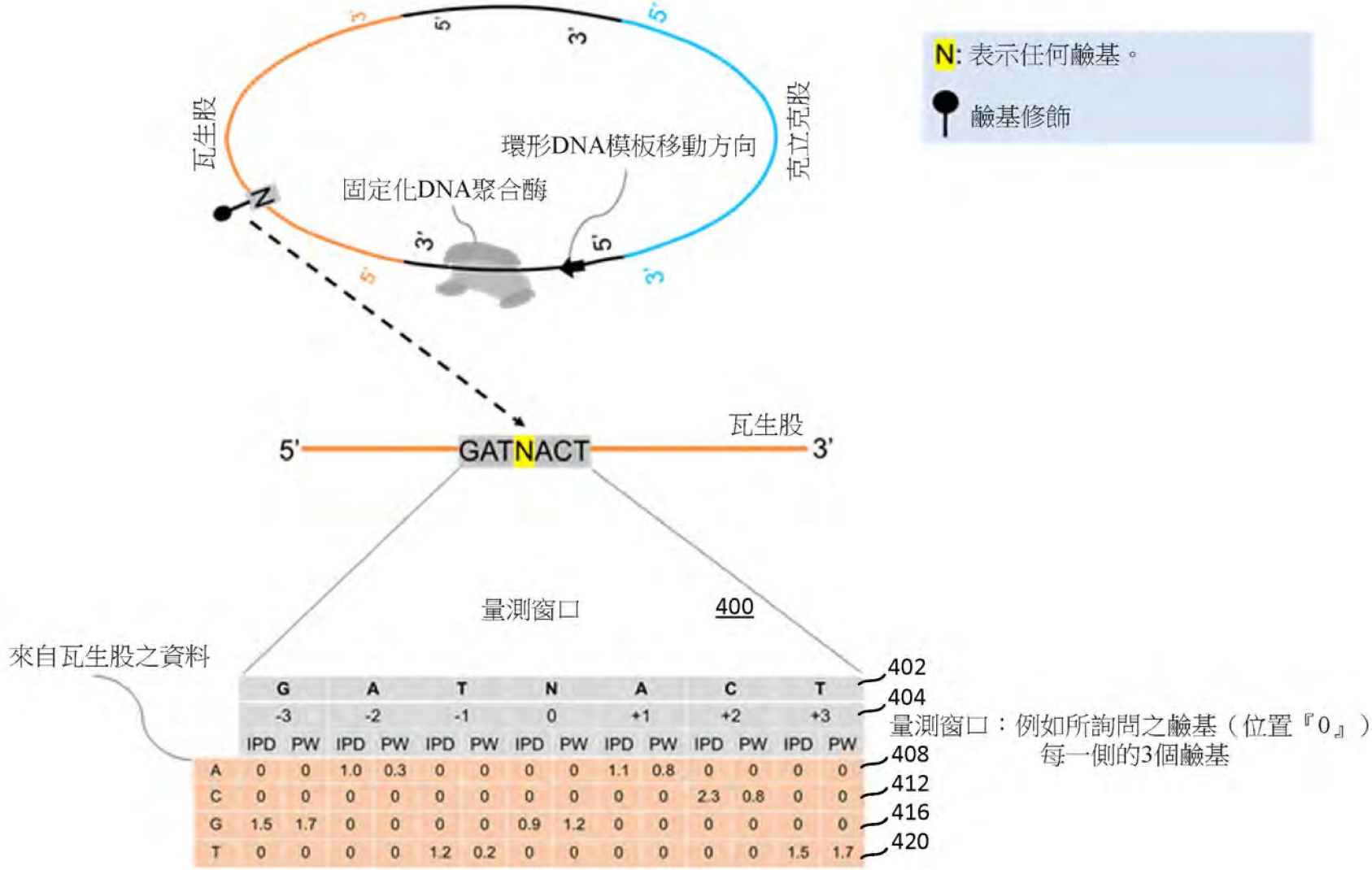


【圖2】

單分子即時（SMRT）環形一致性定序


SMRT定序
染料標記之核苷酸
T
G
C
A
固定化DNA聚合酶
新合成股
5'
環形DNA模板移動方向
脈衝間持續時間（IPD）
IPD
脈衝強度
時間
脈衝寬度（PW）
PW
脈衝強度
時間

【圖3】

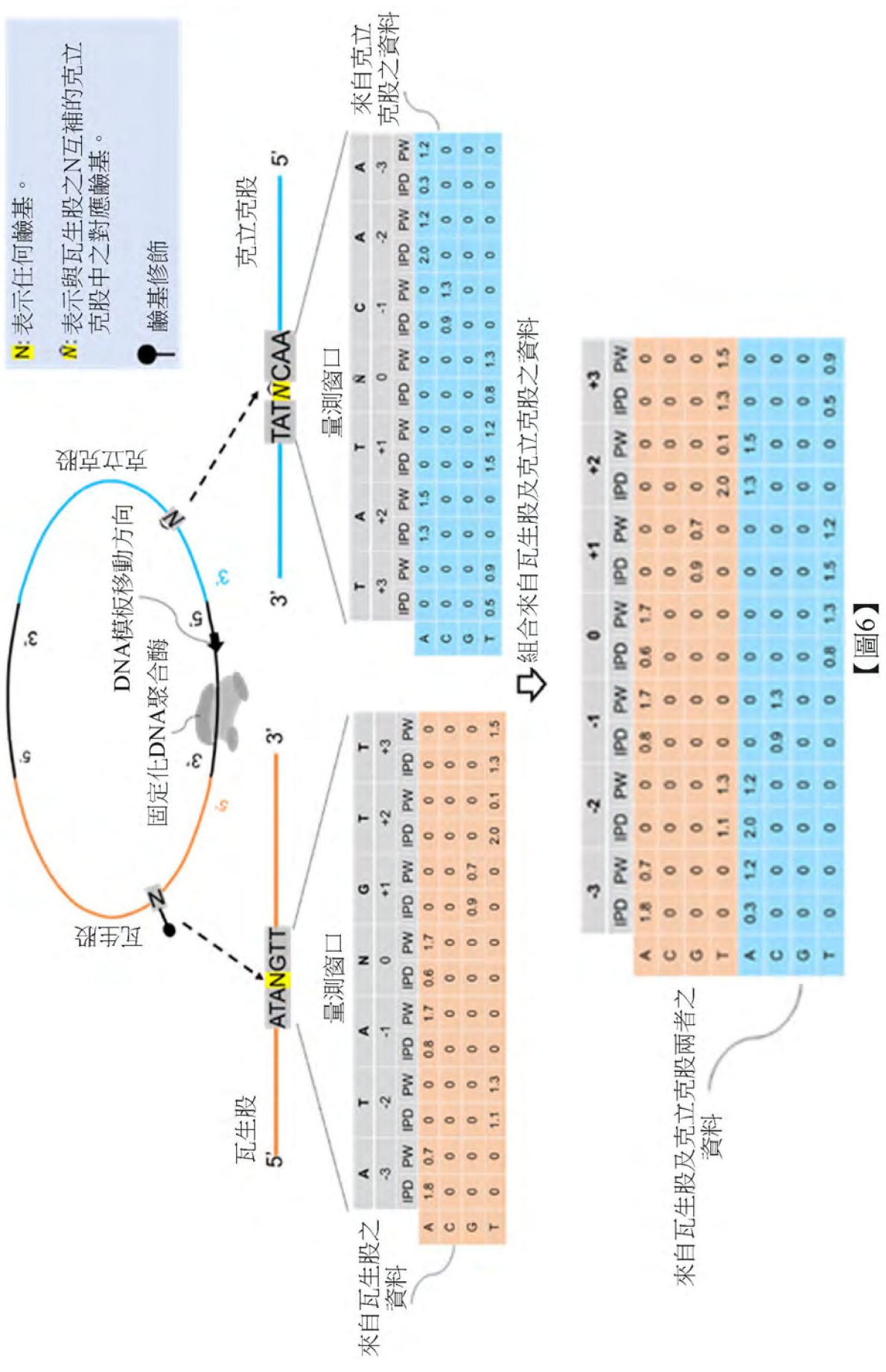


	G		A		T		N		A		C		T	
	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	0	0	1.0	0.3	0	0	0	0	1.1	0.8	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	2.3	0.8	0	0
G	1.5	1.7	0	0	0	0	0.9	1.2	0	0	0	0	0	0
T	0	0	0	0	1.2	0.2	0	0	0	0	0	0	1.5	1.7

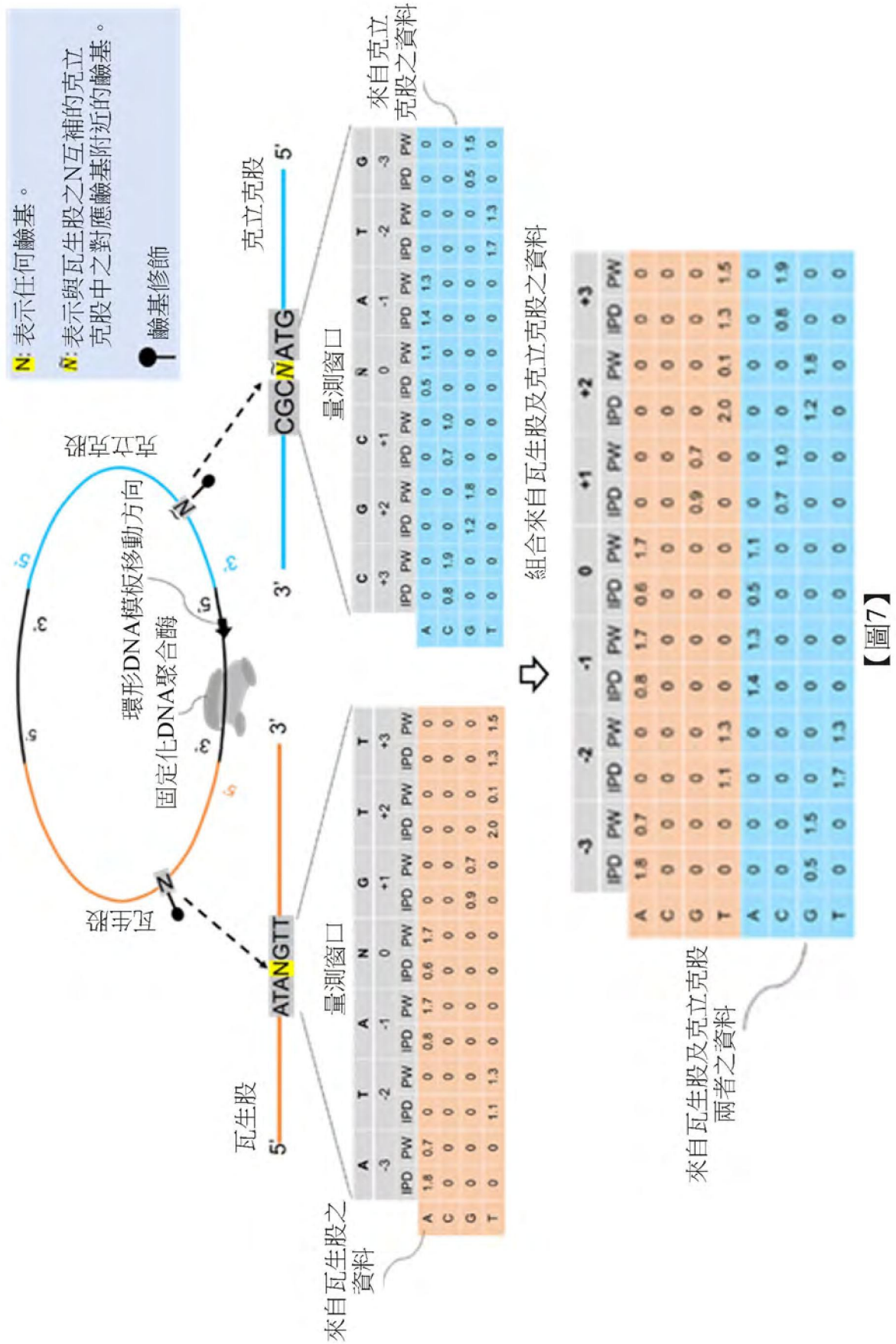
【圖4】

	C			G			A			N			T			C			A		
	+3			+2			+1			0			-1			-2			-3		
	IPD	PW		IPD	PW		IPD	PW		IPD	PW		IPD	PW		IPD	PW		IPD	PW	
A	0	0		0	0		1.2	1.9		0	0		0	0		0	0		0.7	1.2	
C	1.2	0.8		0	0		0	0		0	0		0	0		2.6	1.7		0	0	
G	0	0		0.9	0.5		0	0		0	0		0	0		0	0		0	0	
T	0	0		0	0		0	0		0.9	1.9		0.3	1.8		0	0		0	0	

【圖5】


N: 表示任何鹼基。
N̂: 表示與瓦生股之N互補的克立克股中之對應鹼基。
鹼基修飾
瓦生股
克立克股
DNA模板移動方向
固定化DNA聚合酶
5'
3'
ATANGTT
TATNCAA
量測窗口
來自瓦生股之資料
來自克立克股之資料
組合來自瓦生股及克立克股之資料
來自瓦生股及克立克股兩者之資料

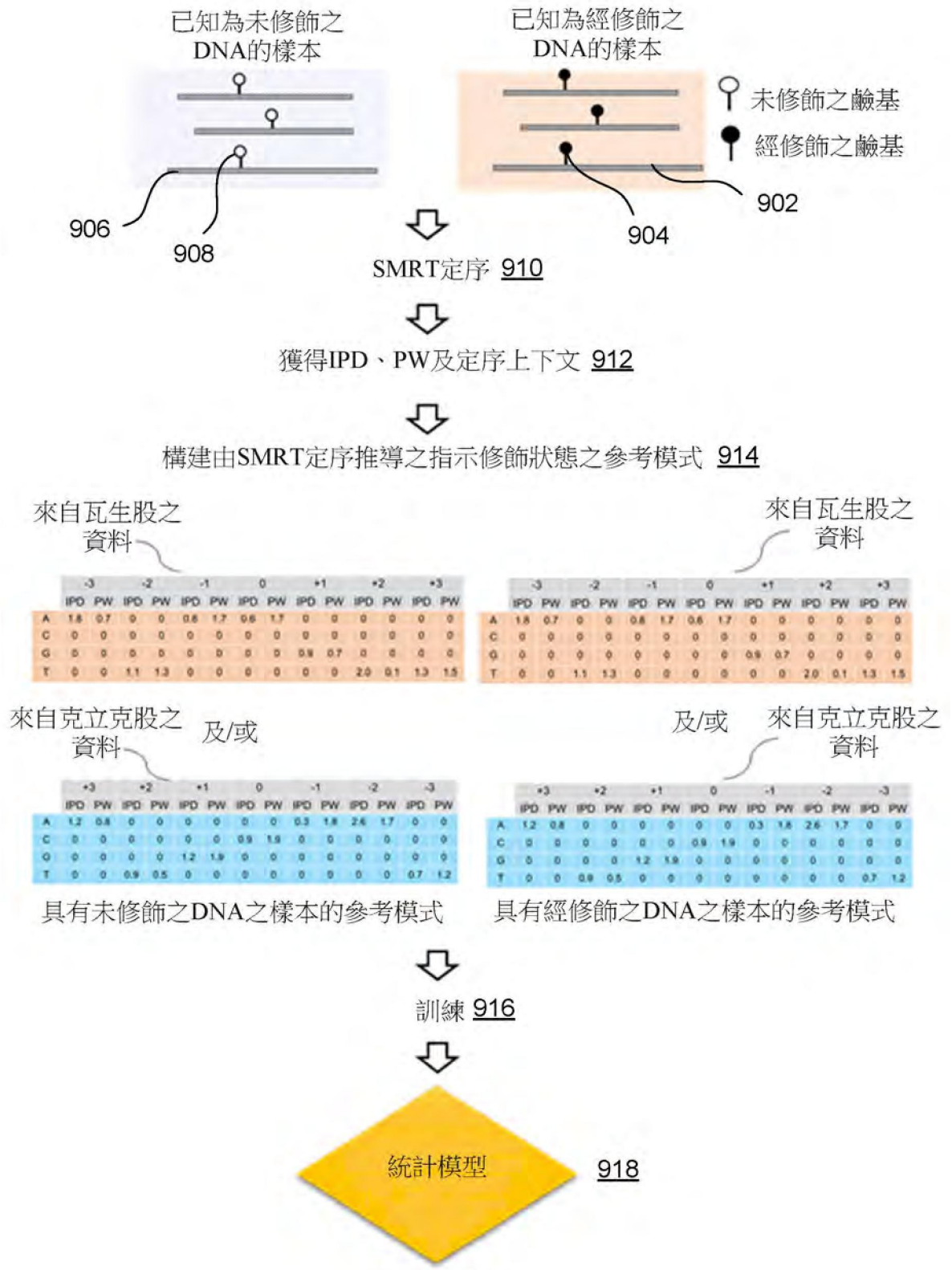
【圖6】


N: 表示任何鹼基。
N̄: 表示與瓦生股之N互補的克立克股中之對應鹼基附近的鹼基。
鹼基修飾
環形DNA模板移動方向
固定化DNA聚合酶
瓦生股
克立克股
5'
3'
ATANGTT
CGCN̄ATG
量測窗口
來自瓦生股之資料
來自克立克股之資料
組合來自瓦生股及克立克股之資料
來自瓦生股及克立克股兩者之資料

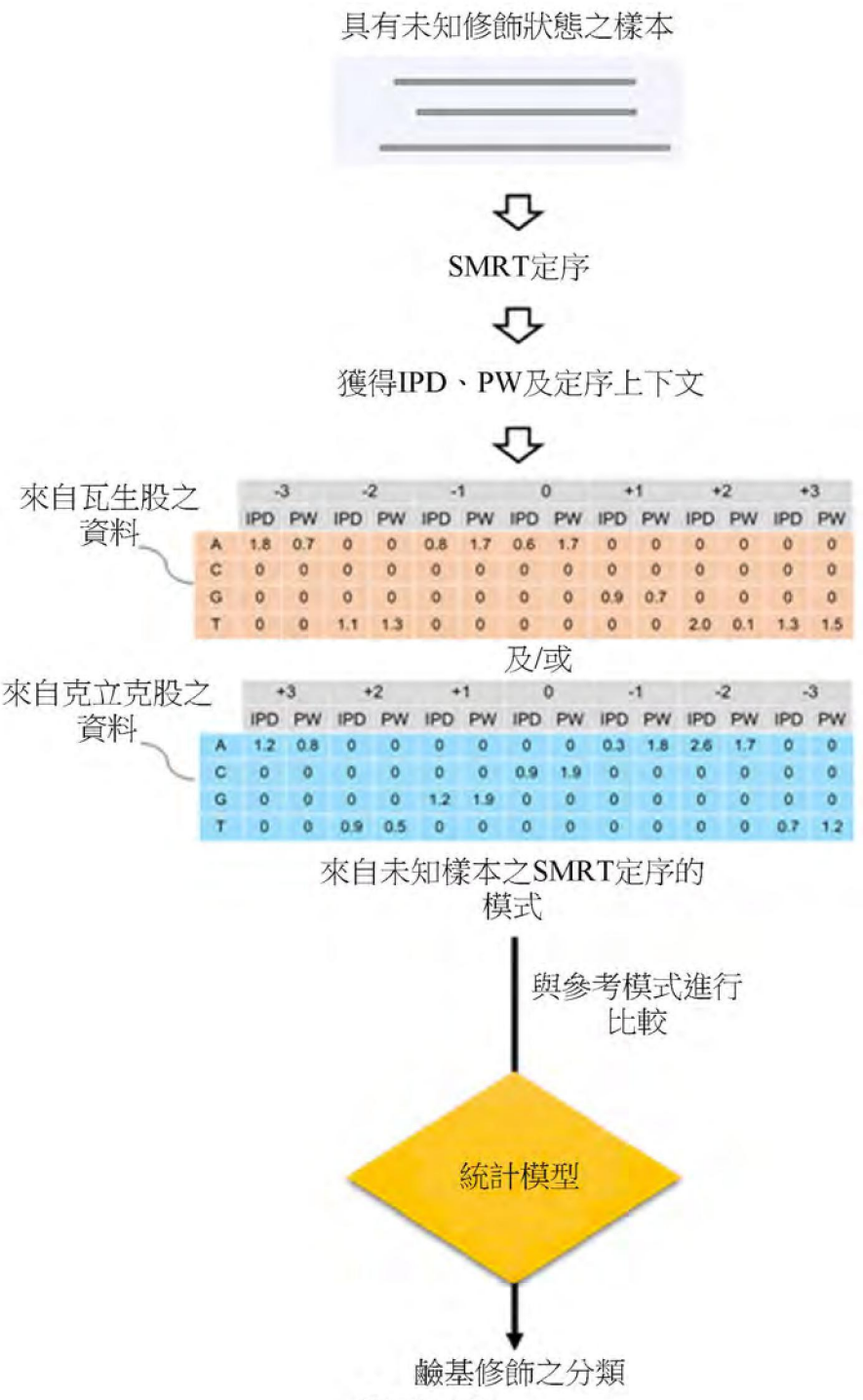
【圖7】

未甲基化之CpG
甲基化之CpG
瓦生股
克立克股
環形DNA模板移動方向
固定化DNA聚合酶
ATACGTT
ATGCAAT
CG
GC
5'
3'
量測窗口
來自瓦生股之資料
來自克立克股之資料
組合來自瓦生股及克立克股之資料
來自瓦生股及克立克股兩者之資料

【圖8】


已知為未修飾之
DNA的樣本
已知為經修飾之
DNA的樣本
未修飾之鹼基
經修飾之鹼基
906
908
902
904
SMRT定序 910
獲得IPD、PW及定序上下文 912
構建由SMRT定序推導之指示修飾狀態之參考模式 914
來自瓦生股之
資料
來自瓦生股之
資料
來自克立克股之
資料
及/或
及/或
來自克立克股之
資料
具有未修飾之DNA之樣本的參考模式
具有經修飾之DNA之樣本的參考模式
訓練 916
統計模型
918

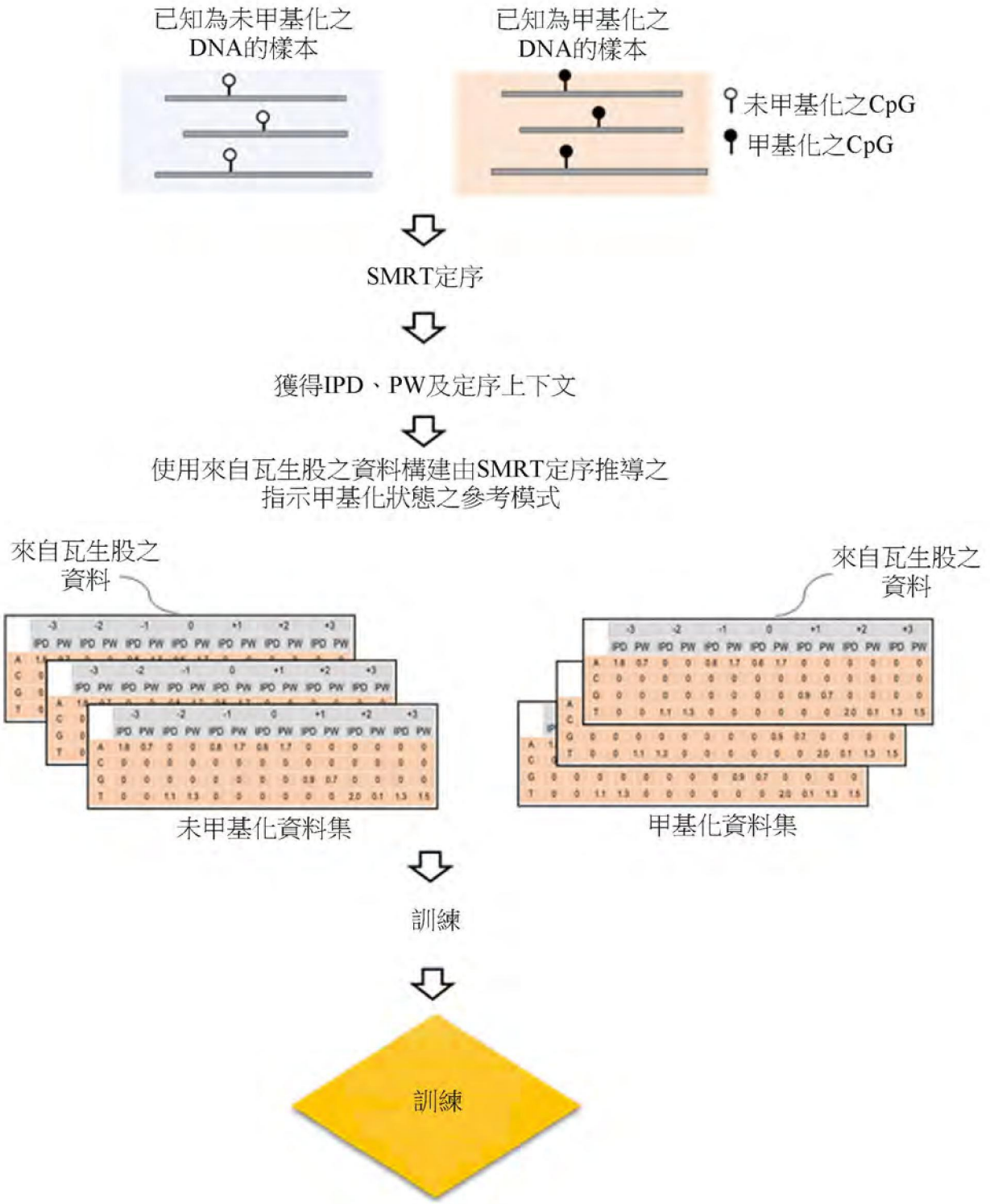
【圖9】



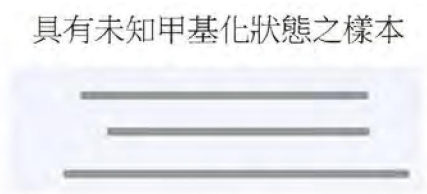
來自瓦生股之資料	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0.6	1.7	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5

來自克立克股之資料	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.2	0.8	0	0	0	0	0	0	0.3	1.8	2.6	1.7	0	0
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	1.2	1.9	0	0	0	0	0	0	0	0
T	0	0	0.9	0.5	0	0	0	0	0	0	0	0	0.7	1.2

【圖10】


已知為未甲基化之
DNA的樣本
已知為甲基化之
DNA的樣本
未甲基化之CpG
甲基化之CpG
SMRT定序
獲得IPD、PW及定序上下文
使用來自瓦生股之資料構建由SMRT定序推導之
指示甲基化狀態之參考模式
來自瓦生股之
資料
來自瓦生股之
資料
未甲基化資料集
甲基化資料集
訓練
訓練

【圖11】

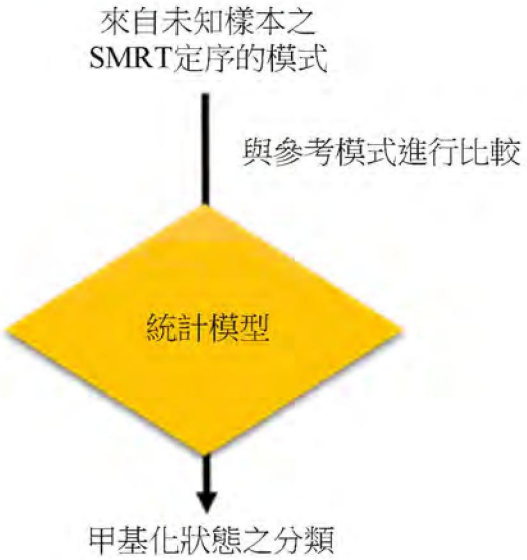


SMRT定序

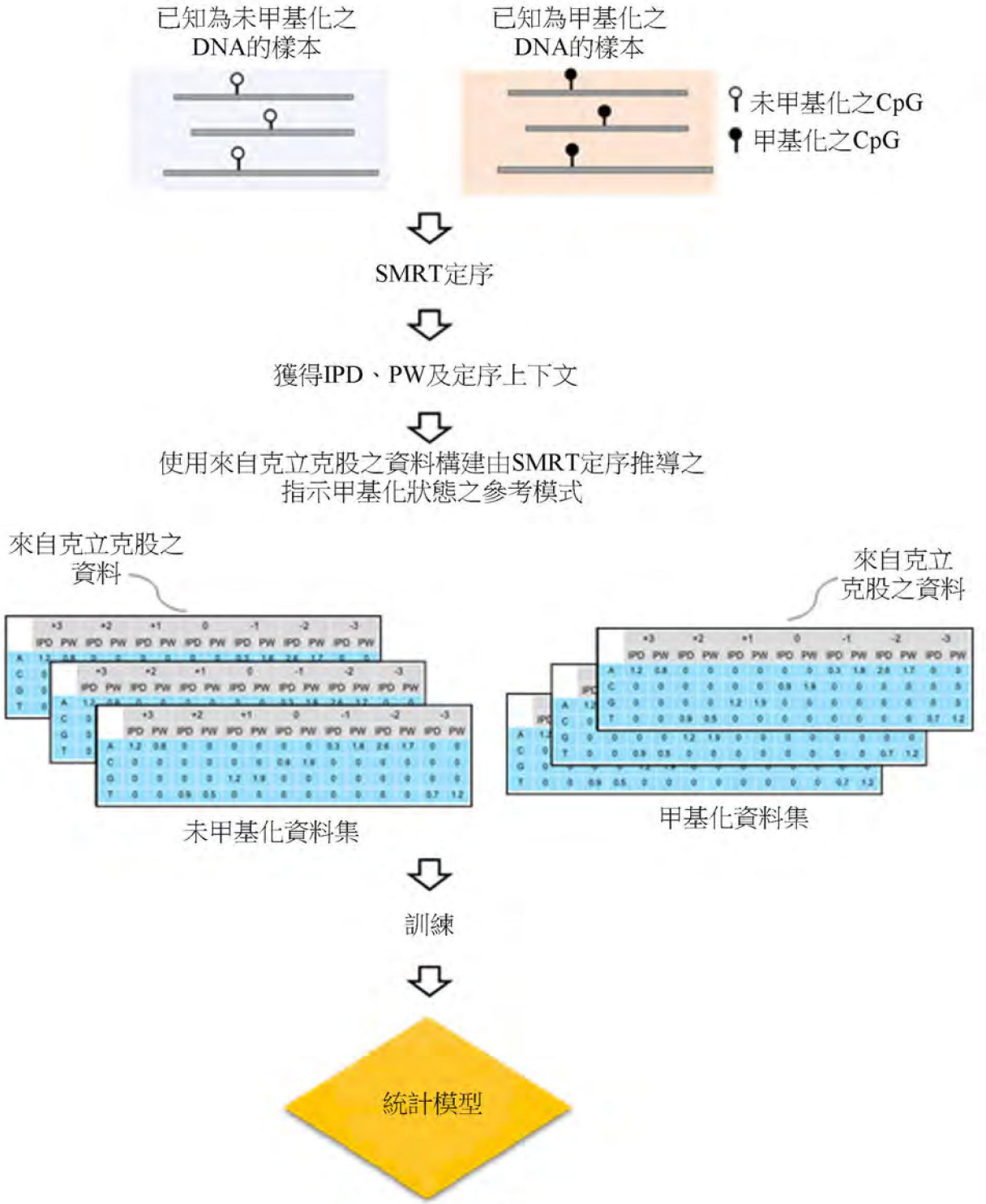
獲得IPD、PW及定序上下文

來自瓦生股之資料

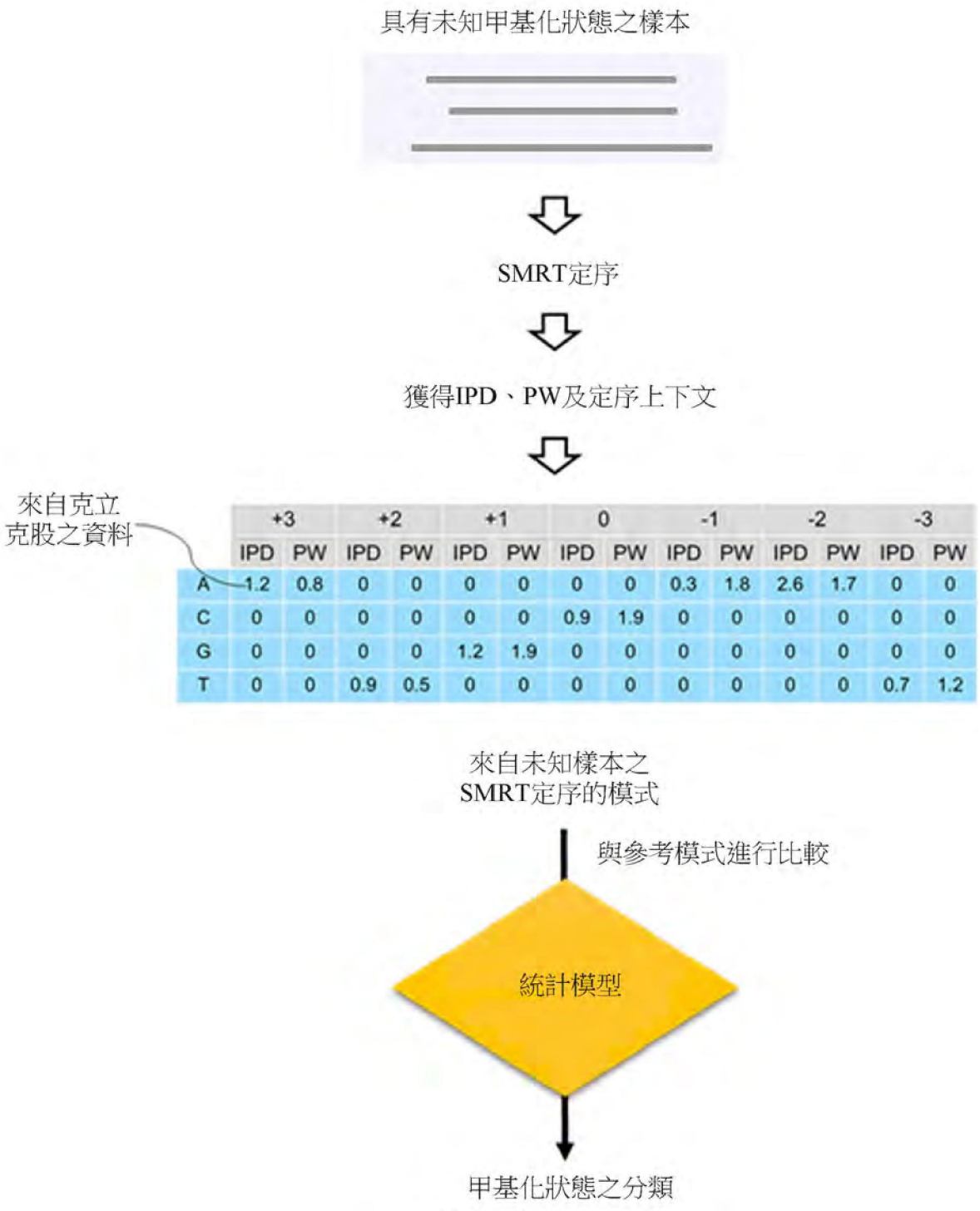
	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0.6	1.7	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5



【圖12】

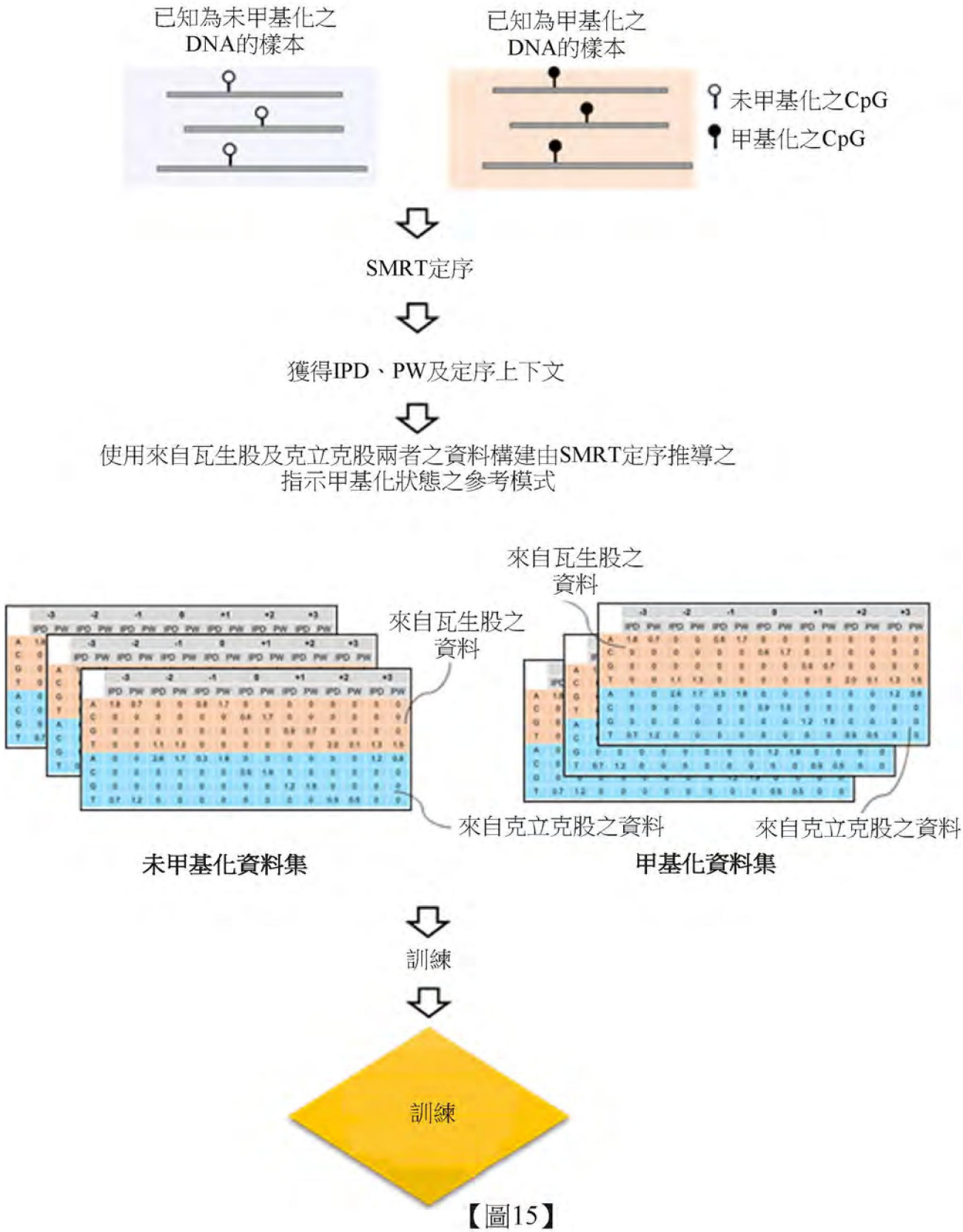

已知為未甲基化之
DNA的樣本
已知為甲基化之
DNA的樣本
未甲基化之CpG
甲基化之CpG
SMRT定序
獲得IPD、PW及定序上下文
使用來自克立克股之資料構建由SMRT定序推導之
指示甲基化狀態之參考模式
來自克立克股之
資料
來自克立
克股之資料
未甲基化資料集
甲基化資料集
訓練
統計模型

【圖13】

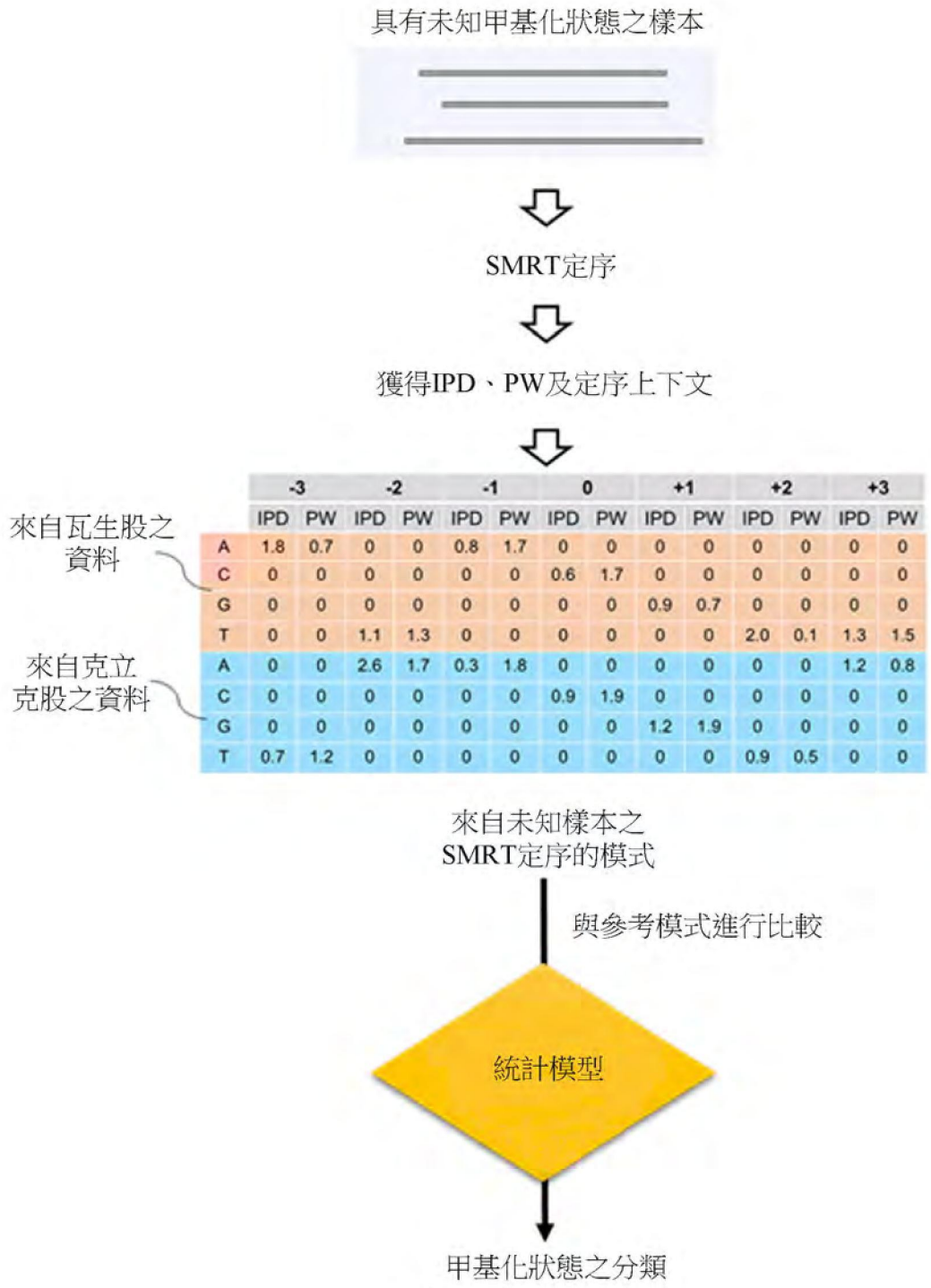


	+3		+2		+1		0		-1		-2		-3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.2	0.8	0	0	0	0	0	0	0.3	1.8	2.6	1.7	0	0
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	1.2	1.9	0	0	0	0	0	0	0	0
T	0	0	0.9	0.5	0	0	0	0	0	0	0	0	0.7	1.2

【圖14】

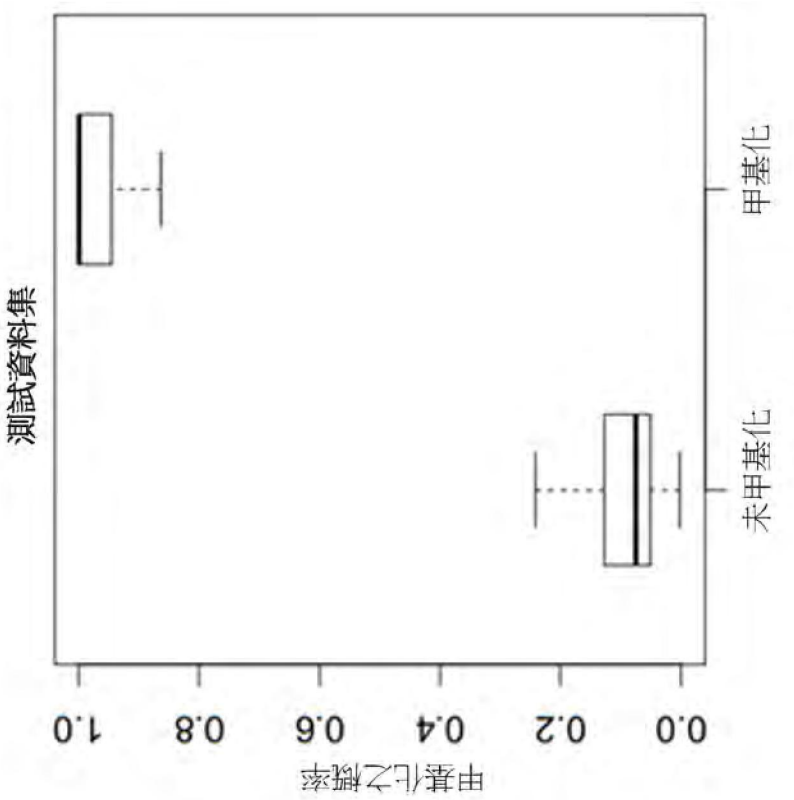

已知為未甲基化之
DNA的樣本
已知為甲基化之
DNA的樣本
未甲基化之CpG
甲基化之CpG
SMRT定序
獲得IPD、PW及定序上下文
使用來自瓦生股及克立克股兩者之資料構建由SMRT定序推導之
指示甲基化狀態之參考模式
來自瓦生股之
資料
來自瓦生股之
資料
來自克立克股之資料
來自克立克股之資料
未甲基化資料集
甲基化資料集
訓練
訓練

【圖15】

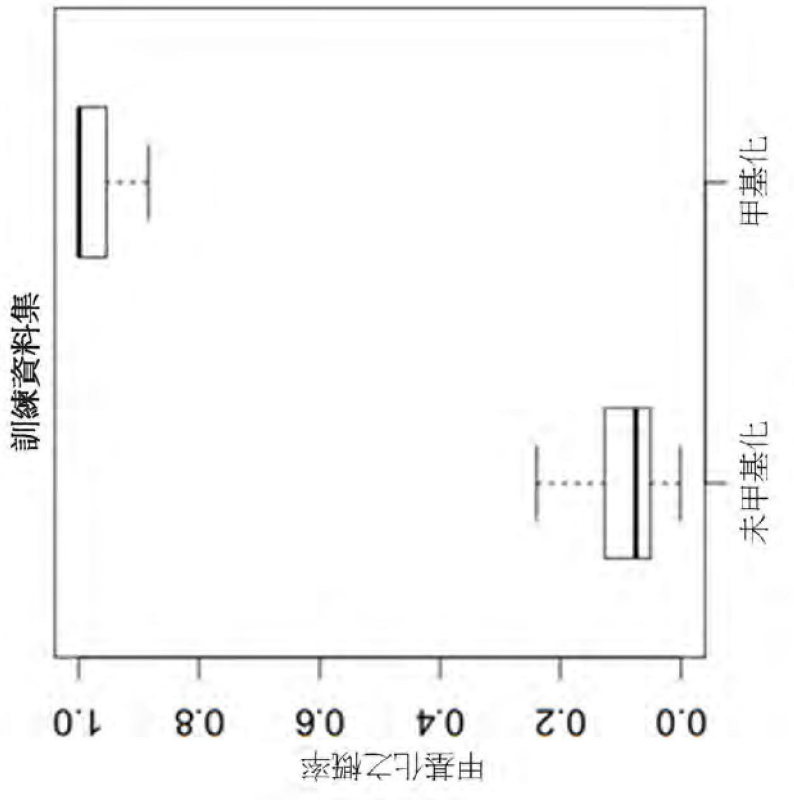


	-3		-2		-1		0		+1		+2		+3	
	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW	IPD	PW
A	1.8	0.7	0	0	0.8	1.7	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0.6	1.7	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0.9	0.7	0	0	0	0
T	0	0	1.1	1.3	0	0	0	0	0	0	2.0	0.1	1.3	1.5
A	0	0	2.6	1.7	0.3	1.8	0	0	0	0	0	0	1.2	0.8
C	0	0	0	0	0	0	0.9	1.9	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	1.2	1.9	0	0	0	0
T	0.7	1.2	0	0	0	0	0	0	0	0	0.9	0.5	0	0

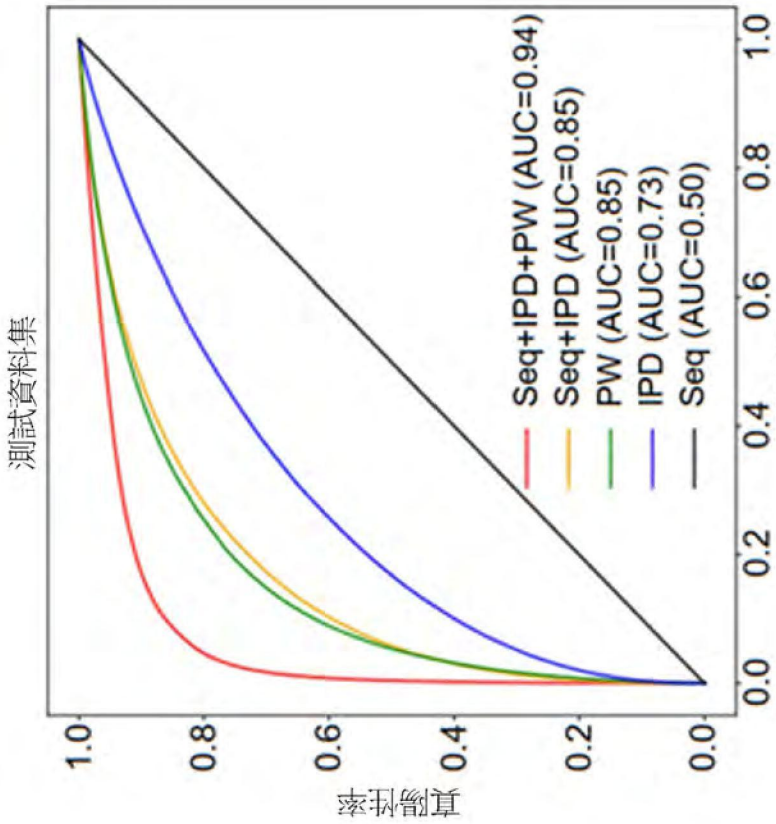
【圖16】


測試資料集
甲基化之機率
0.0
0.2
0.4
0.6
0.8
1.0
未甲基化
甲基化

【圖17B】


訓練資料集
甲基化之機率
0.0
0.2
0.4
0.6
0.8
1.0
未甲基化
甲基化

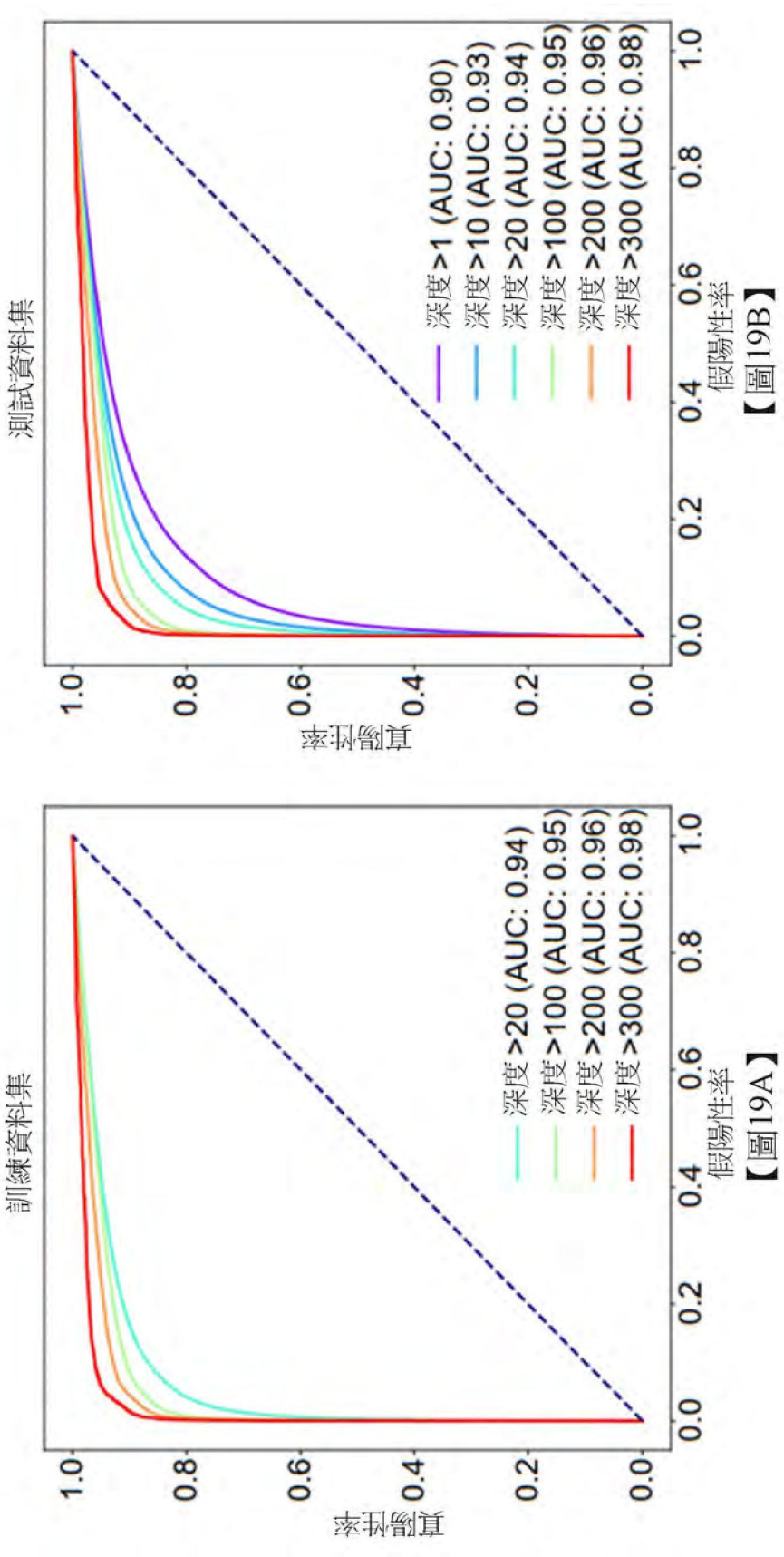
【圖17A】


訓練資料集
真陽性率
假陽性率
0.0
0.2
0.4
0.6
0.8
1.0
Seq+IPD+PW (AUC=0.94)
Seq+IPD (AUC=0.86)
PW (AUC=0.86)
IPD (AUC=0.74)
Seq (AUC=0.50)

【圖18A】

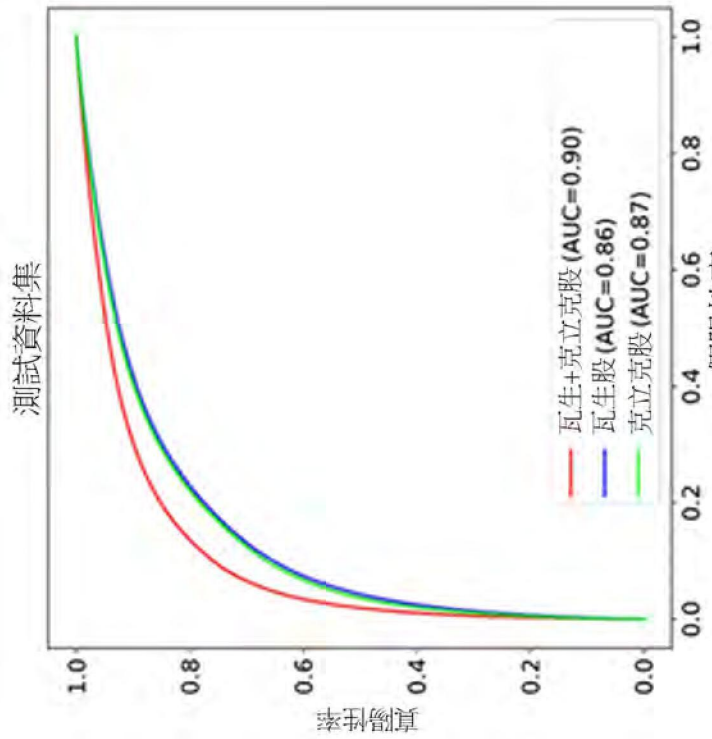
測試資料集
真陽性率
假陽性率
0.0
0.2
0.4
0.6
0.8
1.0
Seq+IPD+PW (AUC=0.94)
Seq+IPD (AUC=0.85)
PW (AUC=0.85)
IPD (AUC=0.73)
Seq (AUC=0.50)

【圖18B】

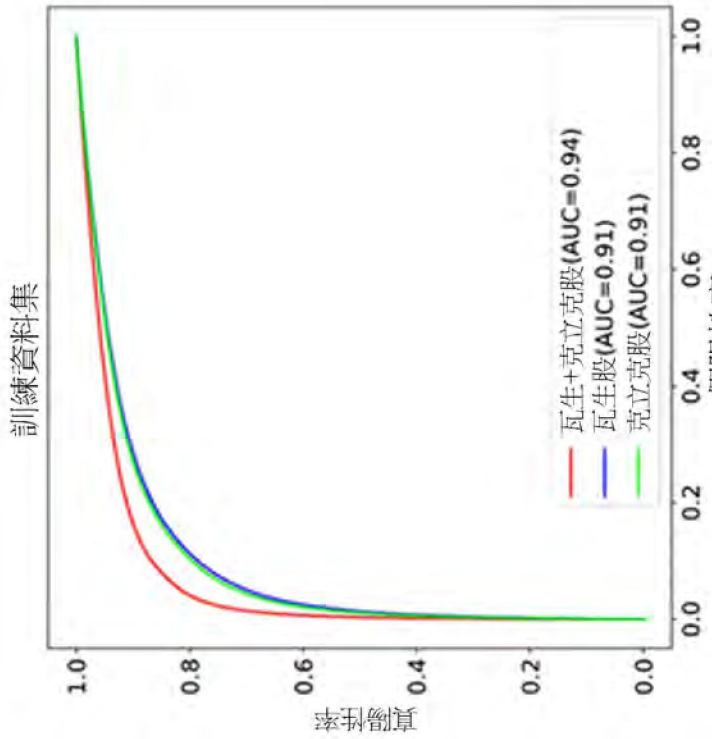

訓練資料集
真陽性率
假陽性率
深度>20 (AUC: 0.94)
深度>100 (AUC: 0.95)
深度>200 (AUC: 0.96)
深度>300 (AUC: 0.98)
測試資料集
真陽性率
假陽性率
深度>1 (AUC: 0.90)
深度>10 (AUC: 0.93)
深度>20 (AUC: 0.94)
深度>100 (AUC: 0.95)
深度>200 (AUC: 0.96)
深度>300 (AUC: 0.98)

【圖19A】

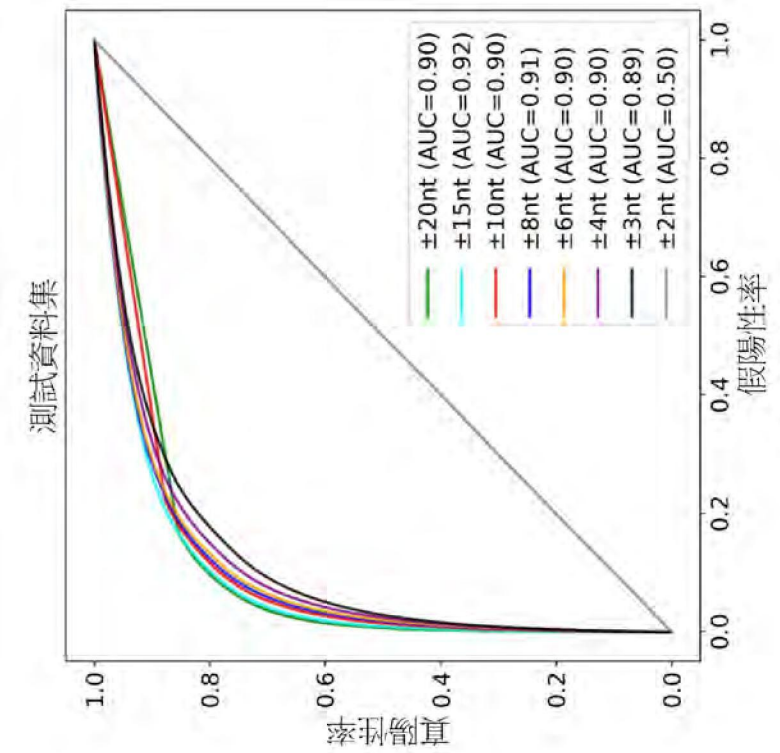
【圖19B】


測試資料集
真陽性率
假陽性率
瓦生+克立克股 (AUC=0.90)
瓦生股 (AUC=0.86)
克立克股 (AUC=0.87)
0.0
0.2
0.4
0.6
0.8
1.0

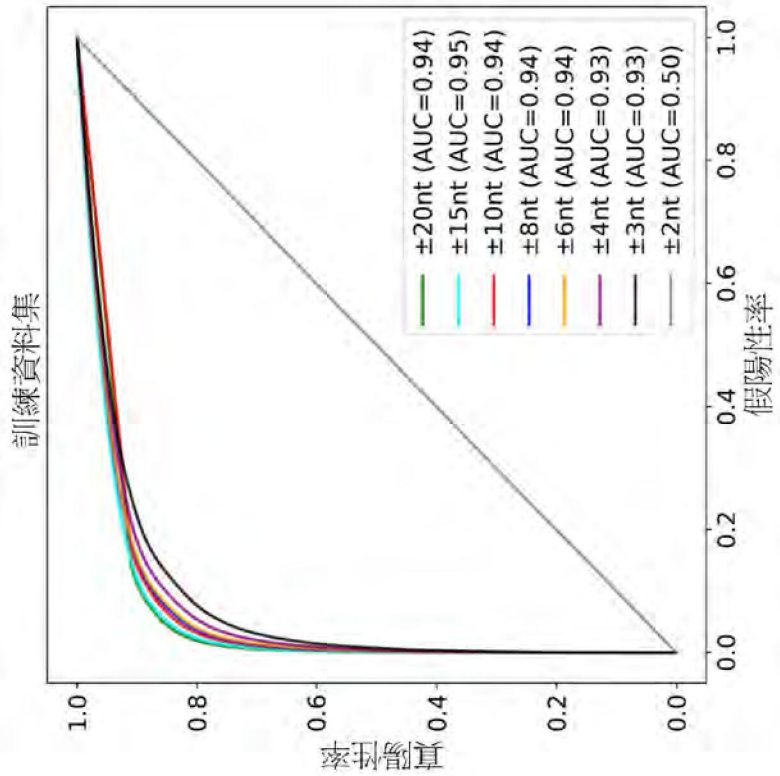
【圖20B】


訓練資料集
真陽性率
假陽性率
瓦生+克立克股(AUC=0.94)
瓦生股(AUC=0.91)
克立克股(AUC=0.91)
0.0
0.2
0.4
0.6
0.8
1.0

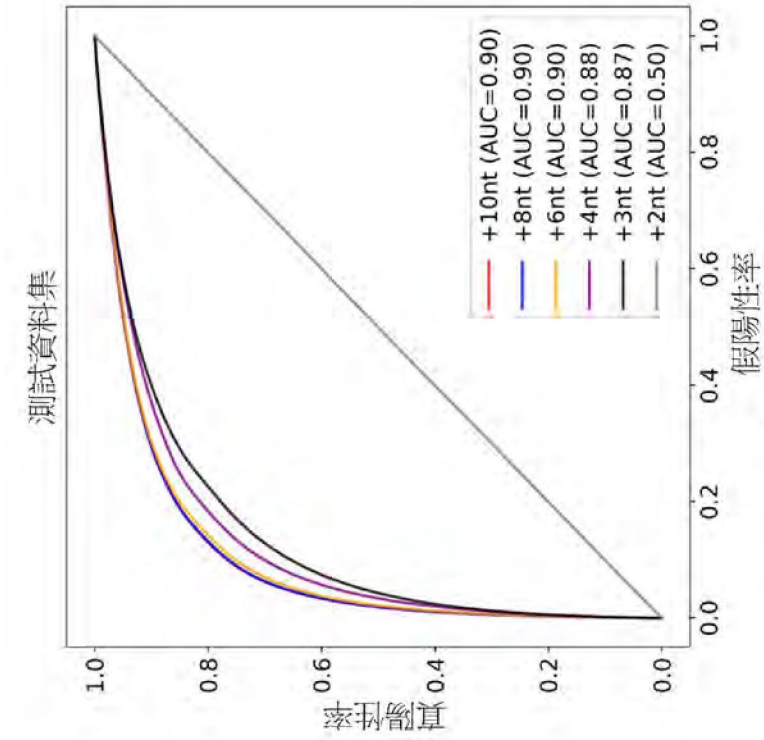
【圖20A】


測試資料集
真陽性率
假陽性率
0.0
0.2
0.4
0.6
0.8
1.0
±20nt (AUC=0.90)
±15nt (AUC=0.92)
±10nt (AUC=0.90)
±8nt (AUC=0.91)
±6nt (AUC=0.90)
±4nt (AUC=0.90)
±3nt (AUC=0.89)
±2nt (AUC=0.50)

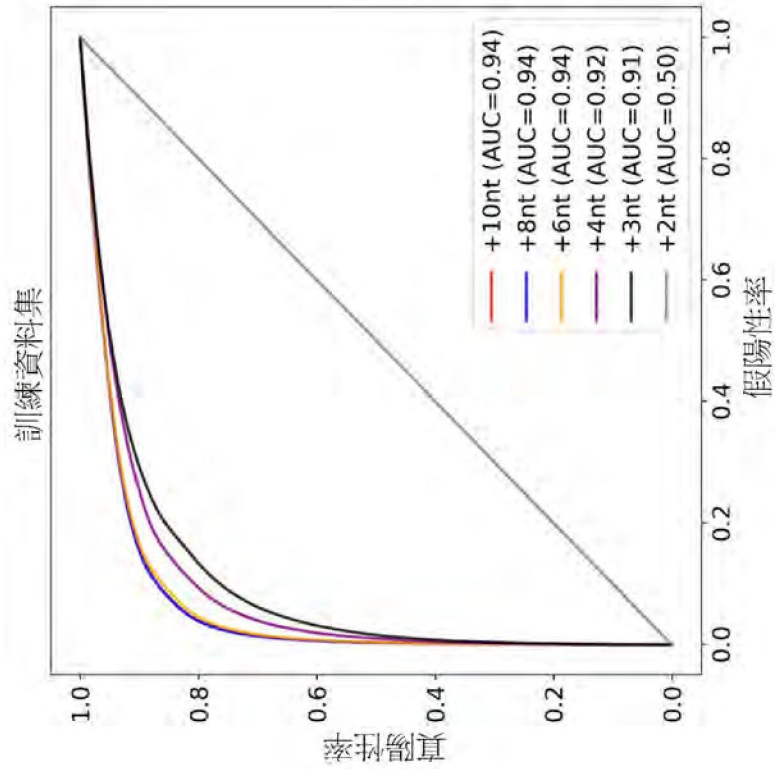
【圖21B】


訓練資料集
真陽性率
假陽性率
0.0
0.2
0.4
0.6
0.8
1.0
±20nt (AUC=0.94)
±15nt (AUC=0.95)
±10nt (AUC=0.94)
±8nt (AUC=0.94)
±6nt (AUC=0.94)
±4nt (AUC=0.93)
±3nt (AUC=0.93)
±2nt (AUC=0.50)

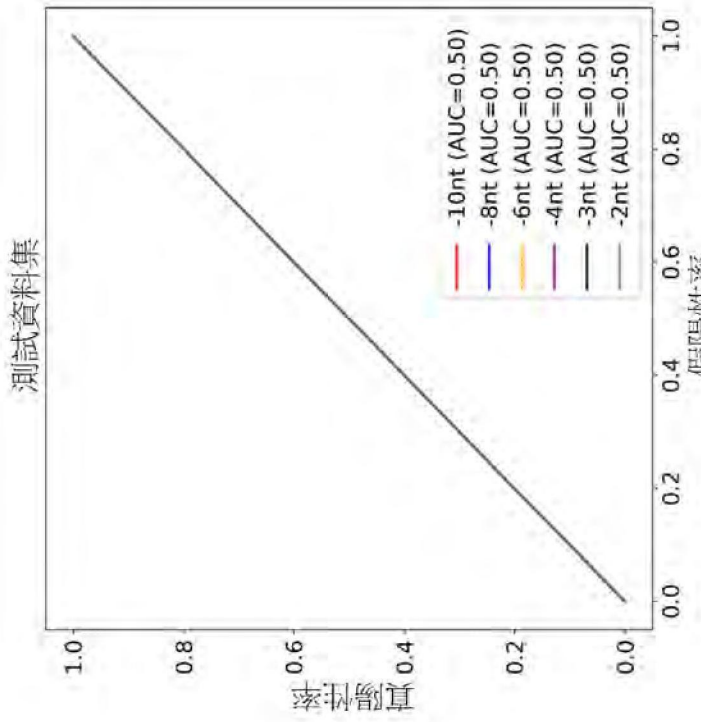
【圖21A】


測試資料集
真陽性率
假陽性率
+10nt (AUC=0.90)
+8nt (AUC=0.90)
+6nt (AUC=0.90)
+4nt (AUC=0.88)
+3nt (AUC=0.87)
+2nt (AUC=0.50)
0.0
0.2
0.4
0.6
0.8
1.0

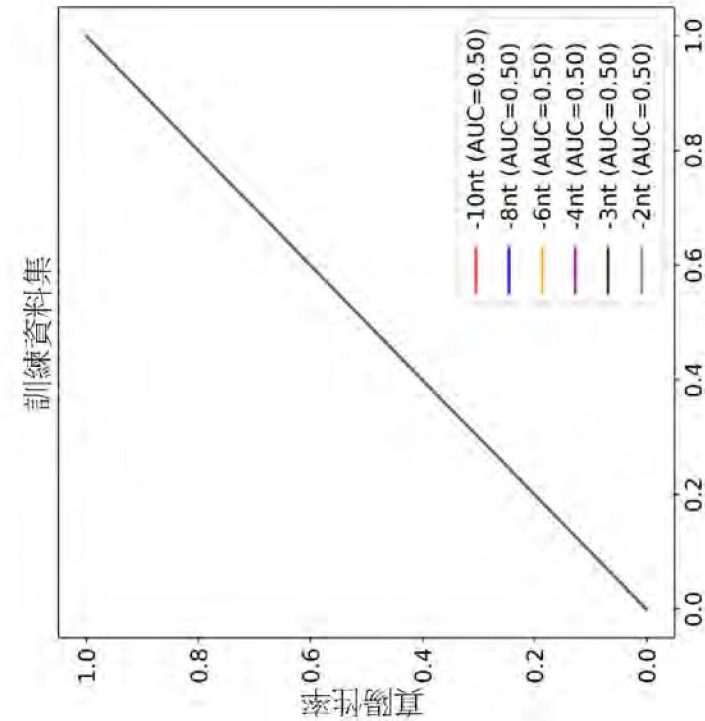
【圖22B】


訓練資料集
真陽性率
假陽性率
+10nt (AUC=0.94)
+8nt (AUC=0.94)
+6nt (AUC=0.94)
+4nt (AUC=0.92)
+3nt (AUC=0.91)
+2nt (AUC=0.50)
0.0
0.2
0.4
0.6
0.8
1.0

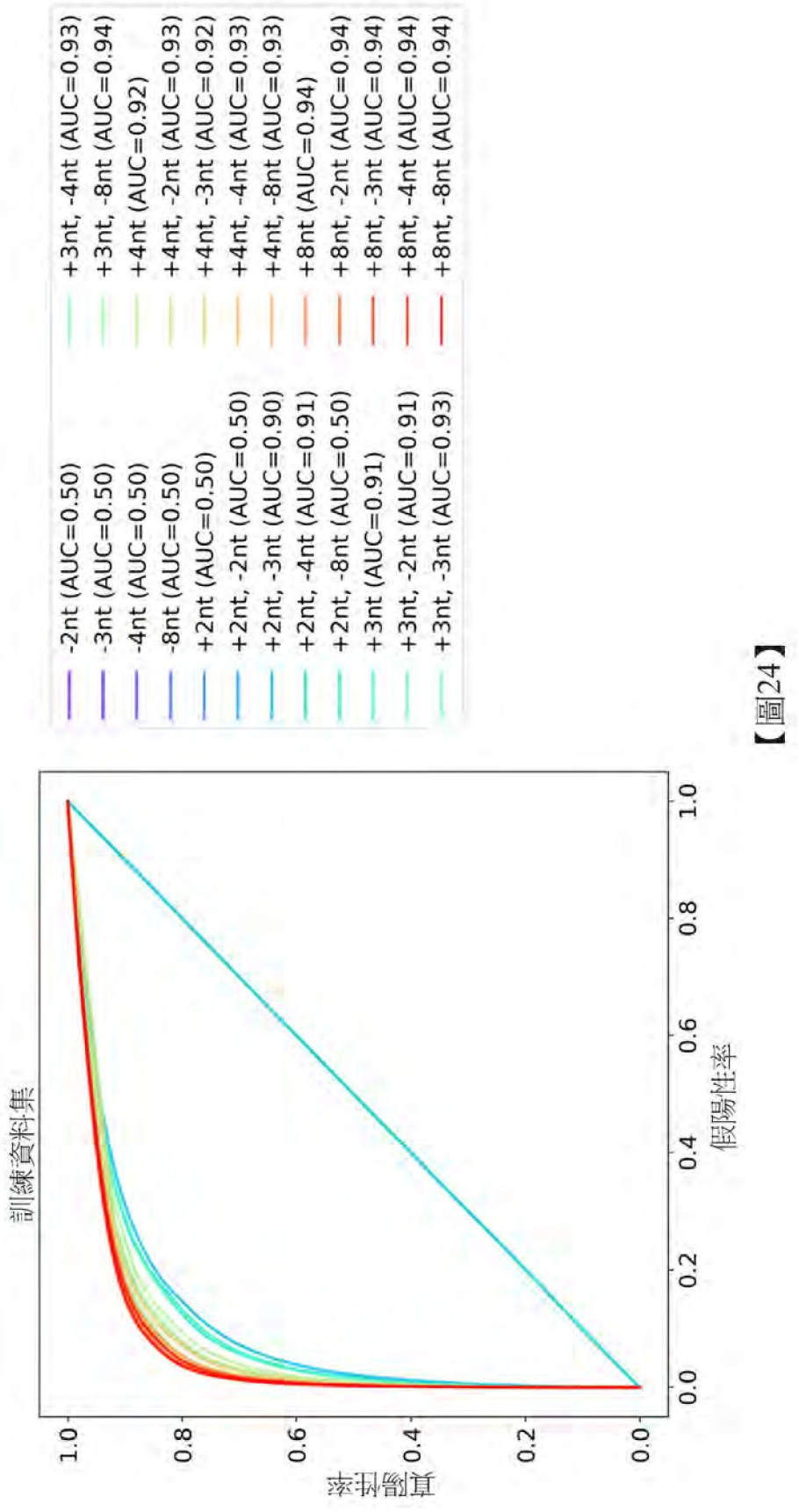
【圖22A】


測試資料集
真陽性率
假陽性率
-10nt (AUC=0.50)
-8nt (AUC=0.50)
-6nt (AUC=0.50)
-4nt (AUC=0.50)
-3nt (AUC=0.50)
-2nt (AUC=0.50)

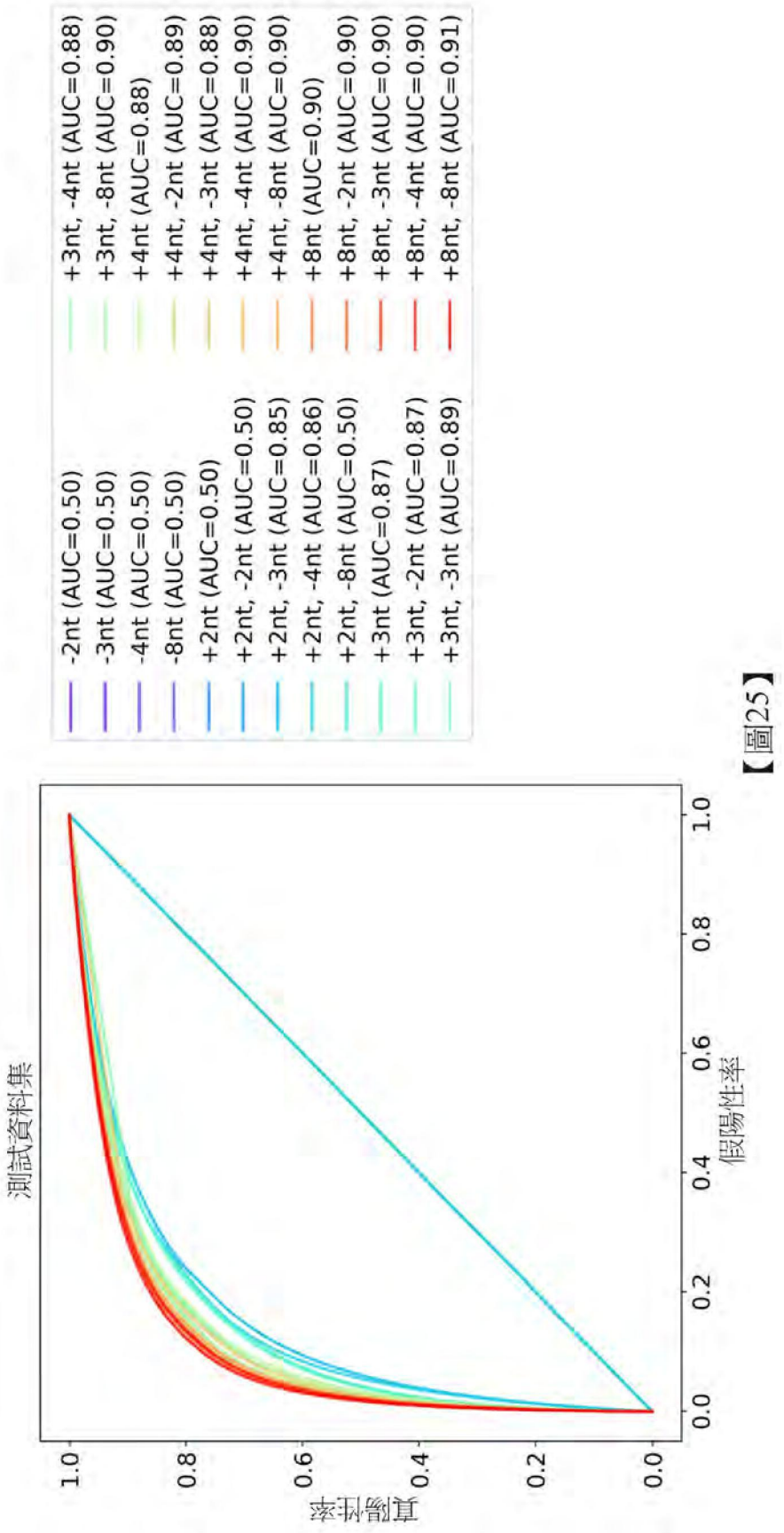
【圖23B】


訓練資料集
真陽性率
假陽性率
-10nt (AUC=0.50)
-8nt (AUC=0.50)
-6nt (AUC=0.50)
-4nt (AUC=0.50)
-3nt (AUC=0.50)
-2nt (AUC=0.50)

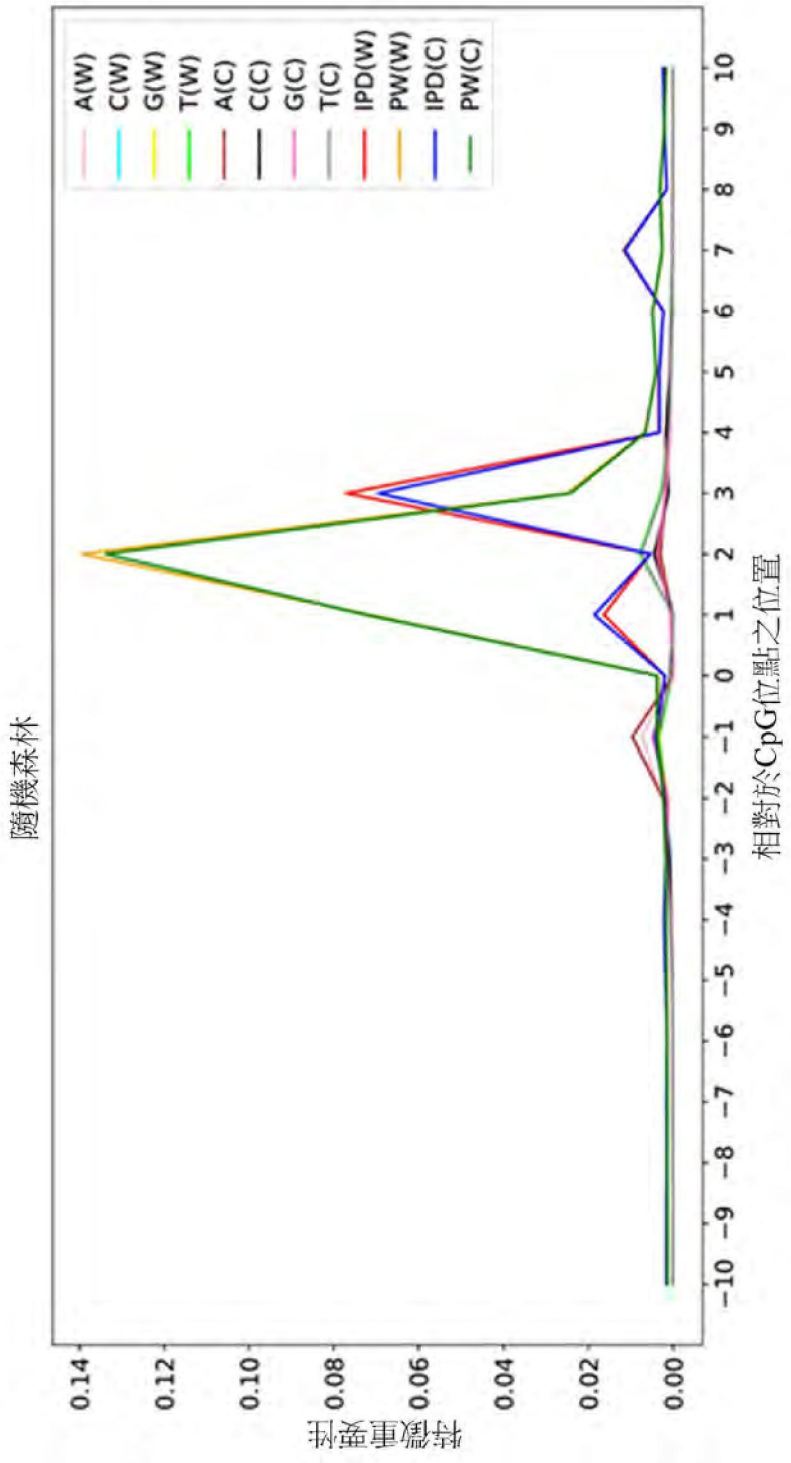
【圖23A】


-2nt (AUC=0.50)
-3nt (AUC=0.50)
-4nt (AUC=0.50)
-8nt (AUC=0.50)
+2nt (AUC=0.50)
+2nt, -2nt (AUC=0.50)
+2nt, -3nt (AUC=0.90)
+2nt, -4nt (AUC=0.91)
+2nt, -8nt (AUC=0.50)
+3nt (AUC=0.91)
+3nt, -2nt (AUC=0.91)
+3nt, -3nt (AUC=0.93)
+3nt, -4nt (AUC=0.93)
+3nt, -8nt (AUC=0.94)
+4nt (AUC=0.92)
+4nt, -2nt (AUC=0.93)
+4nt, -3nt (AUC=0.92)
+4nt, -4nt (AUC=0.93)
+4nt, -8nt (AUC=0.93)
+8nt (AUC=0.94)
+8nt, -2nt (AUC=0.94)
+8nt, -3nt (AUC=0.94)
+8nt, -4nt (AUC=0.94)
+8nt, -8nt (AUC=0.94)
訓練資料集
真陽性率
假陽性率
0.0
0.2
0.4
0.6
0.8
1.0

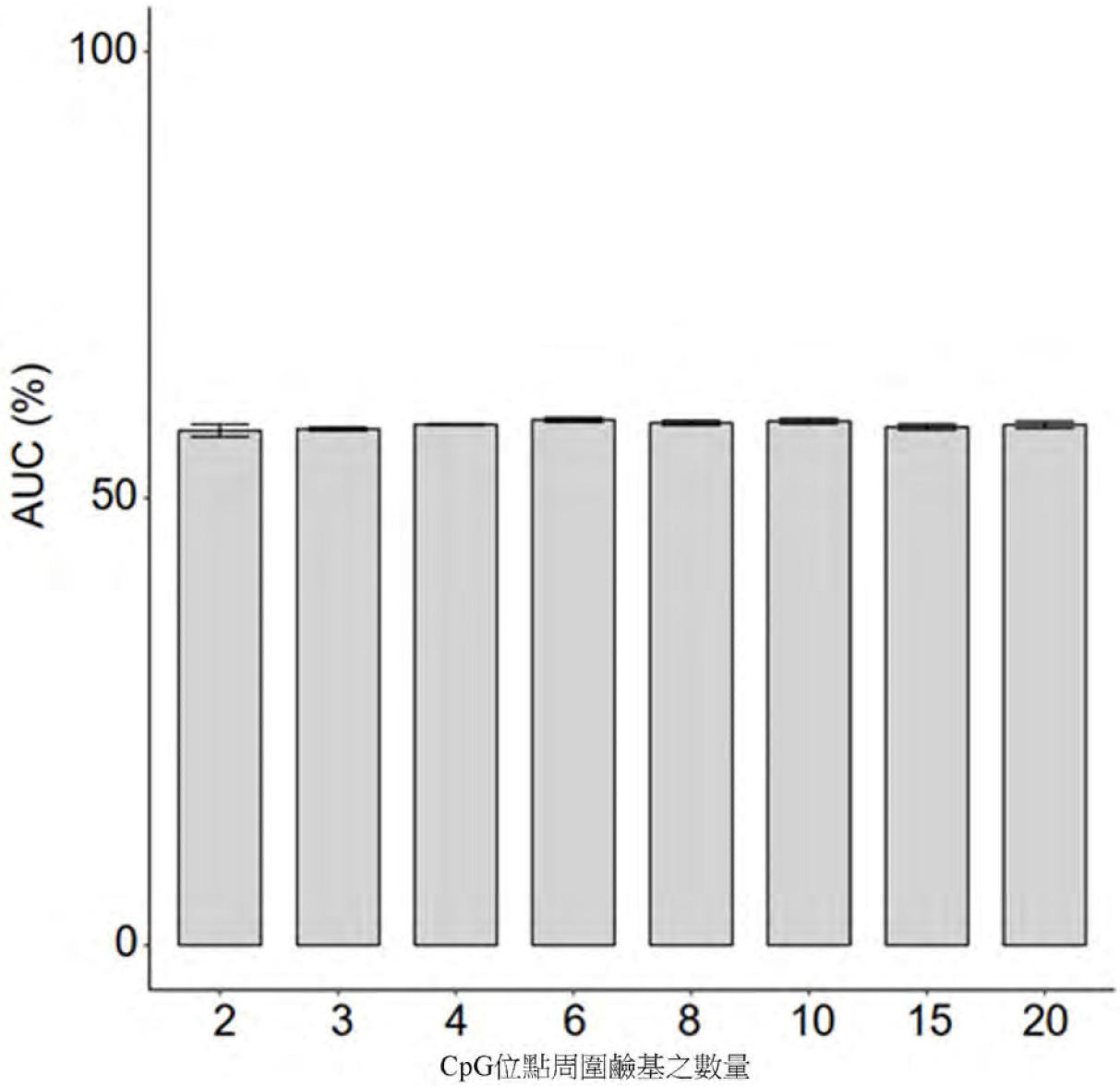
【圖24】


-2nt (AUC=0.50)
-3nt (AUC=0.50)
-4nt (AUC=0.50)
-8nt (AUC=0.50)
+2nt (AUC=0.50)
+2nt, -2nt (AUC=0.50)
+2nt, -3nt (AUC=0.85)
+2nt, -4nt (AUC=0.86)
+2nt, -8nt (AUC=0.50)
+3nt (AUC=0.87)
+3nt, -2nt (AUC=0.87)
+3nt, -3nt (AUC=0.89)
+3nt, -4nt (AUC=0.88)
+3nt, -8nt (AUC=0.90)
+4nt (AUC=0.88)
+4nt, -2nt (AUC=0.89)
+4nt, -3nt (AUC=0.88)
+4nt, -4nt (AUC=0.90)
+4nt, -8nt (AUC=0.90)
+8nt (AUC=0.90)
+8nt, -2nt (AUC=0.90)
+8nt, -3nt (AUC=0.90)
+8nt, -4nt (AUC=0.90)
+8nt, -8nt (AUC=0.91)
測試資料集
真陽性率
假陽性率
0.0
0.2
0.4
0.6
0.8
1.0

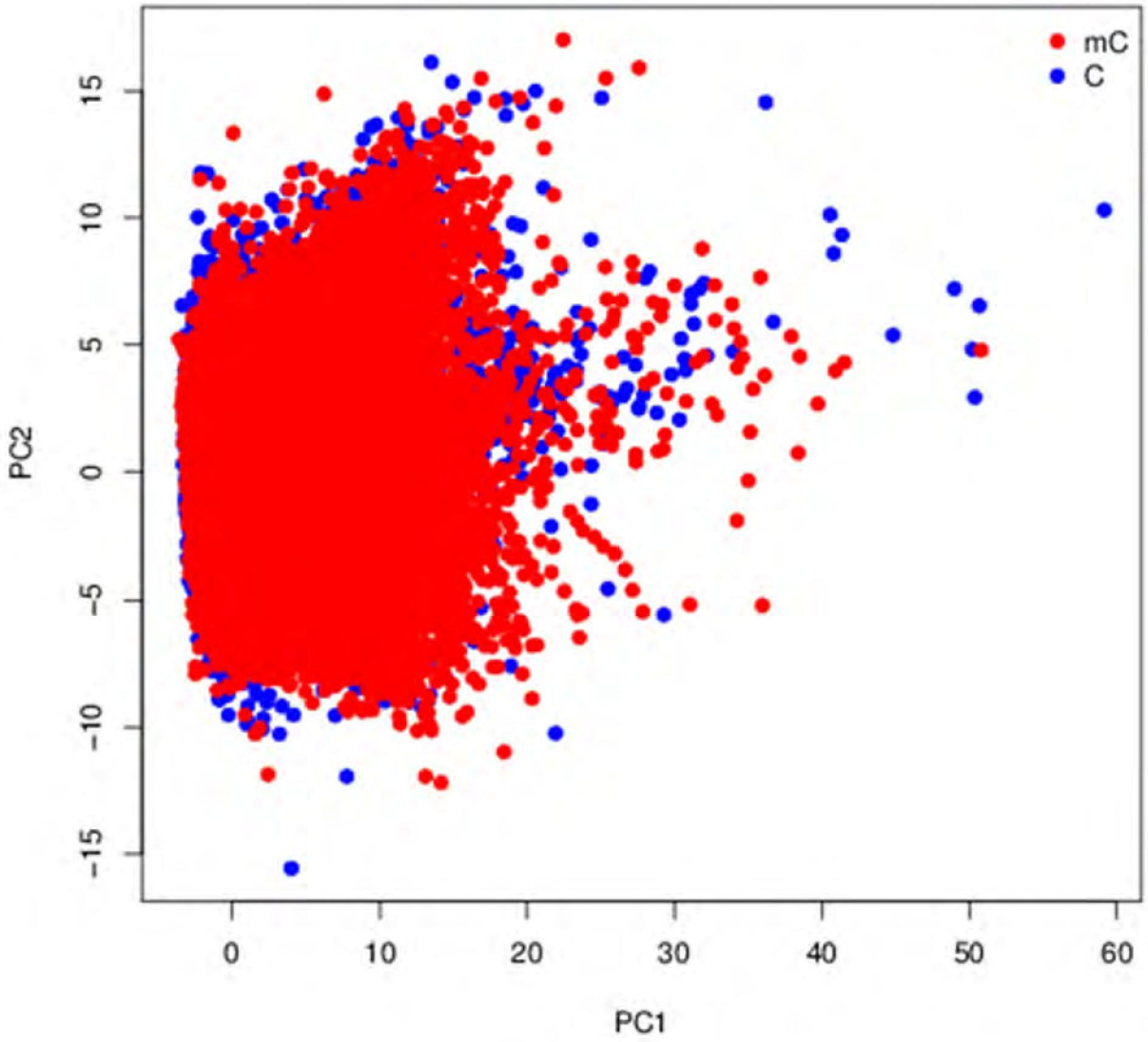
【圖25】


隨機森林
特徵重要性
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
−10 −9 −8 −7 −6 −5 −4 −3 −2 −1 0 1 2 3 4 5 6 7 8 9 10
相對於CpG位點之位置
A(W)
C(W)
G(W)
T(W)
A(C)
C(C)
G(C)
T(C)
IPD(W)
PW(W)
IPD(C)
PW(C)

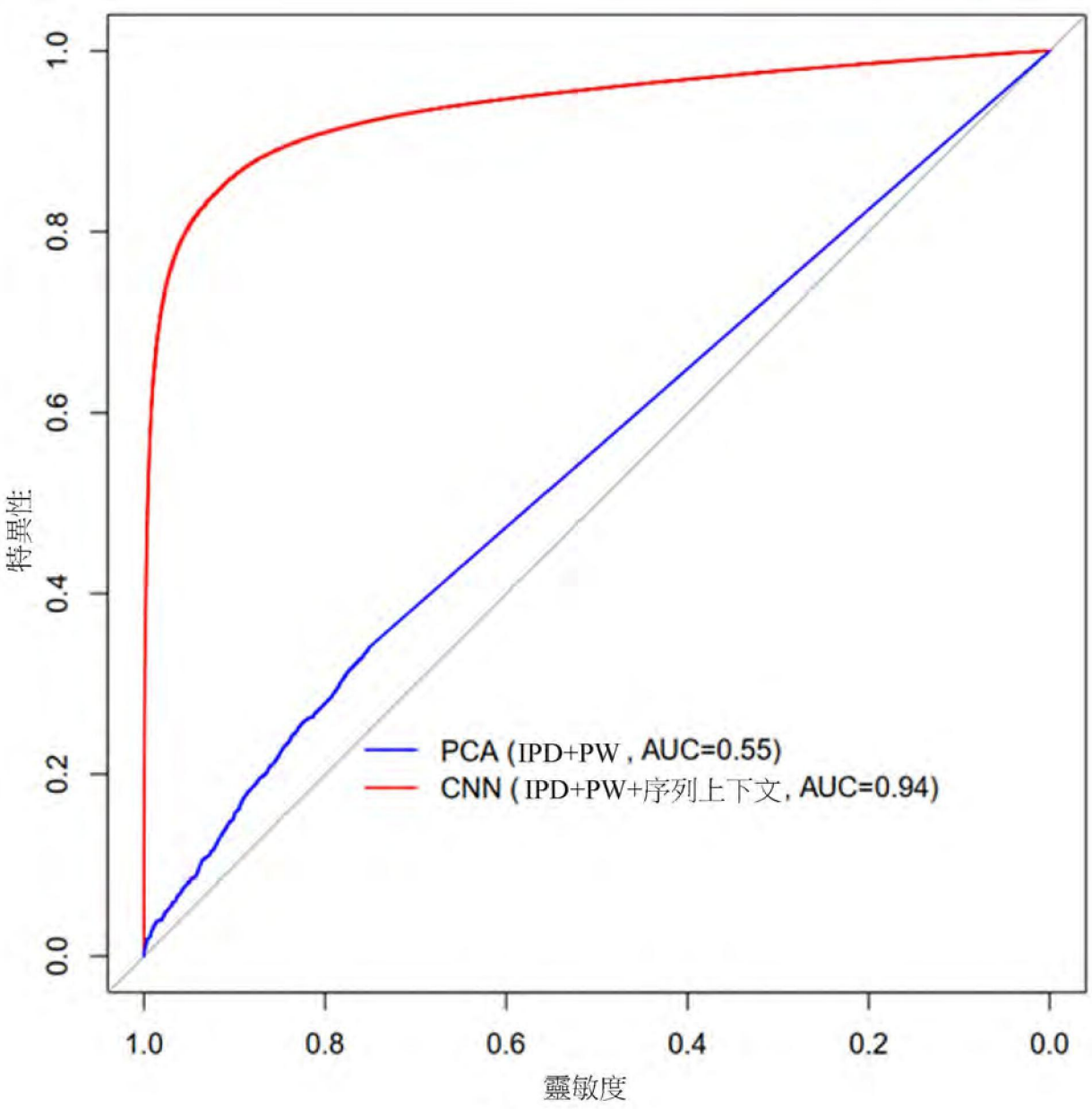
【圖26】



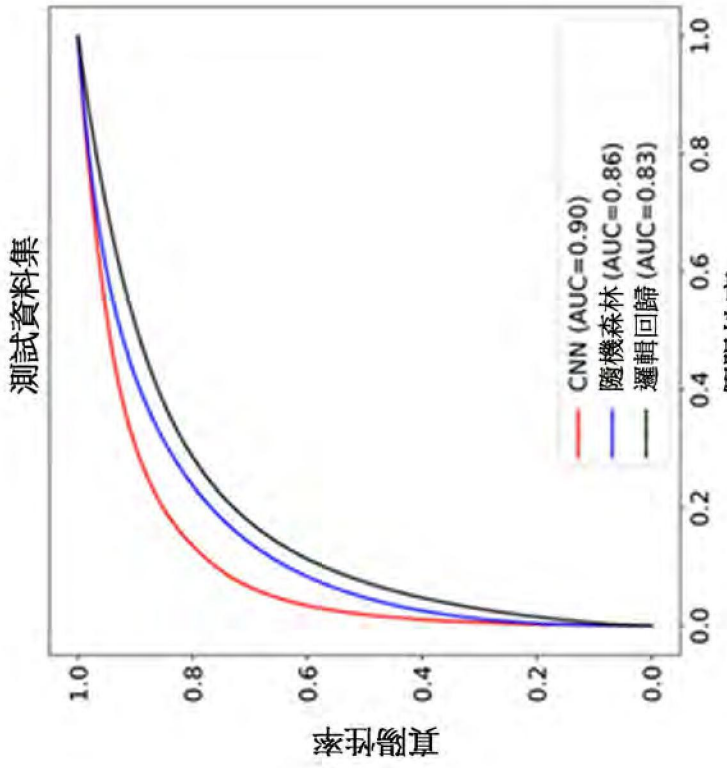
【圖27】



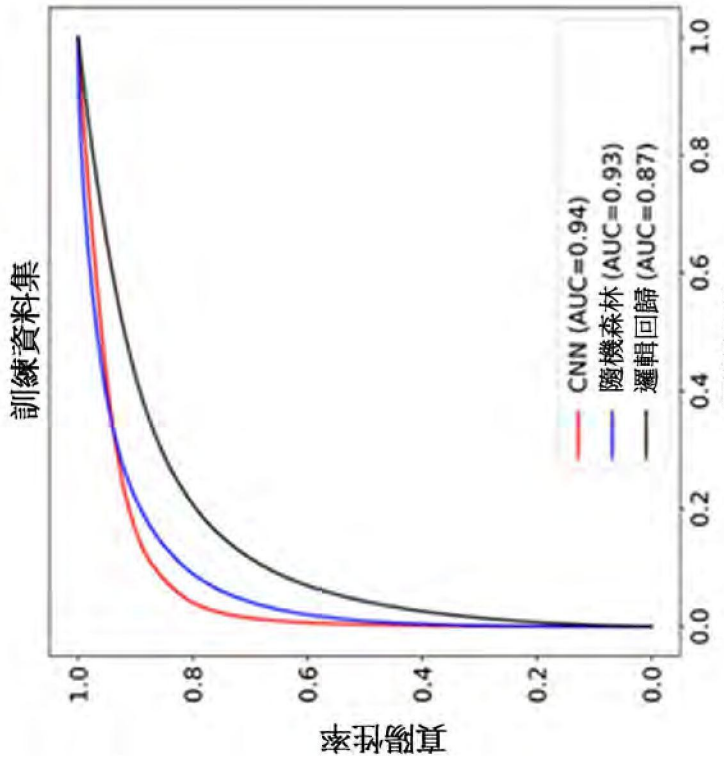
【圖28】


1.0
0.8
0.6
0.4
0.2
0.0
特異性
PCA (IPD+PW , AUC=0.55)
CNN (IPD+PW+序列上下文, AUC=0.94)
1.0
0.8
0.6
0.4
0.2
0.0
靈敏度

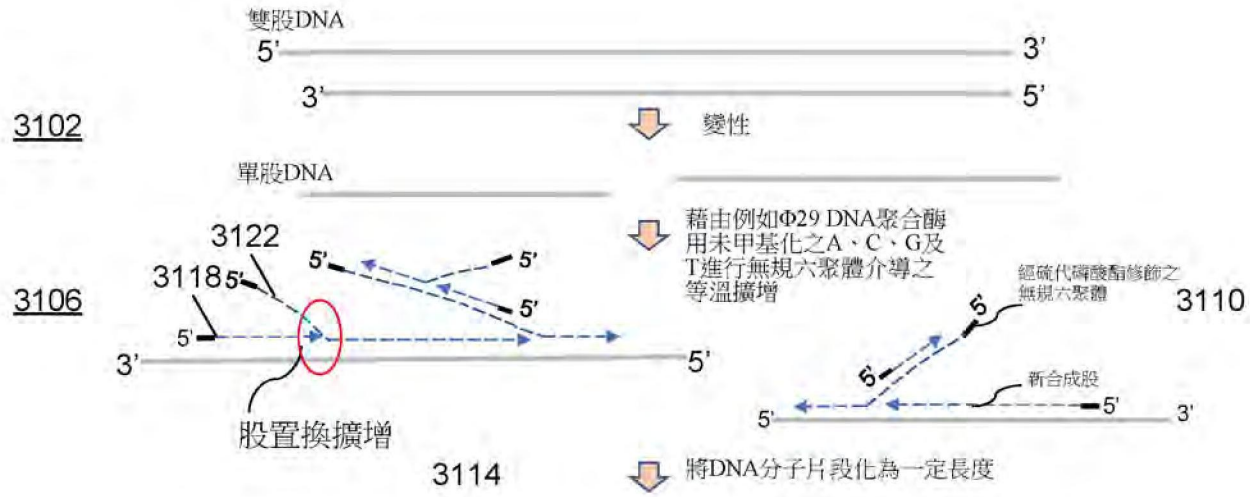
【圖29】

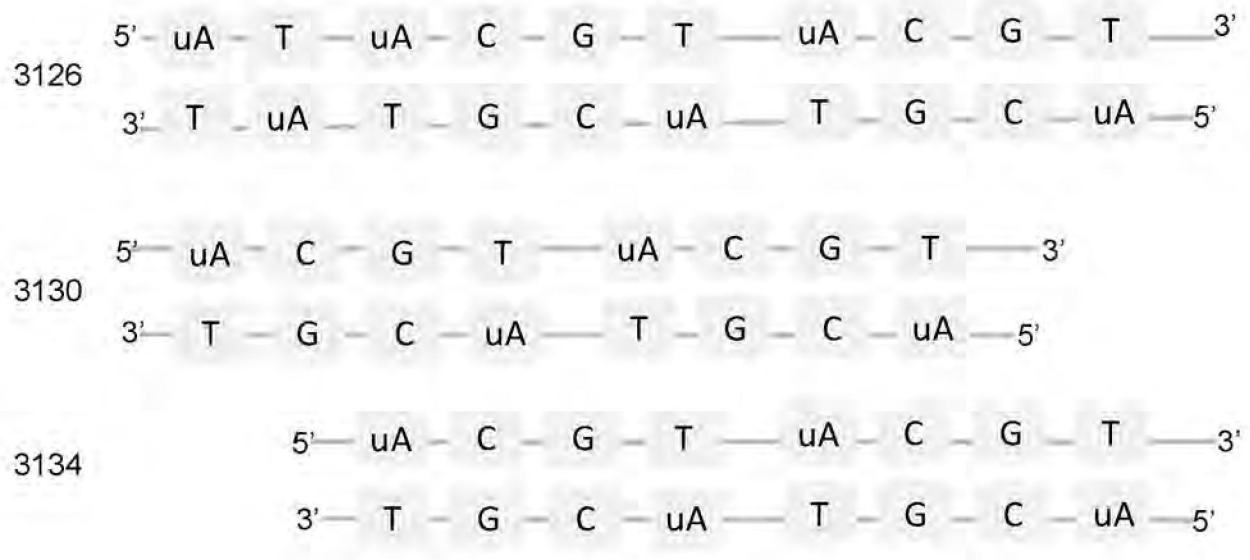

測試資料集
真陽性率
假陽性率
CNN (AUC=0.90)
隨機森林 (AUC=0.86)
邏輯回歸 (AUC=0.83)
0.0
0.2
0.4
0.6
0.8
1.0

【圖30B】


訓練資料集
真陽性率
假陽性率
CNN (AUC=0.94)
隨機森林 (AUC=0.93)
邏輯回歸 (AUC=0.87)
0.0
0.2
0.4
0.6
0.8
1.0

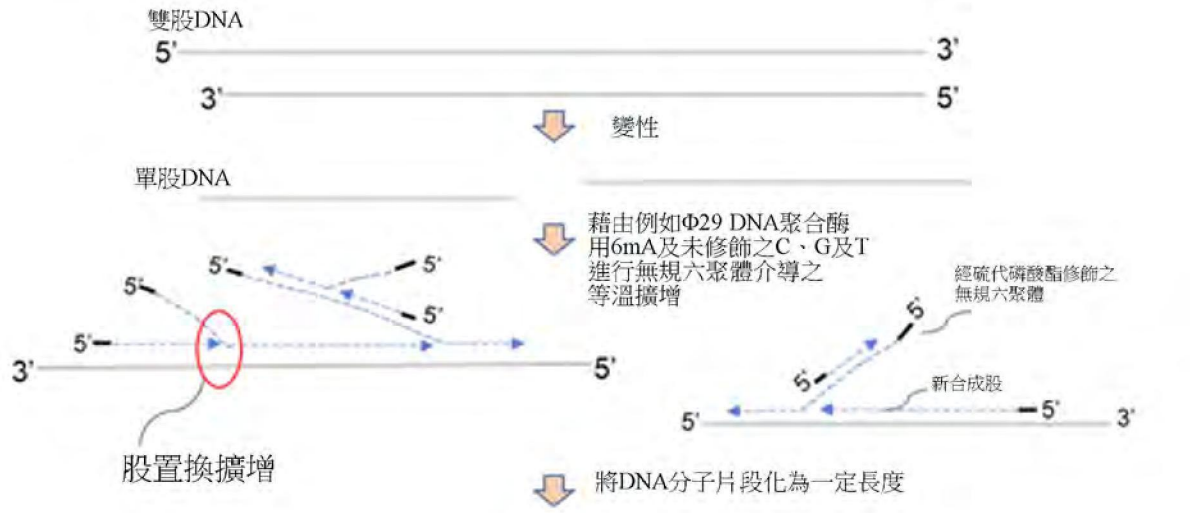
【圖30A】

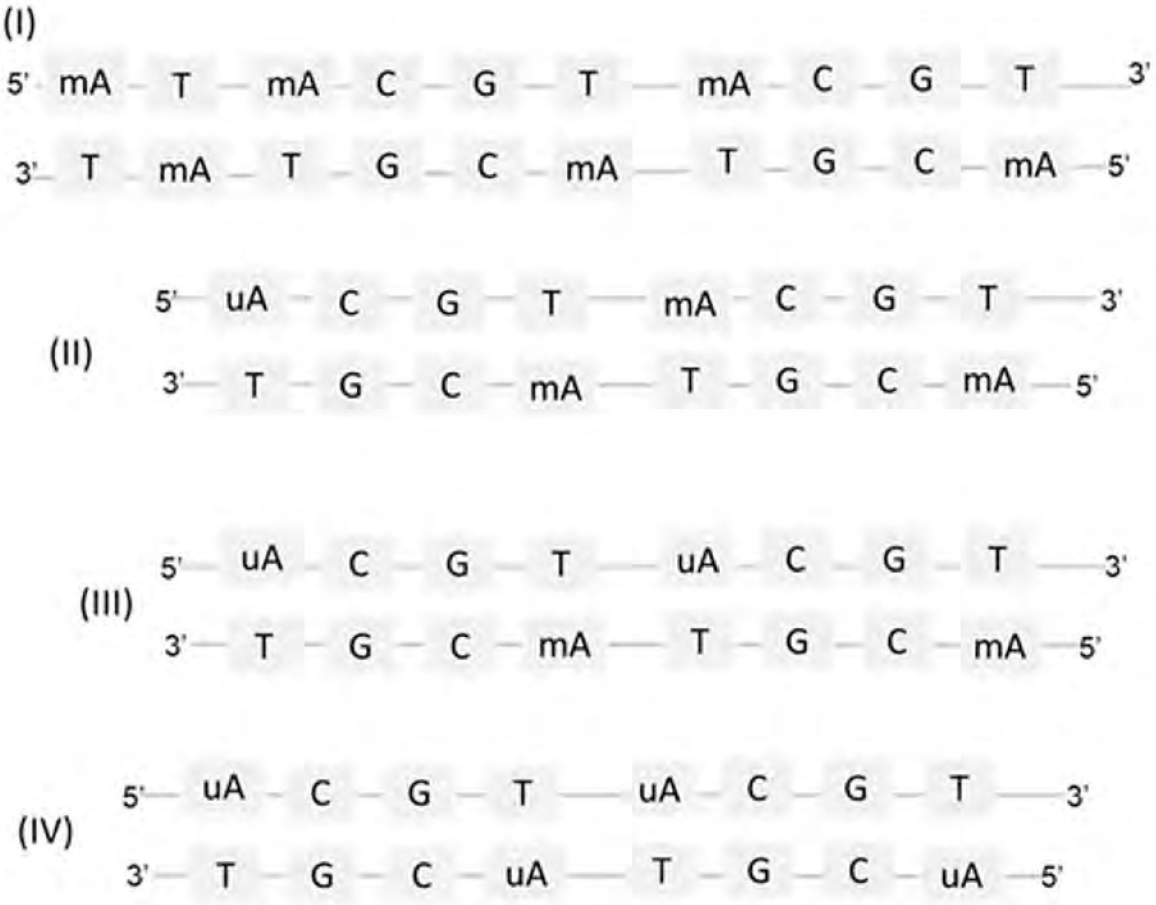




全基因體擴增之DNA產物

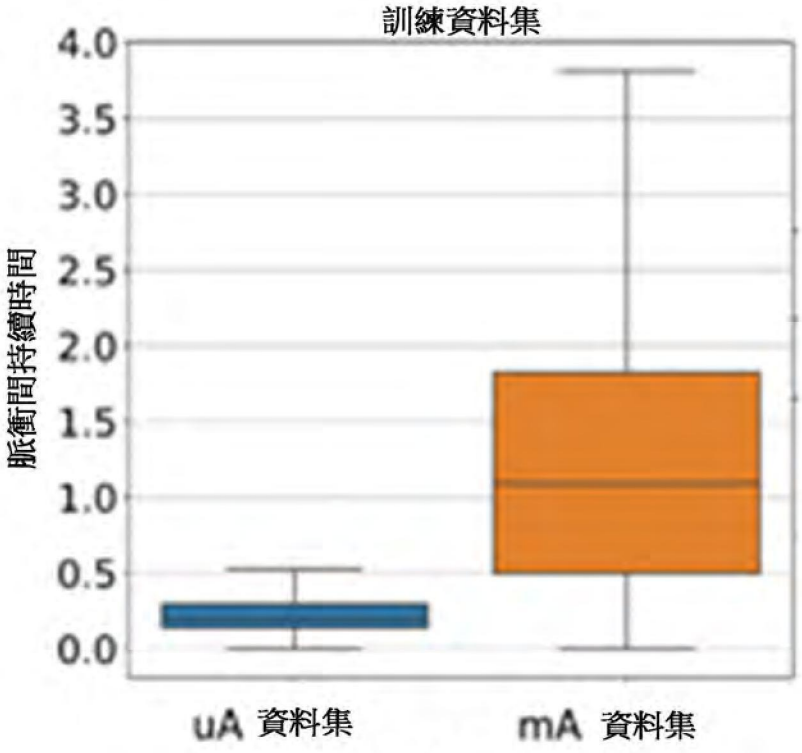
【圖31A】



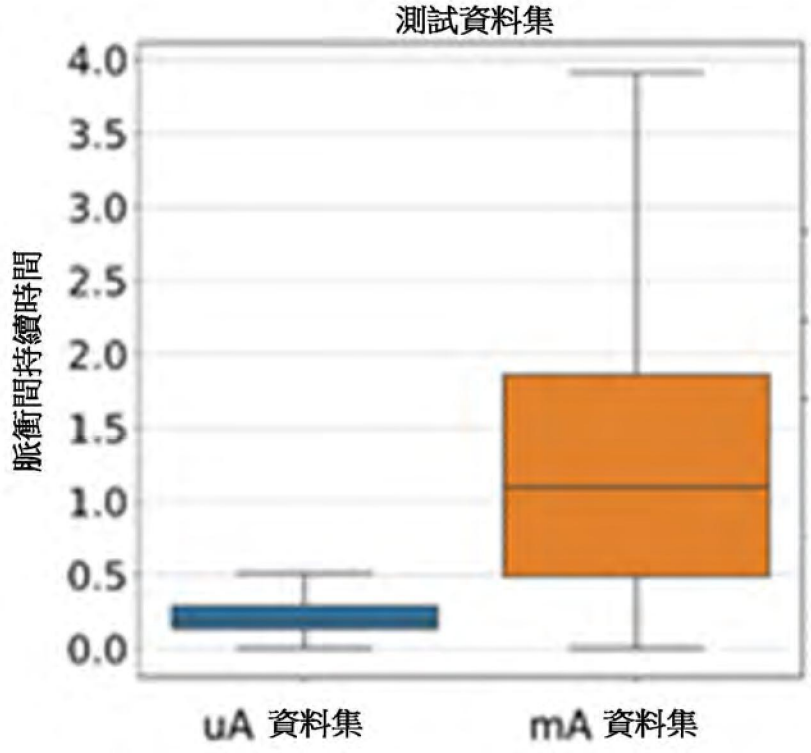


全基因體擴增之DNA產物

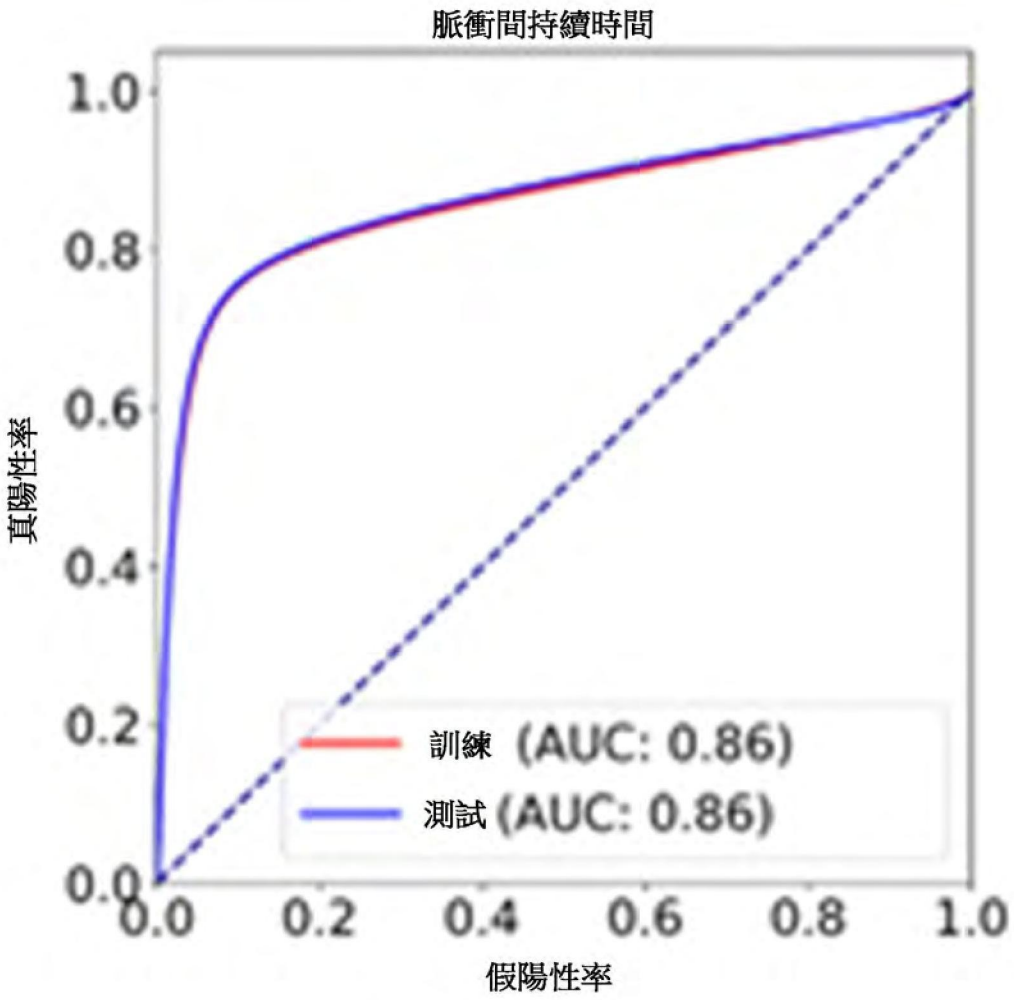
【圖31B】


訓練資料集
脈衝間持續時間
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
uA 資料集
mA 資料集

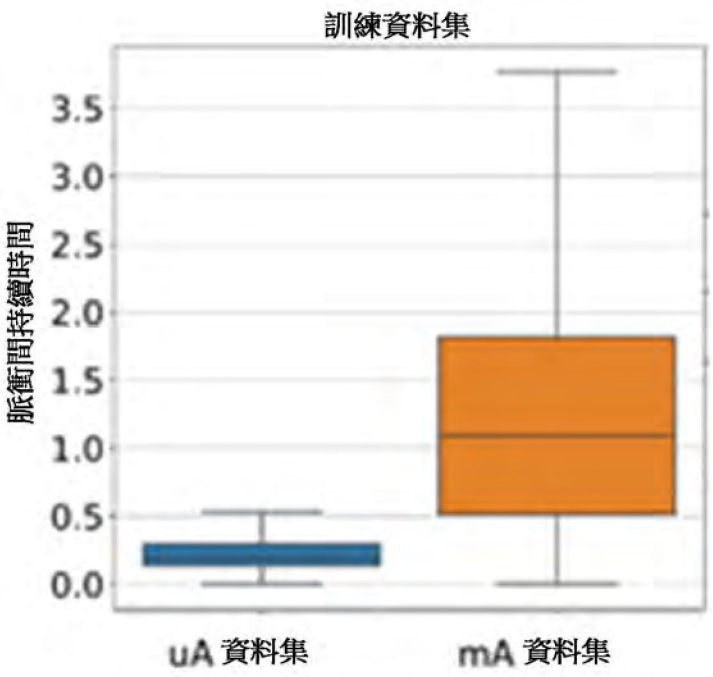
【圖32A】


測試資料集
脈衝間持續時間
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
uA 資料集
mA 資料集

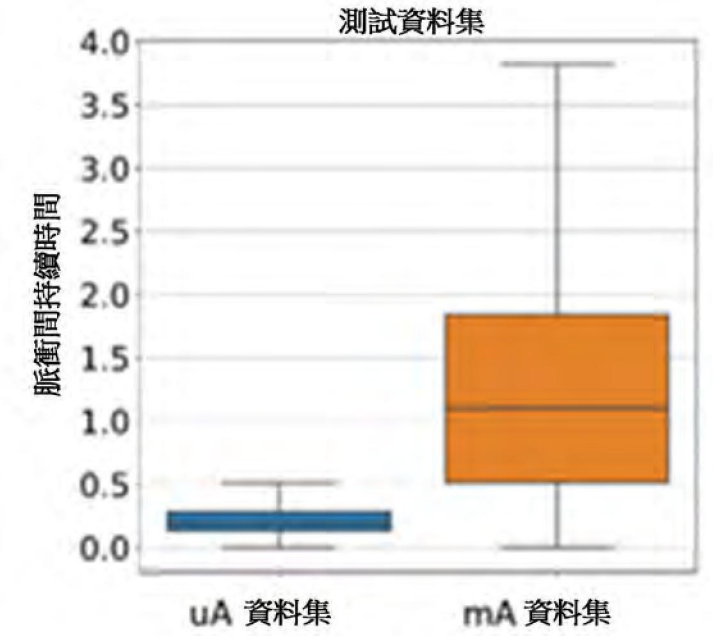
【圖32B】


脈衝間持續時間
1.0
0.8
0.6
0.4
0.2
0.0
真陽性率
0.0
0.2
0.4
0.6
0.8
1.0
假陽性率
訓練 (AUC: 0.86)
測試 (AUC: 0.86)

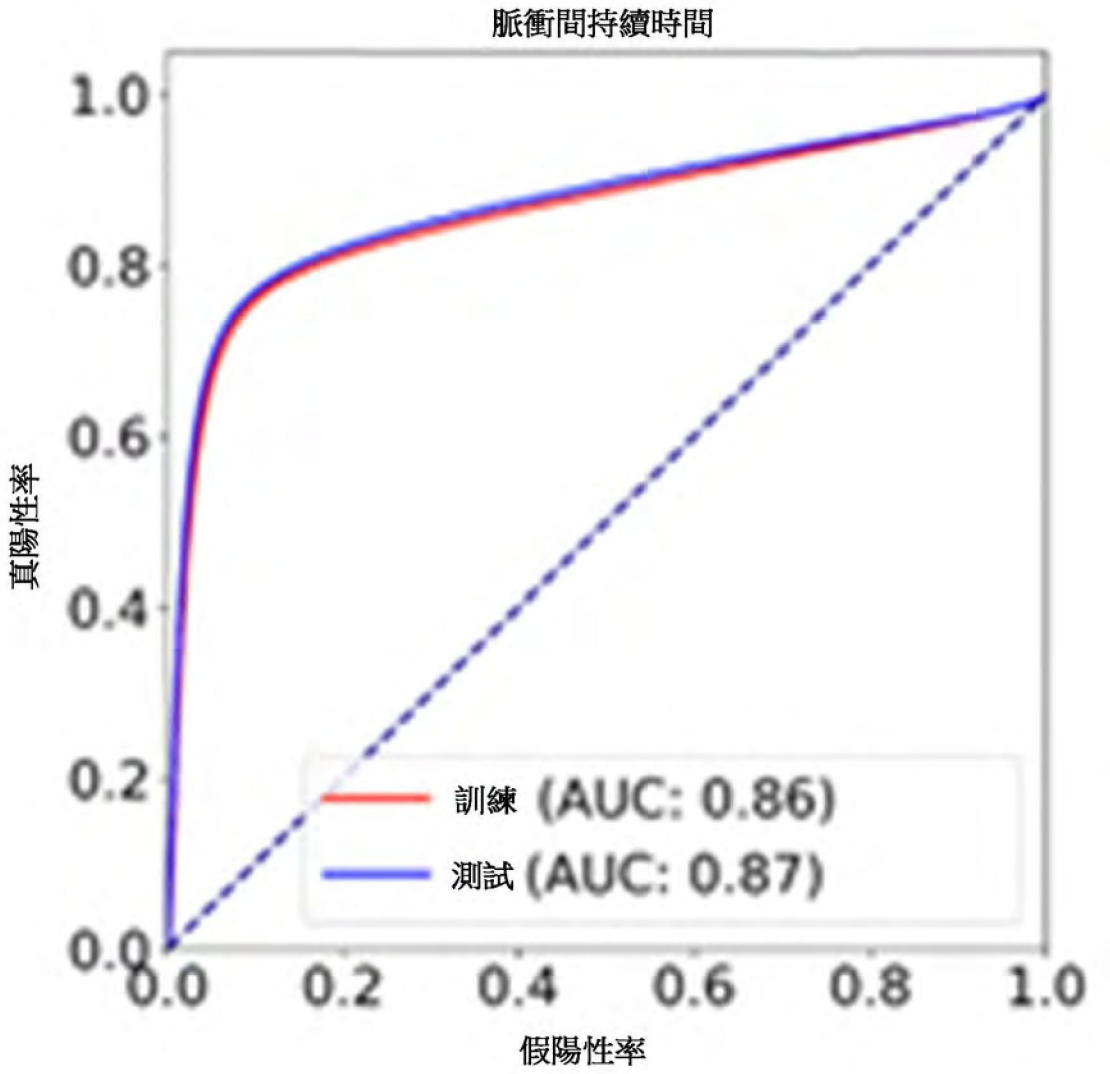
【圖32C】


訓練資料集
脈衝間持續時間
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
uA 資料集
mA 資料集

【圖33A】


測試資料集
脈衝間持續時間
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
uA 資料集
mA 資料集

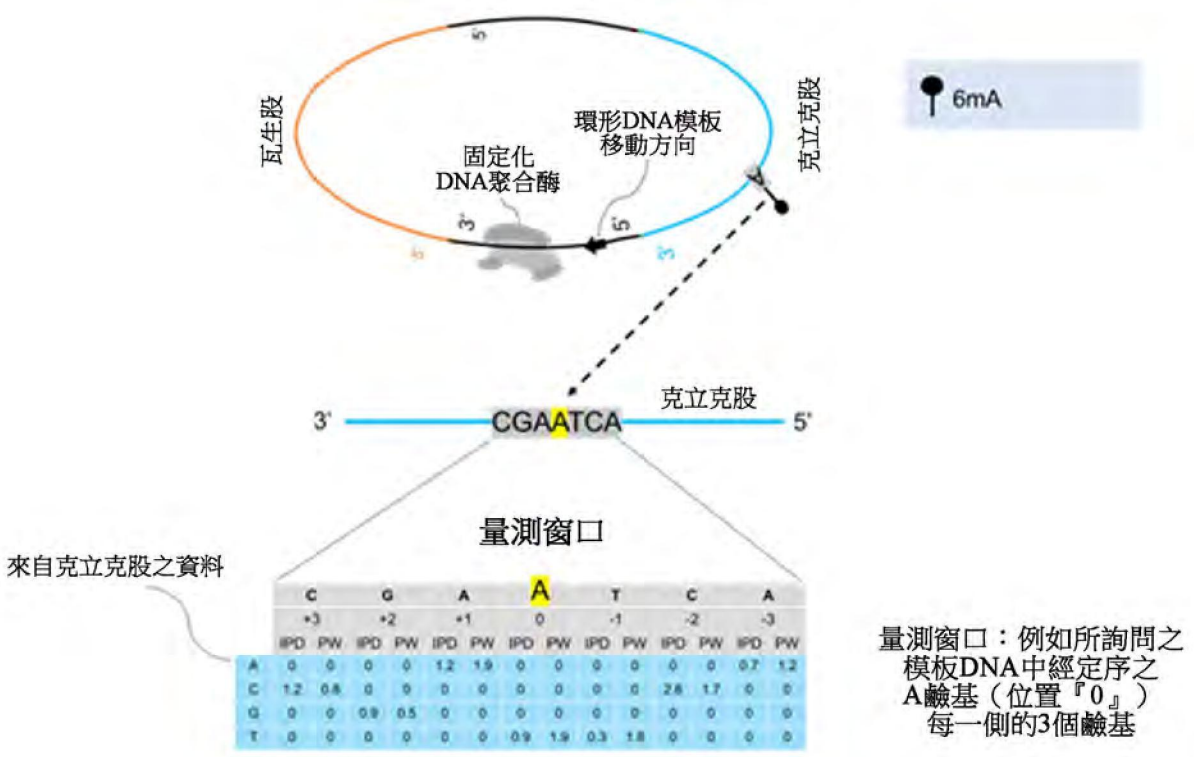
【圖33B】


脈衝間持續時間
真陽性率
假陽性率
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
訓練 (AUC: 0.86)
測試 (AUC: 0.87)

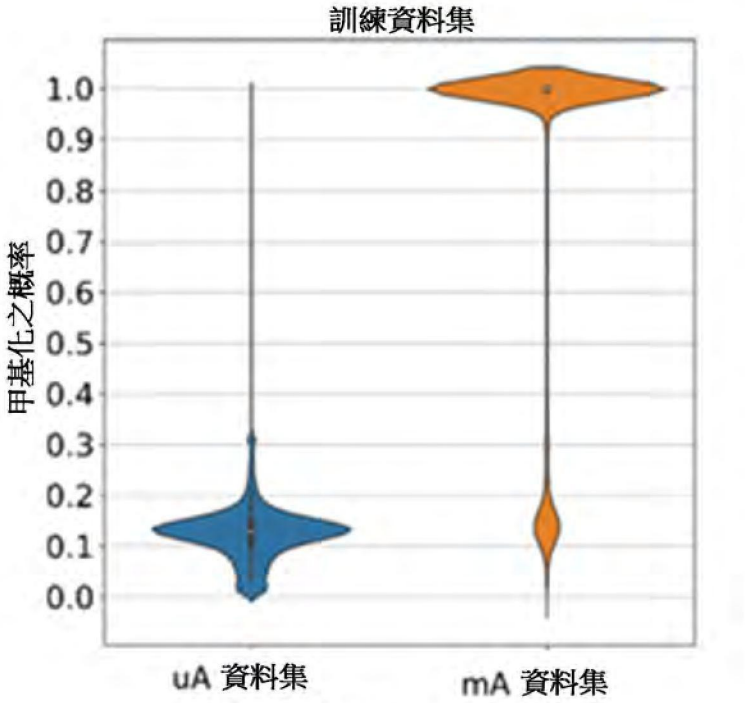
【圖33C】

6mA
瓦生股
克立克股
環形DNA模板
移動方向
固定化
DNA聚合酶
瓦生股
5'
3'
GATAACT
量測窗口
來自瓦生股之資料
量測窗口：例如所詢問之
模板DNA中經定序之
A鹼基（位置『0』）
每一側的3個鹼基

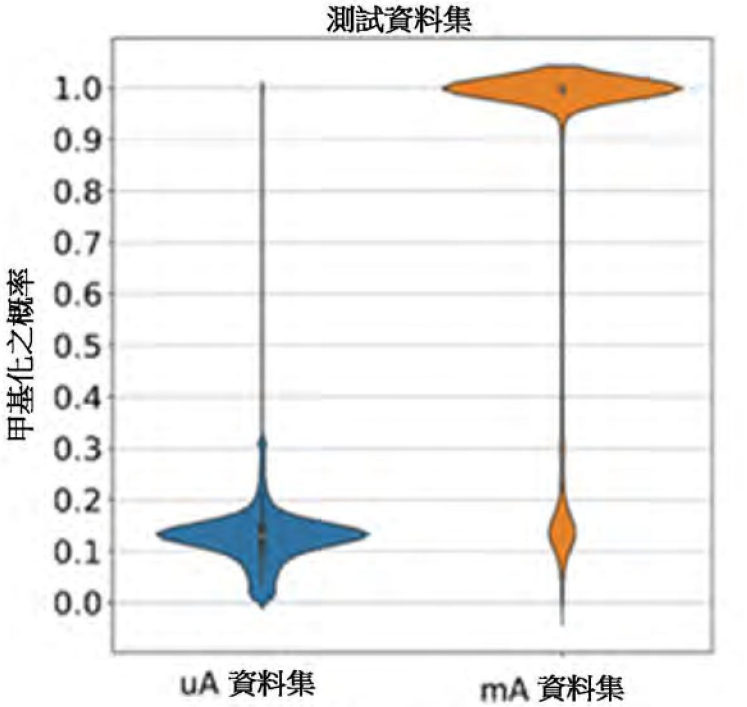
【圖34】


瓦生殷
克立克殷
環形DNA模板
移動方向
固定化
DNA聚合酶
6mA
3'
CGAATCA
5'
克立克殷
量測窗口
來自克立克殷之資料
量測窗口：例如所詢問之
模板DNA中經定序之
A鹼基（位置『0』）
每一側的3個鹼基

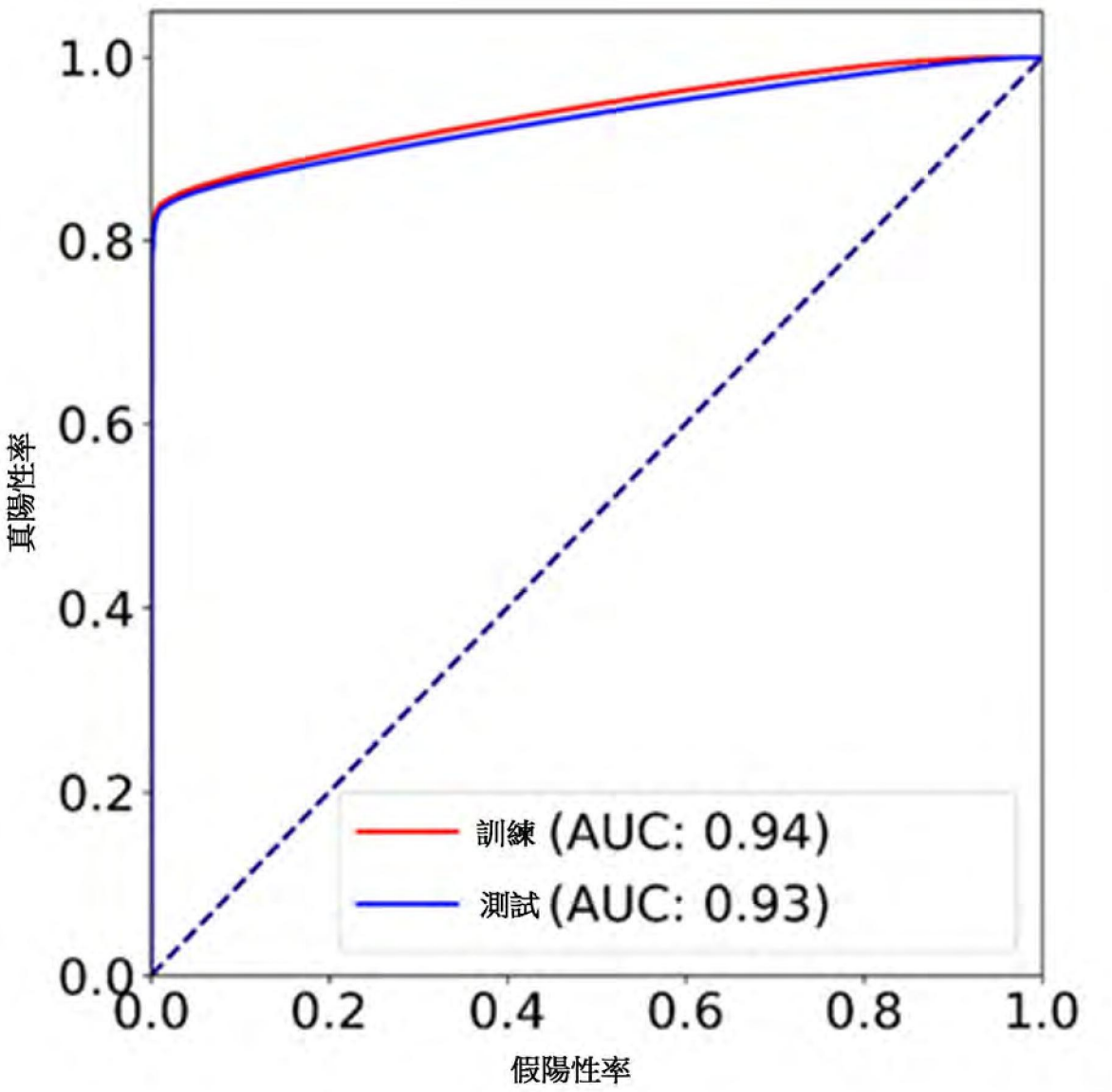
【圖35】


訓練資料集
甲基化之概率
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
uA 資料集
mA 資料集

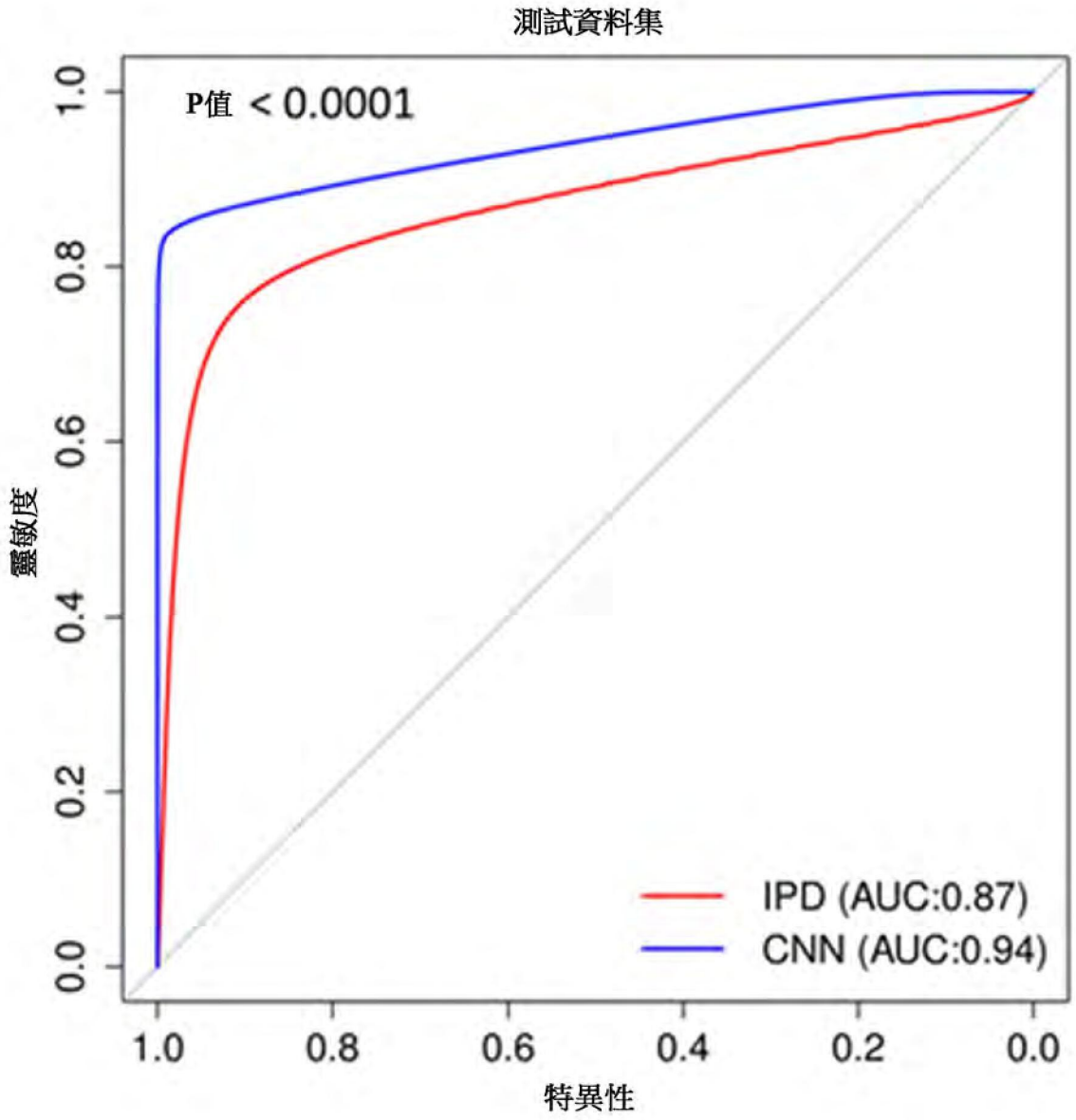
【圖36A】


測試資料集
甲基化之概率
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
uA 資料集
mA 資料集

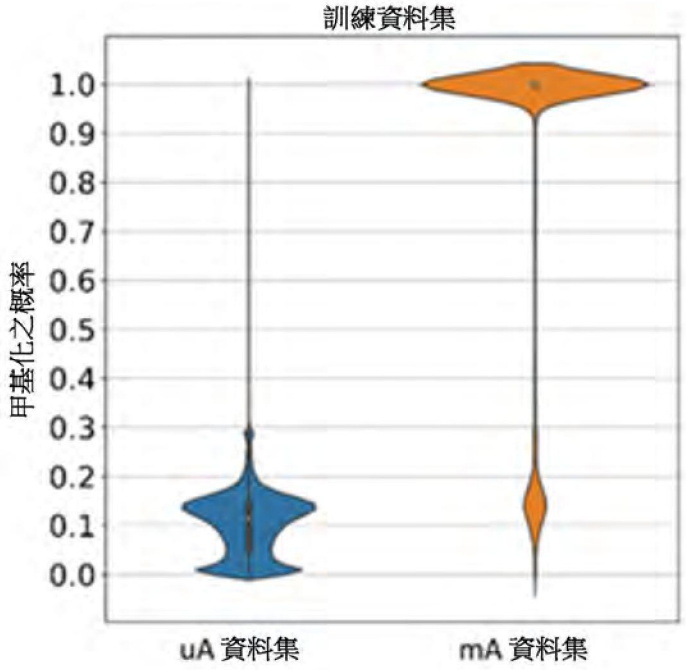
【圖36B】



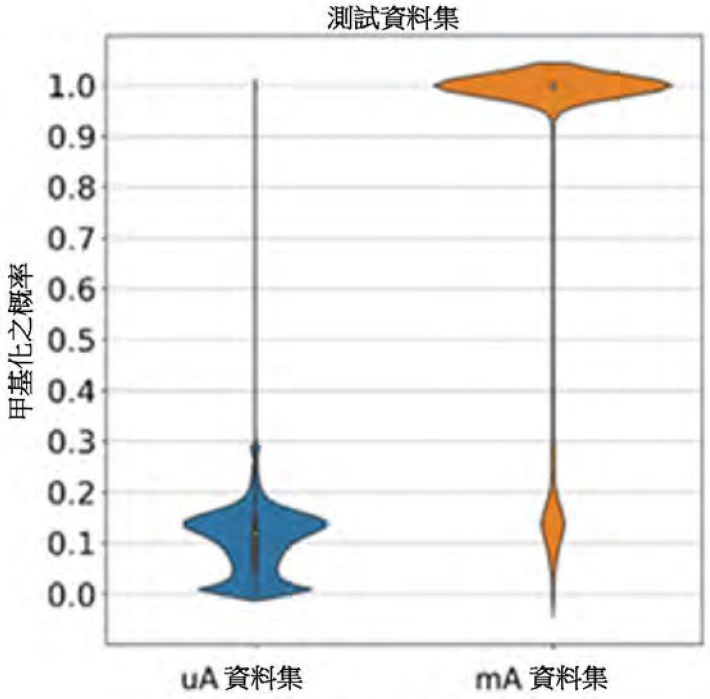
【圖37】


測試資料集
P值 < 0.0001
靈敏度
1.0
0.8
0.6
0.4
0.2
0.0
IPD (AUC:0.87)
CNN (AUC:0.94)
1.0
0.8
0.6
0.4
0.2
0.0
特異性

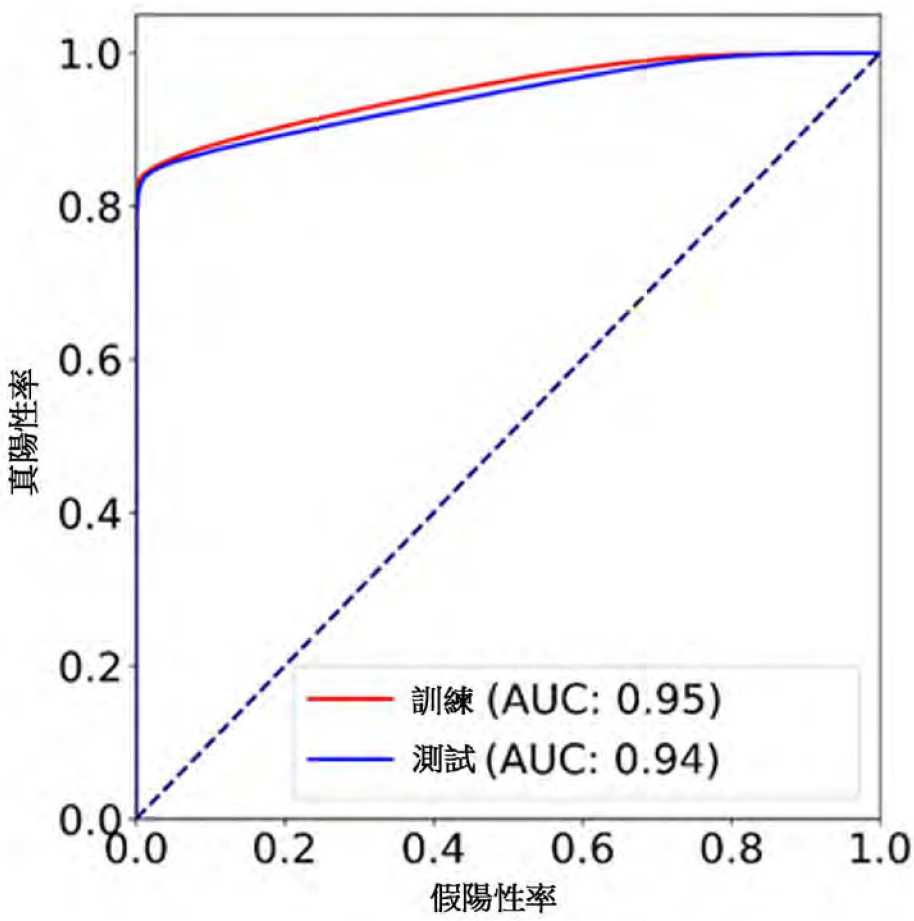
【圖38】


訓練資料集
甲基化之概率
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
uA 資料集
mA 資料集

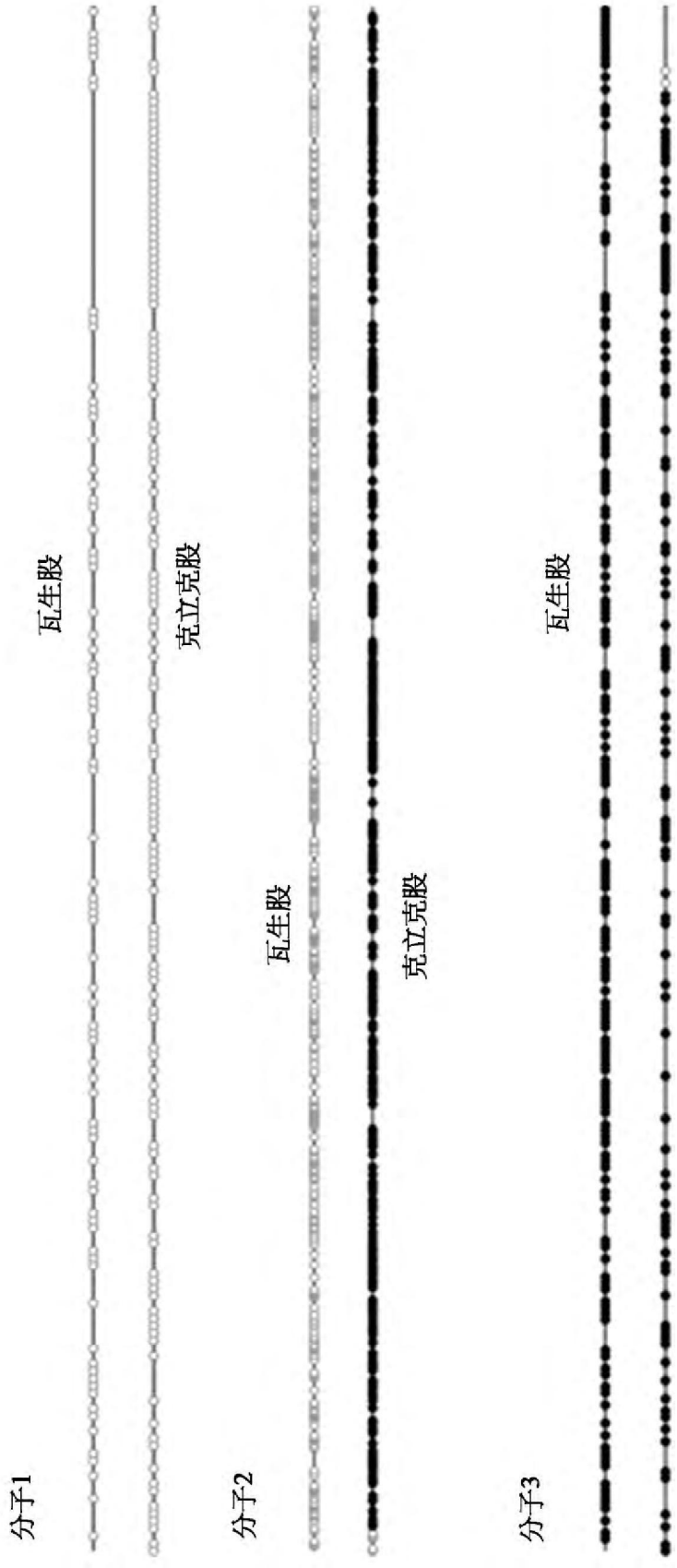
【圖39A】


測試資料集
甲基化之概率
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
uA 資料集
mA 資料集

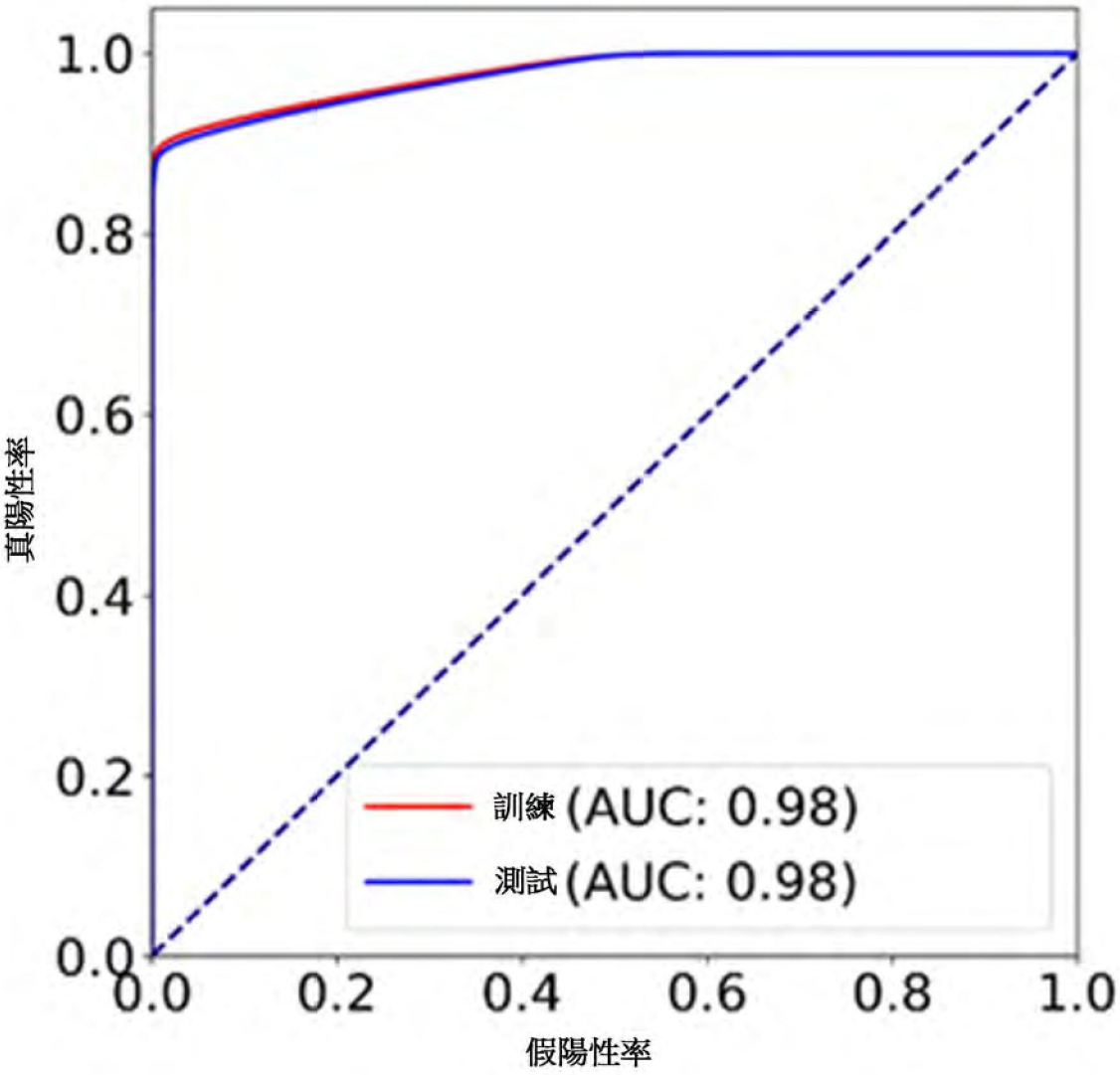
【圖39B】



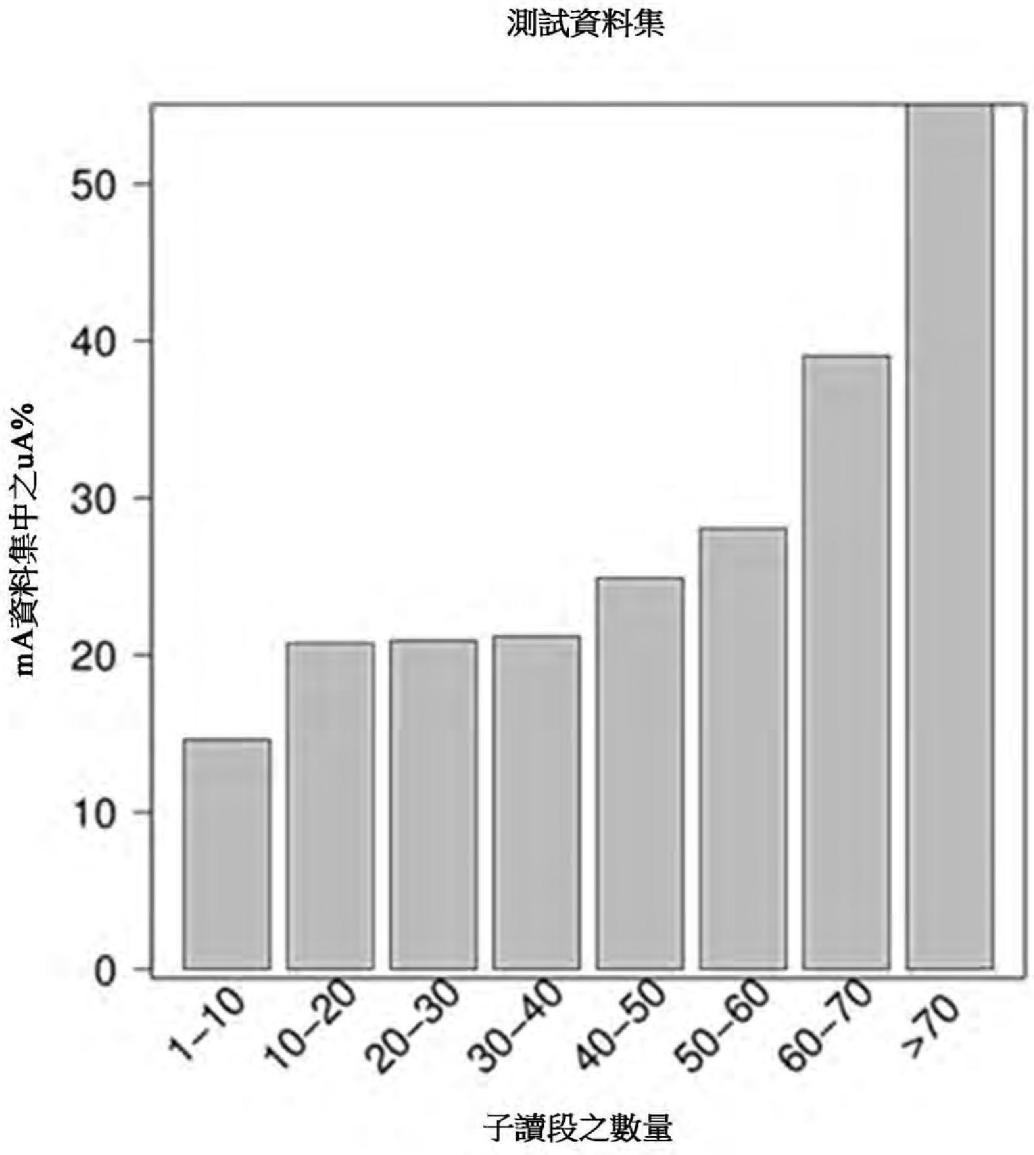
【圖40】



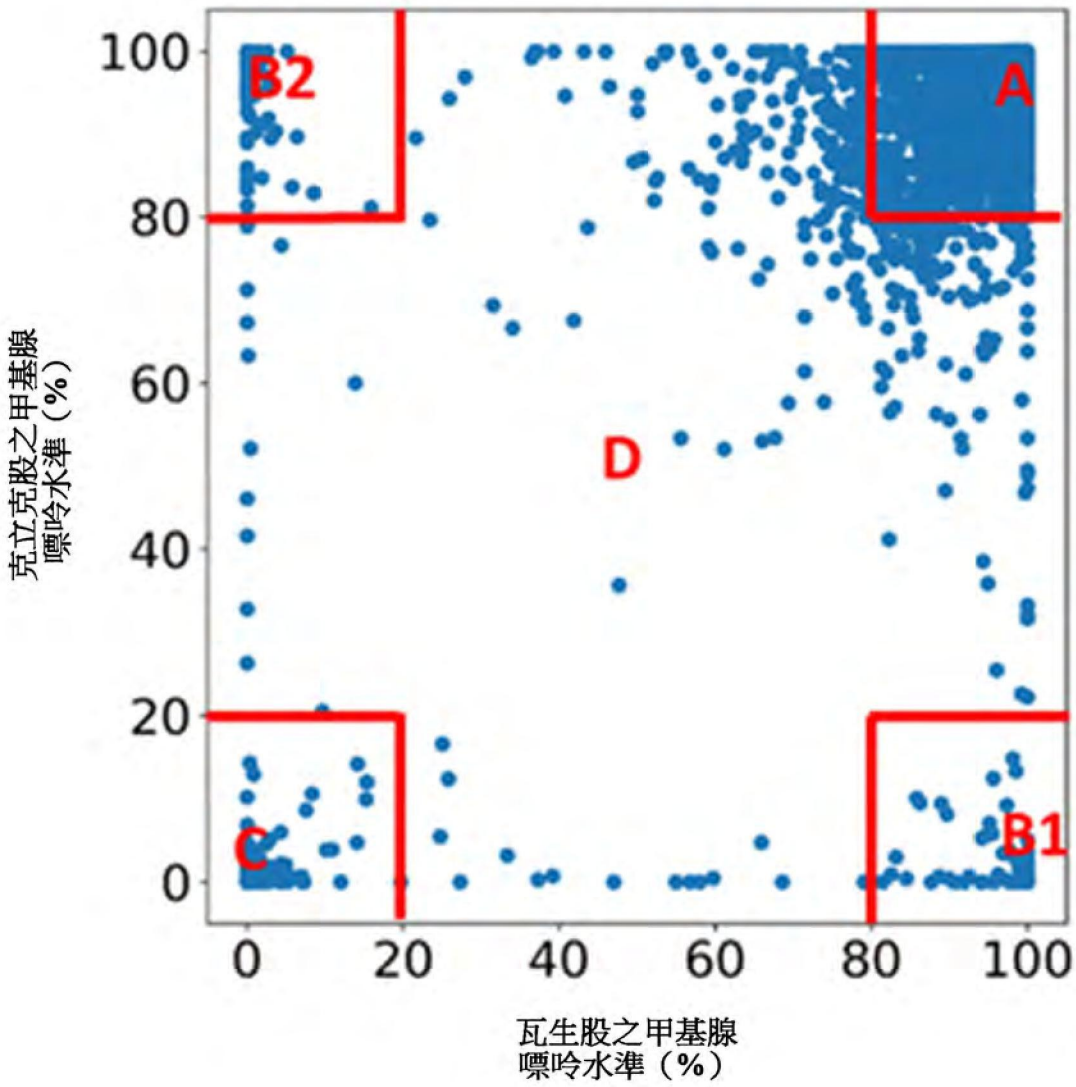
【圖41】



【圖42】


測試資料集
mA資料集中之uA%
0
10
20
30
40
50
1–10
10–20
20–30
30–40
40–50
50–60
60–70
>70
子讀段之數量

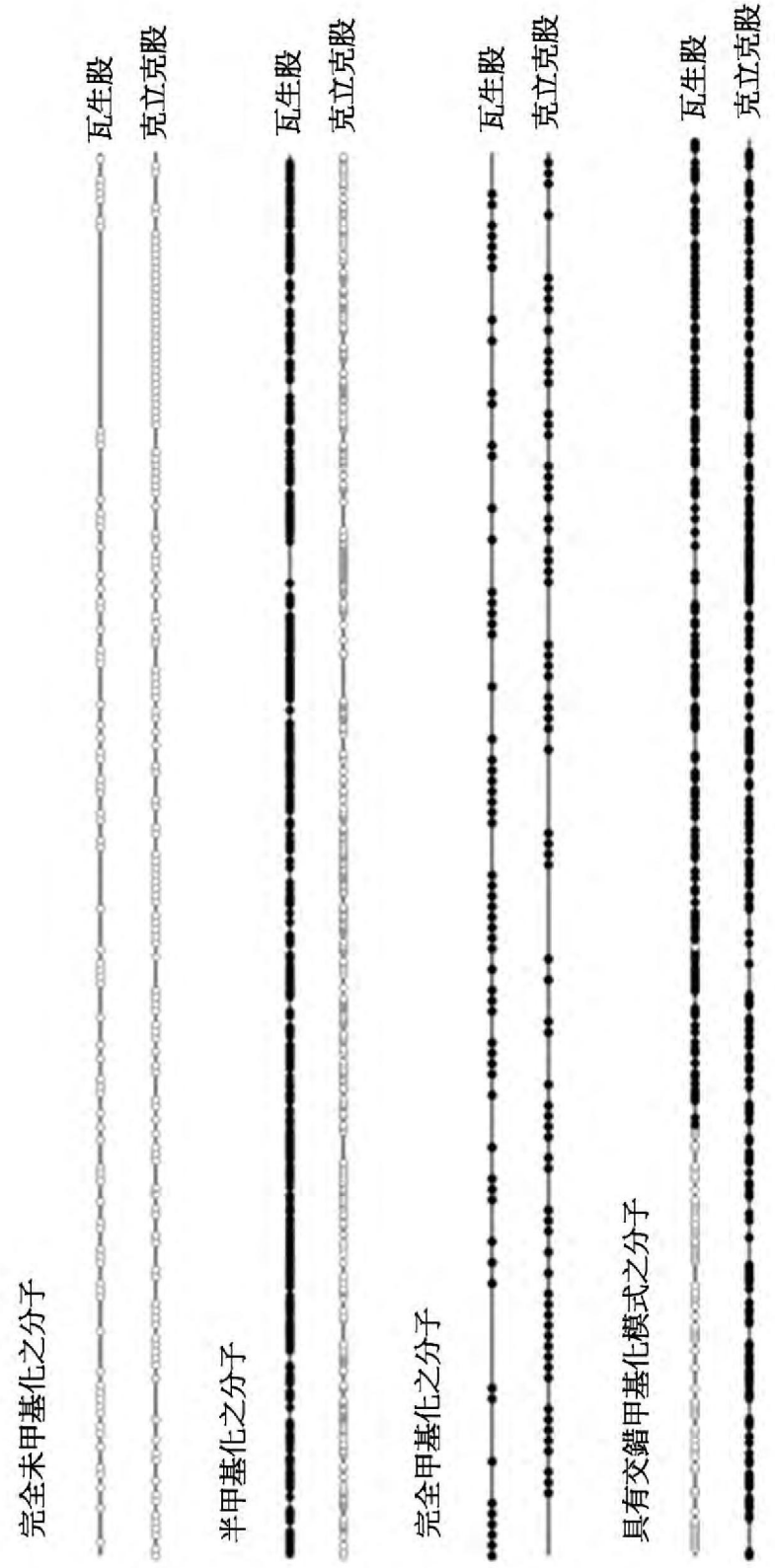
【圖43】


100
80
60
40
20
0
0
20
40
60
80
100
B2
A
D
C
B1
克立克股之甲基腺
嘌呤水準（%）
瓦生股之甲基腺
嘌呤水準（%）

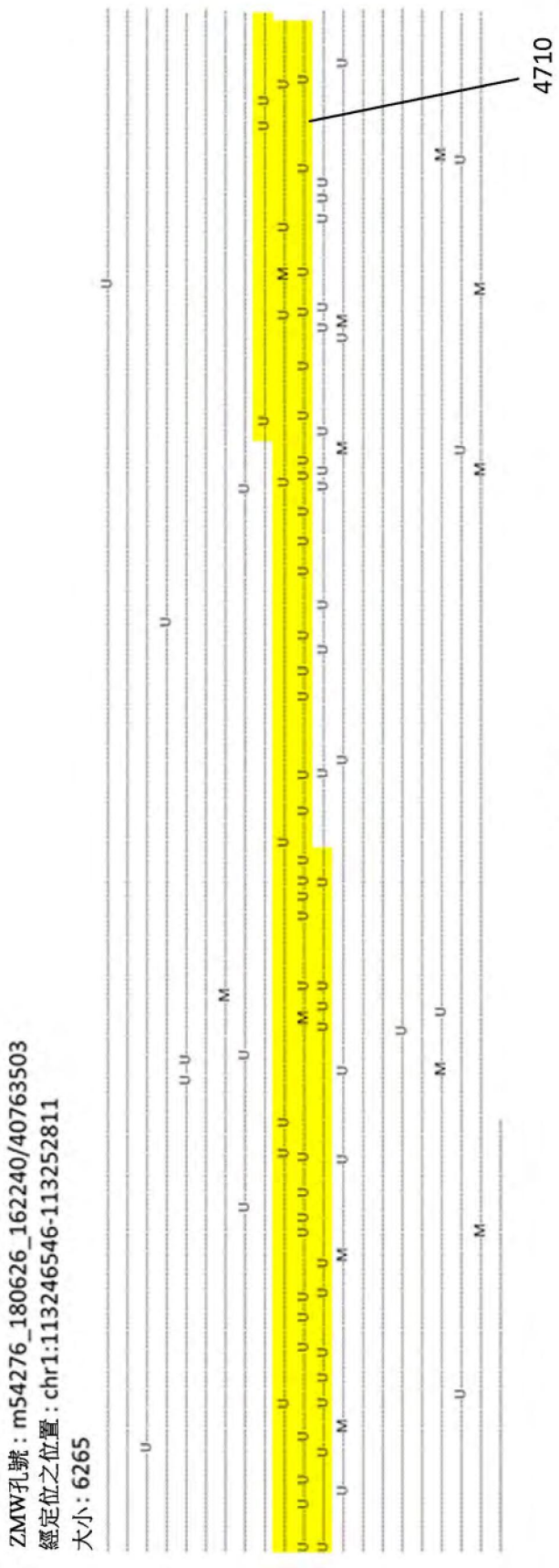
【圖44】

類別	訓練資料集	測試資料集
完全未甲基化	283 (7.0%)	276 (7.0%)
半甲基化	401 (10.0%)	389 (9.8%)
完全甲基化	3194 (79.4%)	3142 (79.4%)
交錯甲基化模式	145 (3.6%)	148 (3.7%)

【圖45】


完全未甲基化之分子
瓦生股
克立克股
半甲基化之分子
瓦生股
克立克股
完全甲基化之分子
瓦生股
克立克股
具有交錯甲基化模式之分子
瓦生股
克立克股

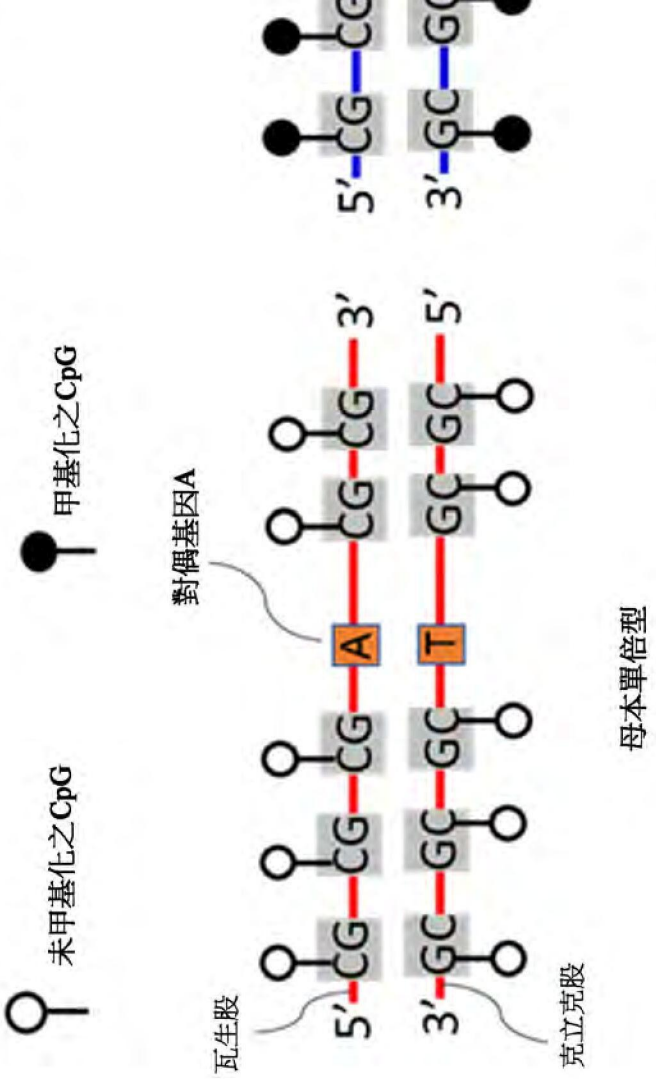
【圖46】


ZMW孔號：m54276_180626_162240/40763503
經定位之位置：chr1:113246546-113252811
大小：6265
4710

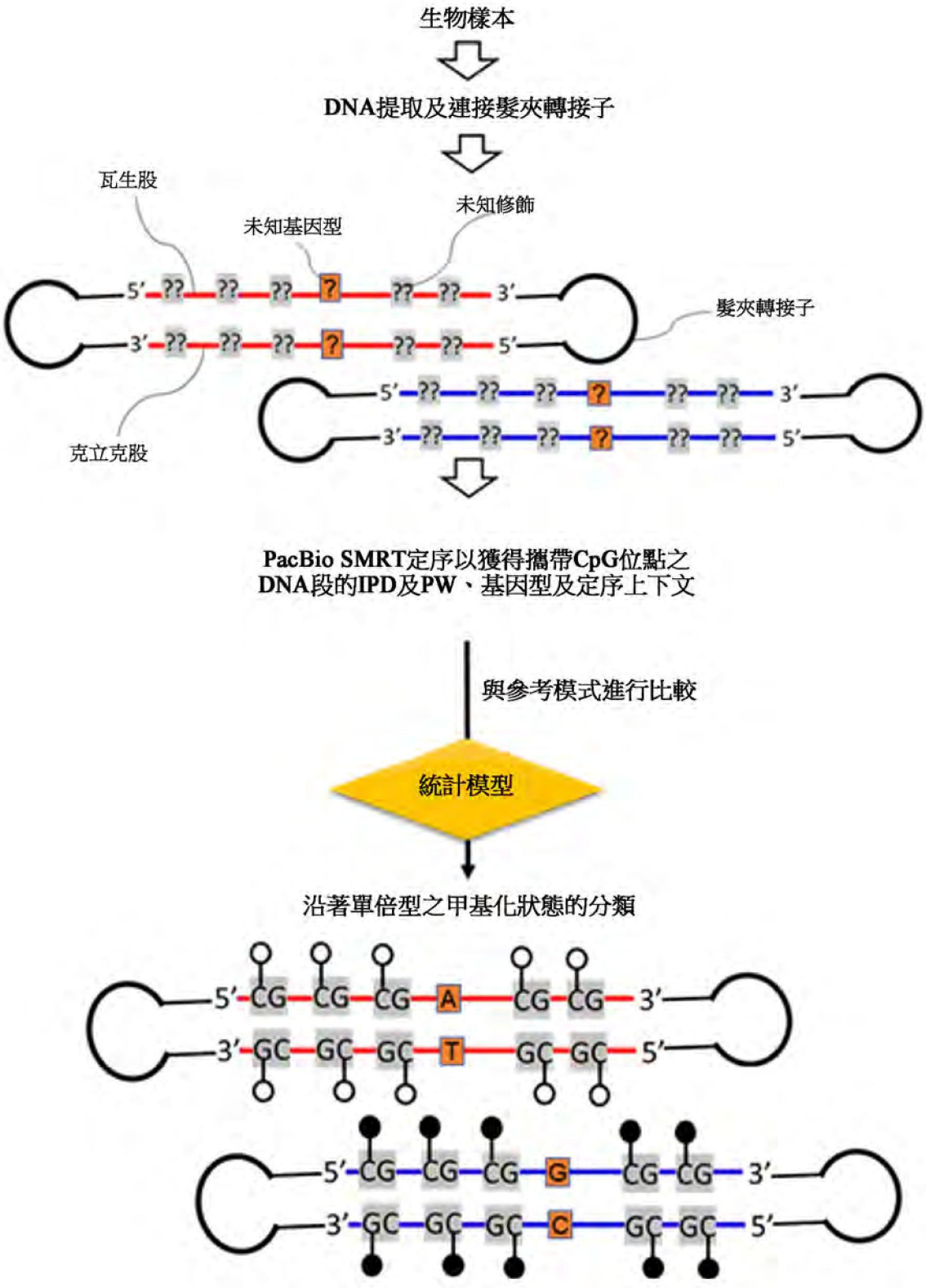
【圖47】

染色體	起始	結束	印記基因名稱	CpG島之長度	藉由PacBio SMRT定序所定序之分子及根據本揭示案中存在之實施例確定的甲基化狀態	分子之甲基化判讀
chr11	2013333	2013617	H19	284	-U----------M----------M-M----------U----U----------M----------M---- --U----------M----------M----------M----------M---- M----------M-M------U----------[T]----------M----------M---- --M--	甲基化
chr11	2019565	2019863	H19	298	-M-M----------M----------M----------[C]----M-M-M----M---------- ----------M----------M----------M-M----------M----------M---- M----M-M--M----------M----M----------M-M----------M---- ----------	甲基化
chr11	32460586	32461004	WT1-AS/WT1	418	-U--U----------U-M----------[C]----------U----------U--U---------- U----------U--U-U----------U----M--M----------U---------- ----U-U-U------U------U----------U----U----------U-U----------U- ---M	未甲基化
chr14	101192851	101193499	DLK1	648	-U----U----------U----------U----------M--M----------U----------U---------- -U----------U----------U-U--U--U-U----------U----------U- --U-U----------U----------U----U-U--U----------U--U--U----U---- M---	未甲基化
chr14	101201559	101201763	DLK1	204	-M--M----------U----------M----M----M----------[T]---------- --------M----------M--M----------M----M----------M---------- ----------M----------M----------M-	甲基化
chr14	101292863	101293101	MEG3	238	----------M----------U----------U----------M- ---M----------M-M----------M-	甲基化
chr15	25981176	25981392	ATP10A	216	----*----------M----------M----------M----------M---------- M--M----------M----------[T]----M-M----------U-	甲基化
chr2	80531367	80531719	LRRTM1	352	---*----[G]----------U--U------M----------M----------U----U--U-- ---U----U----------U----------U----------M----------U---- -U-U----------M----------U----------U--U----------U-U-U-- --	未甲基化
chr7	79082174	79082427	MAGI2	253	--------*-U----------[A]-M-U----------U---U---U-------U-- U----------U----------*--------	未甲基化

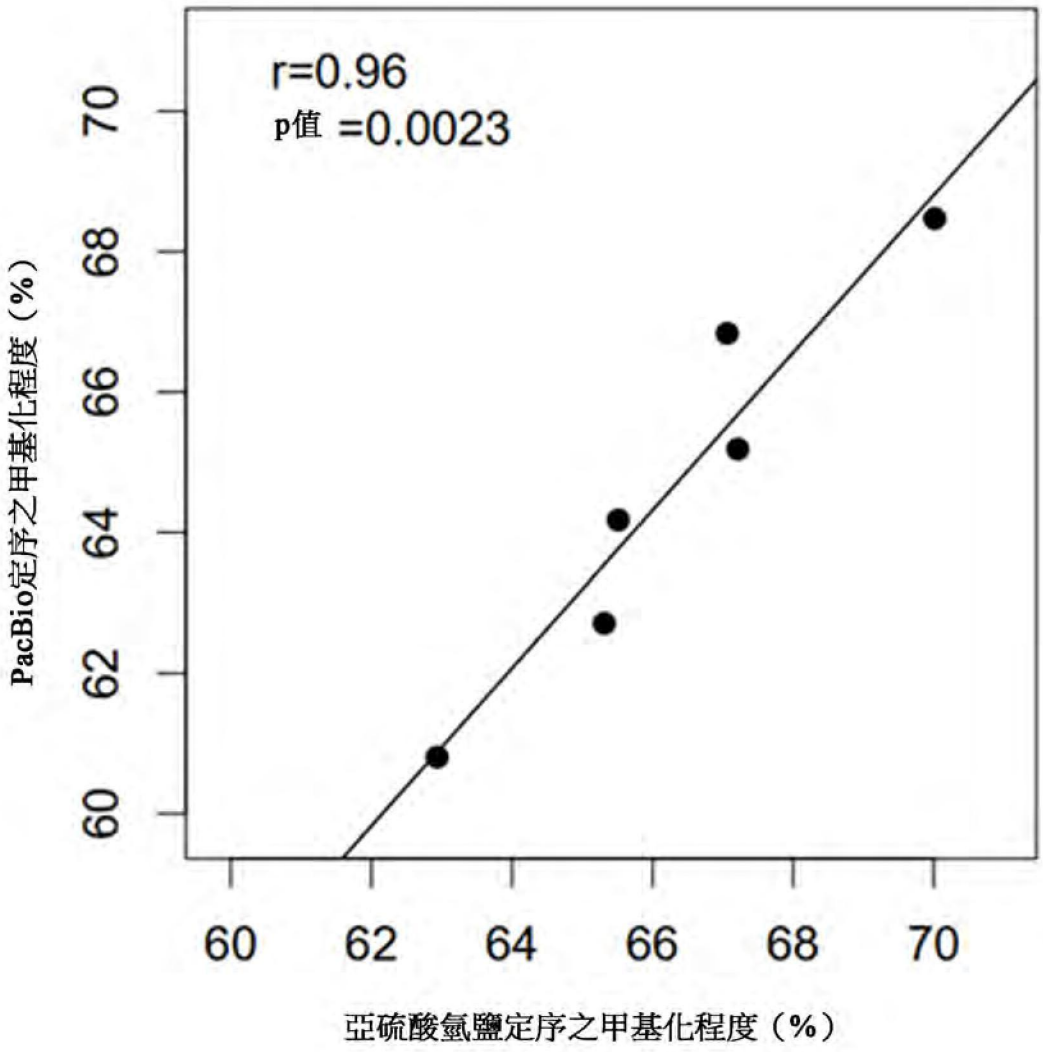
【圖48】


父本印記區域中存在之甲基化模式
甲基化之CpG
未甲基化之CpG
對偶基因A
對偶基因B
瓦生股
克立克股
母本單倍型
父本單倍型
5′
3′
A
T
G
C

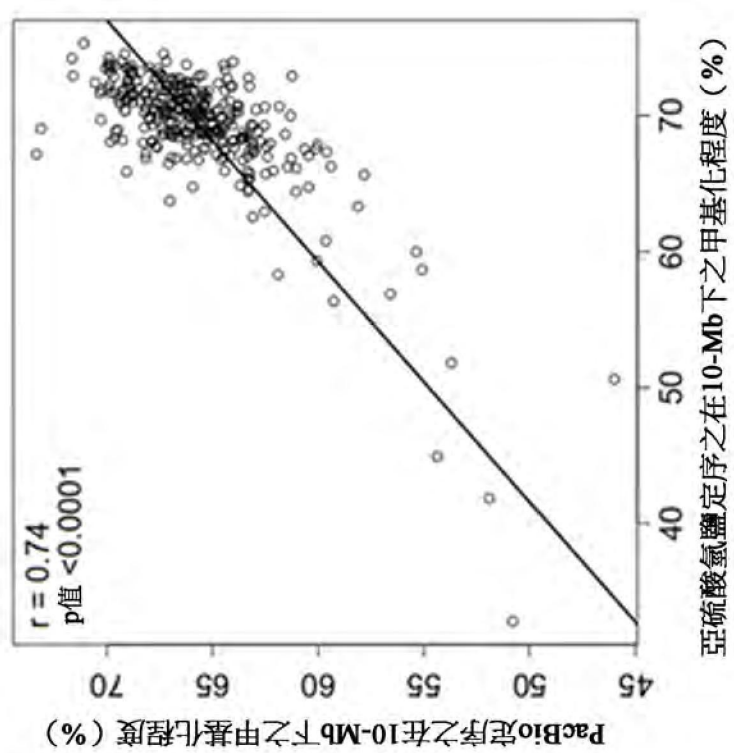
【圖49】


生物樣本
DNA提取及連接髮夾轉接子
瓦生股
未知基因型
未知修飾
髮夾轉接子
克立克股
PacBio SMRT定序以獲得攜帶CpG位點之
DNA段的IPD及PW、基因型及定序上下文
與參考模式進行比較
統計模型
沿著單倍型之甲基化狀態的分類

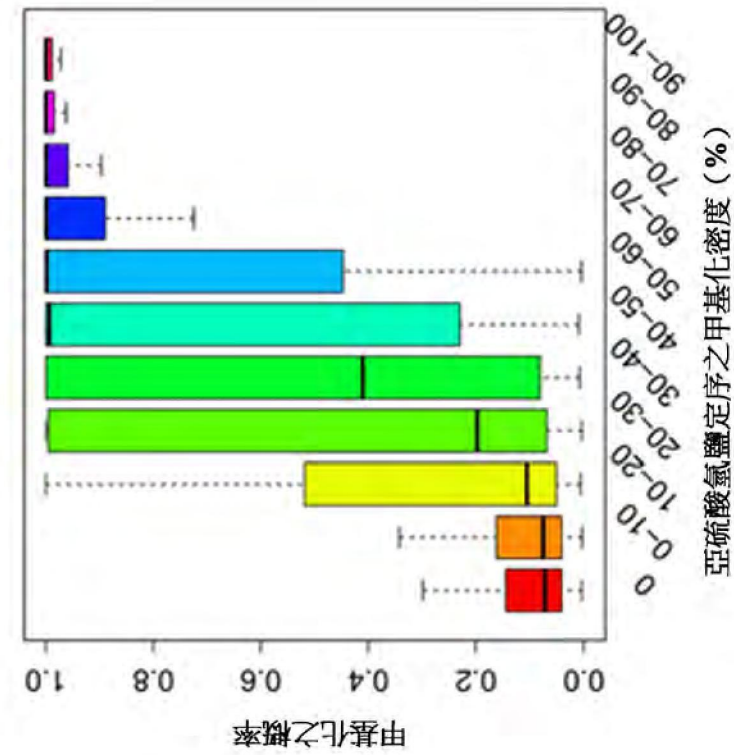
【圖50】


r=0.96
p值 =0.0023
PacBio定序之甲基化程度（%）
60
62
64
66
68
70
亞硫酸氫鹽定序之甲基化程度（%）

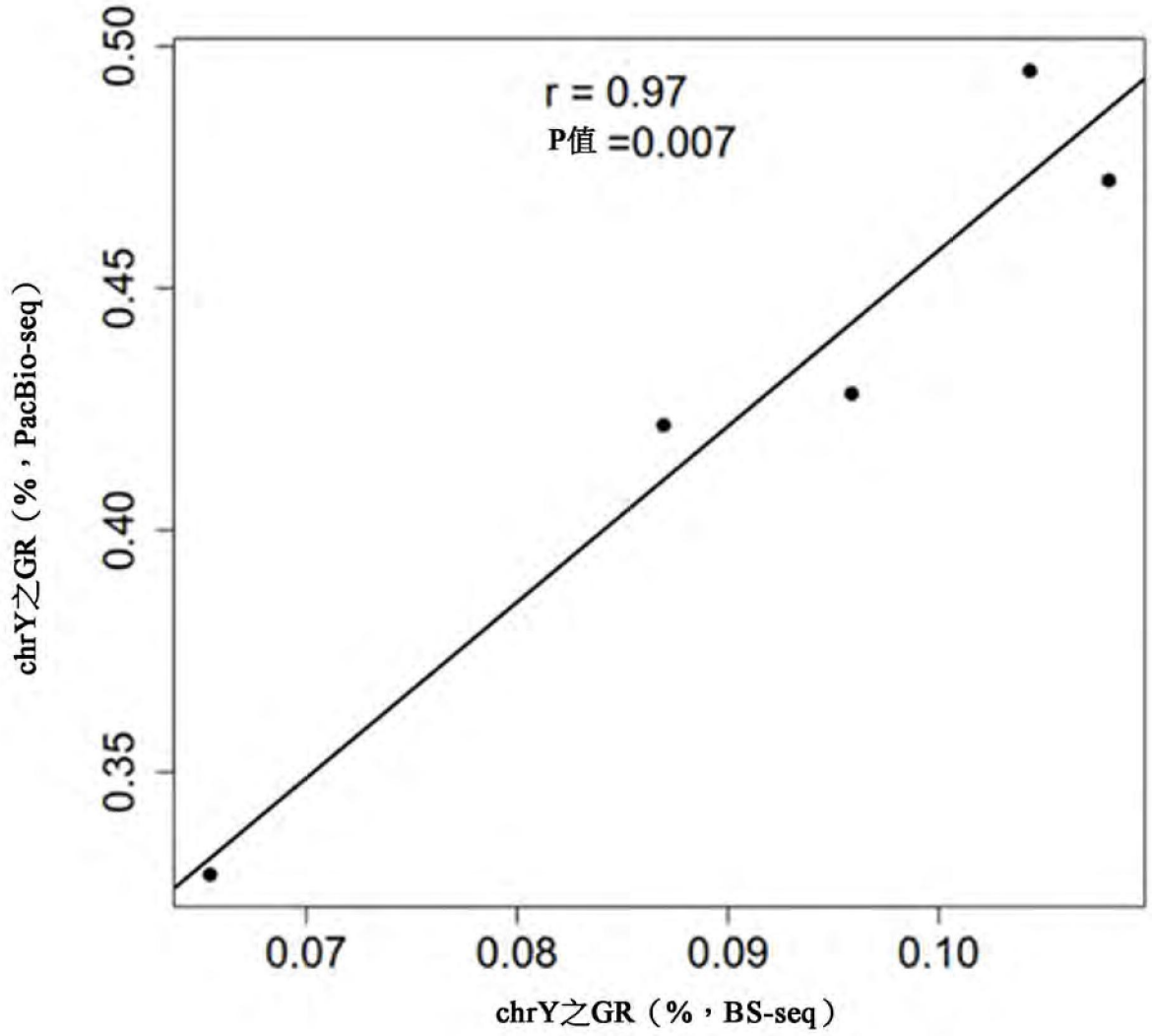
【圖51】


r = 0.74
p值 <0.0001
PacBio定序之在10-Mb下之甲基化程度（%）
亞硫酸氫鹽定序之在10-Mb下之甲基化程度（%）
45
50
55
60
65
70
40
50
60
70

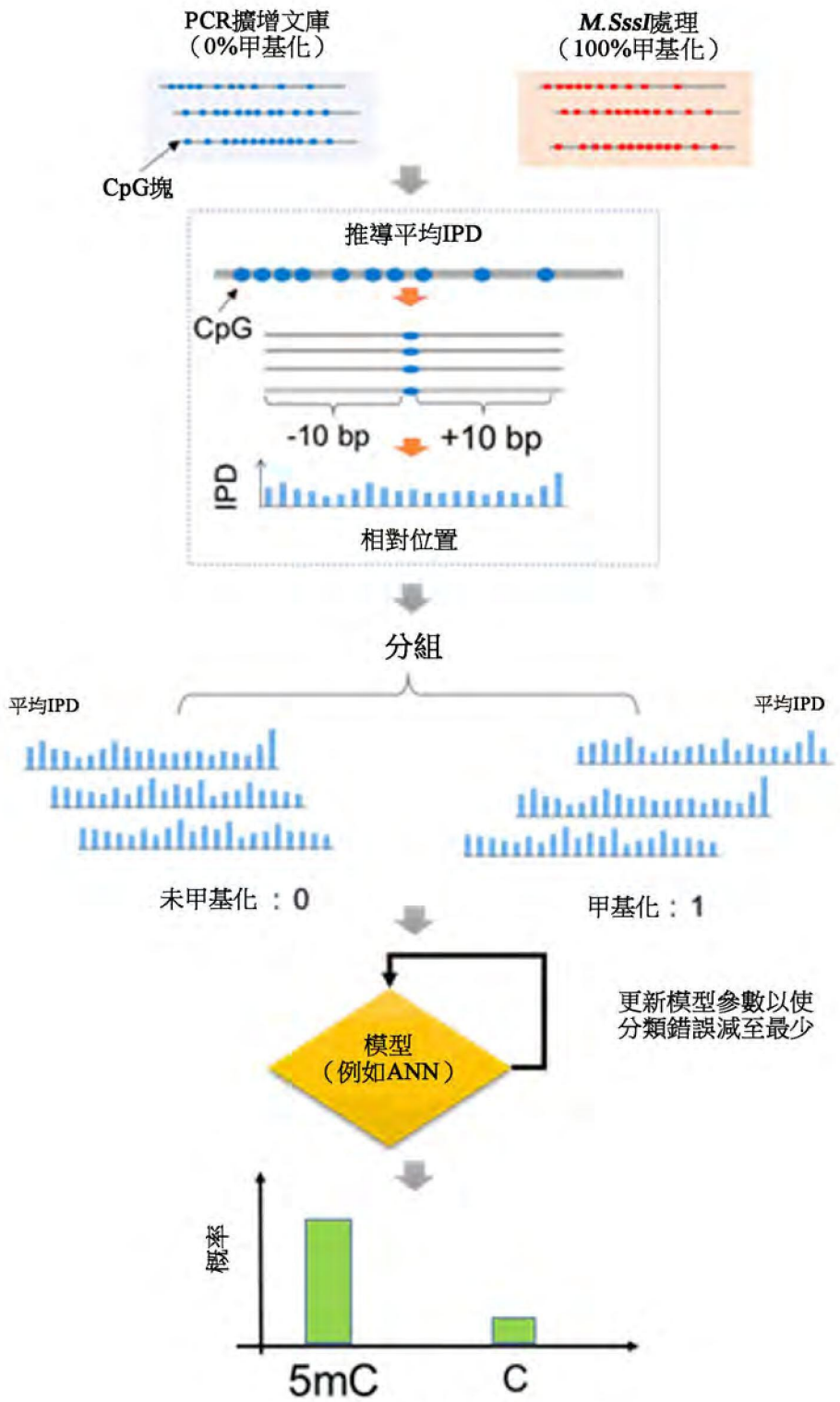
【圖52B】


甲基化之概率
亞硫酸氫鹽定序之甲基化密度（%）
0.0
0.2
0.4
0.6
0.8
1.0
0
0-10
10-20
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100

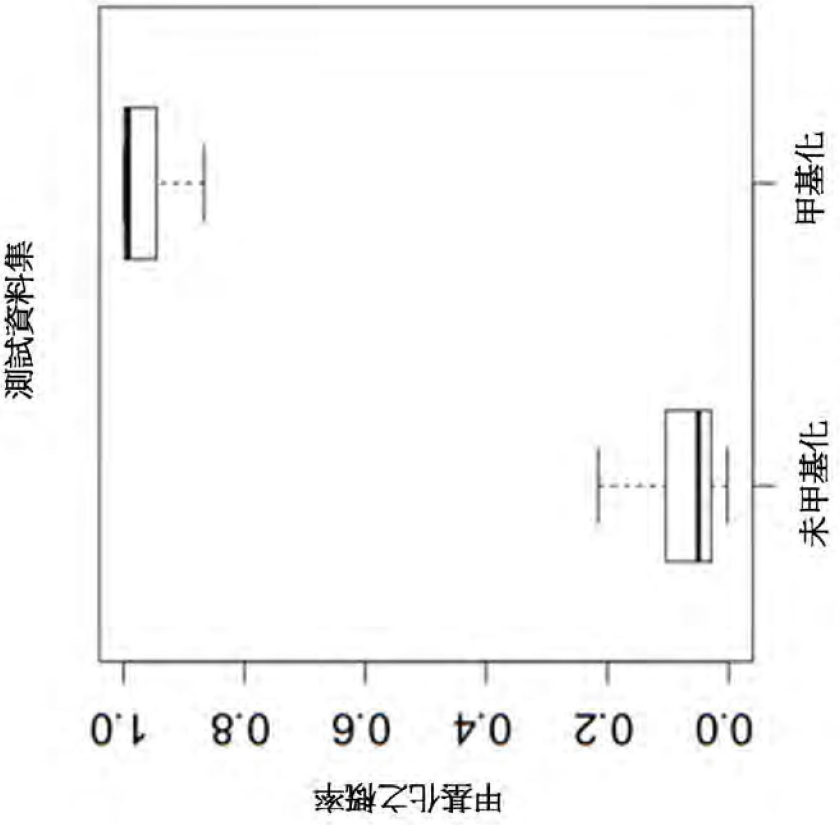
【圖52A】


r = 0.97
P值 =0.007
0.50
0.45
0.40
0.35
0.07
0.08
0.09
0.10
chrY之GR（%，PacBio-seq）
chrY之GR（%，BS-seq）

【圖53】


PCR擴增文庫
（0%甲基化）
M.SssI處理
（100%甲基化）
CpG塊
推導平均IPD
CpG
-10 bp
+10 bp
IPD
相對位置
分組
平均IPD
平均IPD
未甲基化：0
甲基化：1
模型
（例如ANN）
更新模型參數以使
分類錯誤減至最少
概率
5mC
C

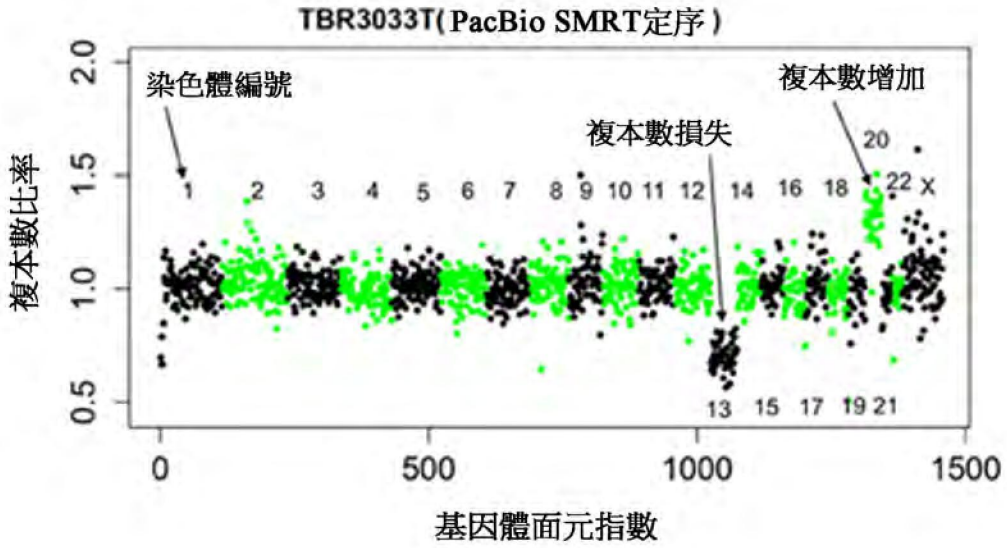
【圖54】


訓練資料集
甲基化之概率
0.0
0.2
0.4
0.6
0.8
1.0
未甲基化
甲基化

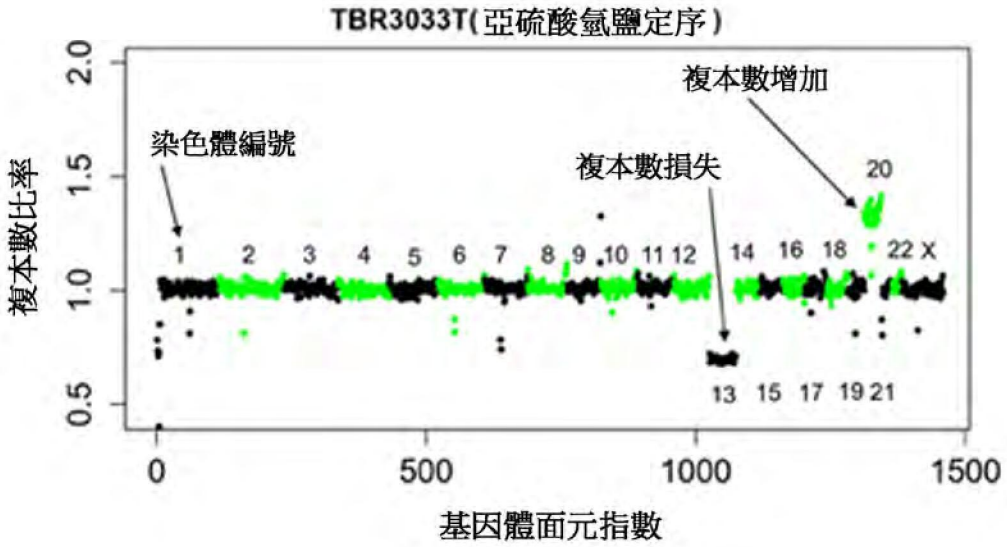
【圖55A】

測試資料集
甲基化之概率
0.0
0.2
0.4
0.6
0.8
1.0
未甲基化
甲基化

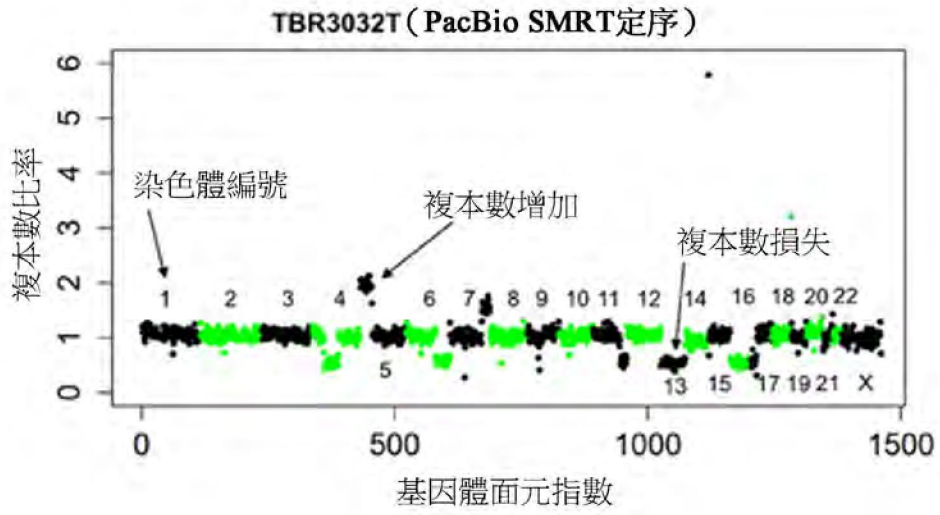
【圖55B】


TBR3033T(PacBio SMRT定序)
染色體編號
複本數損失
複本數增加
複本數比率
基因體面元指數
0.5
1.0
1.5
2.0
0
500
1000
1500
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X

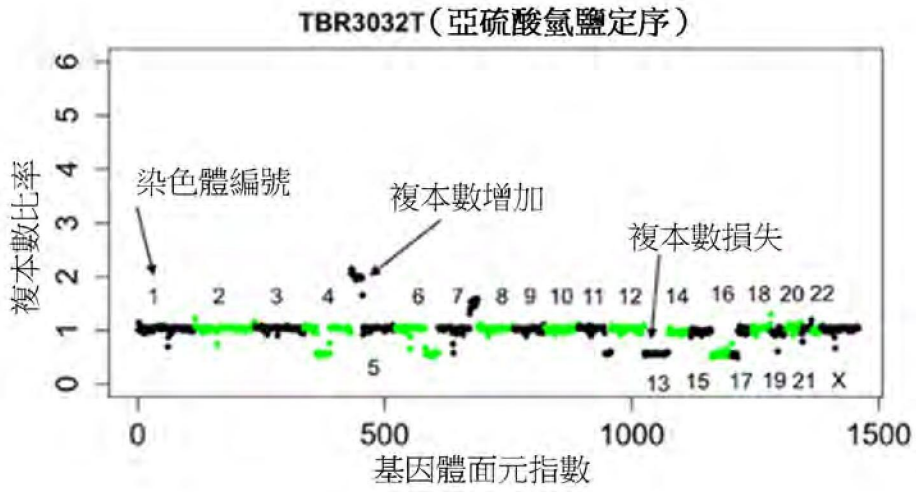
【圖56A】


TBR3033T(亞硫酸氫鹽定序)
染色體編號
複本數損失
複本數增加
複本數比率
基因體面元指數
0.5
1.0
1.5
2.0
0
500
1000
1500
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X

【圖56B】


TBR3032T（PacBio SMRT定序）
複本數比率
染色體編號
複本數增加
複本數損失
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
0 1 2 3 4 5 6
0 500 1000 1500
基因體面元指數

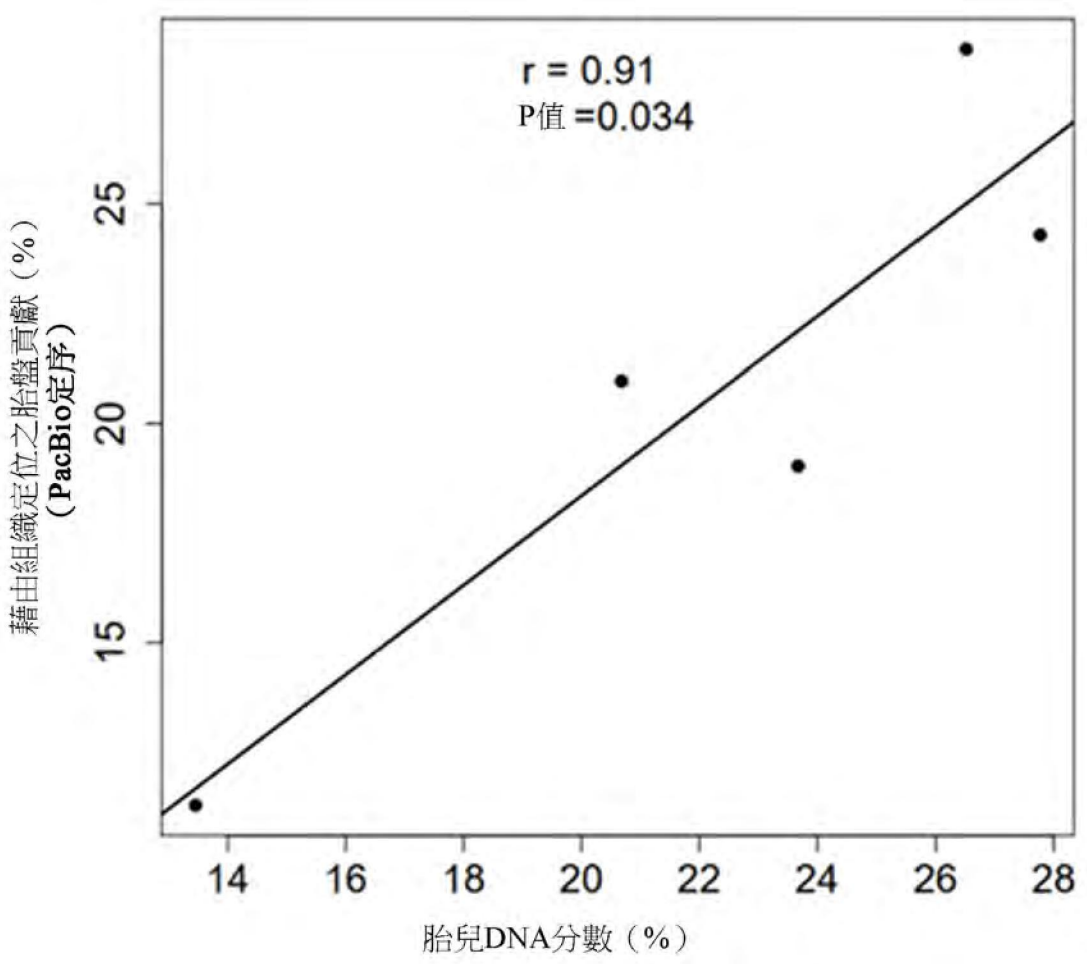
【圖57A】


TBR3032T（亞硫酸氫鹽定序）
複本數比率
染色體編號
複本數增加
複本數損失
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
0 1 2 3 4 5 6
0 500 1000 1500
基因體面元指數

【圖57B】

血漿DNA
循環胎兒DNA
TACGTG
Plasma DNA
PacBio定序
G
T
A
C
亞硫酸氫鹽定序
參考組織甲基化體
T細胞
肝臟
B細胞
肺
嗜中性白血球
結腸
胎盤
血漿DNA組織定位
DNA貢獻%
100
50
0
嗜中性白血球
B細胞
T細胞
肝臟
肺
結腸
胎盤

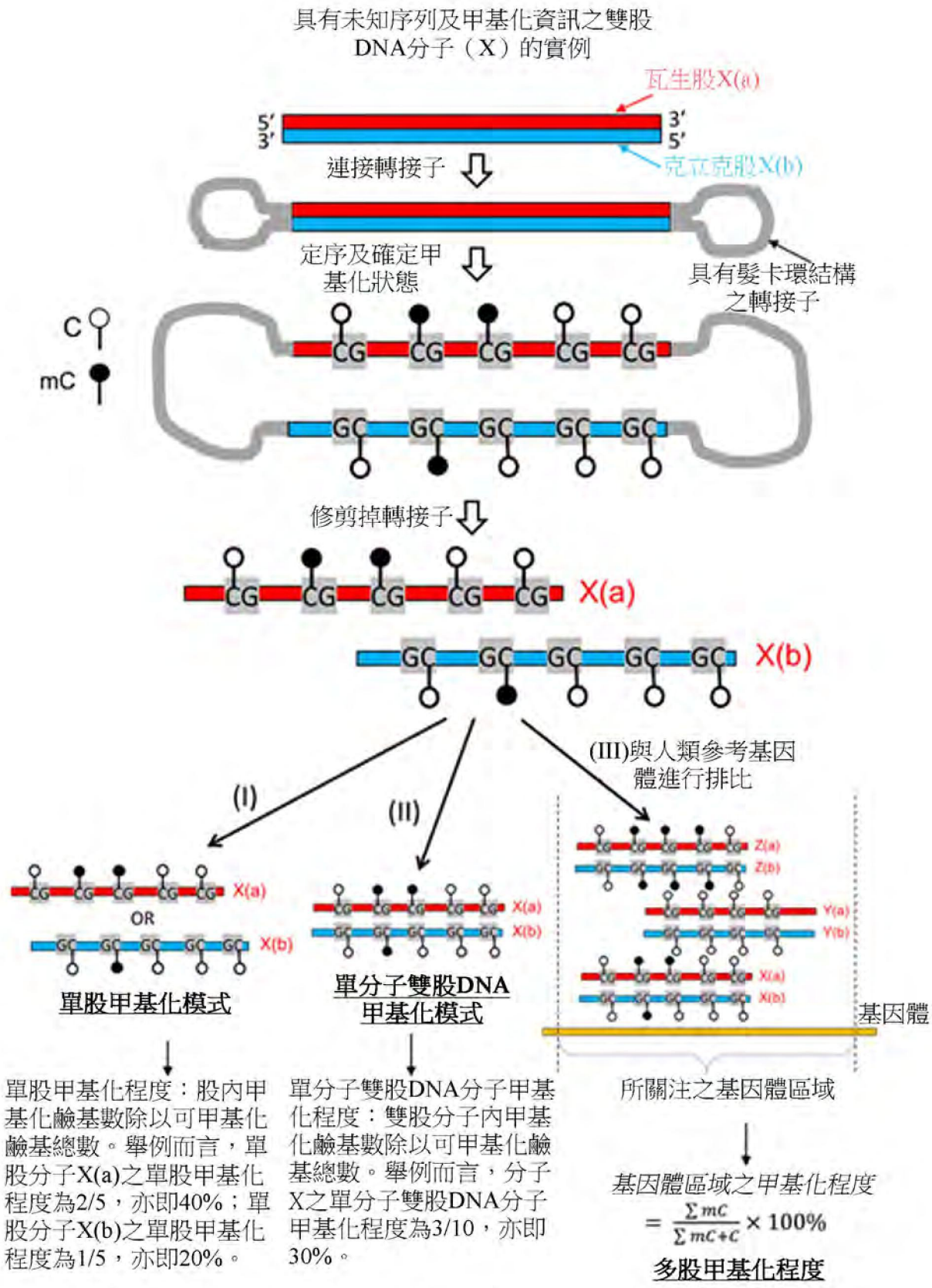
【圖58】


r = 0.91
P值 =0.034
藉由組織定位之胎盤貢獻（%）
（PacBio定序）
15
20
25
14
16
18
20
22
24
26
28
胎兒DNA分數（%）

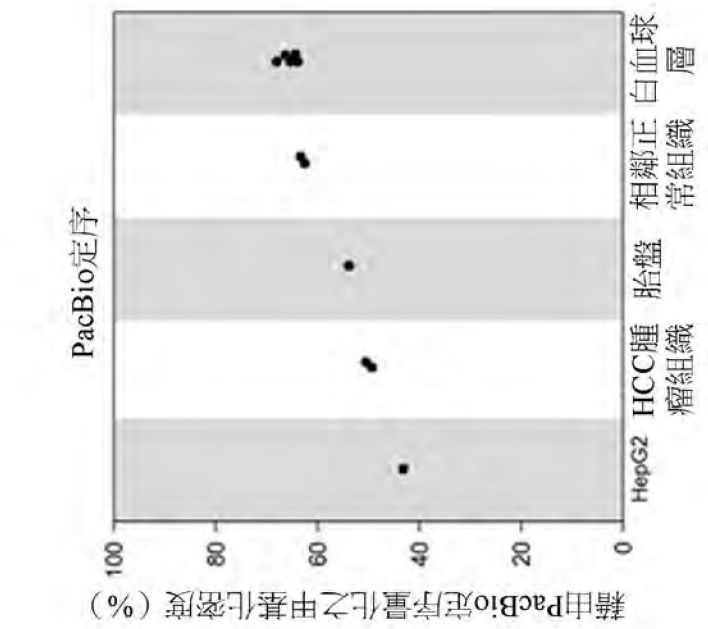
【圖59】

群組	樣本	總子讀段	經定位之子讀段	子讀段可定位性（%）	每個SMRT孔之平均子讀段深度（x）	SMRT孔之數量	可定位孔	可定位孔率（%）
母本白血球層	M13153w	39,006,460	30,673,525	78.6	13.4	3,157,310	2,295,002	72.7
胎盤	N13153	23,013,428	16,374,758	71.2	10.4	2,393,400	1,573,540	65.7
HCC組織	TBR3032T	20,164,513	15,232,744	75.5	13.1	1,742,990	1,147,985	64.8
	TBR3033T	22,639,692	17,479,024	77.2	8.1	2,832,627	2,157,196	76.2
相鄰正常組織	TBR3033N	73,118,110	56,446,202	77.2	12.6	6,881,142	4,471,370	65.0
	TBR3032N	76,852,680	60,145,452	78.3	12.8	6,000,227	4,702,130	78.4
白血球層（健康對照個體）	M1	44,777,423	28,325,587	63.3	7.7	7,316,000	3,659,996	50.0
	F2	49,840,758	32,994,645	66.2	8.6	7,215,112	3,823,329	53.0
	F1	40,012,804	24,717,289	61.8	6.5	7,301,768	3,800,392	52.0
	M2	152,530,411	88,596,520	58.1	7.7	21,794,606	11,563,500	53.1
HCC細胞株	HepG2	47,308,982	34,581,721	73.1	7.3	6,220,000	4,750,581	76.4

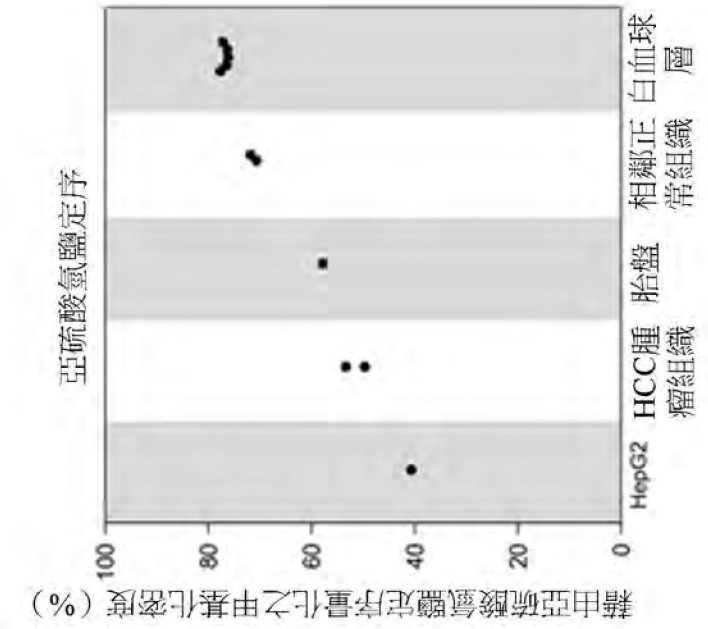
【圖60】


具有未知序列及甲基化資訊之雙股DNA分子（X）的實例
互生股X(a)
克立克股X(b)
5′ 3′
3′ 5′
連接轉接子
定序及確定甲基化狀態
具有髮卡環結構之轉接子
C
mC
CG
GC
修剪掉轉接子
X(a)
X(b)
(I)
(II)
(III)與人類參考基因體進行排比
OR
單股甲基化模式
單分子雙股DNA甲基化模式
Z(a)
Z(b)
Y(a)
Y(b)
基因體
所關注之基因體區域
單股甲基化程度：股內甲基化鹼基數除以可甲基化鹼基總數。舉例而言，單股分子X(a)之單股甲基化程度為2/5，亦即40%；單股分子X(b)之單股甲基化程度為1/5，亦即20%。
單分子雙股DNA分子甲基化程度：雙股分子內甲基化鹼基數除以可甲基化鹼基總數。舉例而言，分子X之單分子雙股DNA分子甲基化程度為3/10，亦即30%。
基因體區域之甲基化程度
$= \frac{\sum mC}{\sum mC+C} \times 100\%$
多股甲基化程度

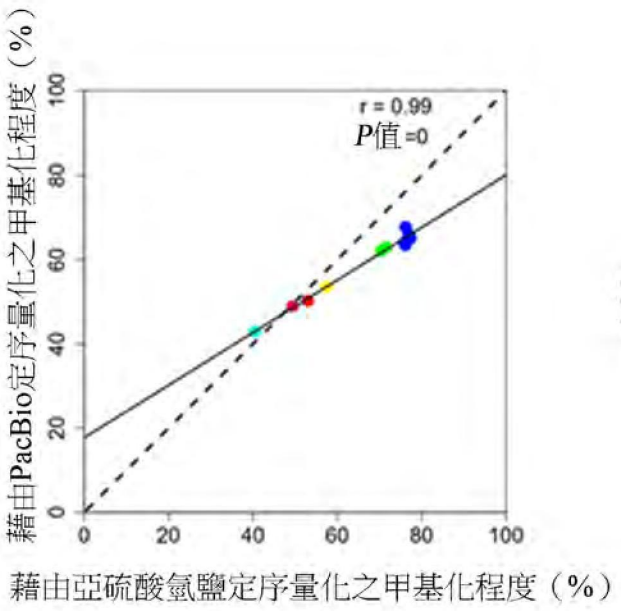
【圖61】


PacBio定序
藉由PacBio定序量化之甲基化密度（%）
100
80
60
40
20
0
HepG2
HCC腫瘤組織
胎盤
相鄰正常組織
白血球層

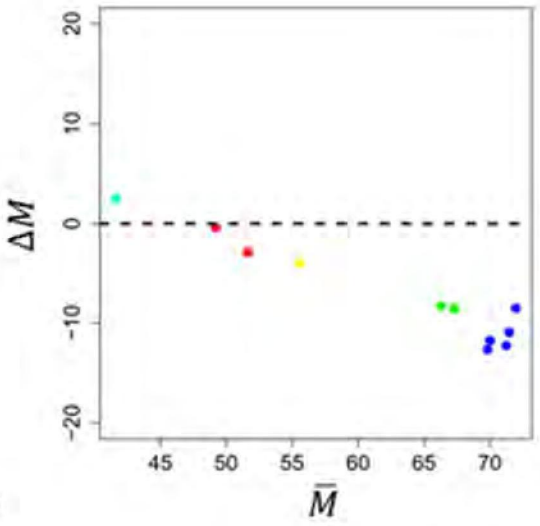
【圖62B】


亞硫酸氫鹽定序
藉由亞硫酸氫鹽定序量化之甲基化密度（%）
100
80
60
40
20
0
HepG2
HCC腫瘤組織
胎盤
相鄰正常組織
白血球層

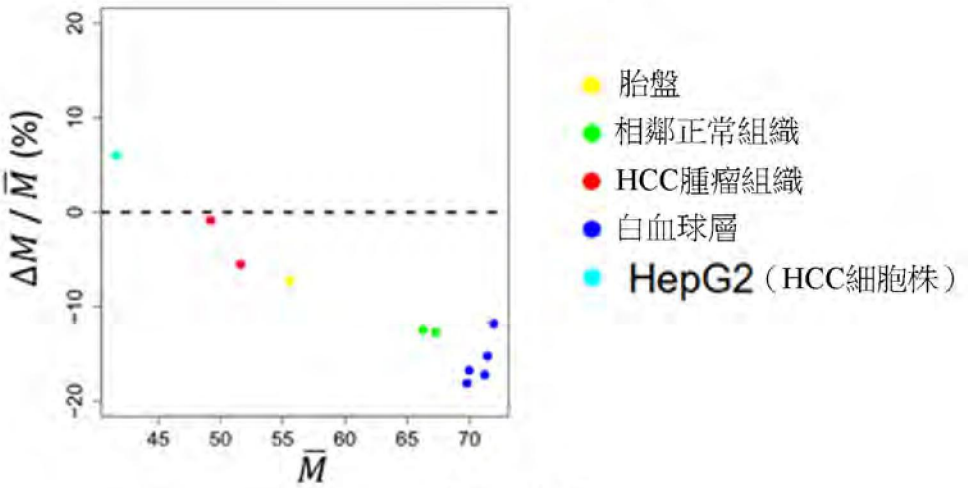
【圖62A】



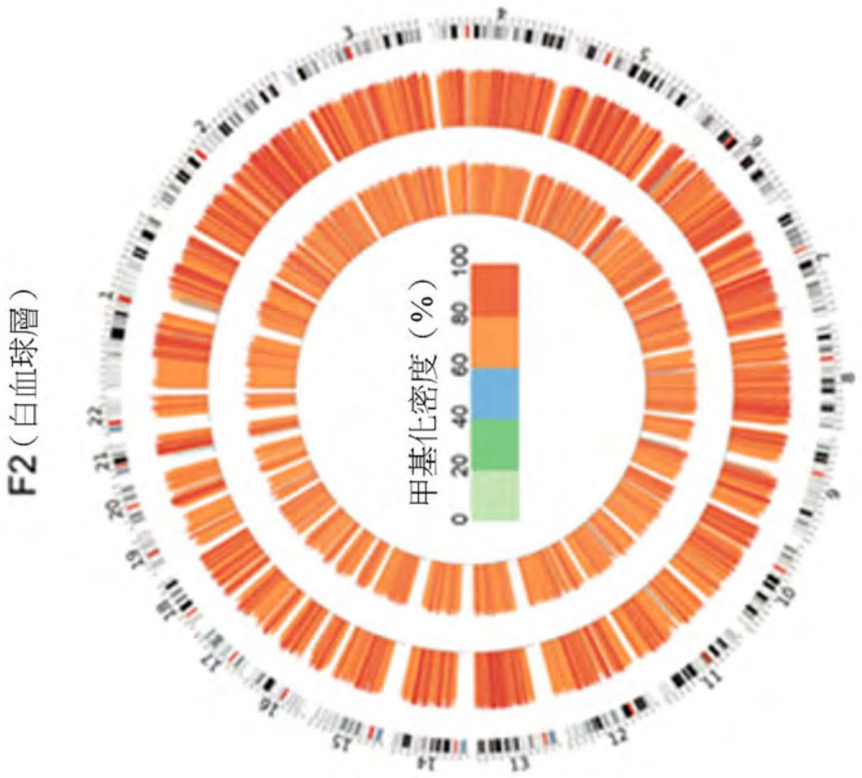
【圖63A】



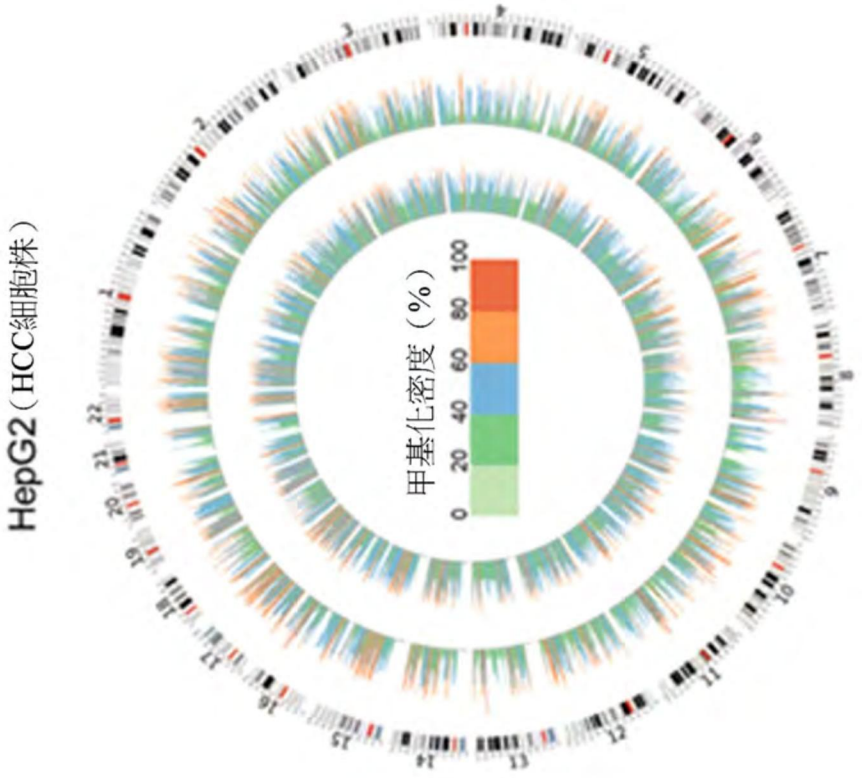
【圖63B】



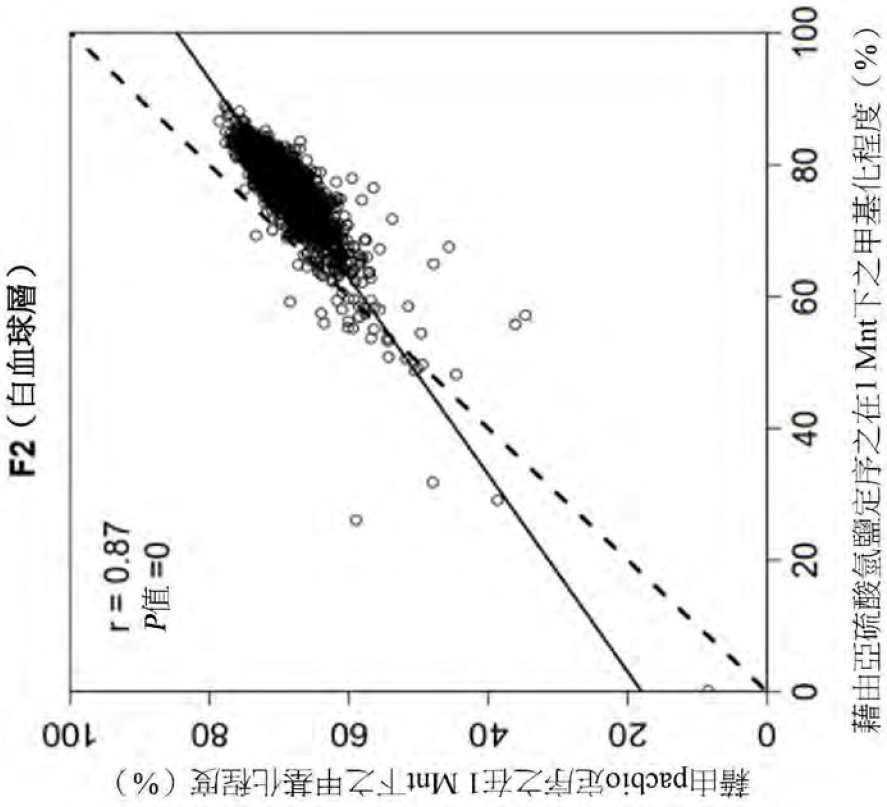
【圖63C】


F2（白血球層）
甲基化密度（%）
0
20
40
60
80
100

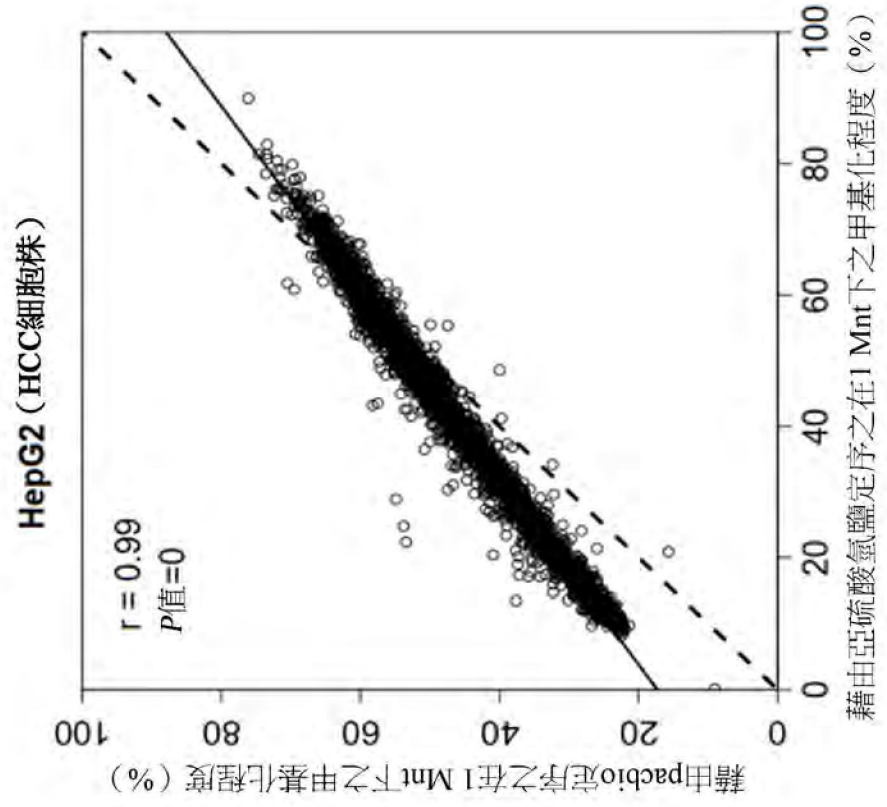
【圖64B】


HepG2（HCC細胞株）
甲基化密度（%）
0
20
40
60
80
100

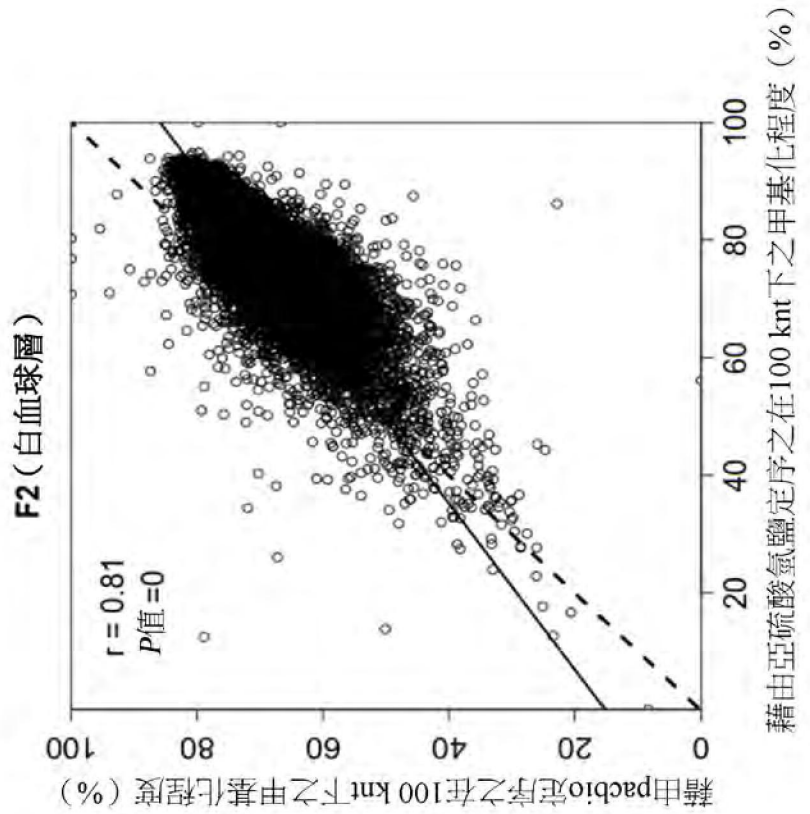
【圖64A】


F2（白血球層）
r = 0.87
P值 =0
藉由pacbio定序之在1 Mnt下之甲基化程度（%）
藉由亞硫酸氫鹽定序之在1 Mnt下之甲基化程度（%）
0
20
40
60
80
100

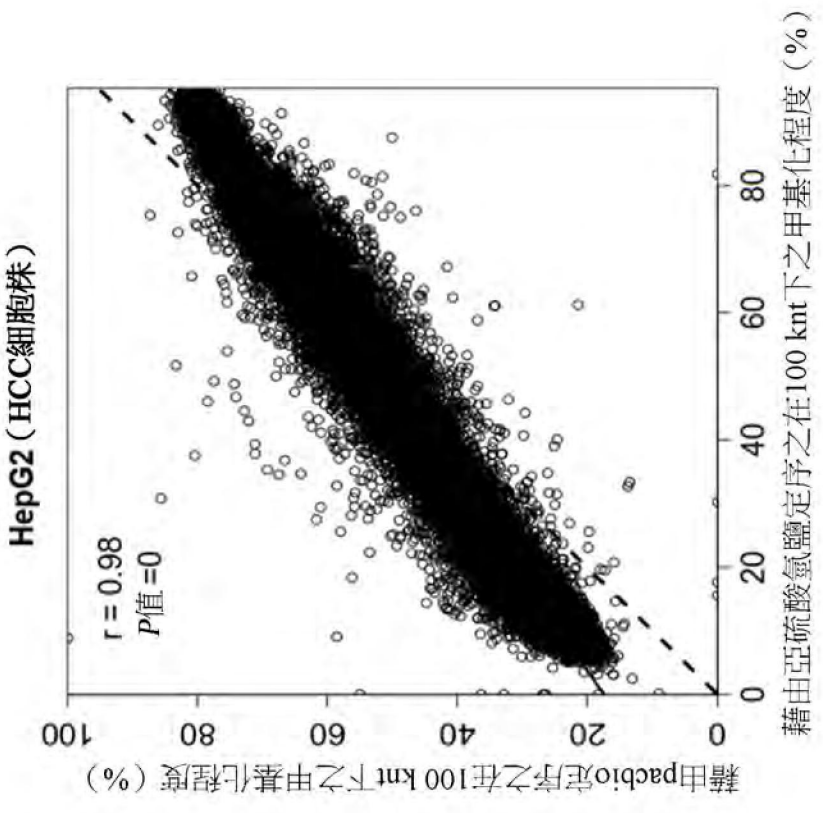
【圖65B】


HepG2（HCC細胞株）
r = 0.99
P值=0
藉由pacbio定序之在1 Mnt下之甲基化程度（%）
藉由亞硫酸氫鹽定序之在1 Mnt下之甲基化程度（%）
0
20
40
60
80
100

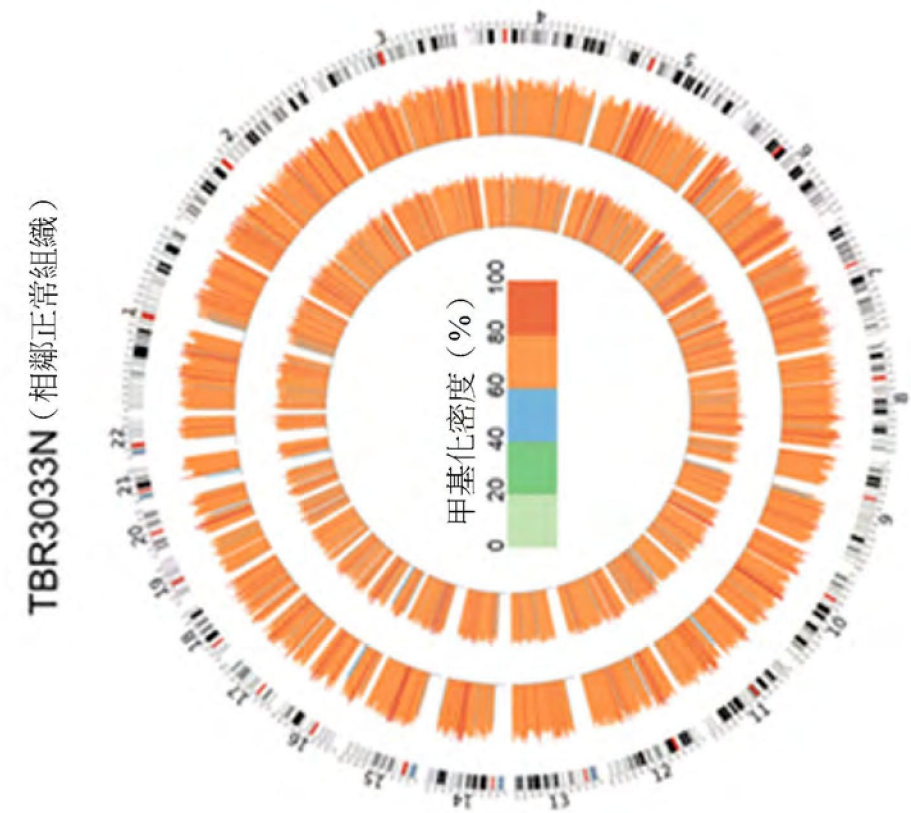
【圖65A】


F2（白血球層）
r = 0.81
P值=0
藉由pacbio定序之在100 knt下之甲基化程度（%）
藉由亞硫酸氫鹽定序之在100 knt下之甲基化程度（%）
0
20
40
60
80
100

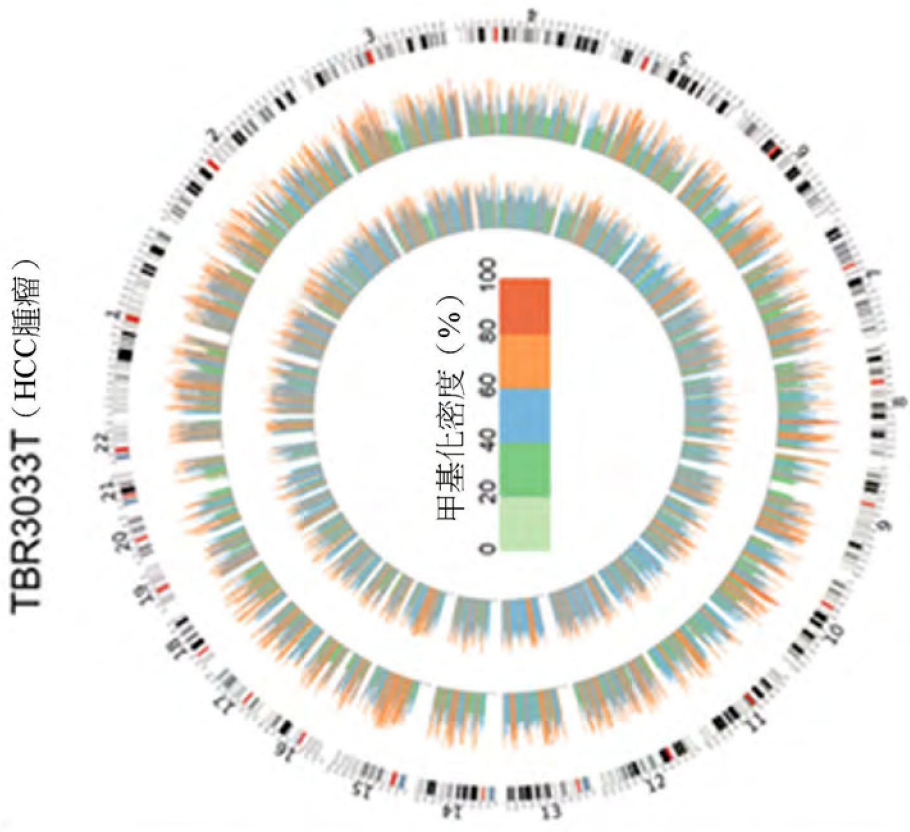
【圖66B】


HepG2（HCC細胞株）
r = 0.98
P值=0
藉由pacbio定序之在100 knt下之甲基化程度（%）
藉由亞硫酸氫鹽定序之在100 knt下之甲基化程度（%）
0
20
40
60
80
100

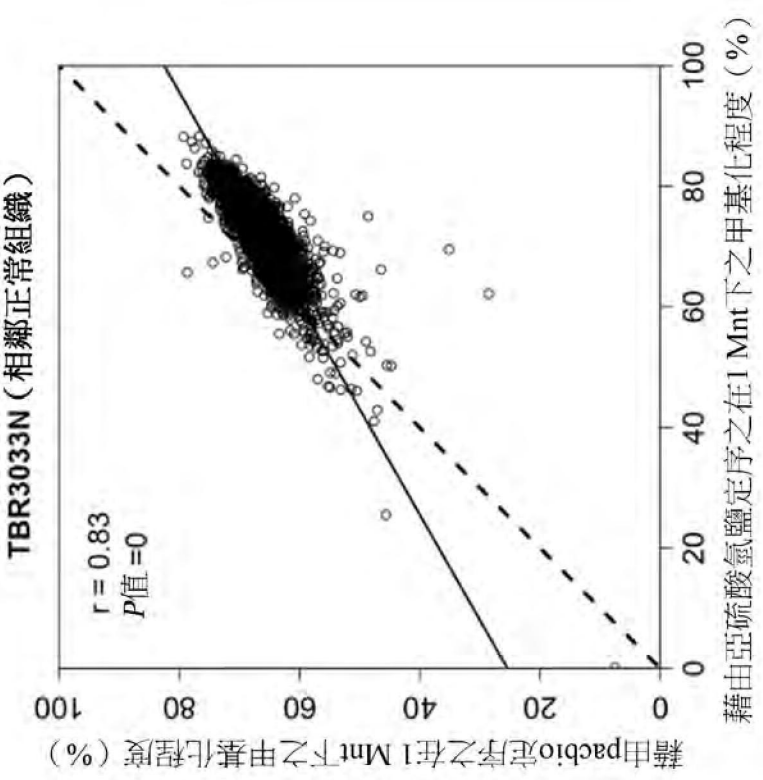
【圖66A】


TBR3033N（相鄰正常組織）
甲基化密度（%）
0
20
40
60
80
100

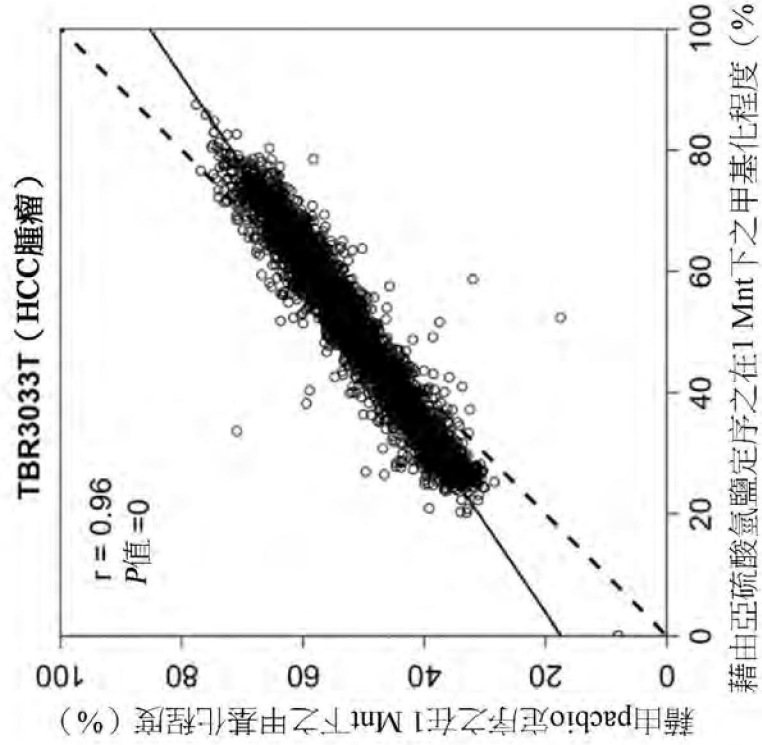
【圖67B】


TBR3033T（HCC腫瘤）
甲基化密度（%）
0
20
40
60
80
100

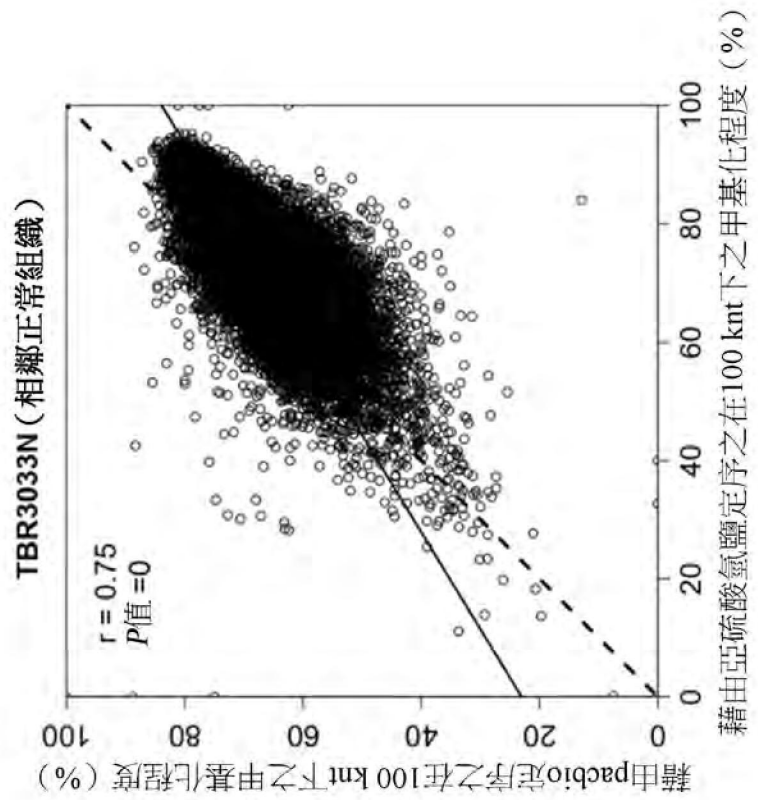
【圖67A】


TBR3033N（相鄰正常組織）
r = 0.83
P值 =0
藉由pacbio定序之在1 Mnt下之甲基化程度（%）
藉由亞硫酸氫鹽定序之在1 Mnt下之甲基化程度（%）
0
20
40
60
80
100

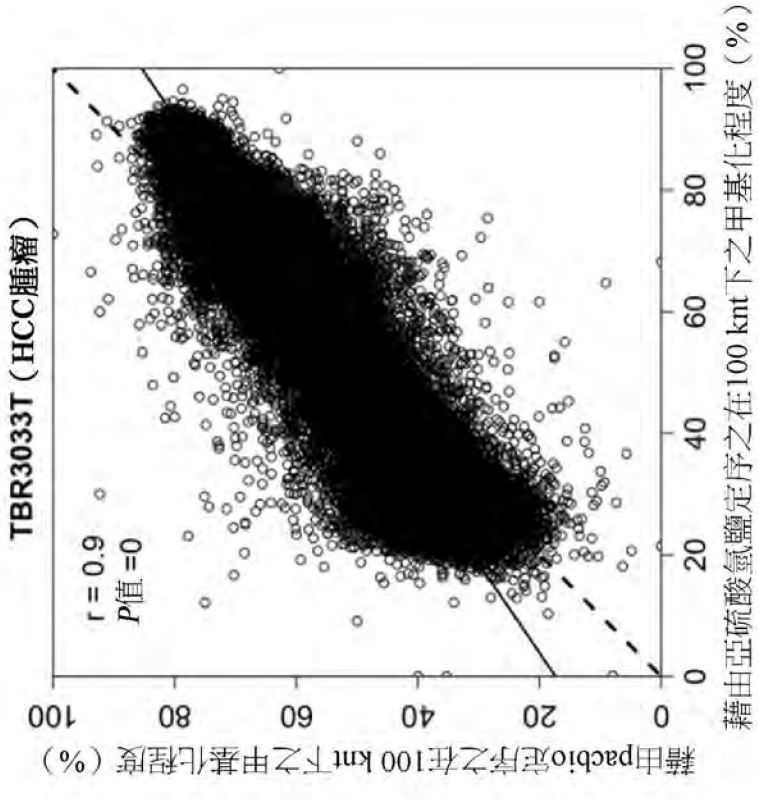
【圖68B】


TBR3033T（HCC腫瘤）
r = 0.96
P值 =0
藉由pacbio定序之在1 Mnt下之甲基化程度（%）
藉由亞硫酸氫鹽定序之在1 Mnt下之甲基化程度（%）
0
20
40
60
80
100

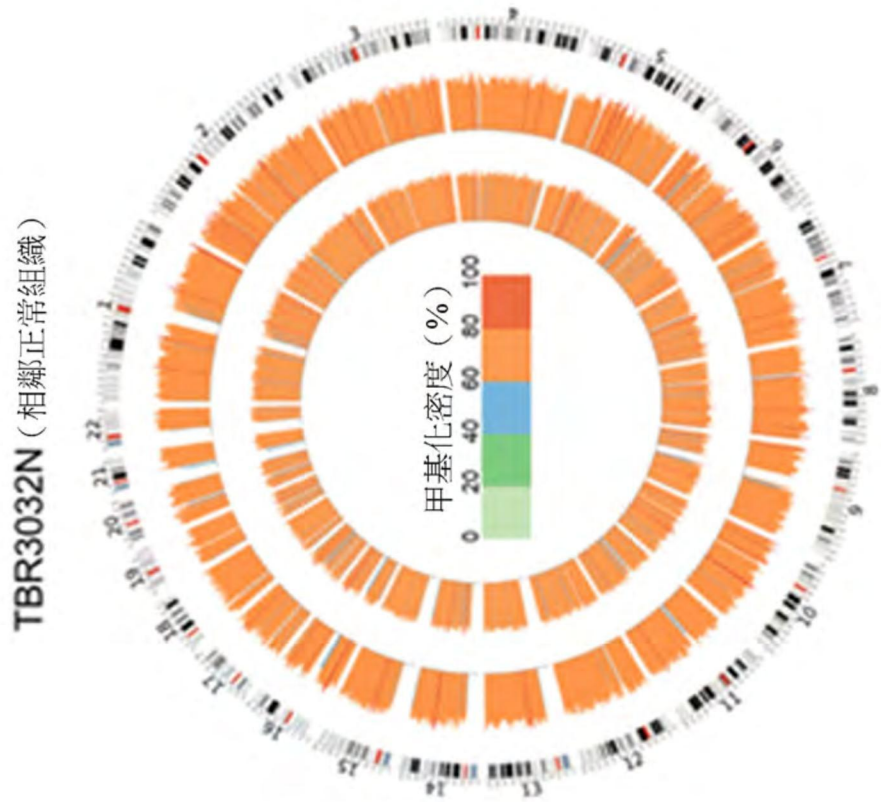
【圖68A】


TBR3033N（相鄰正常組織）
r = 0.75
P值 =0
藉由pacbio定序之在100 knt下之甲基化程度（%）
藉由亞硫酸氫鹽定序之在100 knt下之甲基化程度（%）
0
20
40
60
80
100

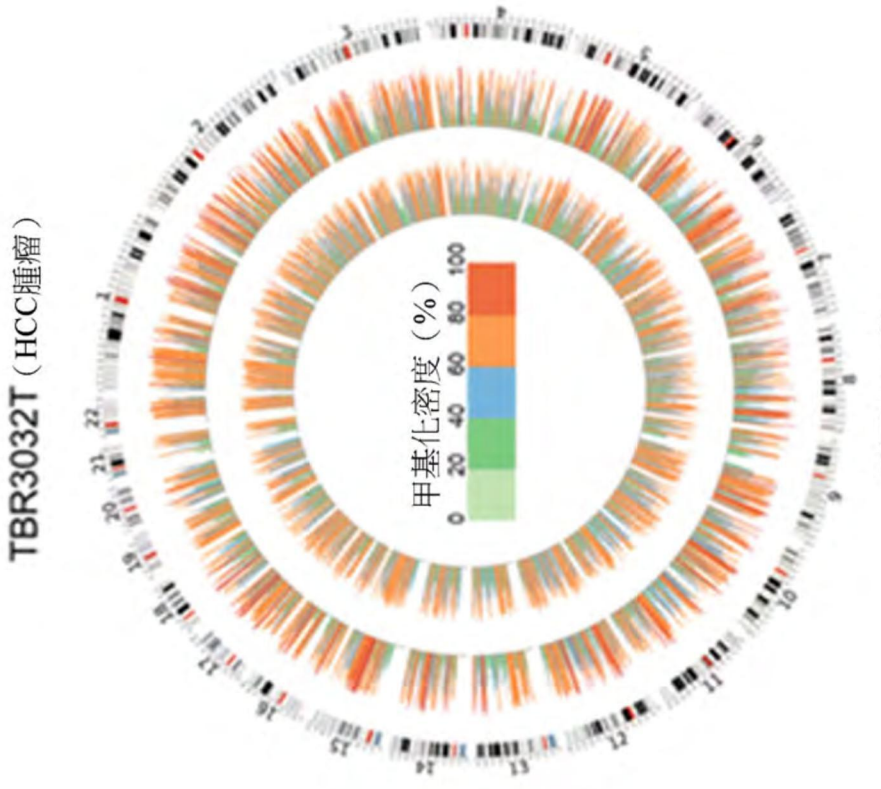
【圖69B】


TBR3033T（HCC腫瘤）
r = 0.9
P值 =0
藉由pacbio定序之在100 knt下之甲基化程度（%）
藉由亞硫酸氫鹽定序之在100 knt下之甲基化程度（%）
0
20
40
60
80
100

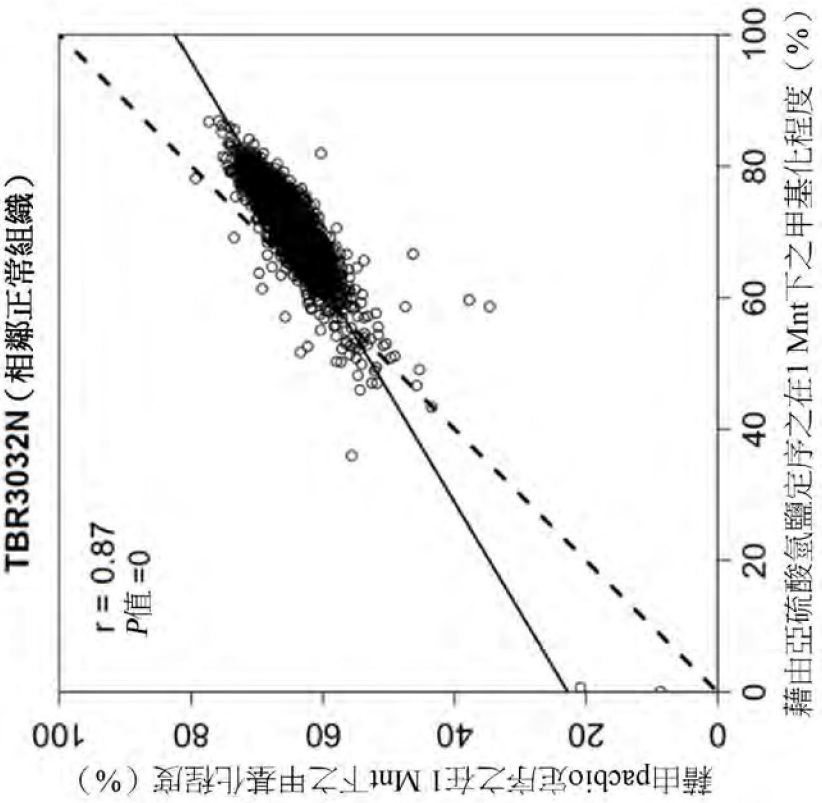
【圖69A】


TBR3032N（相鄰正常組織）
甲基化密度（%）
0
20
40
60
80
100

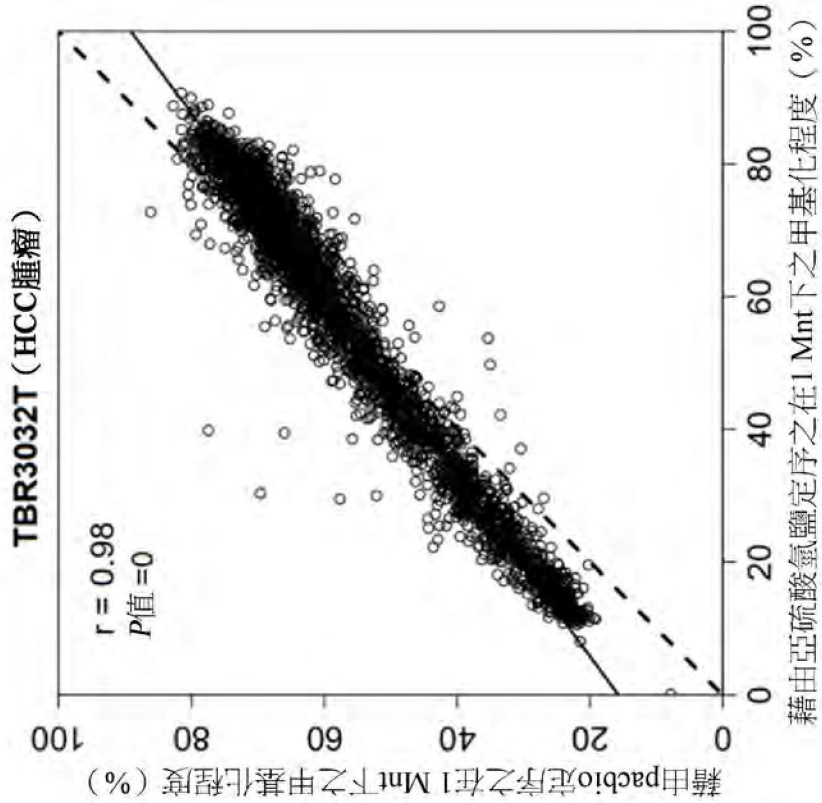
【圖70B】


TBR3032T（HCC腫瘤）
甲基化密度（%）
0
20
40
60
80
100

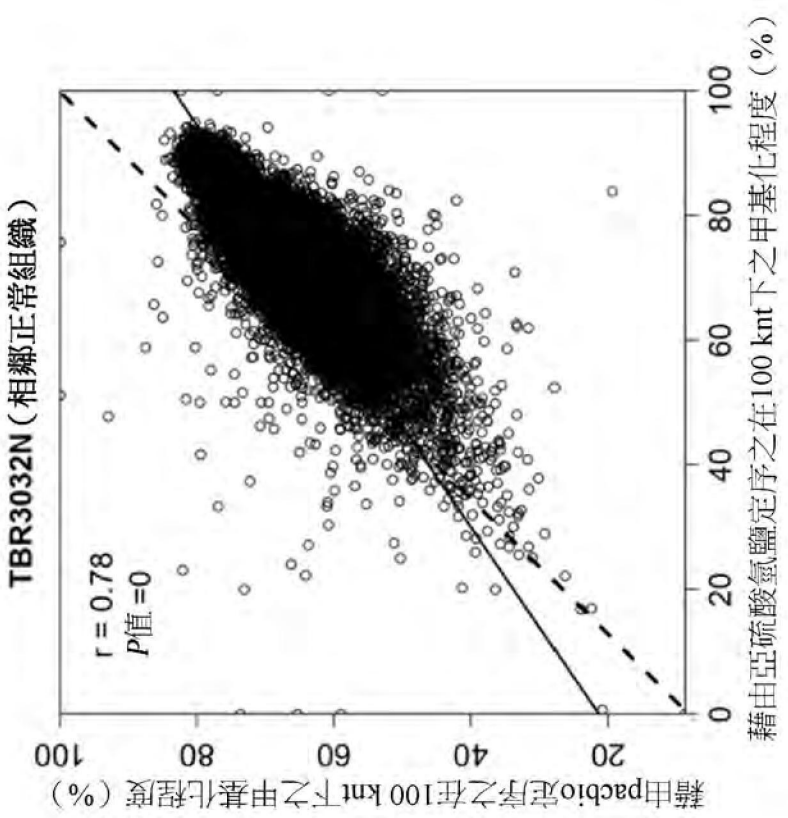
【圖70A】


TBR3032N（相鄰正常組織）
r = 0.87
P值 =0
藉由pacbio定序之在1 Mnt下之甲基化程度（%）
藉由亞硫酸氫鹽定序之在1 Mnt下之甲基化程度（%）
0
20
40
60
80
100

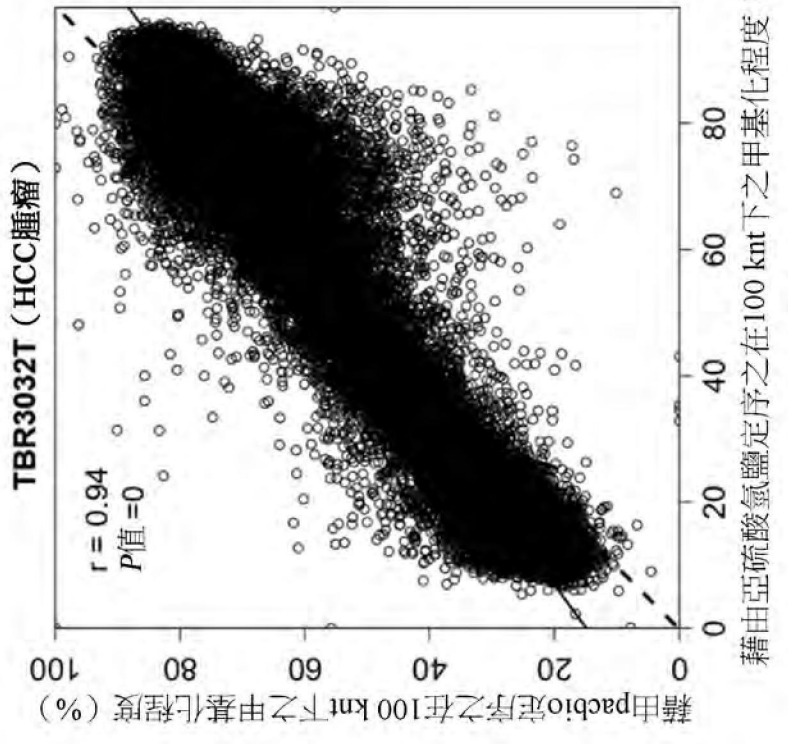
【圖71B】


TBR3032T（HCC腫瘤）
r = 0.98
P值 =0
藉由pacbio定序之在1 Mnt下之甲基化程度（%）
藉由亞硫酸氫鹽定序之在1 Mnt下之甲基化程度（%）
0
20
40
60
80
100

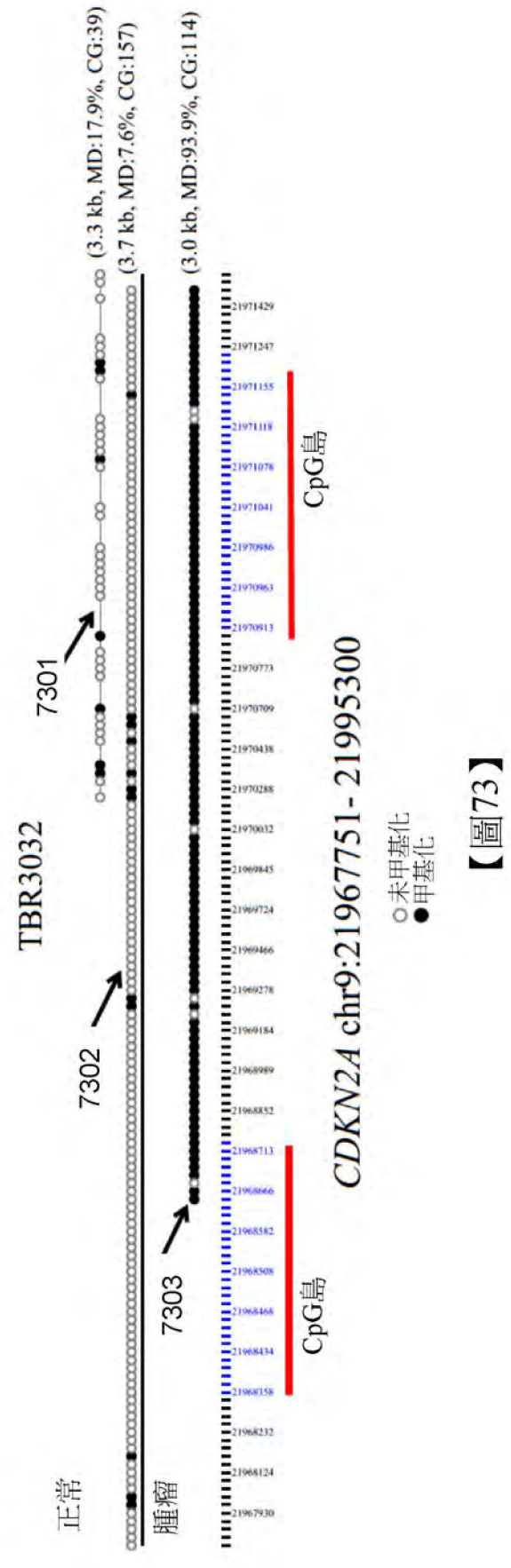
【圖71A】


TBR3032N（相鄰正常組織）
r = 0.78
P值 =0
藉由pacbio定序之在100 knt下之甲基化程度（%）
藉由亞硫酸氫鹽定序之在100 knt下之甲基化程度（%）
0
20
40
60
80
100

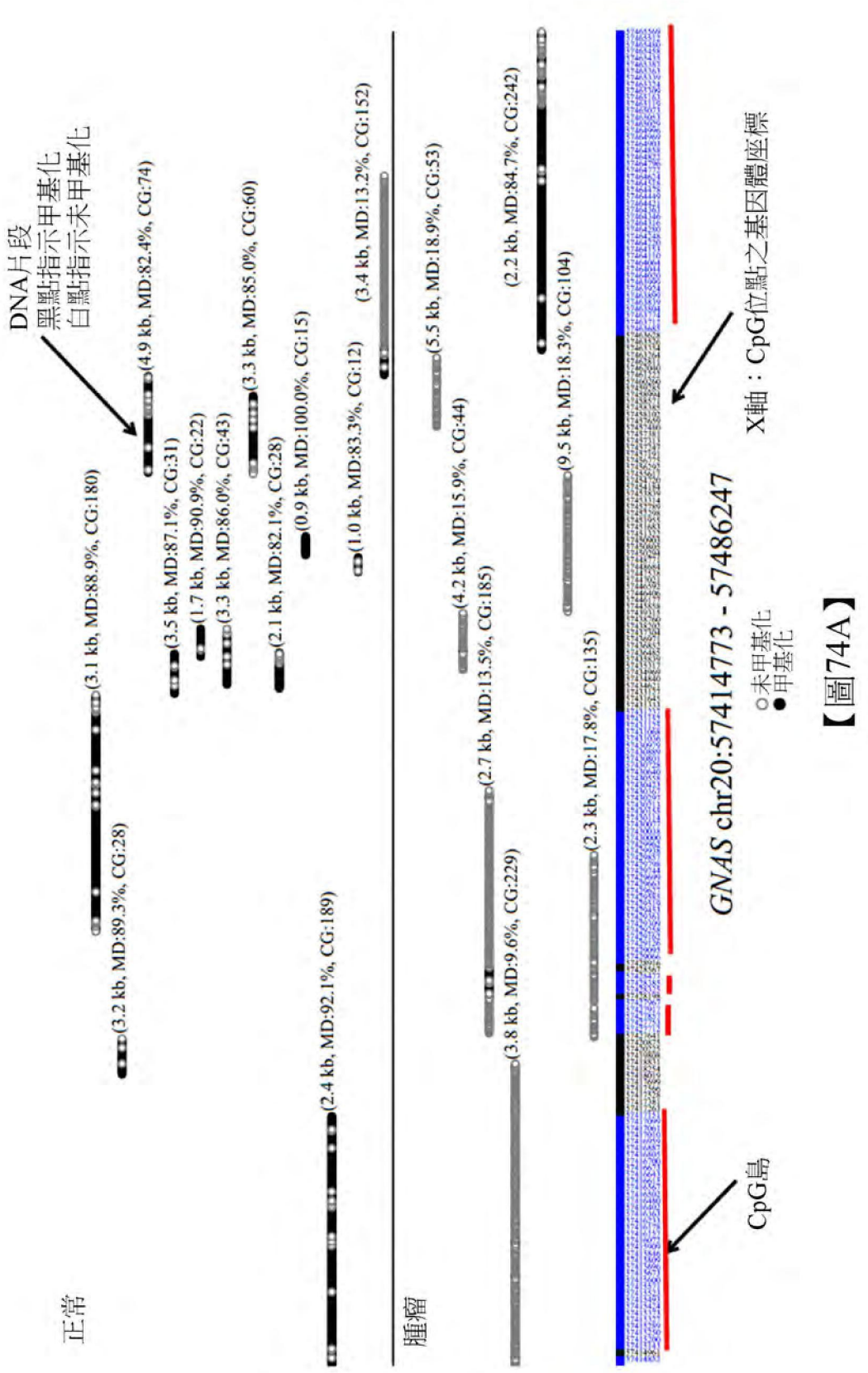
【圖72B】


TBR3032T（HCC腫瘤）
r = 0.94
P值 =0
藉由pacbio定序之在100 knt下之甲基化程度（%）
藉由亞硫酸氫鹽定序之在100 knt下之甲基化程度（%）
0
20
40
60
80
100

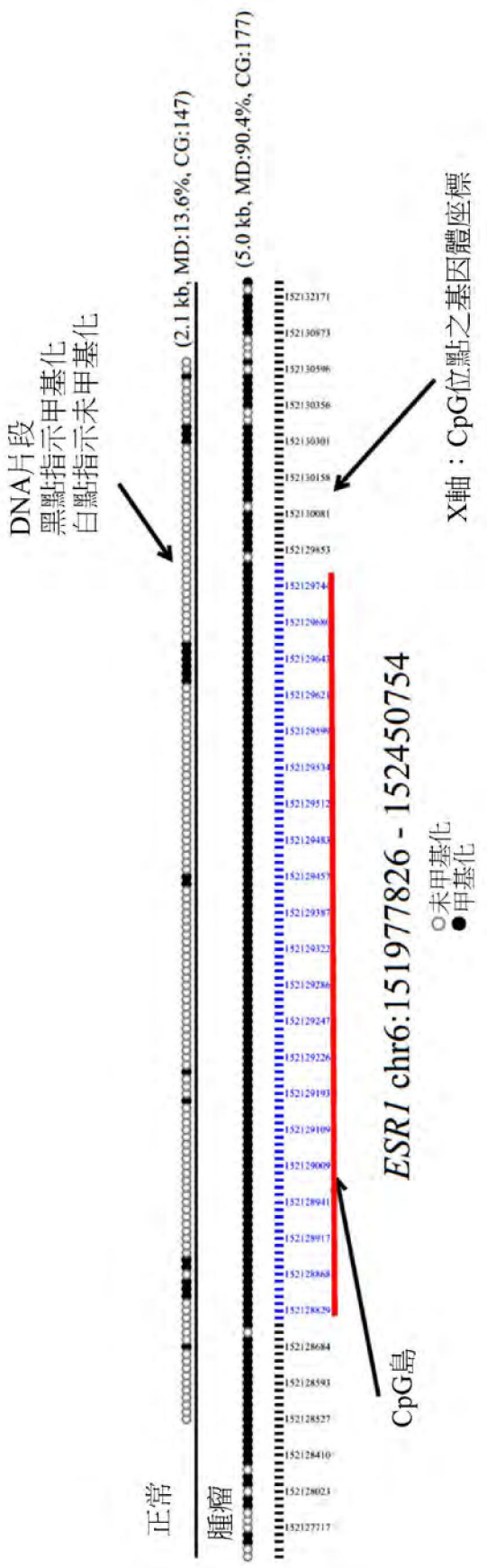
【圖72A】


TBR3032
7301
7302
7303
(3.3 kb, MD:17.9%, CG:39)
(3.7 kb, MD:7.6%, CG:157)
(3.0 kb, MD:93.9%, CG:114)
CpG島
CpG島
CDKN2A chr9:21967751- 21995300
○未甲基化
●甲基化
正常
腫瘤

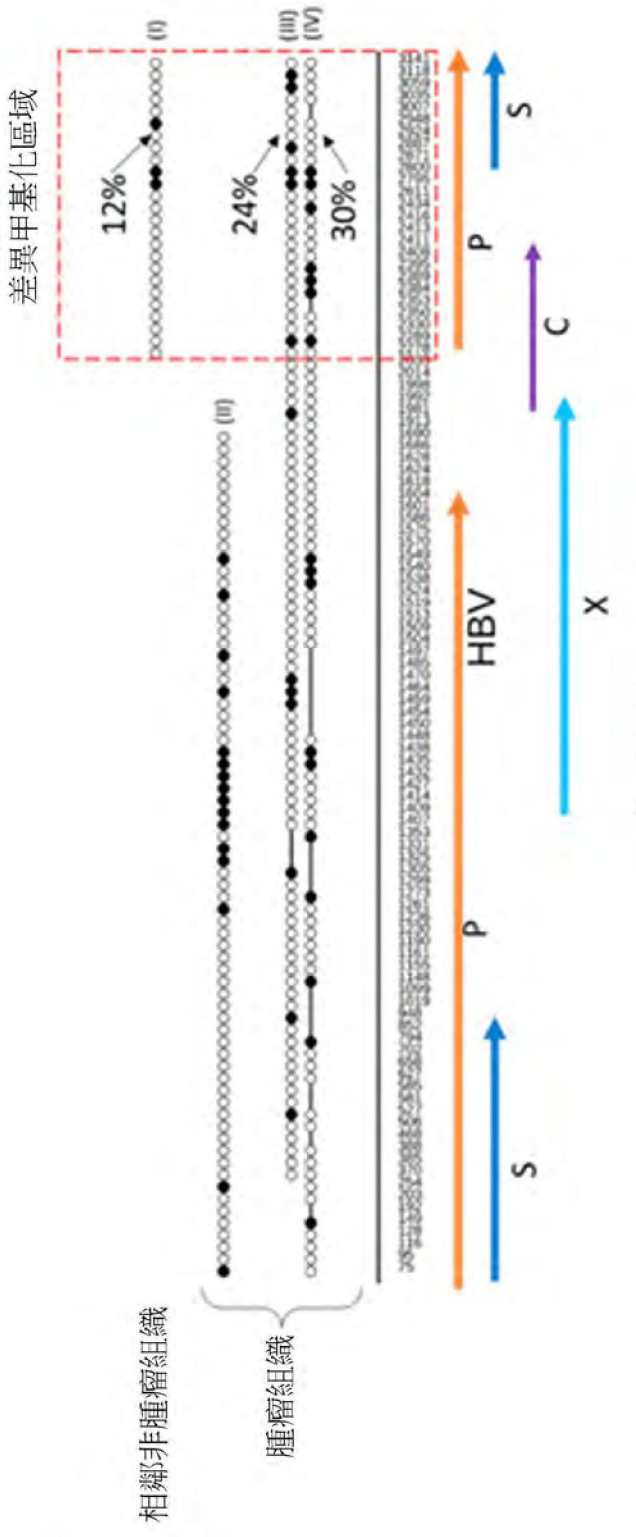
【圖73】


DNA片段
黑點指示甲基化
白點指示未甲基化
正常
腫瘤
(3.1 kb, MD:88.9%, CG:180)
(3.2 kb, MD:89.3%, CG:28)
(4.9 kb, MD:82.4%, CG:74)
(3.5 kb, MD:87.1%, CG:31)
(1.7 kb, MD:90.9%, CG:22)
(3.3 kb, MD:86.0%, CG:43)
(3.3 kb, MD:85.0%, CG:60)
(2.1 kb, MD:82.1%, CG:28)
(0.9 kb, MD:100.0%, CG:15)
(1.0 kb, MD:83.3%, CG:12)
(3.4 kb, MD:13.2%, CG:152)
(2.4 kb, MD:92.1%, CG:189)
(5.5 kb, MD:18.9%, CG:53)
(4.2 kb, MD:15.9%, CG:44)
(2.7 kb, MD:13.5%, CG:185)
(3.8 kb, MD:9.6%, CG:229)
(2.2 kb, MD:84.7%, CG:242)
(9.5 kb, MD:18.3%, CG:104)
(2.3 kb, MD:17.8%, CG:135)
GNAS chr20:57414773 - 57486247
X軸：CpG位點之基因體座標
○未甲基化
●甲基化
CpG島

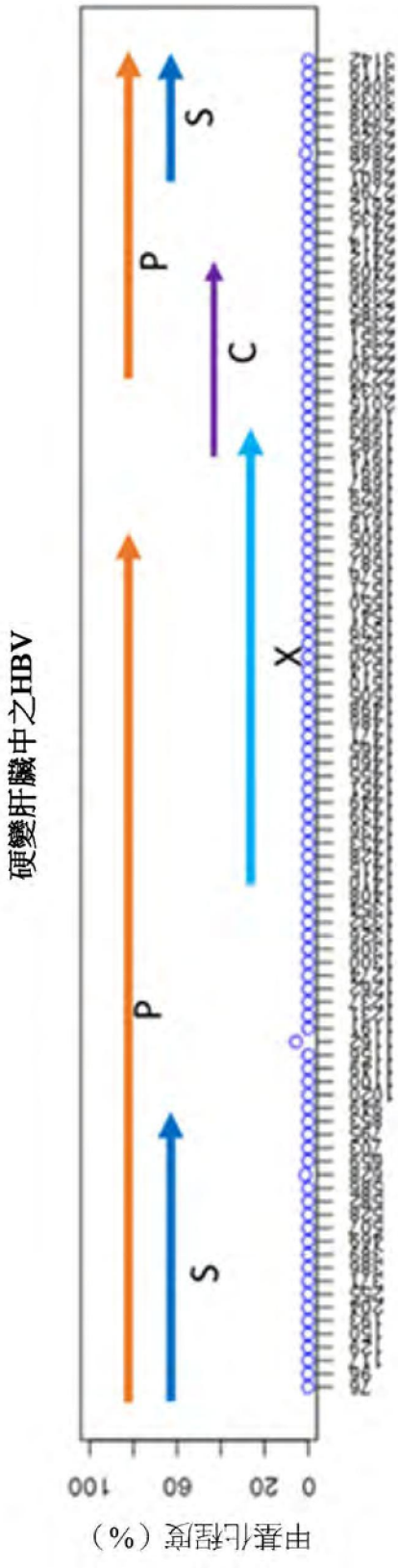
【圖74A】


DNA片段
黑點指示甲基化
白點指示未甲基化
(2.1 kb, MD:13.6%, CG:147)
(5.0 kb, MD:90.4%, CG:177)
X軸：CpG位點之基因體座標
正常
腫瘤
CpG島
ESR1 chr6:151977826 - 152450754
○未甲基化
●甲基化

【圖74B】

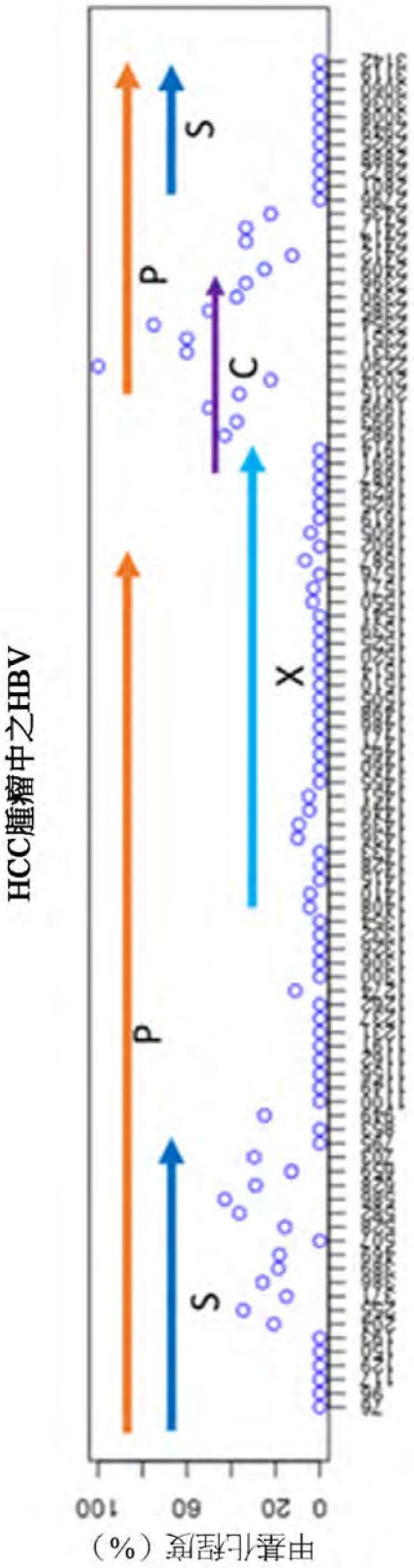


【圖75】



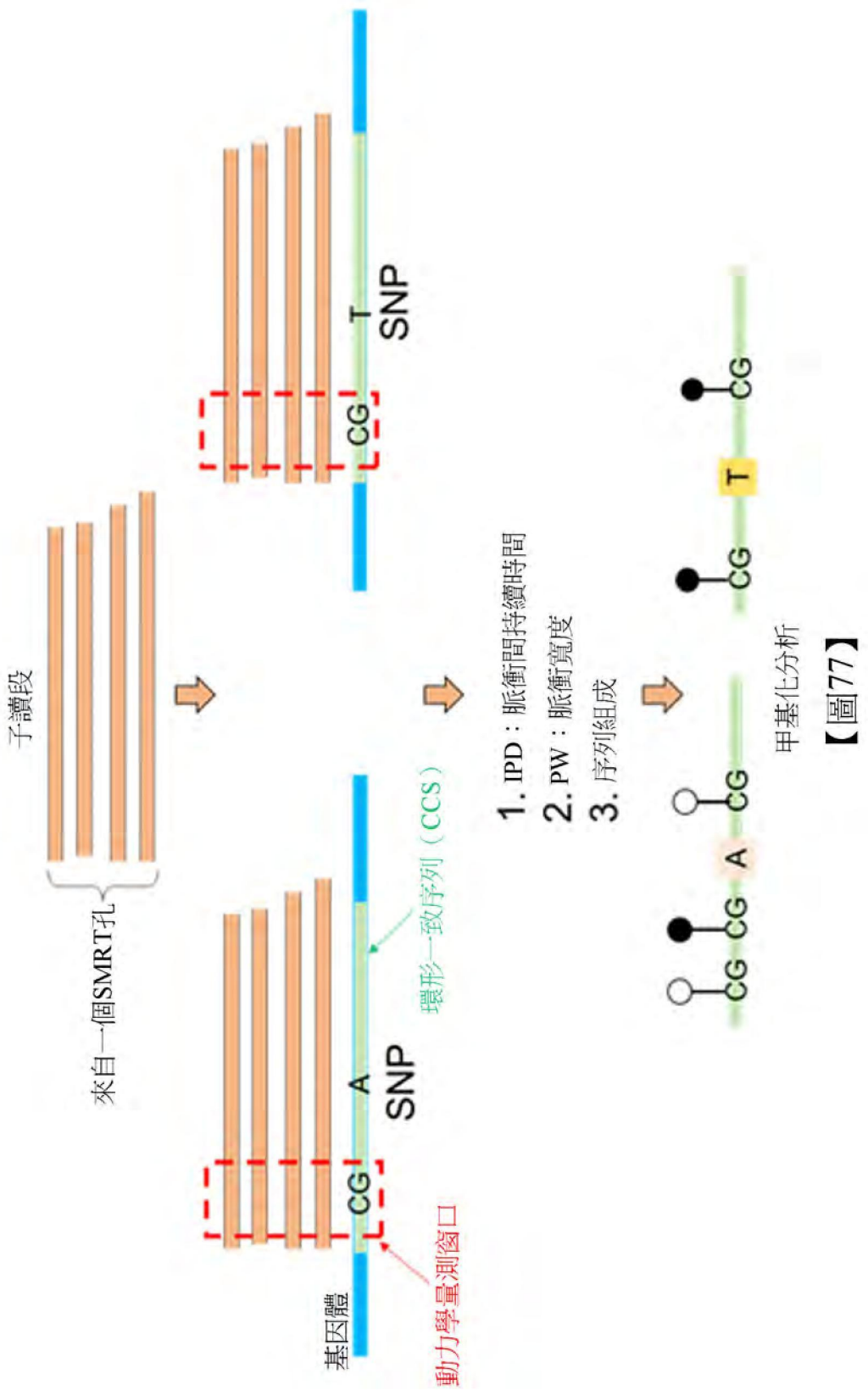
4個硬變肝臟組織中之HBV：25 X深度之中位數

【圖76A】

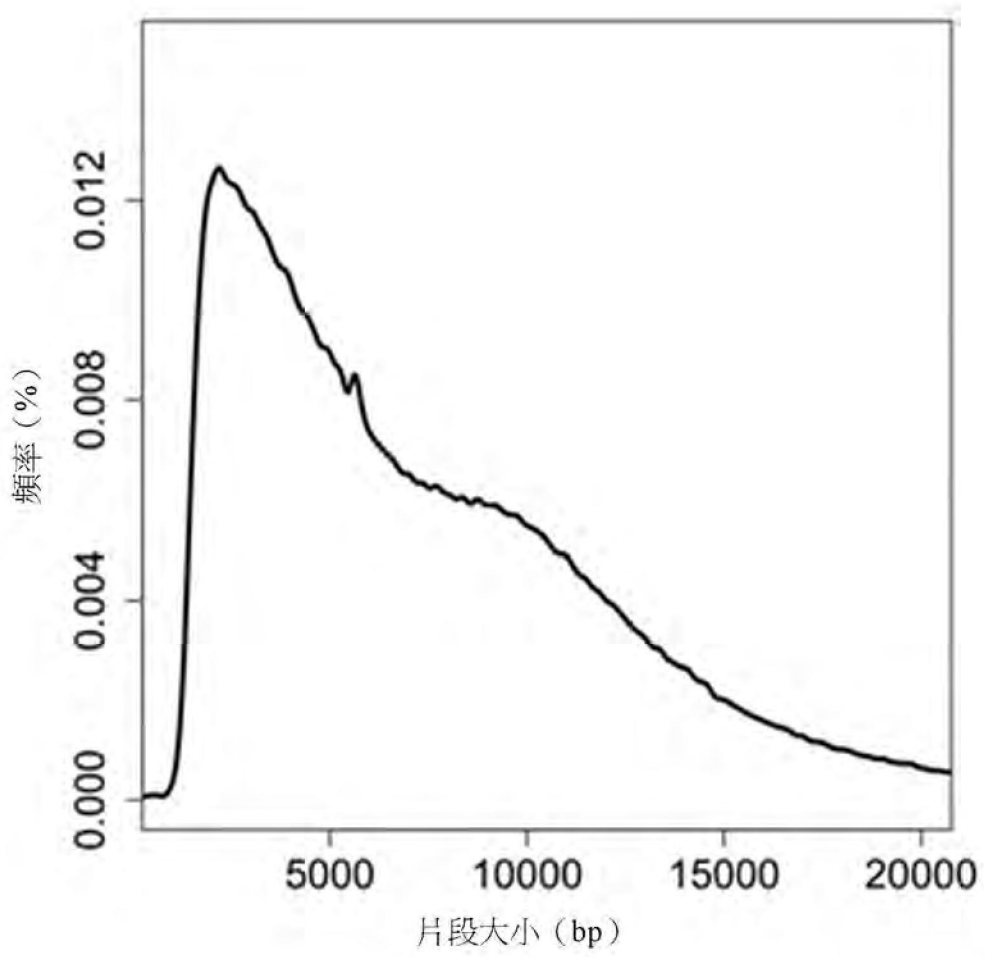


15個HCC腫瘤組織中之HBV：14 X深度之中位數

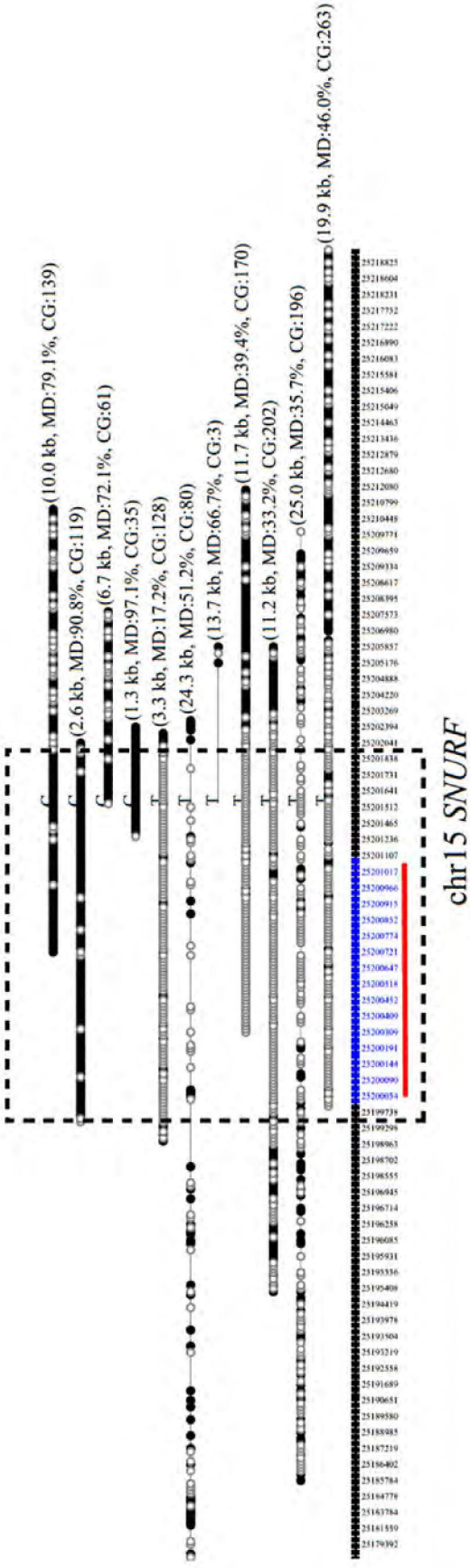
【圖76B】


子讀段
來自一個SMRT孔
基因體
CG
A
SNP
環形一致序列（CCS）
動力學量測窗口
CG
T
SNP
1. IPD：脈衝間持續時間
2. PW：脈衝寬度
3. 序列組成
CG CG A CG
CG T CG
甲基化分析

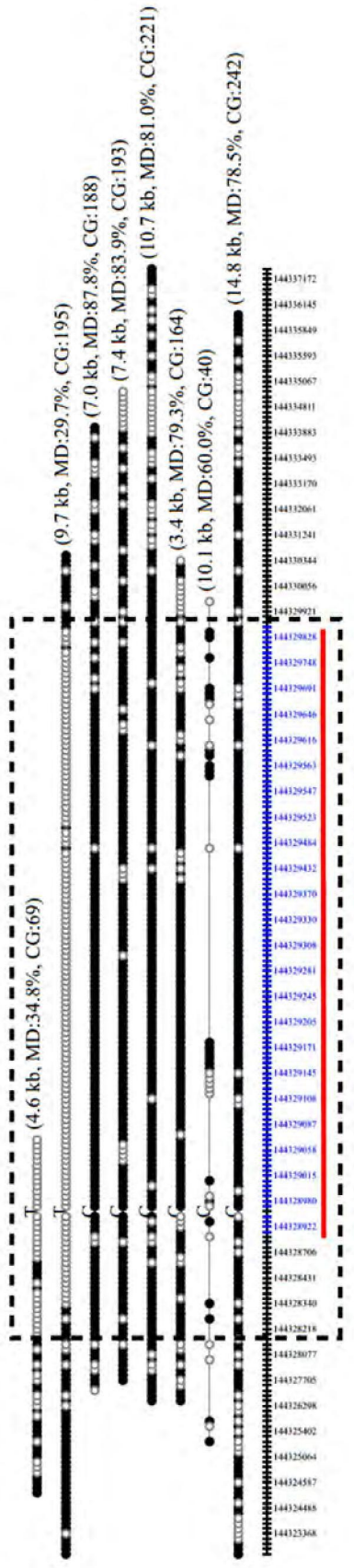
【圖77】



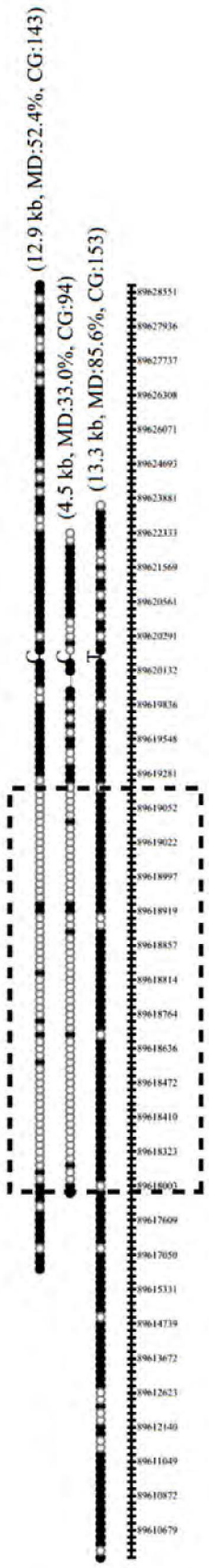
【圖78】


(10.0 kb, MD:79.1%, CG:139)
(2.6 kb, MD:90.8%, CG:119)
(6.7 kb, MD:72.1%, CG:61)
(1.3 kb, MD:97.1%, CG:35)
(3.3 kb, MD:17.2%, CG:128)
(24.3 kb, MD:51.2%, CG:80)
(13.7 kb, MD:66.7%, CG:3)
(11.7 kb, MD:39.4%, CG:170)
(11.2 kb, MD:33.2%, CG:202)
(25.0 kb, MD:35.7%, CG:196)
(19.9 kb, MD:46.0%, CG:263)
chr15 SNURF
○未甲基化
●甲基化

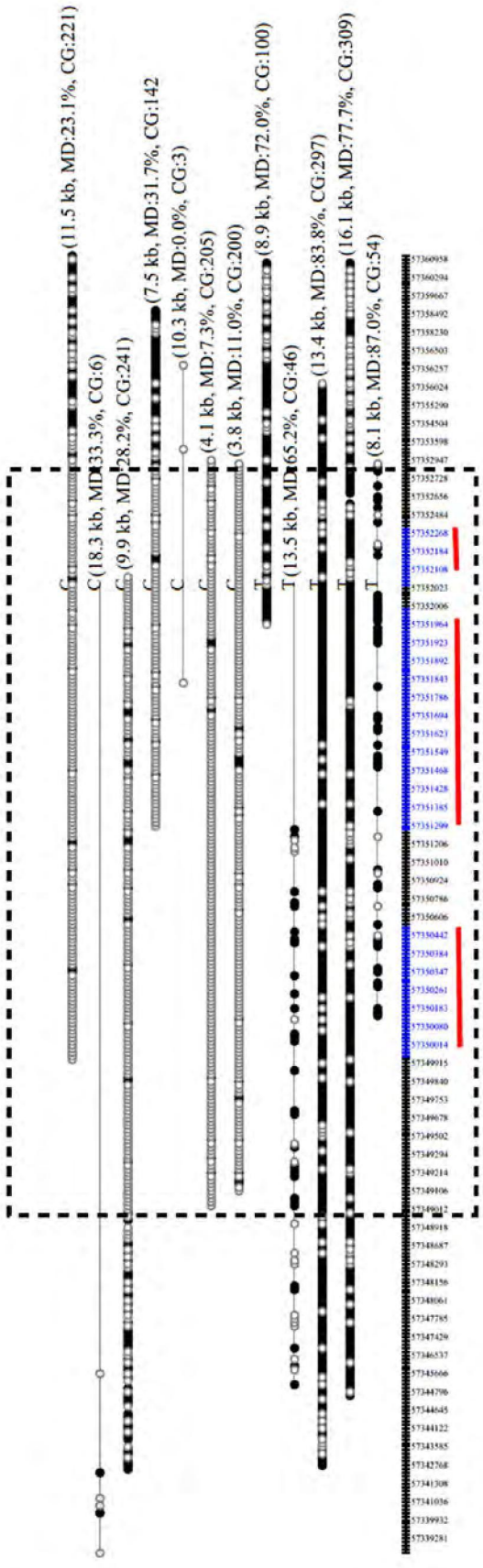
【圖79A】


(4.6 kb, MD:34.8%, CG:69)
(9.7 kb, MD:29.7%, CG:195)
(7.0 kb, MD:87.8%, CG:188)
(7.4 kb, MD:83.9%, CG:193)
(10.7 kb, MD:81.0%, CG:221)
(3.4 kb, MD:79.3%, CG:164)
(10.1 kb, MD:60.0%, CG:40)
(14.8 kb, MD:78.5%, CG:242)
chr6 PLAGL1
○未甲基化
●甲基化

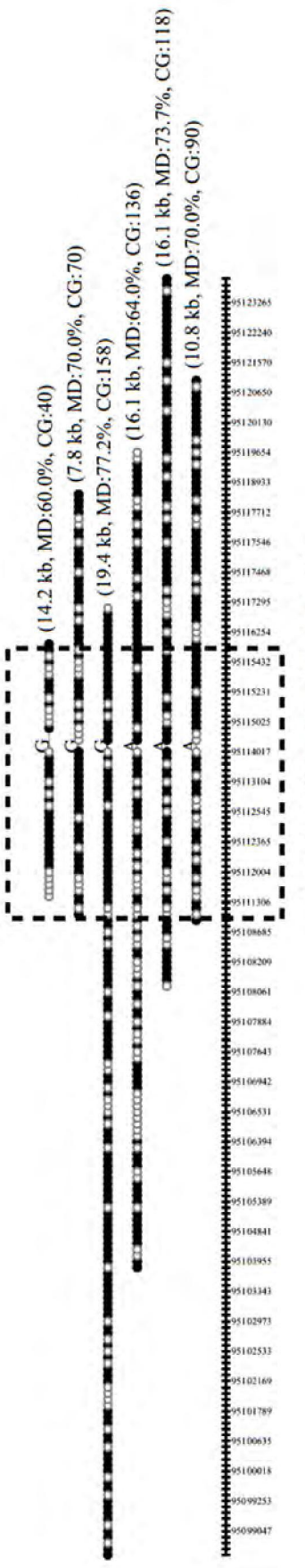
【圖79B】


(12.9 kb, MD:52.4%, CG:143)
(4.5 kb, MD:33.0%, CG:94)
(13.3 kb, MD:85.6%, CG:153)
chr4 NAP1L5
未甲基化
甲基化

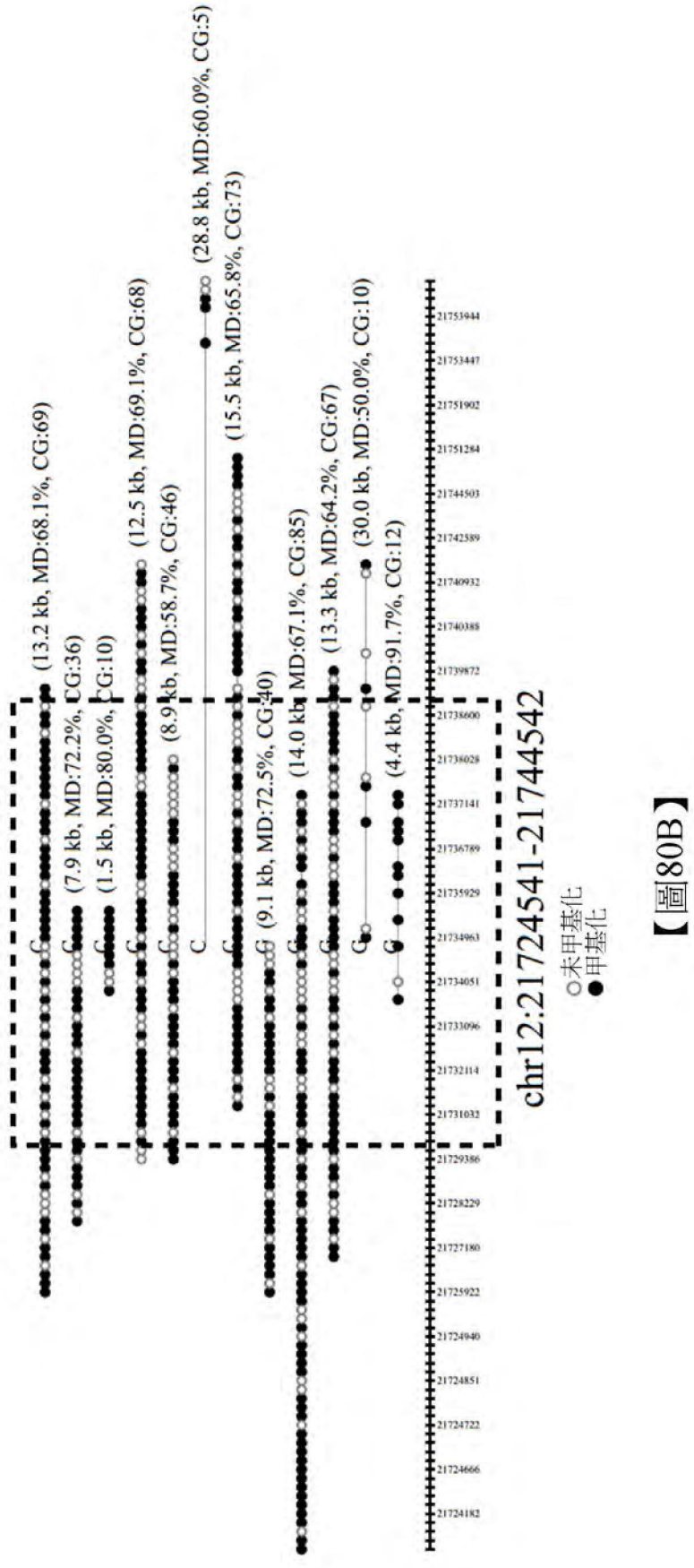
【圖79C】


(11.5 kb, MD:23.1%, CG:221)
C(18.3 kb, MD:33.3%, CG:6)
C(9.9 kb, MD:28.2%, CG:241)
(7.5 kb, MD:31.7%, CG:142
C(10.3 kb, MD:0.0%, CG:3)
C(4.1 kb, MD:7.3%, CG:205)
C(3.8 kb, MD:11.0%, CG:200)
T(8.9 kb, MD:72.0%, CG:100)
T(13.5 kb, MD:65.2%, CG:46)
T(13.4 kb, MD:83.8%, CG:297)
T(16.1 kb, MD:77.7%, CG:309)
T(8.1 kb, MD:87.0%, CG:54)
chr19 ZIM2
○未甲基化
●甲基化

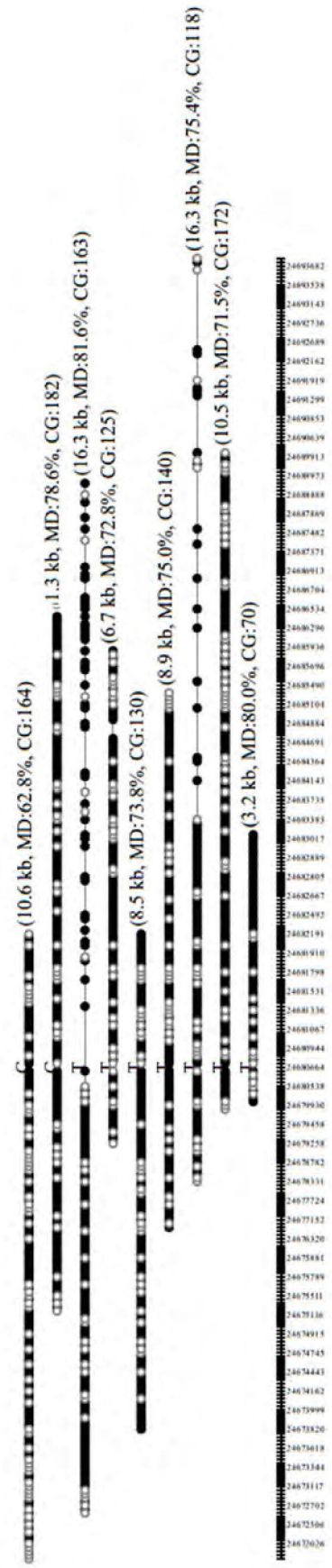
【圖79D】


(14.2 kb, MD:60.0%, CG:40)
(7.8 kb, MD:70.0%, CG:70)
(19.4 kb, MD:77.2%, CG:158)
(16.1 kb, MD:64.0%, CG:136)
(16.1 kb, MD:73.7%, CG:118)
(10.8 kb, MD:70.0%, CG:90)
chr7:95104111-95124112
○未甲基化
●甲基化

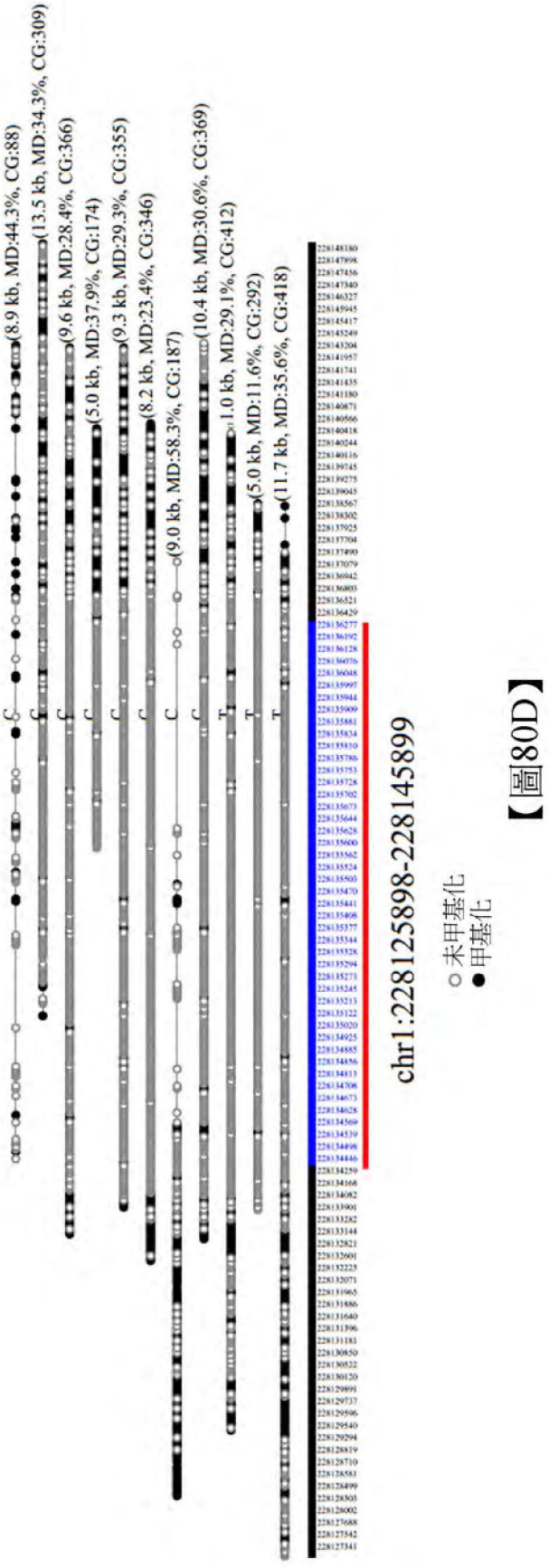
【圖80A】


(13.2 kb, MD:68.1%, CG:69)
(7.9 kb, MD:72.2%, CG:36)
(1.5 kb, MD:80.0%, CG:10)
(12.5 kb, MD:69.1%, CG:68)
(8.9 kb, MD:58.7%, CG:46)
(28.8 kb, MD:60.0%, CG:5)
(15.5 kb, MD:65.8%, CG:73)
(9.1 kb, MD:72.5%, CG:40)
(14.0 kb, MD:67.1%, CG:85)
(13.3 kb, MD:64.2%, CG:67)
(30.0 kb, MD:50.0%, CG:10)
(4.4 kb, MD:91.7%, CG:12)
chr12:21724541-21744542
○未甲基化
●甲基化

【圖80B】


(10.6 kb, MD:62.8%, CG:164)
(1.3 kb, MD:78.6%, CG:182)
(16.3 kb, MD:81.6%, CG:163)
(6.7 kb, MD:72.8%, CG:125)
(8.5 kb, MD:73.8%, CG:130)
(8.9 kb, MD:75.0%, CG:140)
(16.3 kb, MD:75.4%, CG:118)
(10.5 kb, MD:71.5%, CG:172)
(3.2 kb, MD:80.0%, CG:70)
chr1:24670664-24690665
未甲基化
甲基化

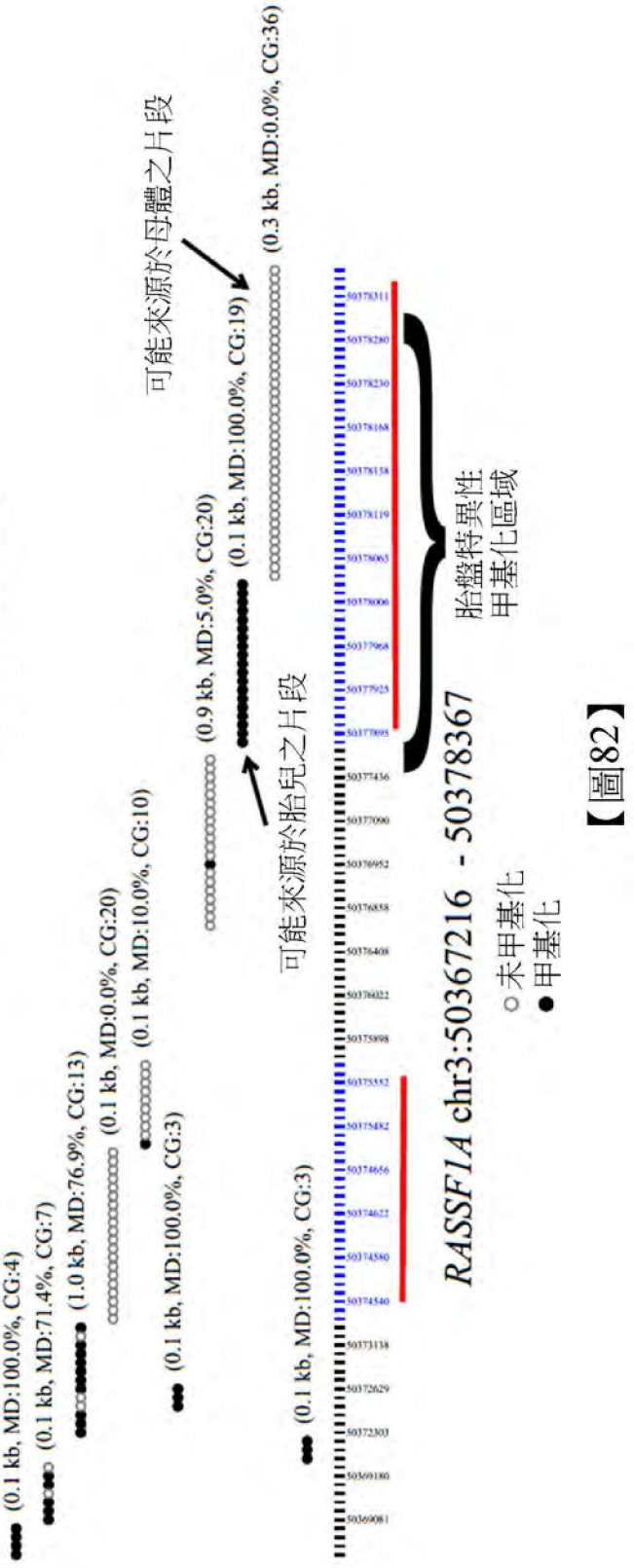
【圖80C】


(8.9 kb, MD:44.3%, CG:88)
(13.5 kb, MD:34.3%, CG:309)
(9.6 kb, MD:28.4%, CG:366)
(5.0 kb, MD:37.9%, CG:174)
(9.3 kb, MD:29.3%, CG:355)
(8.2 kb, MD:23.4%, CG:346)
(9.0 kb, MD:58.3%, CG:187)
(10.4 kb, MD:30.6%, CG:369)
1.0 kb, MD:29.1%, CG:412)
(5.0 kb, MD:11.6%, CG:292)
(11.7 kb, MD:35.6%, CG:418)
chr1:228125898-228145899
未甲基化
甲基化

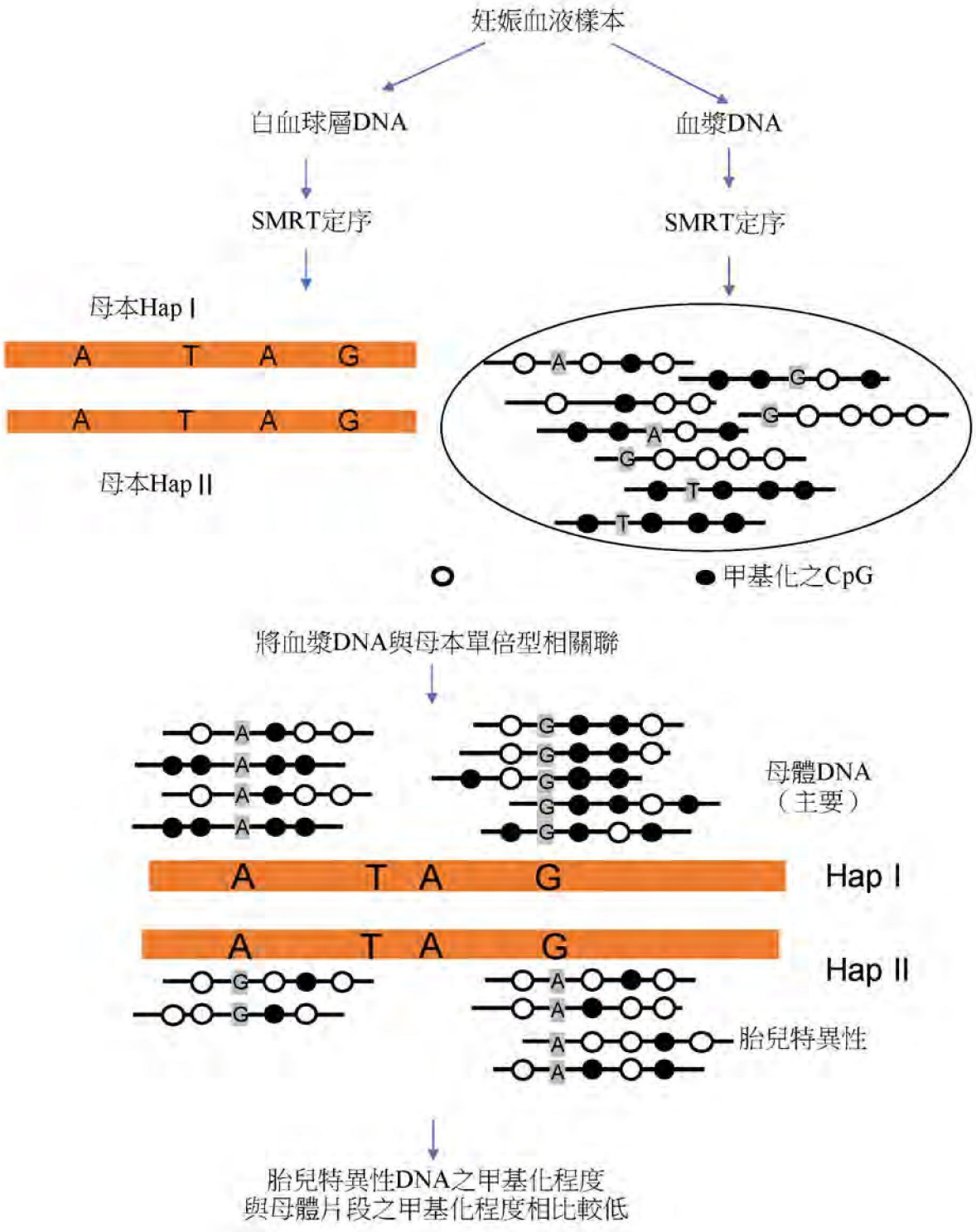
【圖80D】

	基因	對偶基因1	對偶基因2	甲基化程度（%）	
				對偶基因1	對偶基因2
印記基因	*SNURF*	T	C	15.73	89.37
	PLAGL1	T	C	7.56	89.41
	NAP1L5	C	T	12.5	91.07
	ZIM2	C	T	13	84.64
隨機選擇的區域	區域01	G	A	71.79	69.17
	區域02	T	G	63.22	65.22
	區域03	C	T	73.33	74.9
	區域04	C	T	10.83	8.51

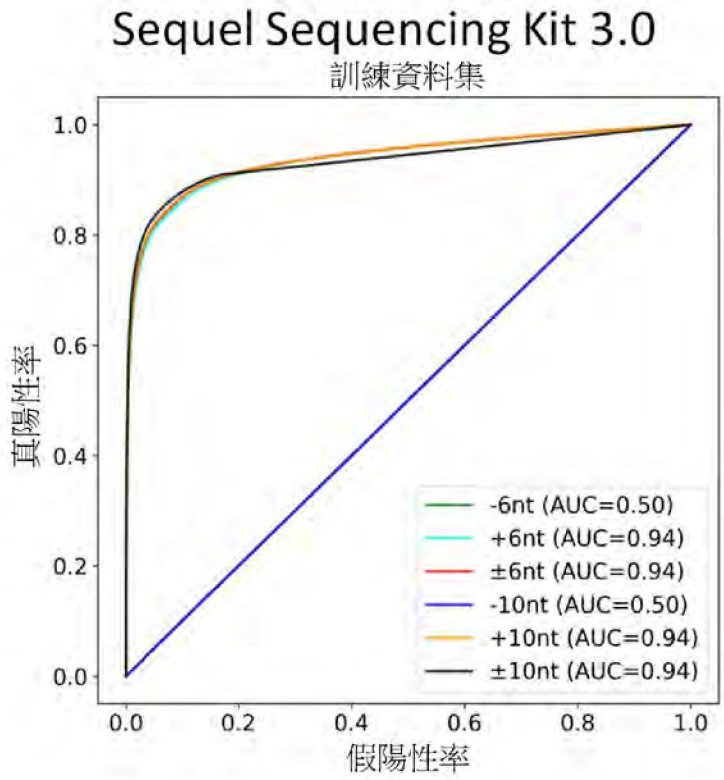
【圖81】


(0.7 kb, MD:20.0%, CG:15)
(0.1 kb, MD:0.0%, CG:2)
(0.1 kb, MD:100.0%, CG:4)
(0.1 kb, MD:71.4%, CG:7)
(1.0 kb, MD:76.9%, CG:13)
(0.1 kb, MD:0.0%, CG:20)
(0.1 kb, MD:10.0%, CG:10)
(0.1 kb, MD:100.0%, CG:3)
(0.9 kb, MD:5.0%, CG:20)
(0.1 kb, MD:100.0%, CG:19)
(0.3 kb, MD:0.0%, CG:36)
(0.1 kb, MD:100.0%, CG:3)
可能來源於母體之片段
可能來源於胎兒之片段
RASSF1A chr3:50367216 - 50378367
胎盤特異性甲基化區域
○未甲基化
●甲基化

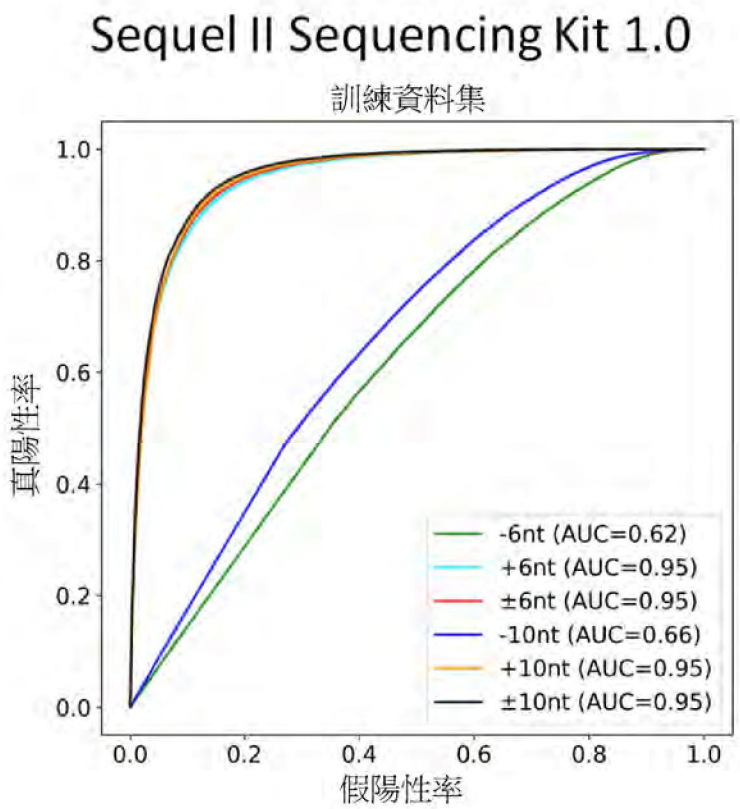
【圖82】


妊娠血液樣本
白血球層DNA
血漿DNA
SMRT定序
SMRT定序
母本Hap I
A T A G
A T A G
母本Hap II
甲基化之CpG
將血漿DNA與母本單倍型相關聯
母體DNA
（主要）
A T A G
Hap I
A T A G
Hap II
胎兒特異性
胎兒特異性DNA之甲基化程度
與母體片段之甲基化程度相比較低

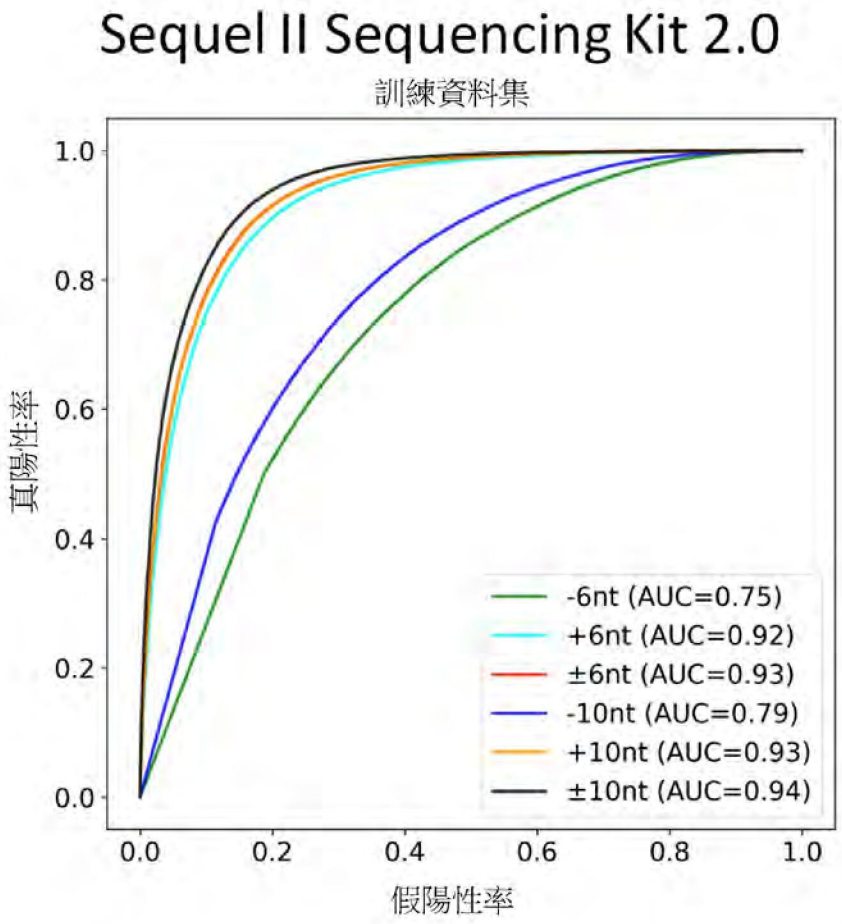
【圖83】


Sequel Sequencing Kit 3.0
訓練資料集
真陽性率
假陽性率
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
-6nt (AUC=0.50)
+6nt (AUC=0.94)
±6nt (AUC=0.94)
-10nt (AUC=0.50)
+10nt (AUC=0.94)
±10nt (AUC=0.94)

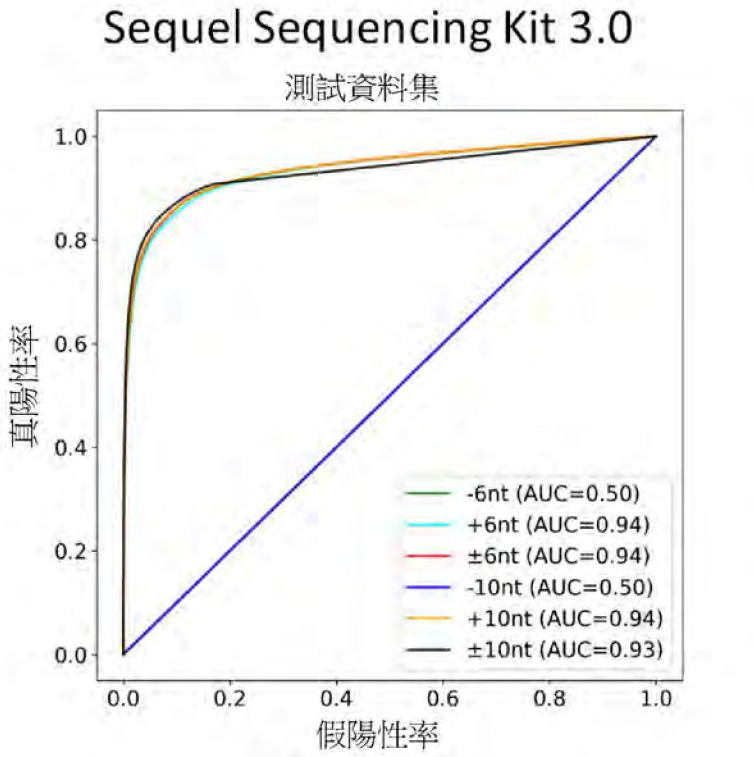
【圖84A】


Sequel II Sequencing Kit 1.0
訓練資料集
真陽性率
假陽性率
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
-6nt (AUC=0.62)
+6nt (AUC=0.95)
±6nt (AUC=0.95)
-10nt (AUC=0.66)
+10nt (AUC=0.95)
±10nt (AUC=0.95)

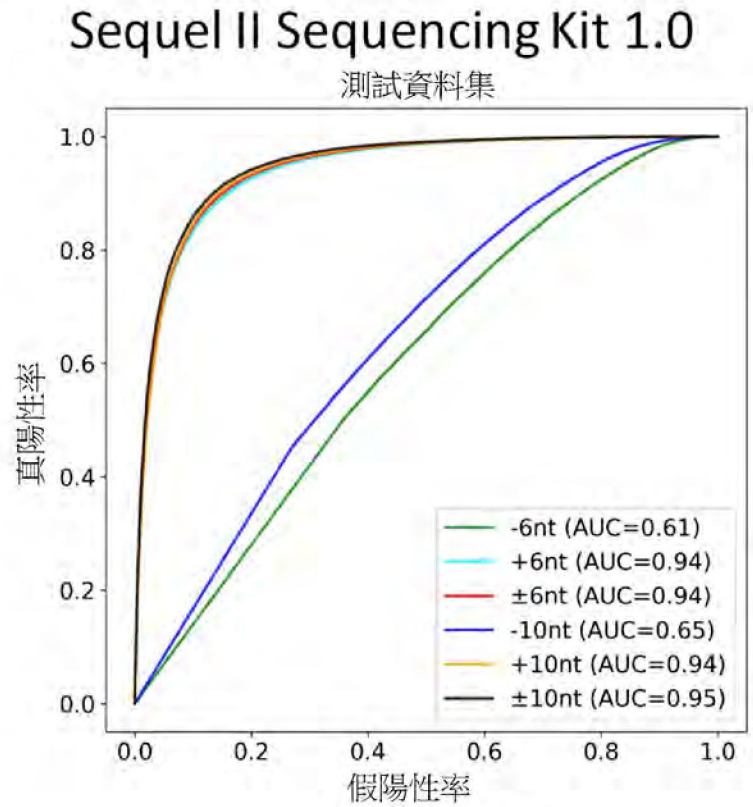
【圖84B】


Sequel II Sequencing Kit 2.0
訓練資料集
真陽性率
假陽性率
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
-6nt (AUC=0.75)
+6nt (AUC=0.92)
±6nt (AUC=0.93)
-10nt (AUC=0.79)
+10nt (AUC=0.93)
±10nt (AUC=0.94)

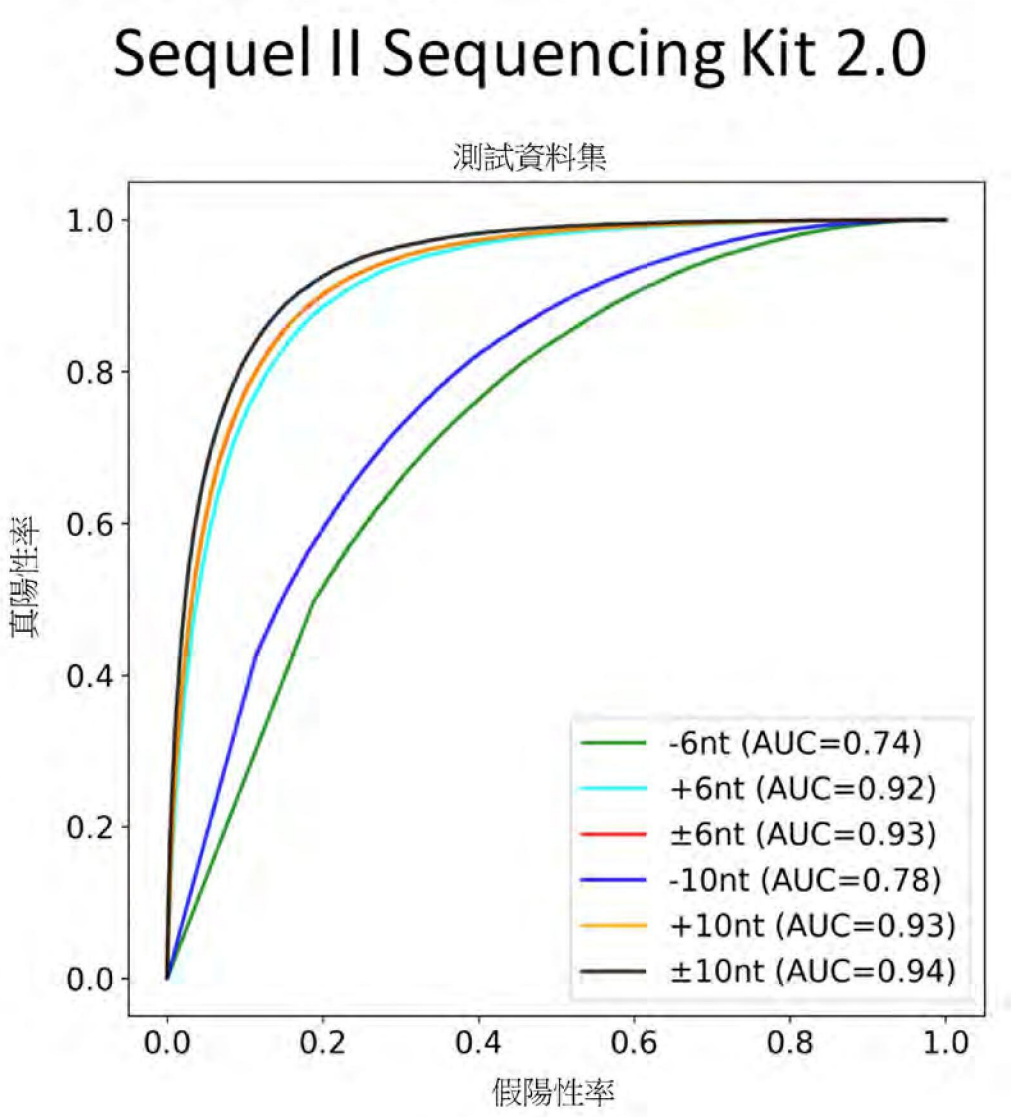
【圖84C】


Sequel Sequencing Kit 3.0
測試資料集
真陽性率
假陽性率
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
-6nt (AUC=0.50)
+6nt (AUC=0.94)
±6nt (AUC=0.94)
-10nt (AUC=0.50)
+10nt (AUC=0.94)
±10nt (AUC=0.93)

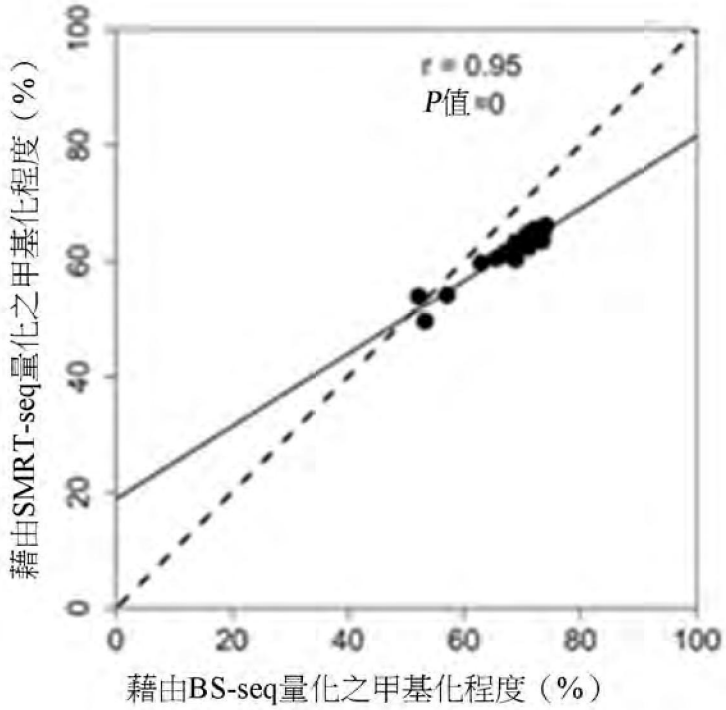
【圖85A】


Sequel II Sequencing Kit 1.0
測試資料集
真陽性率
假陽性率
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
-6nt (AUC=0.61)
+6nt (AUC=0.94)
±6nt (AUC=0.94)
-10nt (AUC=0.65)
+10nt (AUC=0.94)
±10nt (AUC=0.95)

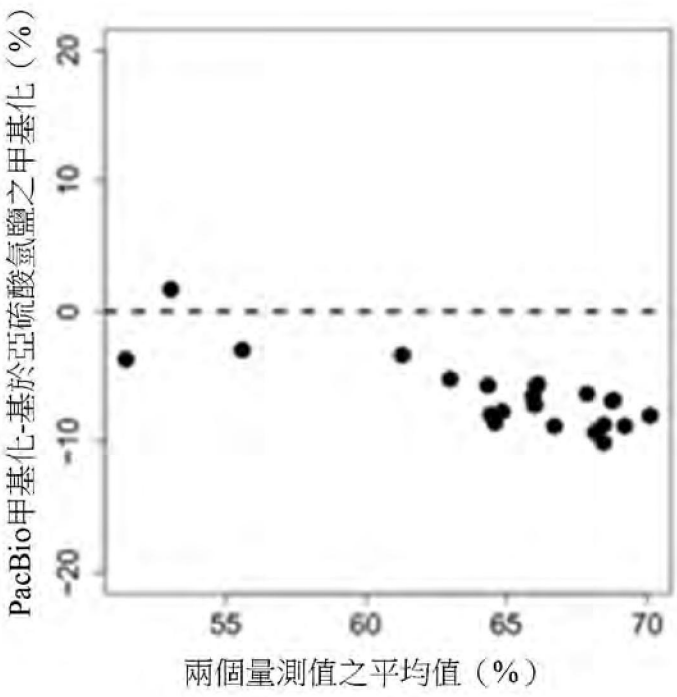
【圖85B】


Sequel II Sequencing Kit 2.0
測試資料集
真陽性率
假陽性率
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
-6nt (AUC=0.74)
+6nt (AUC=0.92)
±6nt (AUC=0.93)
-10nt (AUC=0.78)
+10nt (AUC=0.93)
±10nt (AUC=0.94)

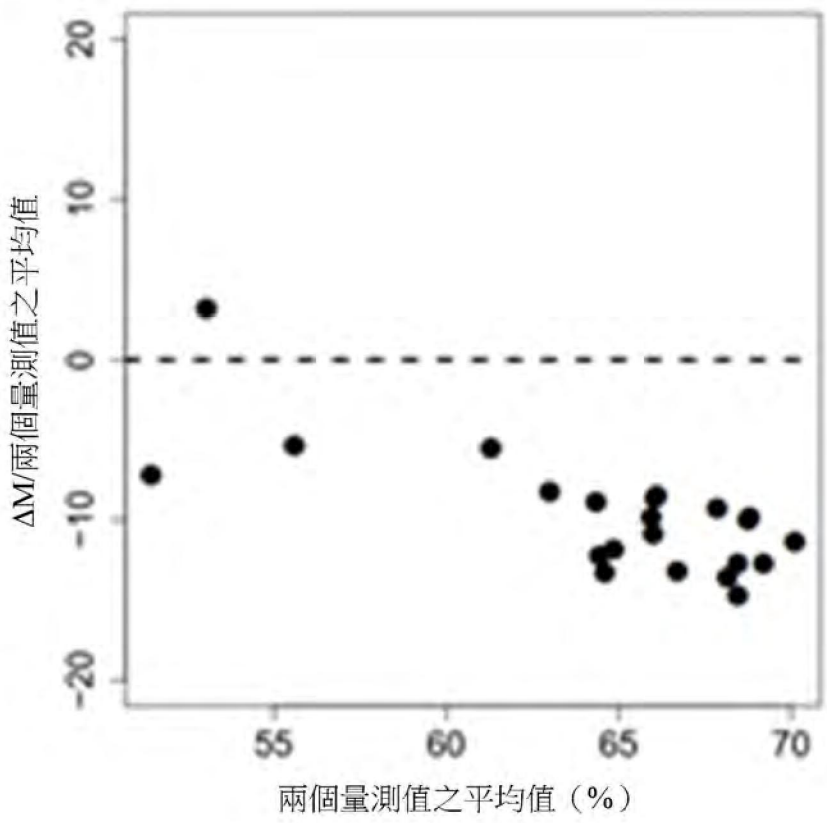
【圖85C】


r = 0.95
P值=0
藉由SMRT-seq量化之甲基化程度（%）
藉由BS-seq量化之甲基化程度（%）
0
20
40
60
80
100

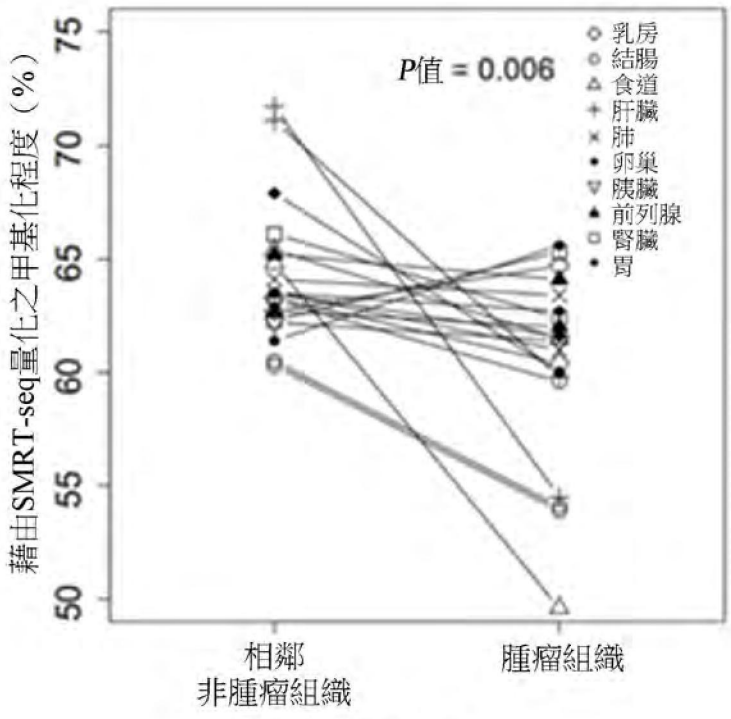
【圖86A】


PacBio甲基化-基於亞硫酸氫鹽之甲基化（%）
兩個量測值之平均值（%）
-20
-10
0
10
20
55
60
65
70

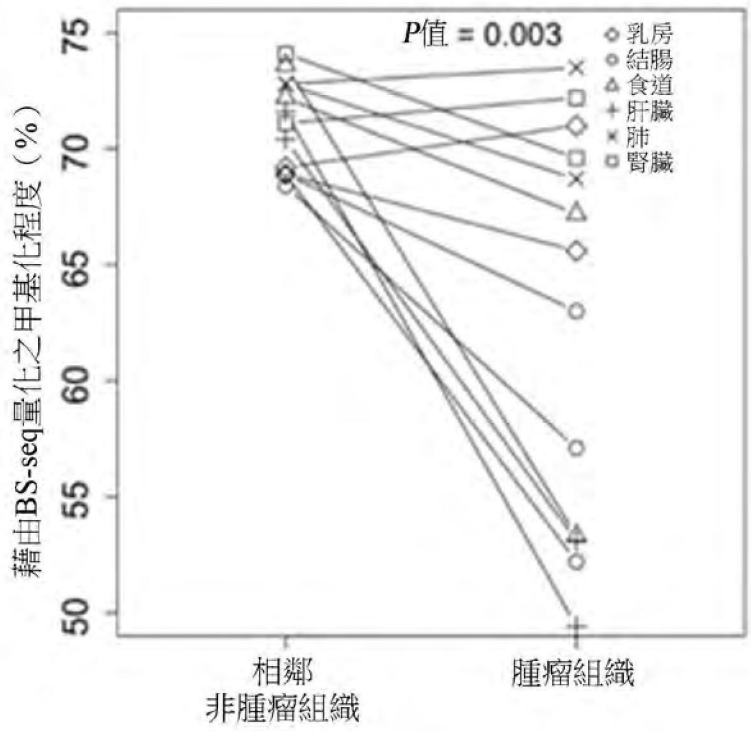
【圖86B】



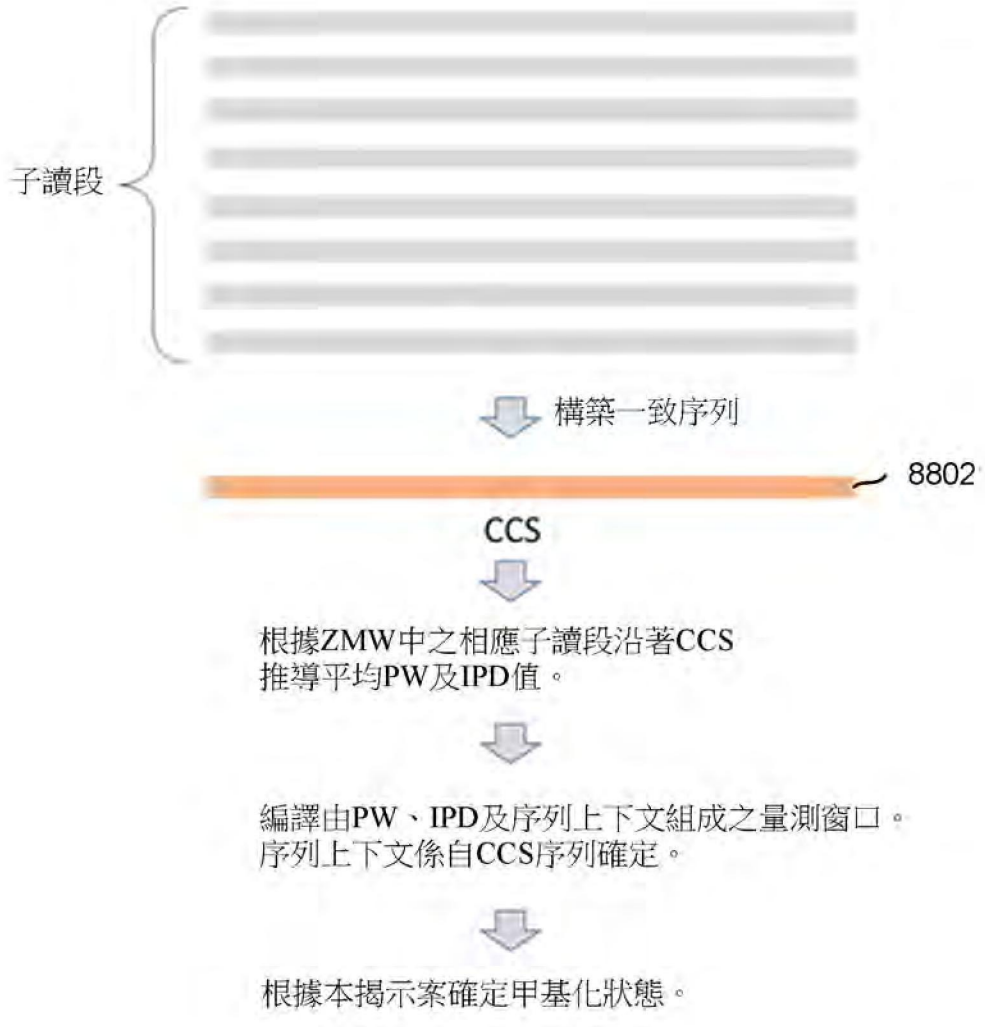
【圖86C】


藉由SMRT-seq量化之甲基化程度（%）
50
55
60
65
70
75
P值 = 0.006
乳房
結腸
食道
肝臟
肺
卵巢
胰臟
前列腺
腎臟
胃
相鄰
非腫瘤組織
腫瘤組織

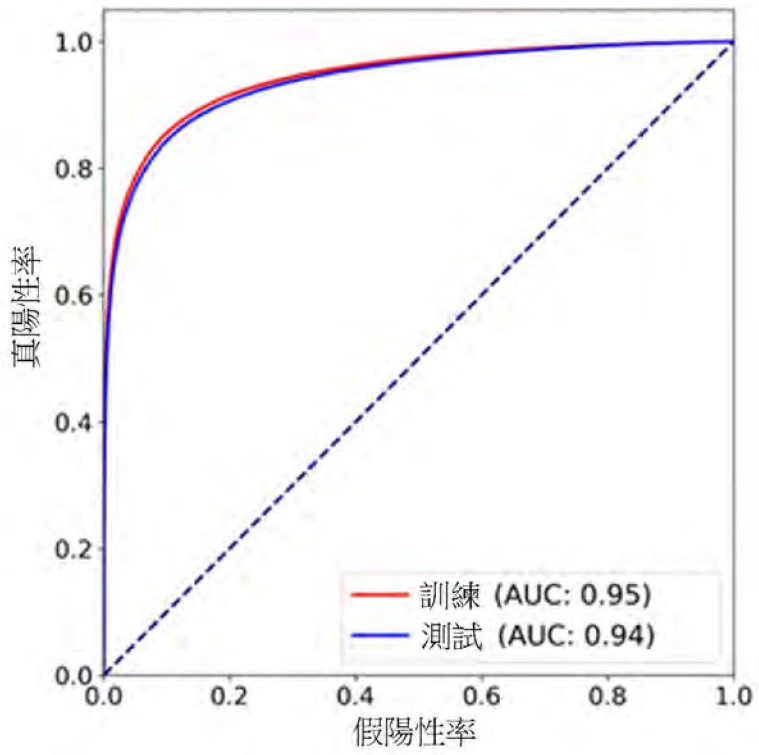
【圖87A】


藉由BS-seq量化之甲基化程度（%）
50
55
60
65
70
75
P值 = 0.003
乳房
結腸
食道
肝臟
肺
腎臟
相鄰
非腫瘤組織
腫瘤組織

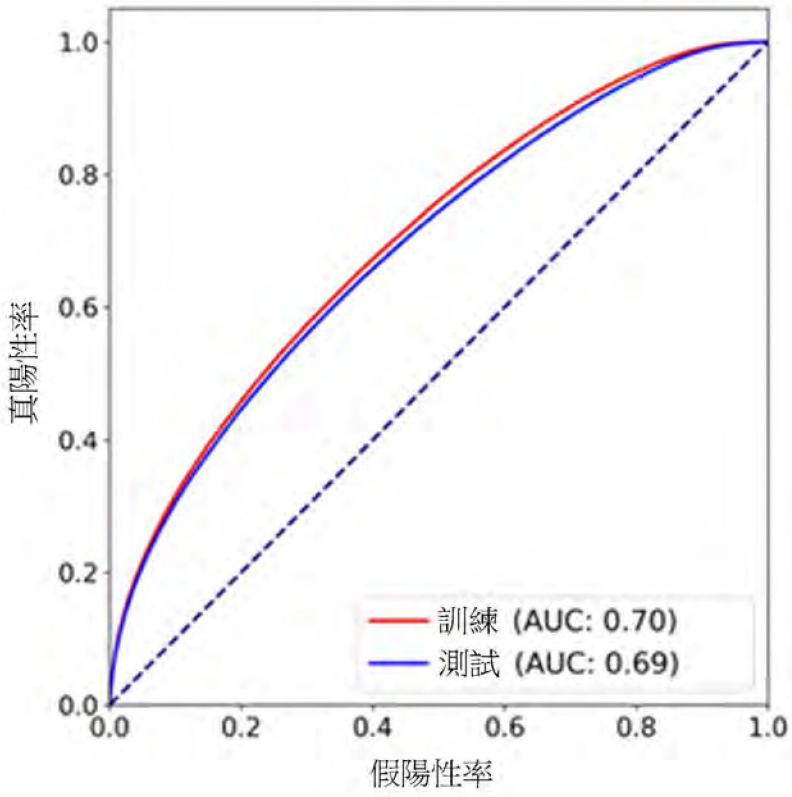
【圖87B】


子讀段
構築一致序列
8802
CCS
根據ZMW中之相應子讀段沿著CCS
推導平均PW及IPD值。
編譯由PW、IPD及序列上下文組成之量測窗口。
序列上下文係自CCS序列確定。
根據本揭示案確定甲基化狀態。

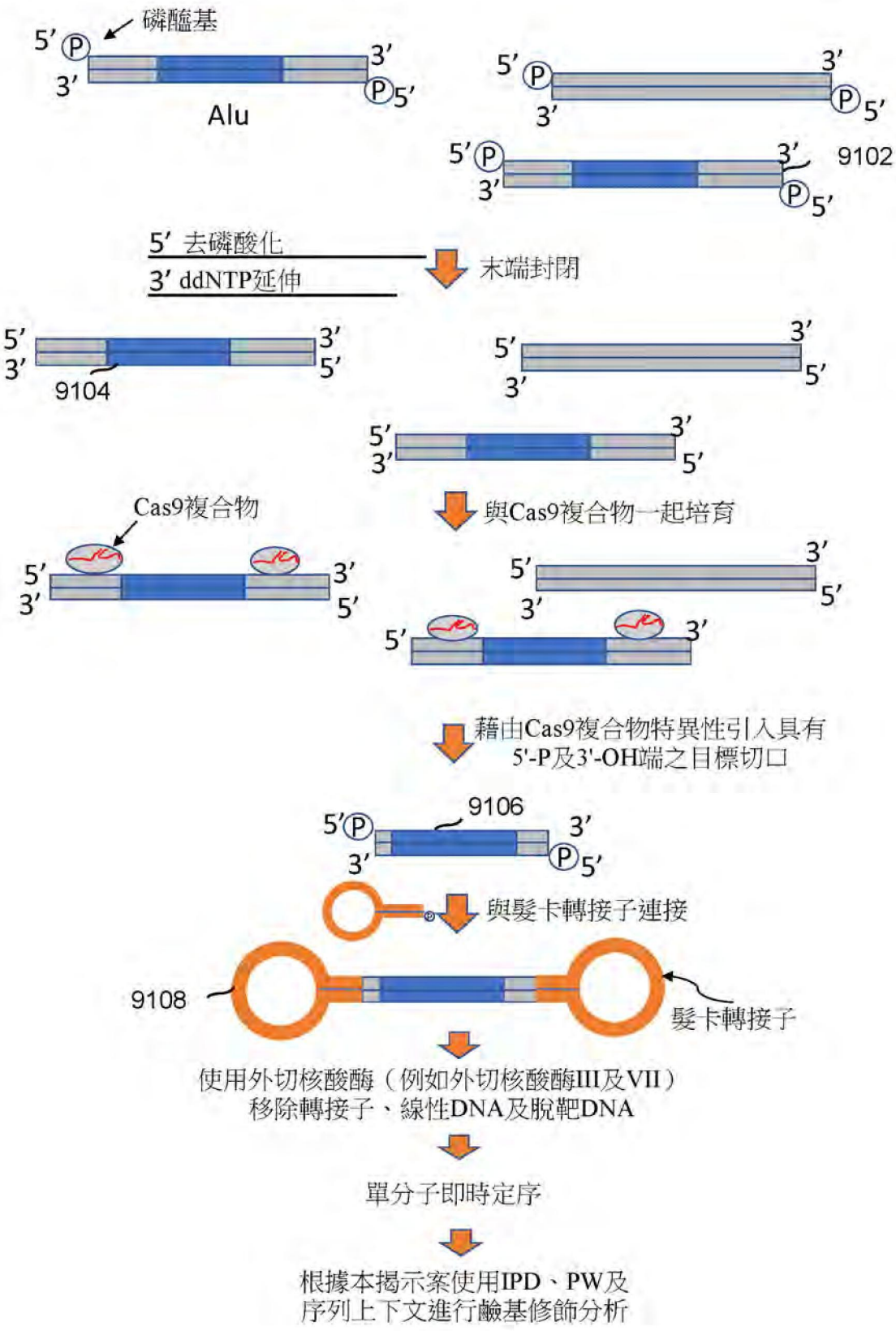
【圖88】


1.0
0.8
0.6
0.4
0.2
0.0
真陽性率
0.0
0.2
0.4
0.6
0.8
1.0
假陽性率
訓練 (AUC: 0.95)
測試 (AUC: 0.94)

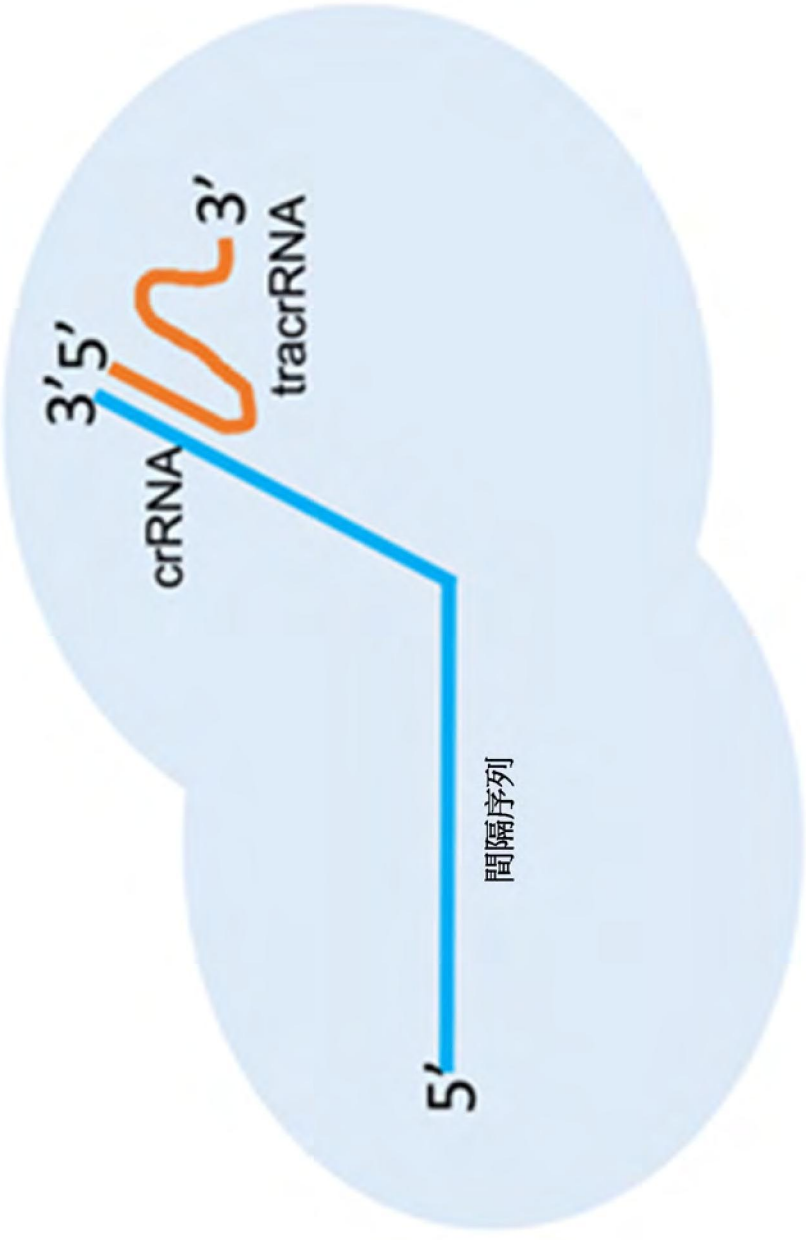
【圖89】


1.0
0.8
0.6
0.4
0.2
0.0
真陽性率
0.0
0.2
0.4
0.6
0.8
1.0
假陽性率
訓練 (AUC: 0.70)
測試 (AUC: 0.69)

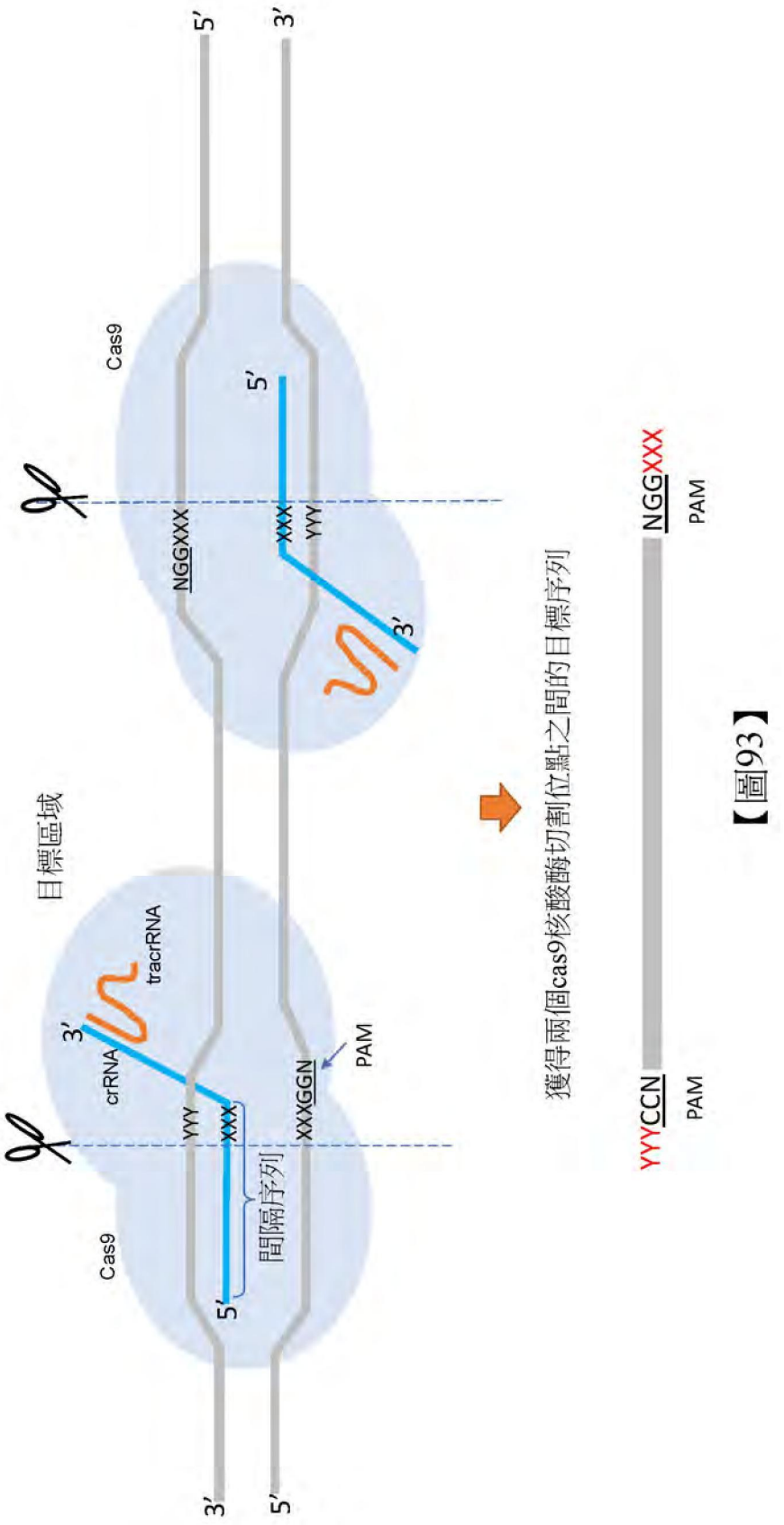
【圖90】


磷醯基
5′
3′
Alu
9102
5′ 去磷酸化
3′ ddNTP延伸
末端封閉
9104
Cas9複合物
與Cas9複合物一起培育
藉由Cas9複合物特異性引入具有
5'-P及3'-OH端之目標切口
9106
與髮卡轉接子連接
9108
髮卡轉接子
使用外切核酸酶（例如外切核酸酶III及VII）
移除轉接子、線性DNA及脫靶DNA
單分子即時定序
根據本揭示案使用IPD、PW及
序列上下文進行鹼基修飾分析

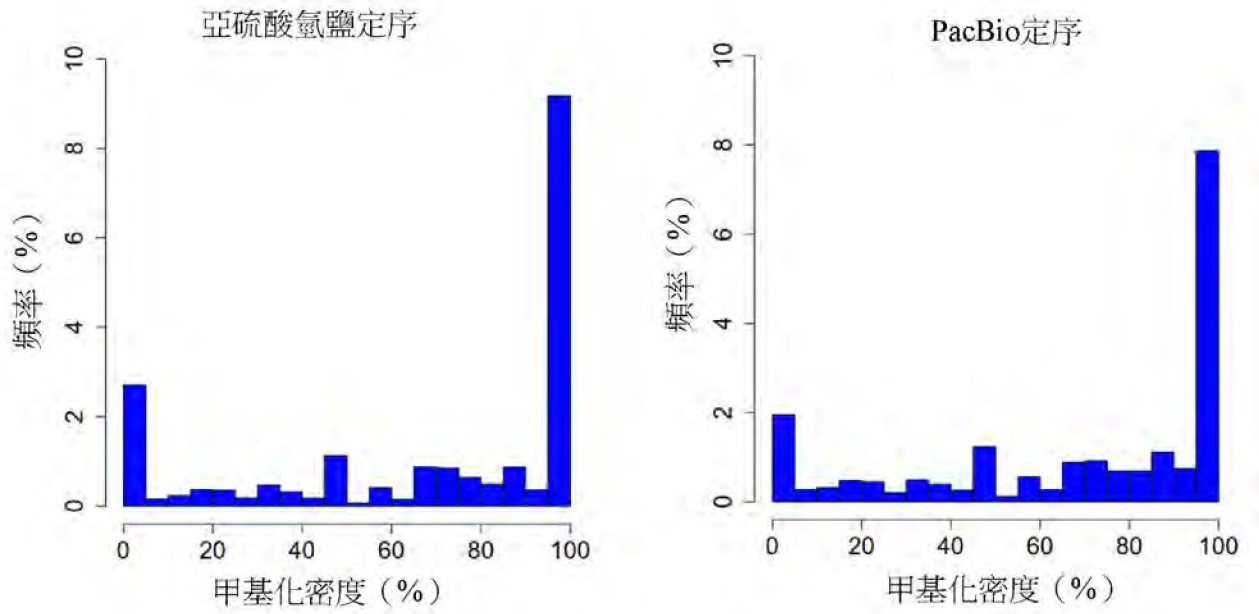
【圖91】



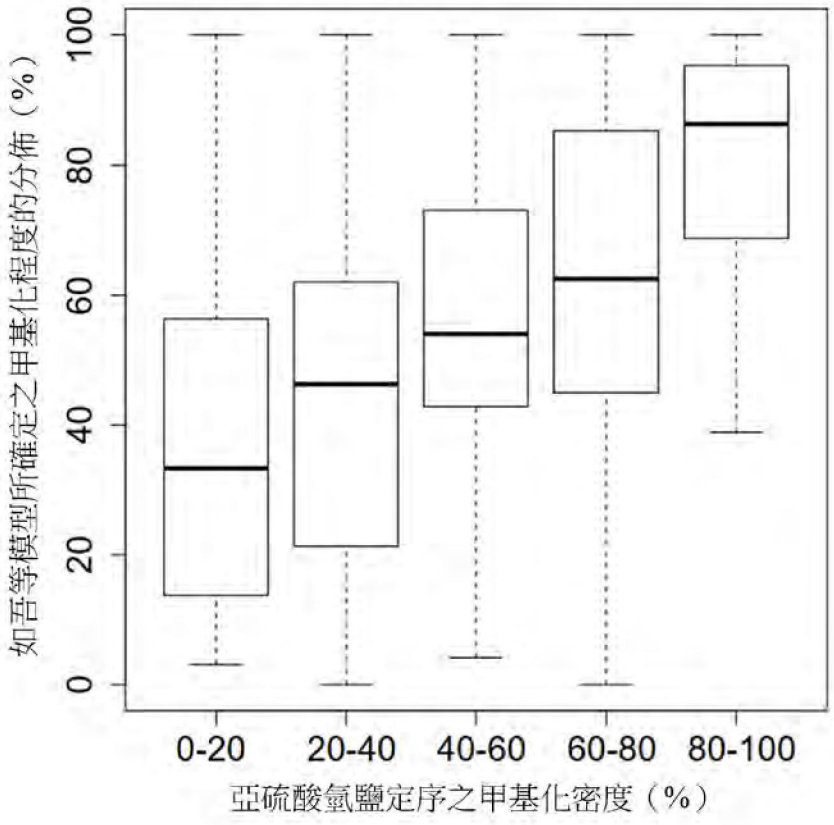
【圖92】


5′
3′
Cas9
NGGXXX
5′
XXX
YYY
3′
目標區域
3′
crRNA
tracrRNA
YYY
XXX
XXXGGN
PAM
Cas9
間隔序列
5′
3′
5′
獲得兩個cas9核酸酶切割位點之間的目標序列
NGGXXX
PAM
YYYCCN
PAM

【圖93】


亞硫酸氫鹽定序
PacBio定序
頻率（%）
甲基化密度（%）
0
2
4
6
8
10
0
20
40
60
80
100

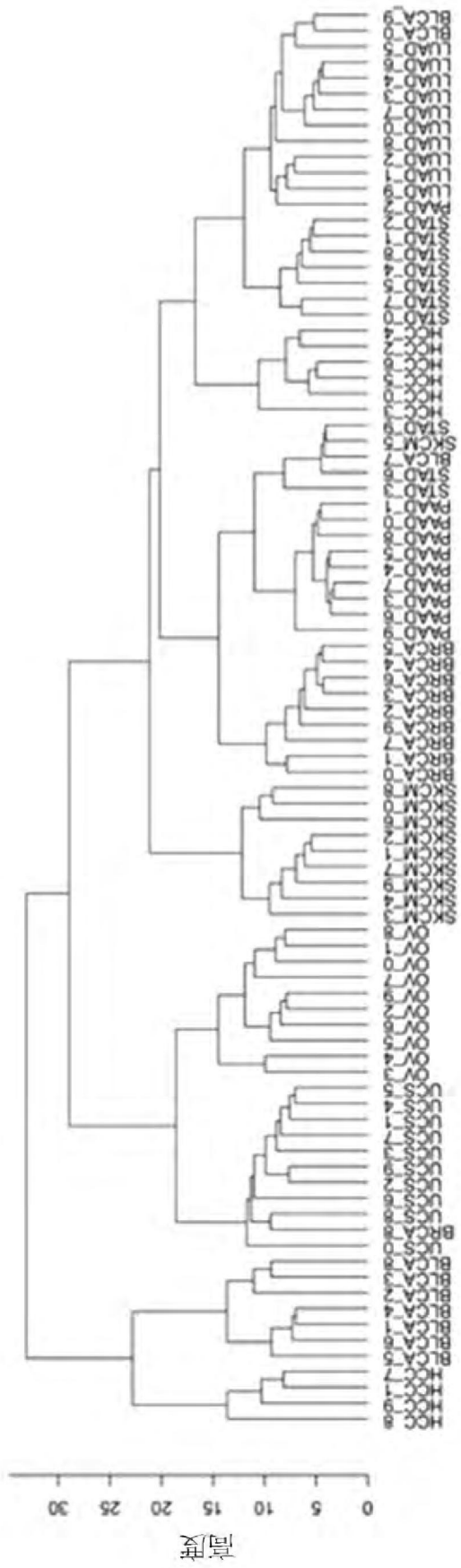
【圖94】


如吾等模型所確定之甲基化程度的分佈（%）
亞硫酸氫鹽定序之甲基化密度（%）
0
20
40
60
80
100
0-20
20-40
40-60
60-80
80-100

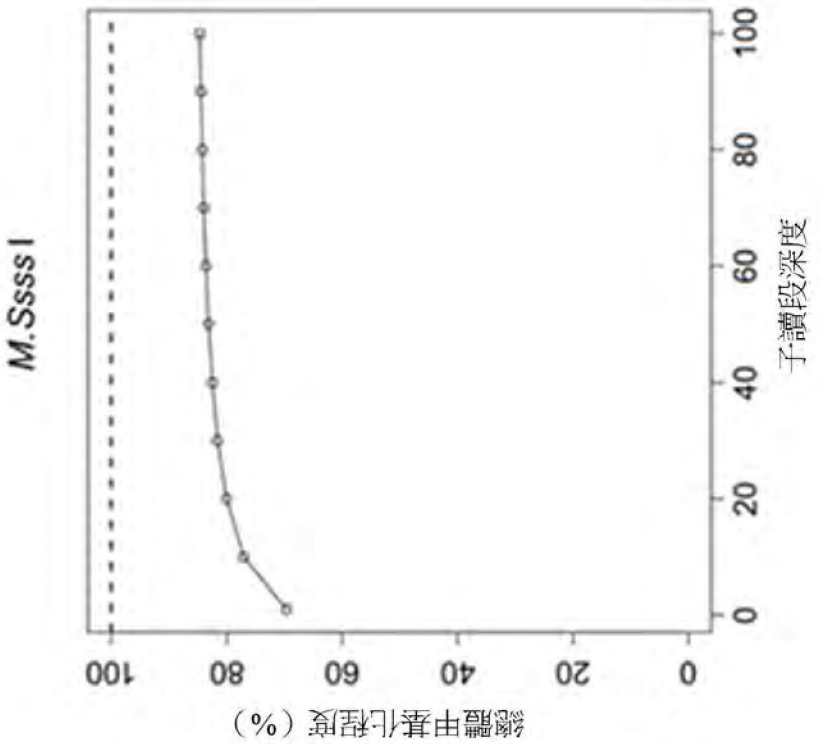
【圖95】

組織	Alu之甲基化程度（%）
白血球層	89.54
肝臟	88.18
結腸	89.56
肺	91.52
小腸	86.56
腎上腺	89.07
脂肪	91.44
胰臟	85.82
腦	91.79
HCC	76.74
胎盤	73.04

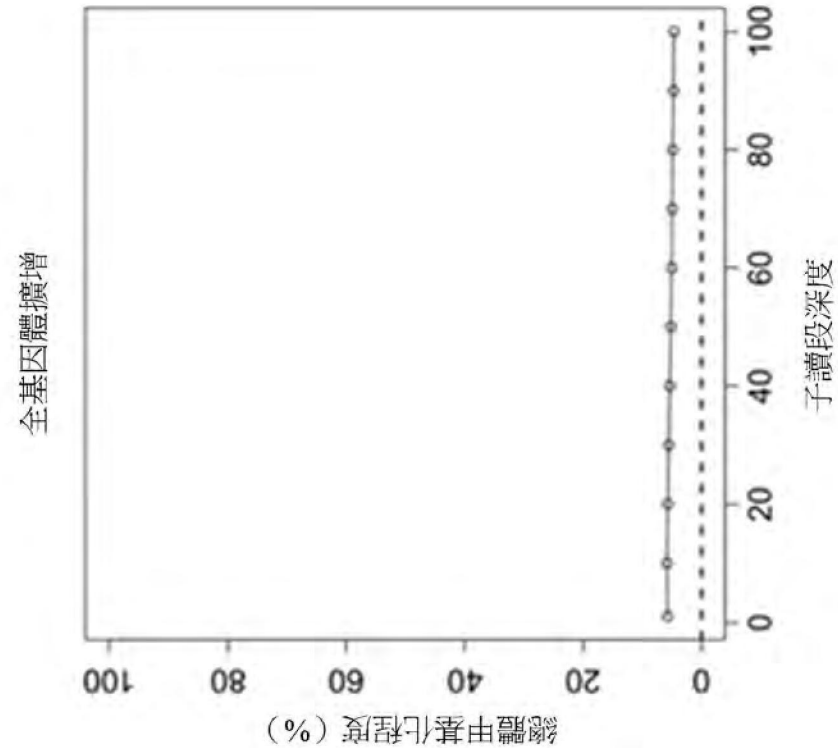
【圖96】


高度
0
5
10
15
20
25
30
癌症類型
BLCA: 膀胱尿道上皮癌
BRCA: 乳房侵襲性癌
OV: 卵巢漿液性囊腺癌
PAAD: 胰臟腺癌
HCC: 肝細胞癌
LUAD: 肺腺癌
STAD: 胃腺癌
SKCM: 皮膚黑素瘤
UCS: 子宮癌肉瘤

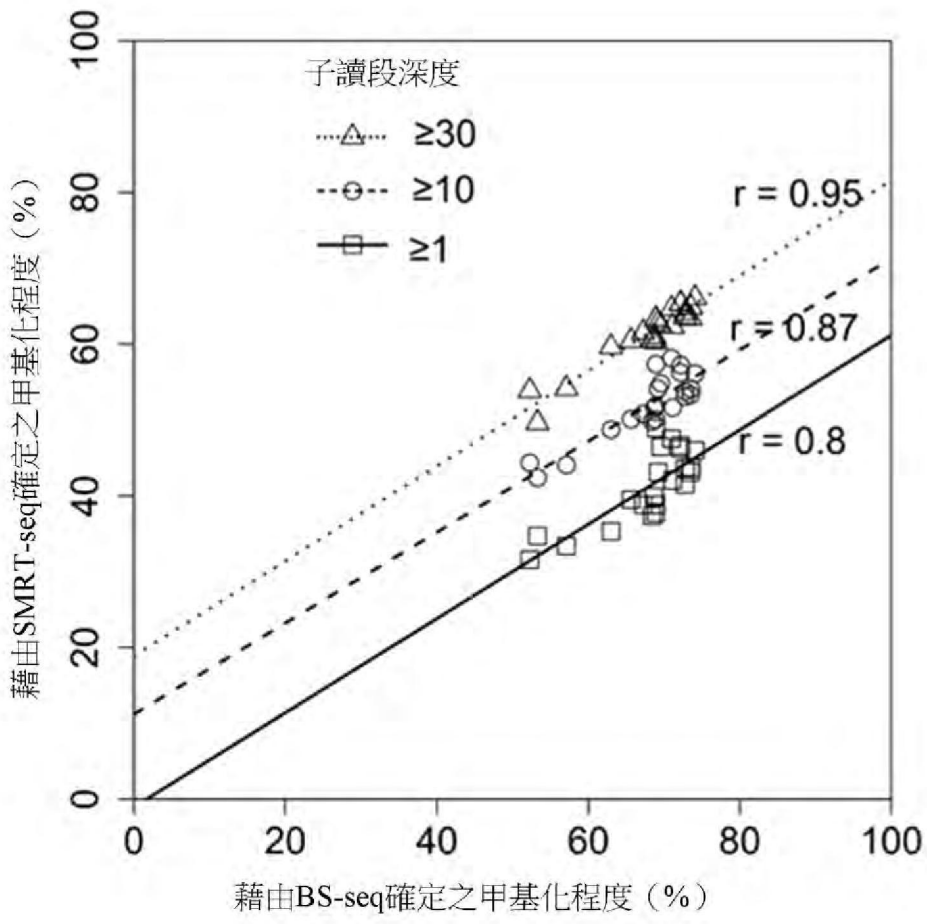
【圖97】


M.Ssss I
總體甲基化程度（%）
0
20
40
60
80
100
子讀段深度
0
20
40
60
80
100

【圖98B】


全基因體擴增
總體甲基化程度（%）
0
20
40
60
80
100
子讀段深度
0
20
40
60
80
100

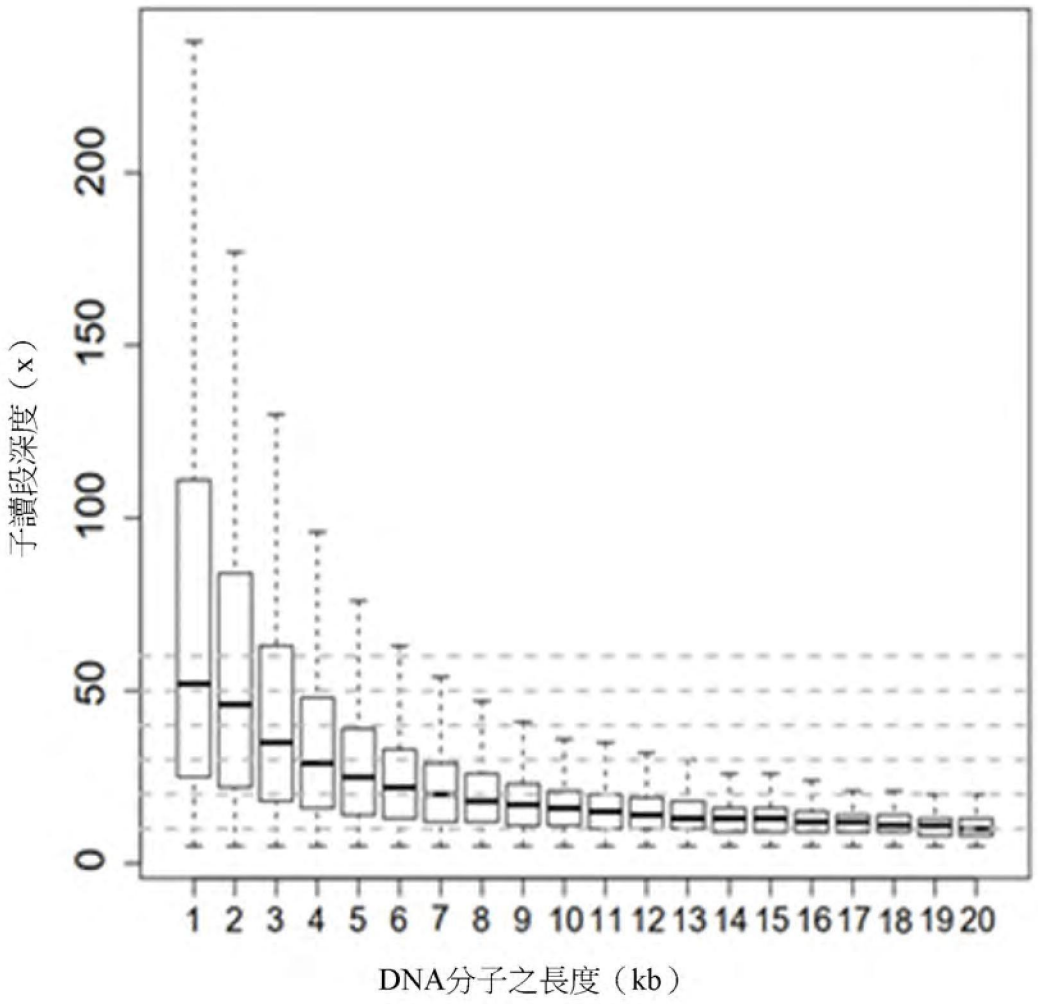
【圖98A】


子讀段深度
≥30
≥10
≥1
r = 0.95
r = 0.87
r = 0.8
藉由SMRT-seq確定之甲基化程度（%）
藉由BS-seq確定之甲基化程度（%）
0
20
40
60
80
100

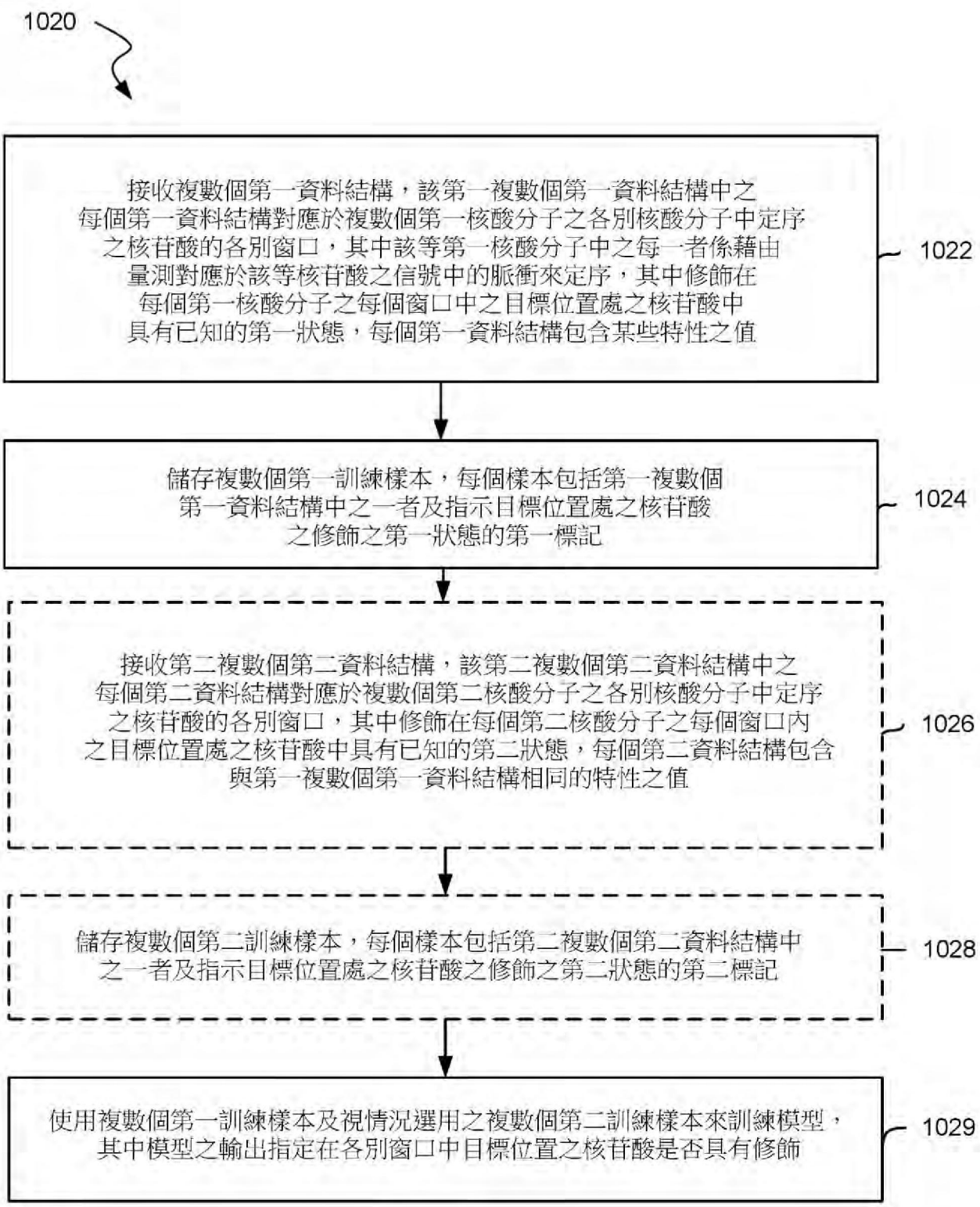
【圖99】

子讀段深度 截止值≥	Pearson's r (SMRT-seq vs BS-seq)	CpG位點之數量
1	0.797	25,606,068 (23,949,832-27,008,582)
10	0.873	21,668,418 (18,263,886-23,515,147)
20	0.933	14,276,212 (10,526,406-16,736,887)
30	0.952	6,736,890 (4,255,452-10,449,814)
40	0.948	3,420,790 (2,232,511-5,792,825)
50	0.941	1,684,871 (1,278,475-3,055,876)
60	0.929	911,961 (707,295-1,581,313)
70	0.917	532,422 (350,001-866,045)
80	0.907	284,375 (177,698-534,540)
90	0.906	150,974 (98,000-333,933)
100	0.875	89,788 (58,552-182,861)

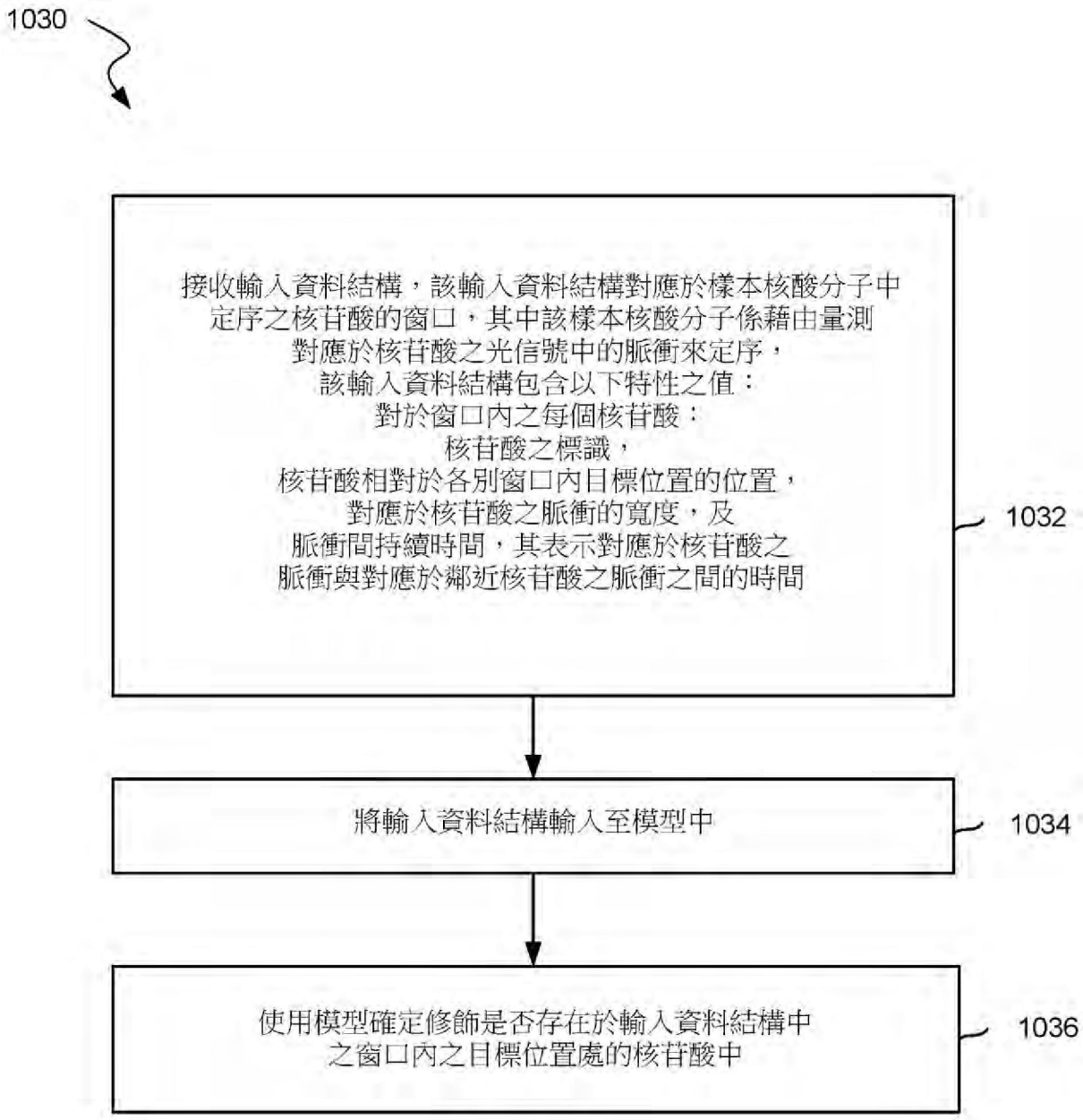
【圖100】


子讀段深度（x）
0
50
100
150
200
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
DNA分子之長度（kb）

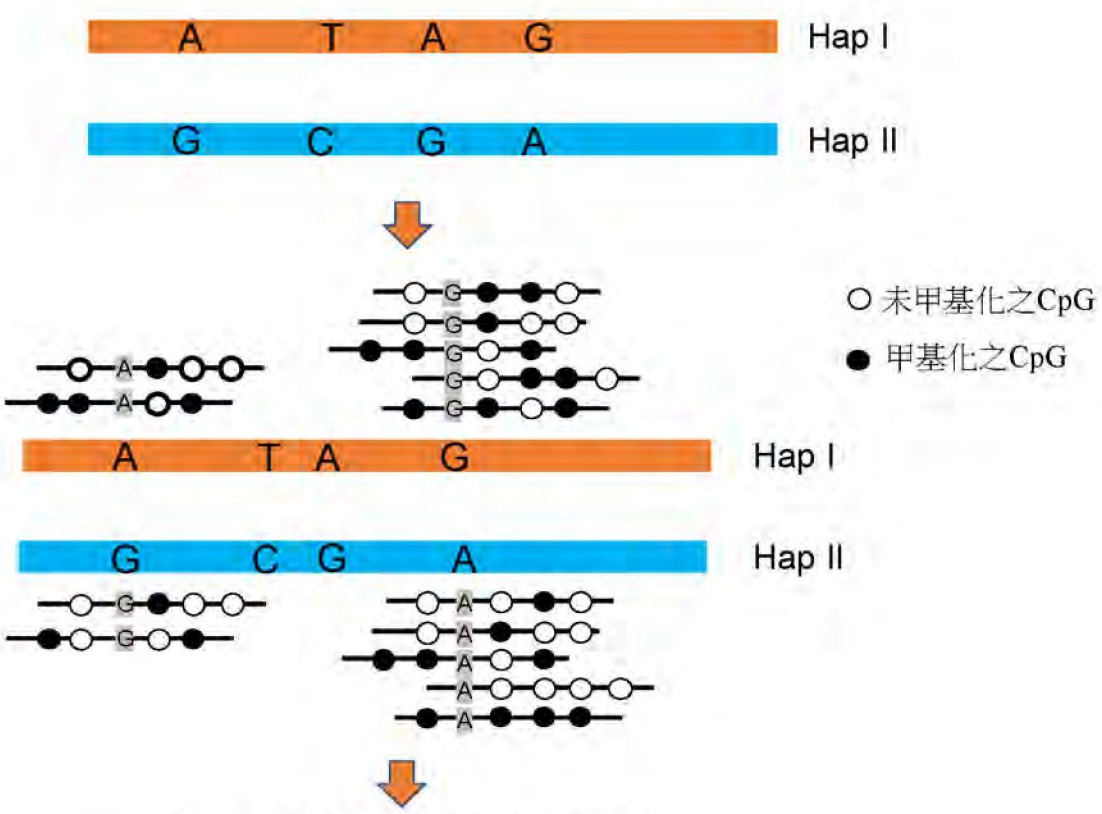
【圖101】


1020
接收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別核酸分子中定序之核苷酸的各別窗口，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信號中的脈衝來定序，其中修飾在每個第一核酸分子之每個窗口中之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含某些特性之值
1022
儲存複數個第一訓練樣本，每個樣本包括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之第一狀態的第一標記
1024
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別核酸分子中定序之核苷酸的各別窗口，其中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含與第一複數個第一資料結構相同的特性之值
1026
儲存複數個第二訓練樣本，每個樣本包括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之第二狀態的第二標記
1028
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來訓練模型，其中模型之輸出指定在各別窗口中目標位置之核苷酸是否具有修飾
1029

【圖102】


1030
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定序之核苷酸的窗口，其中該樣本核酸分子係藉由量測對應於核苷酸之光信號中的脈衝來定序，該輸入資料結構包含以下特性之值：對於窗口內之每個核苷酸：核苷酸之標識，核苷酸相對於各別窗口內目標位置的位置，對應於核苷酸之脈衝的寬度，及脈衝間持續時間，其表示對應於核苷酸之脈衝與對應於鄰近核苷酸之脈衝之間的時間
1032
將輸入資料結構輸入至模型中
1034
使用模型確定修飾是否存在於輸入資料結構中之窗口內之目標位置處的核苷酸中
1036

【圖103】



基於單倍型之相對甲基化不平衡指數

例如Hap I之甲基化程度—Hap II之甲基化程度

【圖104】

Chr	起始	結束	長度	單倍型區塊id	PacBio定序			
					相鄰非腫瘤組織中之甲基化程度		腫瘤組織中之甲基化程度	
					Hap I	Hap II	Hap I	Hap II
chr1	56312395	56347696	35301	hap1927	68.2	67.4	60.3	23.5
chr1	194413819	194424806	10987	hap5953	52.8	49.5	48.8	9.3
chr1	220674478	220699011	24533	hap6863	63.0	64.5	50.4	17.3
chr10	113088792	113124248	35456	hap11838	62.7	63.4	38.1	5.7
chr11	5482746	5498801	16055	hap12904	70.3	75.0	16.3	51.7
chr11	42819351	42852772	33421	hap14385	54.6	54.9	65.3	17.8
chr11	57983961	58051078	67117	hap14930	67.3	66.4	58.2	18.6
chr11	60174708	60204209	29501	hap14990	58.4	59.8	49.6	10.8
chr12	128079419	128114656	35237	hap22249	60.0	58.3	12.1	45.2
chr15	20480575	20533464	52889	hap29631	64.7	69.1	27.7	59.3
chr15	94902853	94946231	43378	hap32161	74.1	74.5	74.9	15.8
chr15	96526684	96549225	22541	hap32221	70.8	68.8	28.9	64.4
chr16	31595372	31613277	17905	hap33499	55.9	59.3	46.3	14.4
chr16	80151778	80182097	30319	hap34821	71.1	71.0	11.5	51.8
chr16	82519715	82554191	34476	hap34920	71.3	66.5	47.4	13.0
chr17	21668593	21685572	16979	hap36049	50.3	47.8	67.4	19.6
chr17	44999177	45012087	12910	hap36640	47.1	45.2	81.6	35.1
chr17	69911623	69926625	15002	hap37435	67.3	63.0	37.8	5.2
chr18	11441122	11458521	17399	hap38335	65.5	66.8	65.9	22.4
chr18	23405569	23423387	17818	hap38673	66.3	61.7	3.3	48.1
chr18	68887284	68925031	37747	hap40390	63.0	61.0	22.0	53.4
chr18	69487809	69505470	17661	hap40414	74.5	74.1	33.3	72.2
chr2	41480394	41514135	33741	hap43972	54.0	54.0	14.9	77.8
chr2	114171214	114182880	11666	hap46226	72.4	68.8	79.7	16.7
chr2	123762541	123797629	35088	hap46589	66.7	68.1	24.0	54.5
chr2	125236882	125241950	5068	hap46673	58.9	59.2	10.7	46.4
chr2	130016110	130040331	24221	hap46835	54.6	50.8	5.6	41.6
chr2	137757638	137783716	26078	hap47090	61.8	61.4	13.5	69.2
chr2	144128597	144160845	32248	hap47343	65.8	66.6	9.3	50.3
chr20	15736792	15753459	16667	hap51505	78.9	74.3	45.8	77.3
chr20	26167979	26177235	9256	hap51868	55.0	52.2	38.5	68.6
chr20	44255808	44264190	8382	hap52246	57.4	56.1	9.7	50.6
chr20	59518410	59559273	40863	hap52761	61.0	62.4	30.0	72.8
chr21	21402034	21424129	22095	hap53197	63.5	67.3	25.0	75.5
chr21	24750027	24768793	18766	hap53333	68.2	64.6	3.4	38.9
chr21	26666833	26701575	34742	hap53418	62.1	66.5	47.6	16.7
chr3	2364024	2387896	23872	hap55539	67.4	67.8	54.9	10.9
chr3	21036965	21049451	12486	hap56223	54.8	51.4	53.1	21.1
chr3	56011690	56046642	34952	hap57346	64.2	61.2	71.2	22.6

【圖105A】

chr3	73330942	73371216	40274	hap57939	60.9	62.9	9.4	42.9
chr3	106372440	106401301	28861	hap59077	67.8	67.9	13.8	53.2
chr3	107772994	107807482	34488	hap59122	69.6	73.5	30.4	66.4
chr3	116742501	116776747	34246	hap59493	64.3	69.1	14.1	51.6
chr3	171076306	171100102	23796	hap61495	68.0	66.0	80.6	48.8
chr3	193058272	193080344	22072	hap62231	65.5	64.7	54.6	20.0
chr4	30411613	30432317	20704	hap63589	59.3	60.6	53.4	14.6
chr4	31304718	31338193	33475	hap63633	60.2	60.0	7.2	55.0
chr4	92003467	92030505	27038	hap65794	65.3	65.1	54.1	21.7
chr4	155224697	155250915	26218	hap68104	60.5	57.5	57.3	25.0
chr5	2281802	2299281	17479	hap69632	71.5	66.9	69.9	6.6
chr5	4624948	4664704	39756	hap69739	62.8	61.0	14.0	52.0
chr5	89593236	89606080	12844	hap72628	76.6	74.0	20.3	78.4
chr5	119214026	119233058	19032	hap73698	62.8	61.2	57.6	13.1
chr5	119940397	119972658	32261	hap73720	59.1	54.7	53.8	12.2
chr5	132859668	132877415	17747	hap74150	62.5	66.6	59.5	28.3
chr6	26914610	26936918	22308	hap76887	41.9	40.9	71.9	32.6
chr6	66879106	66957243	78137	hap78266	61.6	59.6	25.4	62.0
chr6	77349083	77377529	28446	hap78674	64.5	66.4	27.0	62.9
chr6	159738794	159751033	12239	hap81616	79.6	79.0	21.2	59.8
chr7	26585255	26641907	56652	hap83161	66.2	64.7	49.4	13.3
chr7	48214640	48248036	33396	hap84003	76.0	76.7	78.0	32.3
chr7	88558182	88575482	17300	hap85335	63.8	59.6	63.8	22.9
chr7	96588562	96607580	19018	hap85620	60.4	63.1	19.7	50.0
chr7	122942180	122956897	14717	hap86454	42.3	39.0	19.2	50.0
chr7	132321970	132344802	22832	hap86807	61.4	60.7	52.5	11.5
chr7	153296219	153302441	6222	hap87487	48.7	53.7	64.4	19.3
chr7	156356247	156371897	15650	hap87631	74.9	71.6	87.5	56.6
chr7	159091986	159119486	27500	hap87738	54.0	49.1	52.0	13.2
chr8	51530582	51550889	20307	hap89477	66.4	65.7	68.0	19.9
chr8	63513932	63537543	23611	hap89942	62.0	63.3	11.6	48.4
chr8	72373321	72398122	24801	hap90226	58.0	54.9	71.6	32.0
chr8	94100451	94141855	41404	hap90991	65.2	65.7	36.2	68.7
chr8	109300499	109326404	25905	hap91510	63.6	67.7	29.5	65.8

【圖105B】

Chr	起始	結束	長度	單倍型區塊id	PacBio定序			
					相鄰非腫瘤組織中之甲基化程度		腫瘤組織中之甲基化程度	
					Hap I	Hap II	Hap I	Hap II
chr9	27803548	27888202	84654	hap58508	64.2	60.9	20.6	75.4
chr6	242149	386636	144487	hap47880	62.3	63.3	77.4	32.2
chr5	28219159	28302858	83699	hap44666	59.3	58.0	16.8	58.2
chr5	18119943	18153743	33800	hap44475	61.6	65.0	53.2	21.7
chr7	24906307	25046195	139888	hap52069	69.3	68.7	44.0	76.2
chr15	27689897	27752573	62676	hap18337	65.9	61.9	64.8	20.5
chr12	42183870	42212433	28563	hap12045	63.5	68.4	19.4	51.2
chr21	9825597	9935752	110155	hap34175	54.3	53.5	60.9	29.1
chr2	118813055	118893366	80311	hap30060	62.6	62.3	77.0	38.6
chr6	90307702	90344869	37167	hap49779	69.1	66.4	84.7	53.9
chr7	107932914	108049376	116462	hap53838	67.2	62.9	43.8	76.4
chr7	137039327	137160933	121606	hap54447	59.5	60.9	22.9	72.0
chr17	21193754	21254930	61176	hap22633	59.2	54.3	69.7	31.6
chr12	11473697	11644714	171017	hap11451	62.8	66.4	35.5	75.9
chr5	129212299	129353349	141050	hap46632	50.9	54.5	45.5	14.0
chr11	93910738	94028887	118149	hap10288	67.6	63.6	36.6	74.2
chr3	131707434	132003636	296202	hap38642	57.8	55.9	17.9	60.2
chr3	43024004	43161785	137781	hap36769	69.1	66.5	46.1	80.2
chr3	190403156	190606658	203502	hap39947	60.9	61.6	36.9	72.7
chr15	40218970	40279780	60810	hap18606	53.4	57.5	79.1	47.4

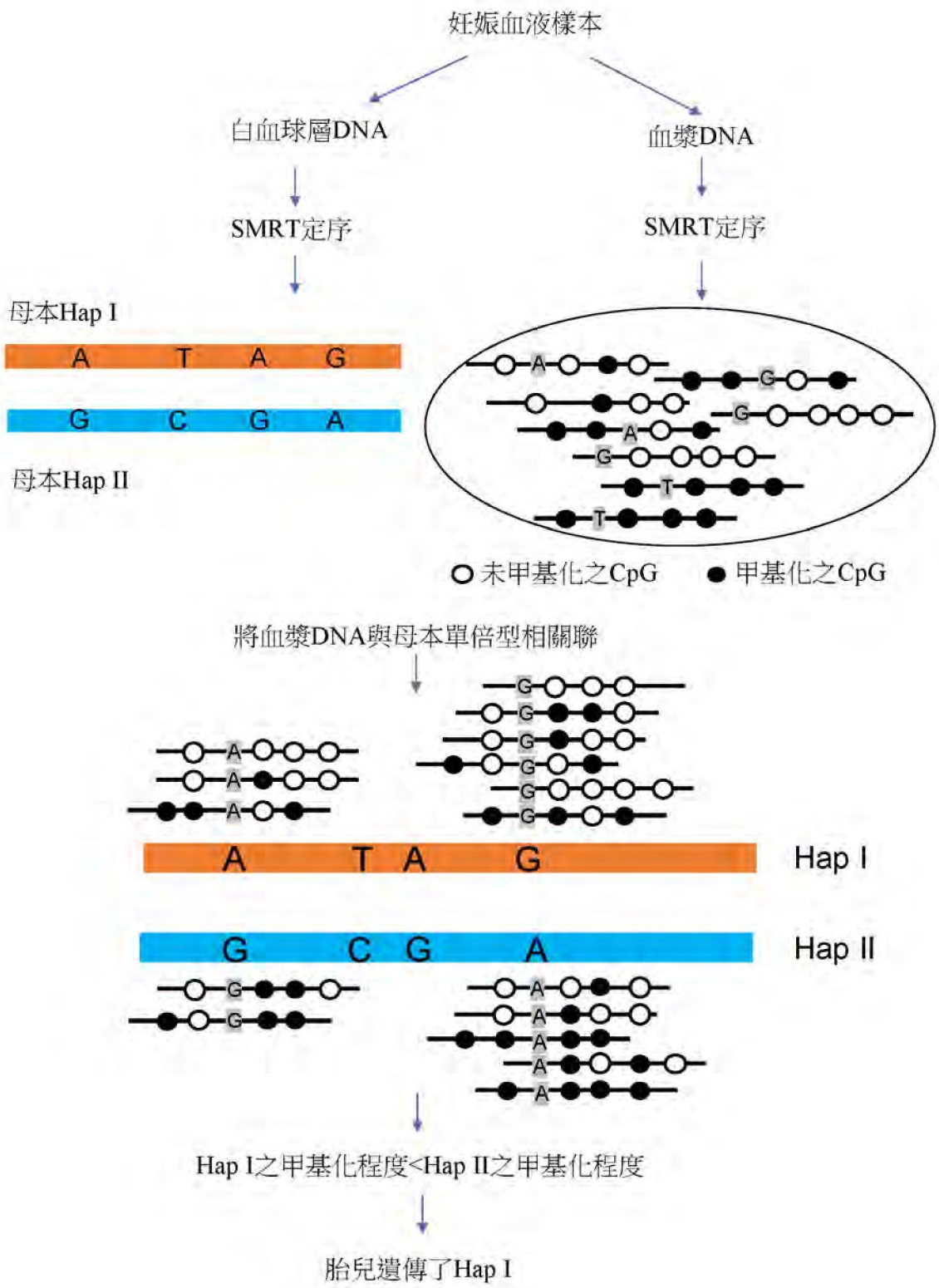
【圖106】

組織類型	顯示腫瘤組織中兩個單倍型之間的甲基化不平衡的單倍型區塊的數量	顯示配對的相鄰非腫瘤組織中兩個單倍型之間的甲基化不平衡的單倍型區塊的數量
結腸	92	47
乳房	57	13
腎臟	68	18
肺	31	21
前列腺	26	19
胃	2	0

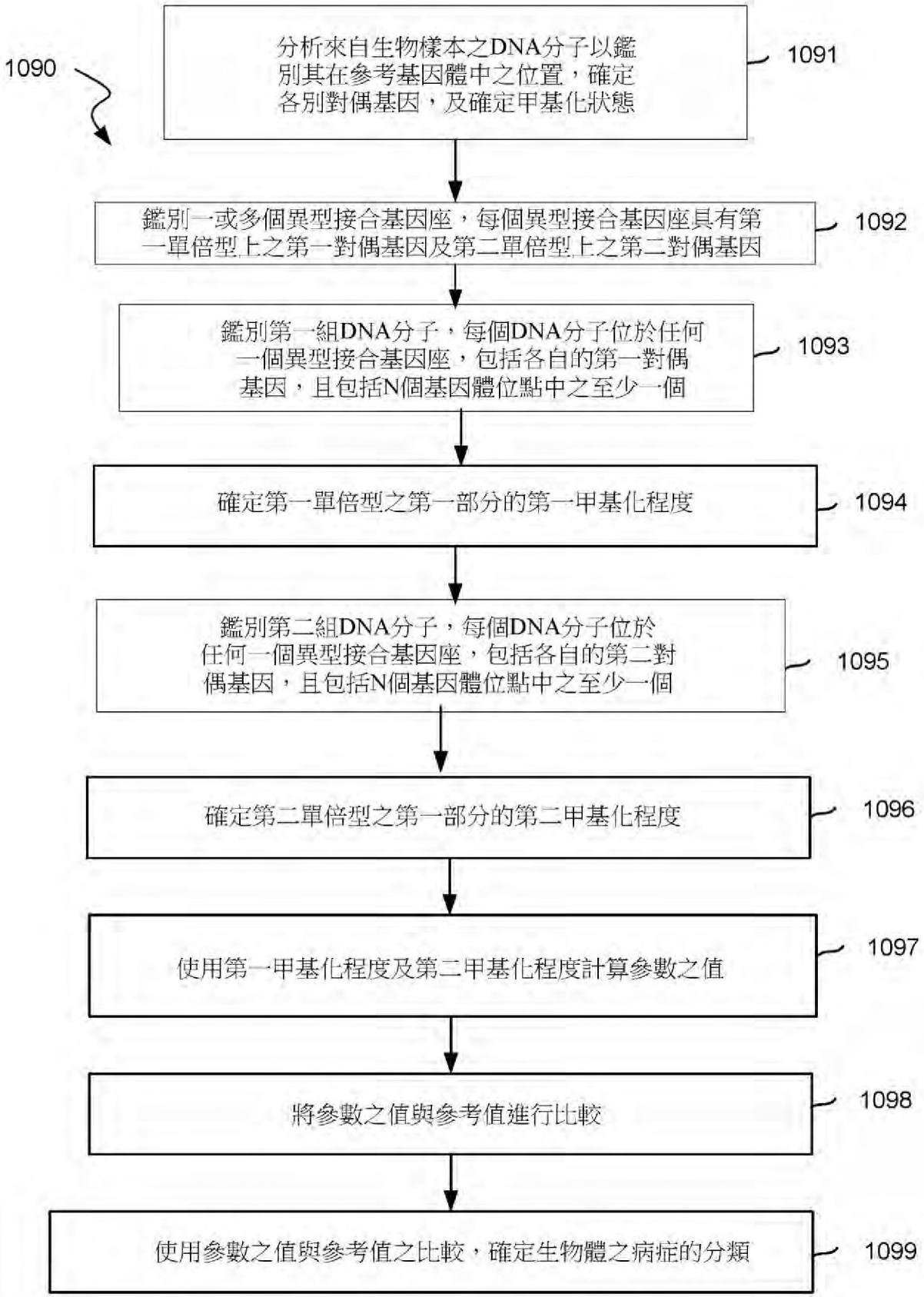
【圖107A】

組織類型	顯示腫瘤組織中兩個單倍型之間的甲基化不平衡的單倍型區塊的數量	可獲得的腫瘤分期資訊（TNM）
乳房	18	T2
	57	T3
腎臟	68	T3a
	0	T2

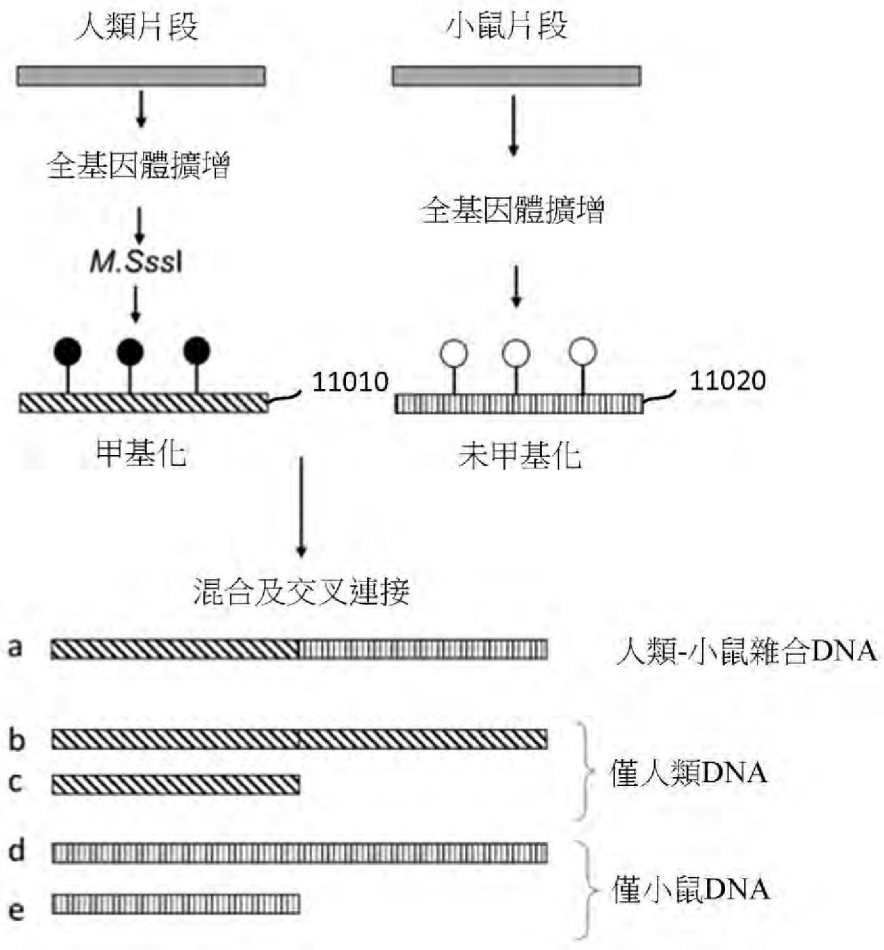
【圖107B】


妊娠血液樣本
白血球層DNA
血漿DNA
SMRT定序
SMRT定序
母本Hap I
A T A G
G C G A
母本Hap II
未甲基化之CpG
甲基化之CpG
將血漿DNA與母本單倍型相關聯
A T A G
Hap I
G C G A
Hap II
Hap I之甲基化程度<Hap II之甲基化程度
胎兒遺傳了Hap I

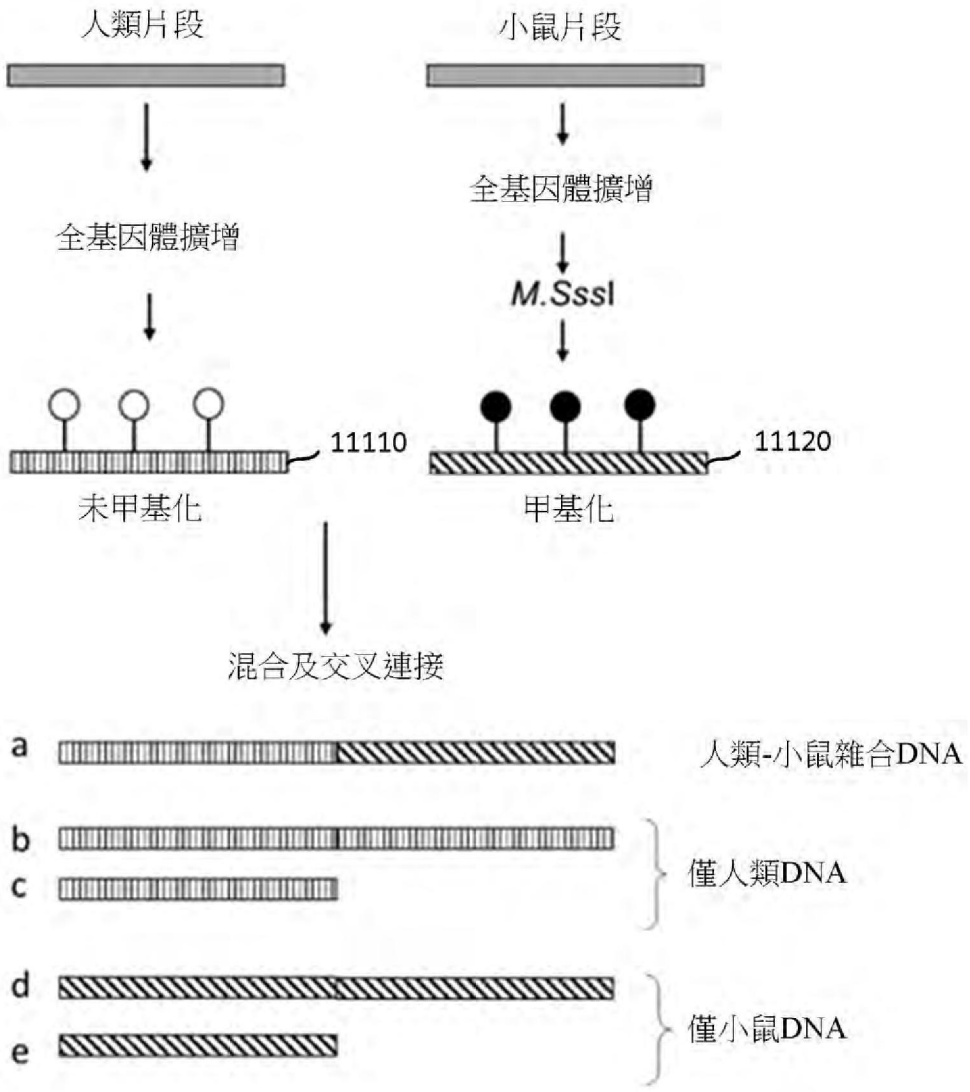
【圖108】


1090
分析來自生物樣本之DNA分子以鑑別其在參考基因體中之位置，確定各別對偶基因，及確定甲基化狀態
1091
鑑別一或多個異型接合基因座，每個異型接合基因座具有第一單倍型上之第一對偶基因及第二單倍型上之第二對偶基因
1092
鑑別第一組DNA分子，每個DNA分子位於任何一個異型接合基因座，包括各自的第一對偶基因，且包括N個基因體位點中之至少一個
1093
確定第一單倍型之第一部分的第一甲基化程度
1094
鑑別第二組DNA分子，每個DNA分子位於任何一個異型接合基因座，包括各自的第二對偶基因，且包括N個基因體位點中之至少一個
1095
確定第二單倍型之第一部分的第二甲基化程度
1096
使用第一甲基化程度及第二甲基化程度計算參數之值
1097
將參數之值與參考值進行比較
1098
使用參數之值與參考值之比較，確定生物體之病症的分類
1099

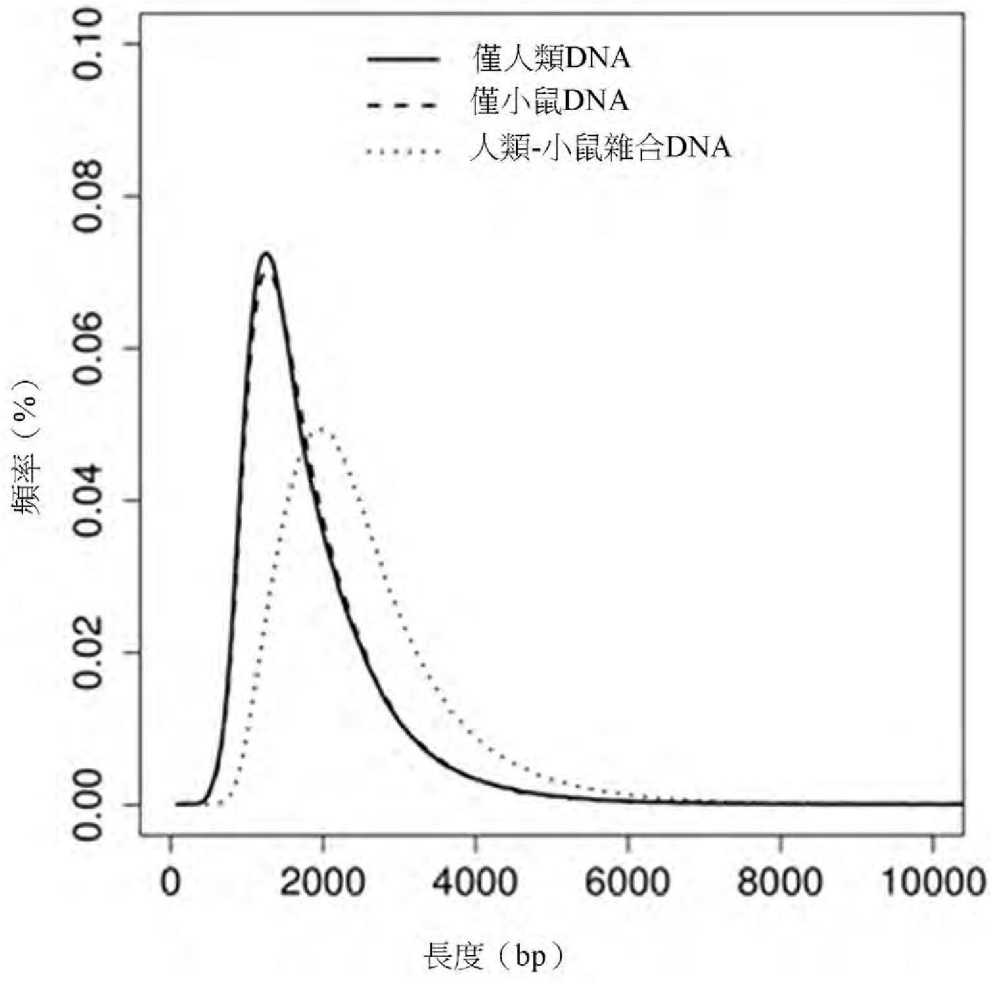
【圖109】


人類片段
小鼠片段
全基因體擴增
全基因體擴增
M.SssI
11010
11020
甲基化
未甲基化
混合及交叉連接
a
人類-小鼠雜合DNA
b
c
僅人類DNA
d
e
僅小鼠DNA

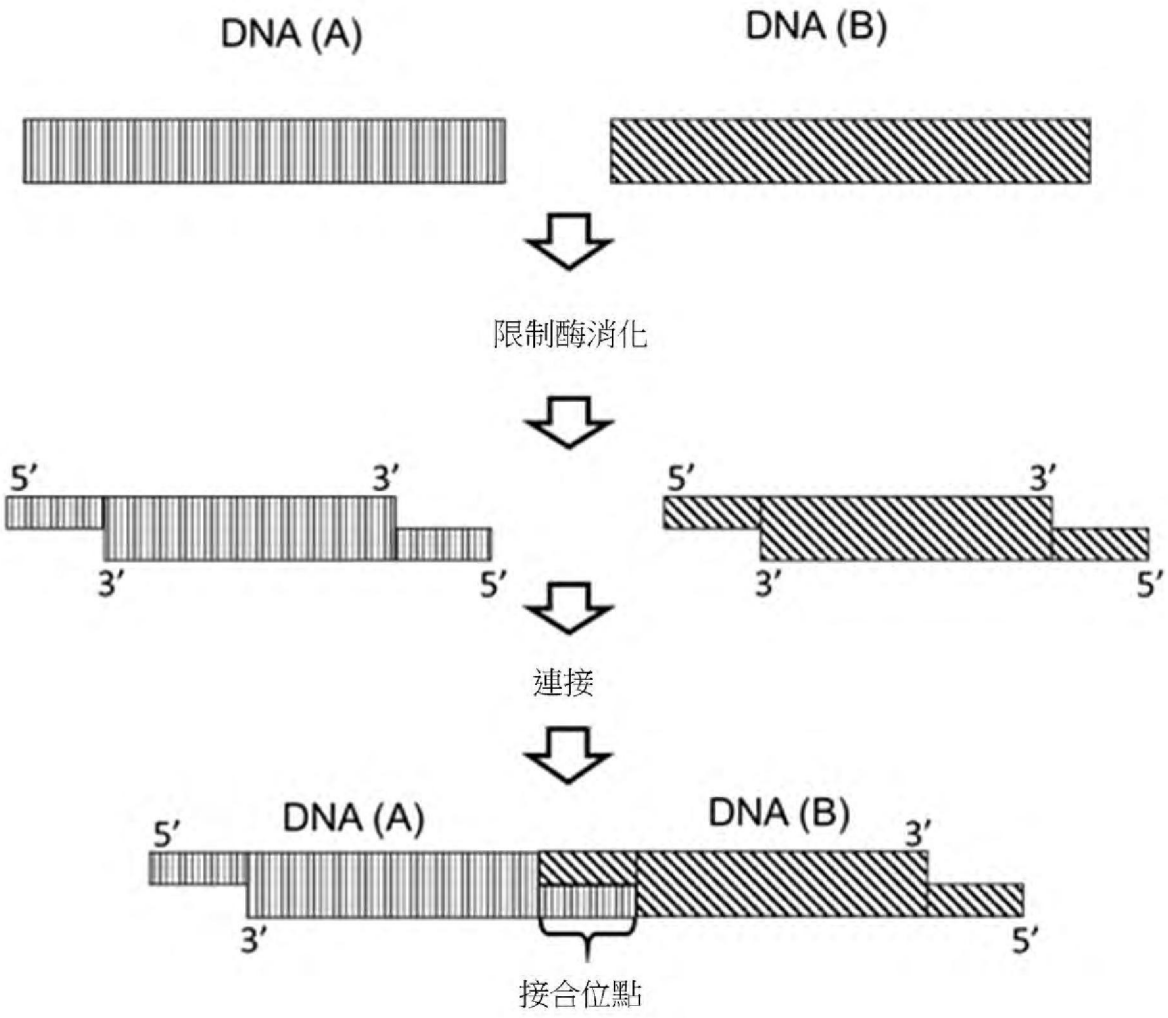
【圖110】


人類片段
小鼠片段
全基因體擴增
全基因體擴增
M.SssI
11110
11120
未甲基化
甲基化
混合及交叉連接
a
人類-小鼠雜合DNA
b
c
僅人類DNA
d
e
僅小鼠DNA

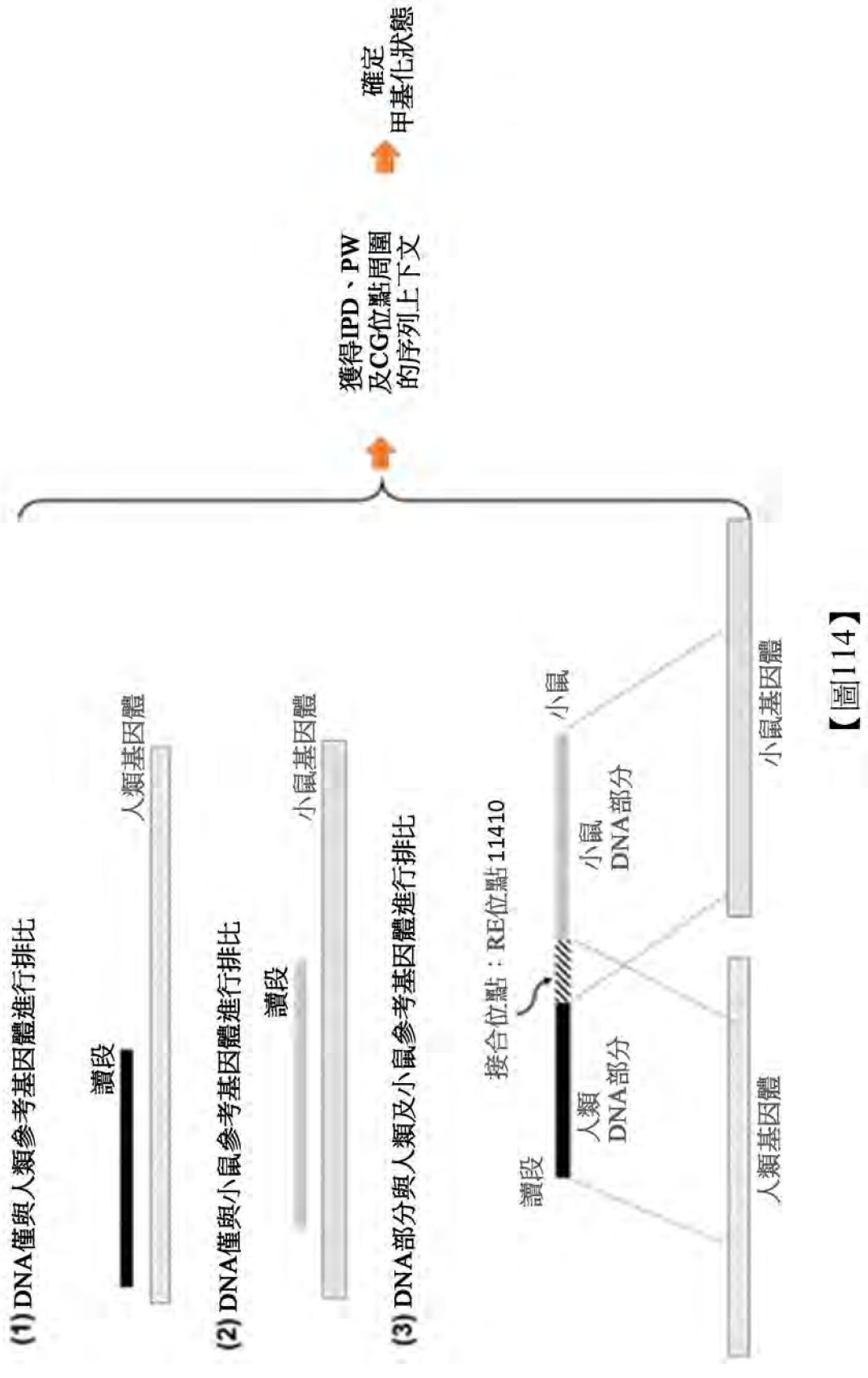
【圖111】


僅人類DNA
僅小鼠DNA
人類-小鼠雜合DNA
頻率（%）
0.00
0.02
0.04
0.06
0.08
0.10
0
2000
4000
6000
8000
10000
長度（bp）

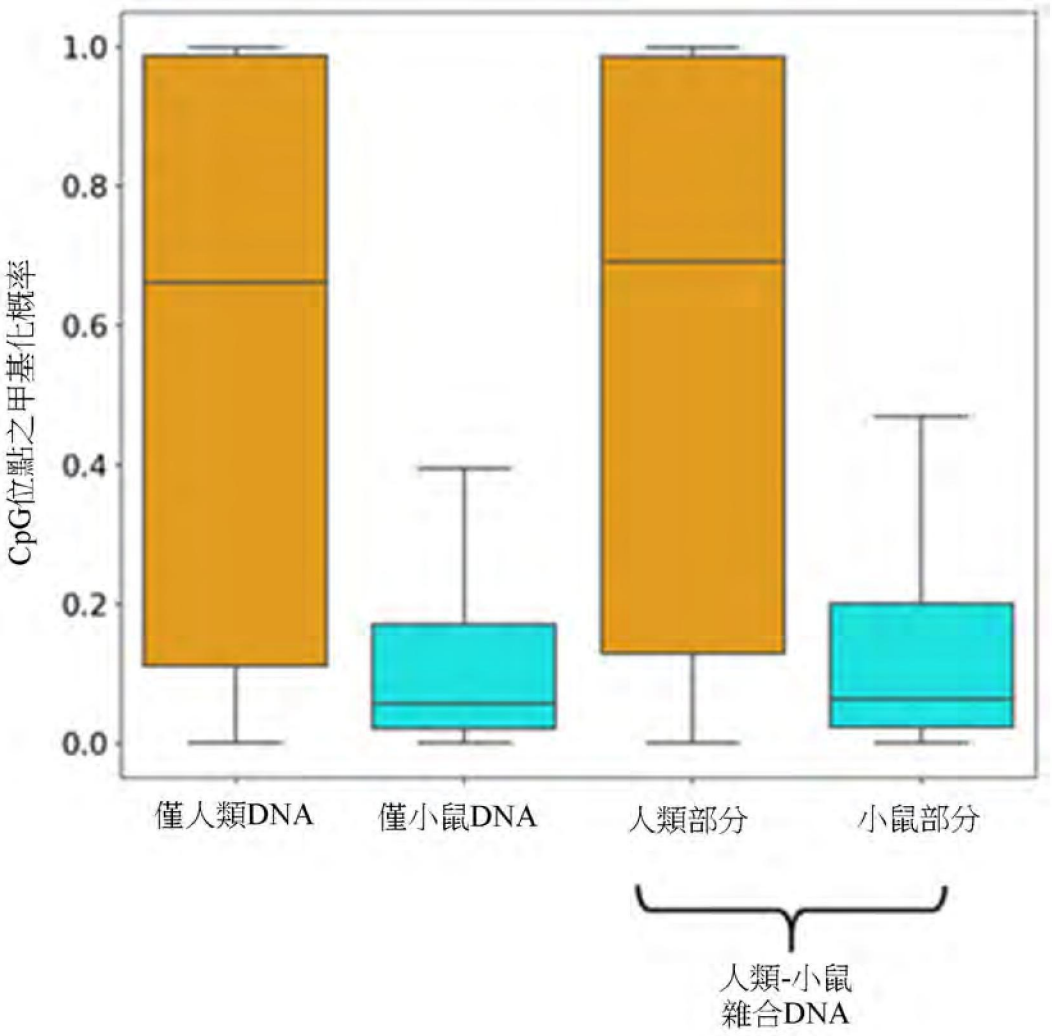
【圖112】



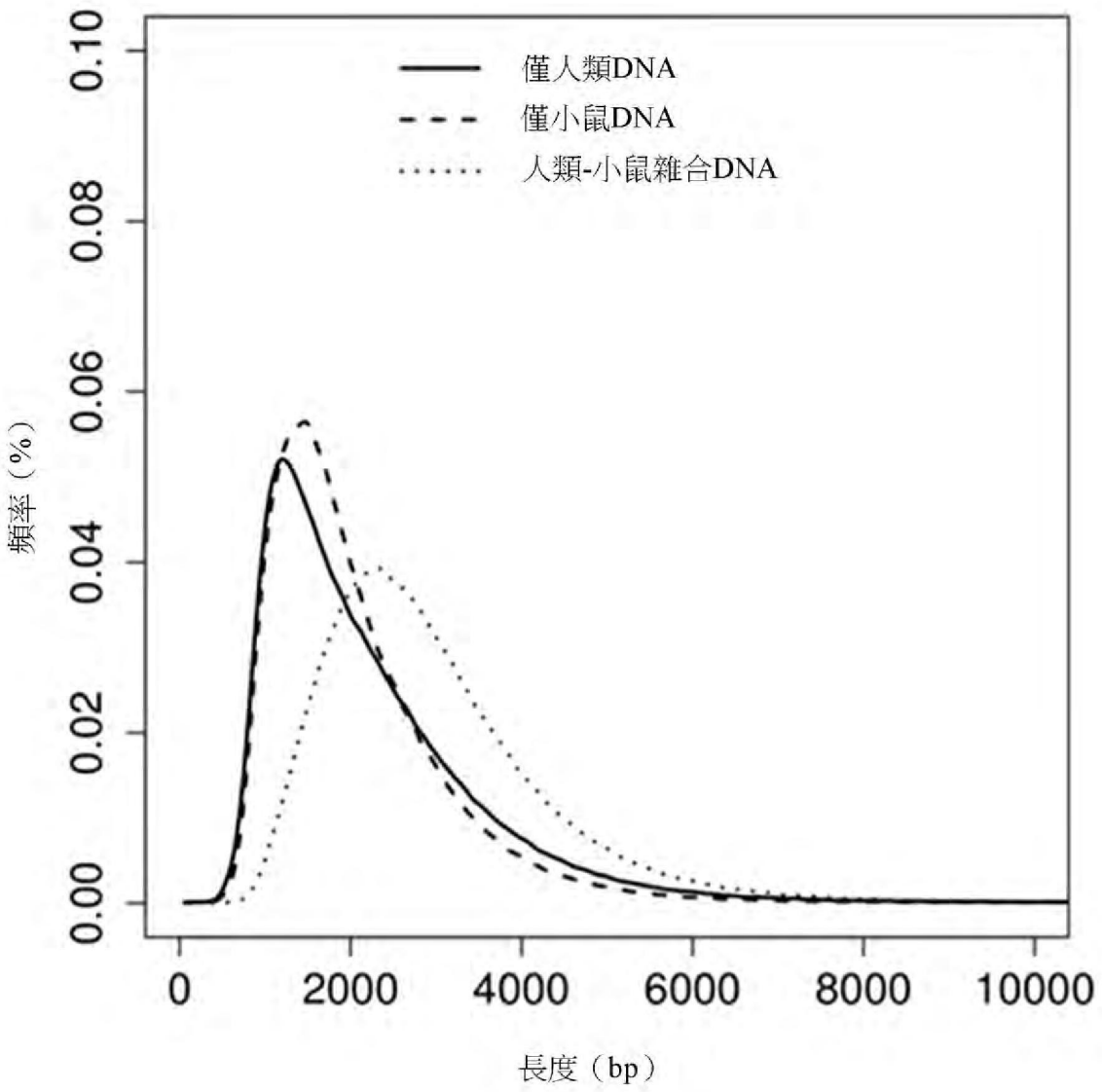
【圖113】


(1) DNA僅與人類參考基因體進行排比
讀段
人類基因體
(2) DNA僅與小鼠參考基因體進行排比
讀段
小鼠基因體
(3) DNA部分與人類及小鼠參考基因體進行排比
接合位點：RE位點11410
讀段
人類
DNA部分
小鼠
小鼠
DNA部分
人類基因體
小鼠基因體
獲得IPD、PW
及CG位點周圍
的序列上下文
確定
甲基化狀態

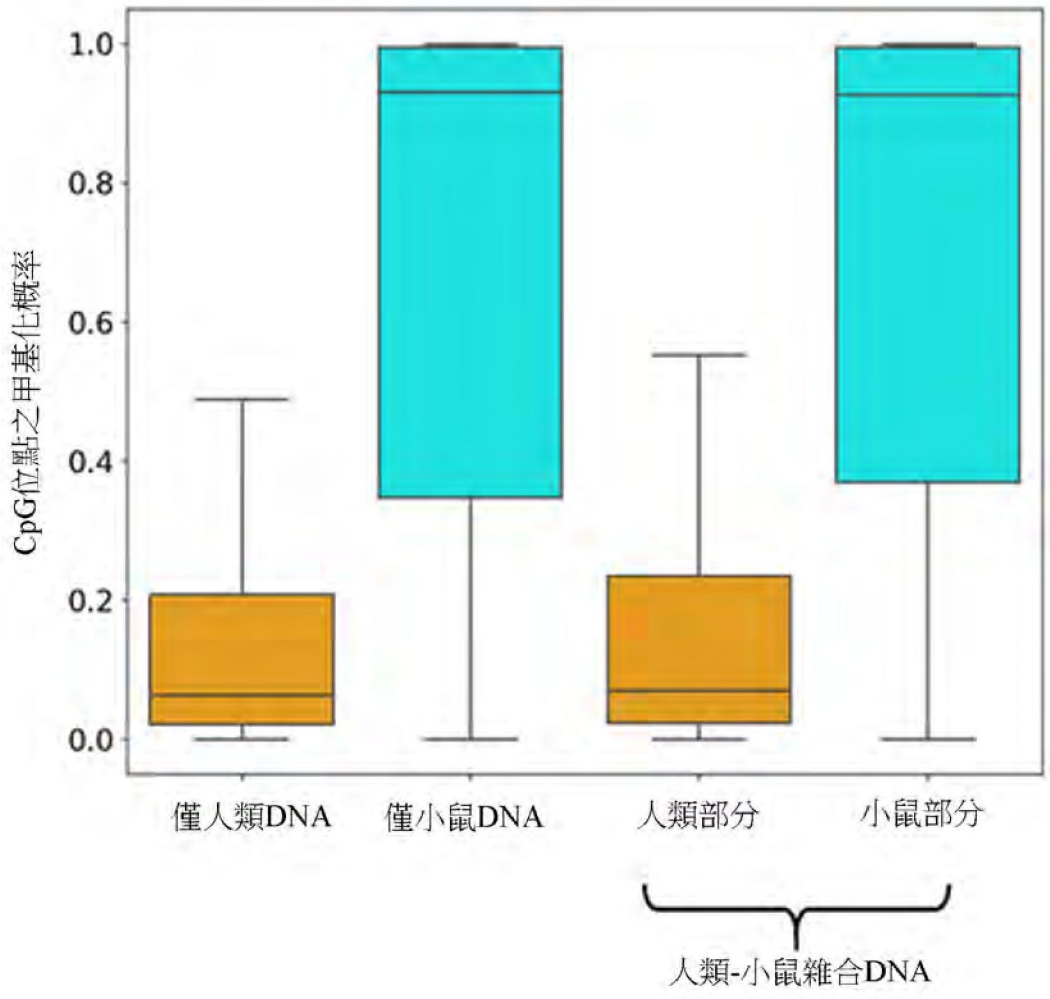
【圖114】


1.0
0.8
0.6
0.4
0.2
0.0
CpG位點之甲基化概率
僅人類DNA
僅小鼠DNA
人類部分
小鼠部分
人類-小鼠
雜合DNA

【圖115】


僅人類DNA
僅小鼠DNA
人類-小鼠雜合DNA
頻率（%）
0.00
0.02
0.04
0.06
0.08
0.10
0
2000
4000
6000
8000
10000
長度（bp）

【圖116】


1.0
0.8
0.6
0.4
0.2
0.0
CpG位點之甲基化概率
僅人類DNA
僅小鼠DNA
人類部分
小鼠部分
人類-小鼠雜合DNA

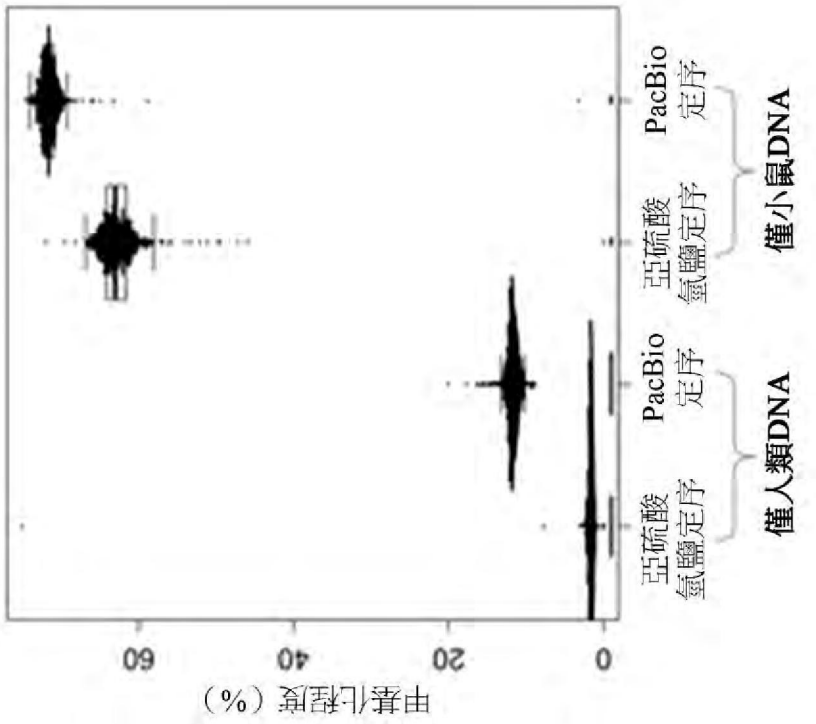
【圖117】

		亞硫酸氫鹽定序		PacBio定序	
		CG位點之數量	甲基化密度（%）	CG位點之數量	甲基化密度（%）
1) 僅人類		2,230,407	41.4	16,226,014	56.0
2) 僅小鼠		2,726,499	1.6	9,398,340	10.7
3) 人類-小鼠雜合 DNA	人類部分	73,780	46.8	4,838,454	57.4
	小鼠部分	76,312	2.3	4,385,046	12.1

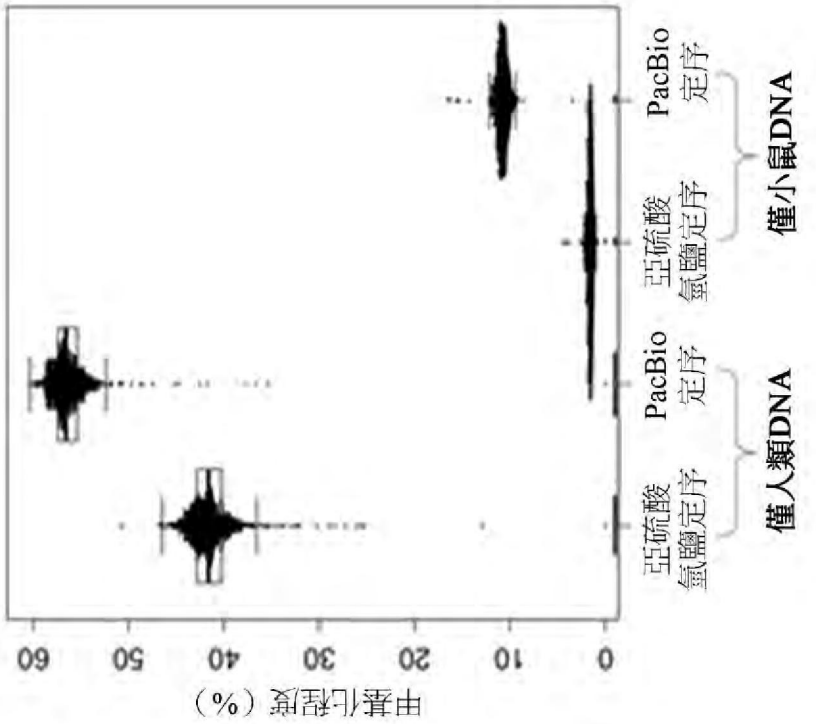
【圖118】

		亞硫酸氫鹽定序		PacBio定序	
		CG位點之數量	甲基化密度（%）	CG位點之數量	甲基化密度（%）
1) 僅人類		2,938,088	1.6	14,503,548	11.6
2) 僅小鼠		1,513,971	62.4	11,348,555	71.5
3) 人類-小鼠雜合 DNA	人類部分	67,371	1.8	5,824,379	13.1
	小鼠部分	58,242	67.4	5,093,097	72.2

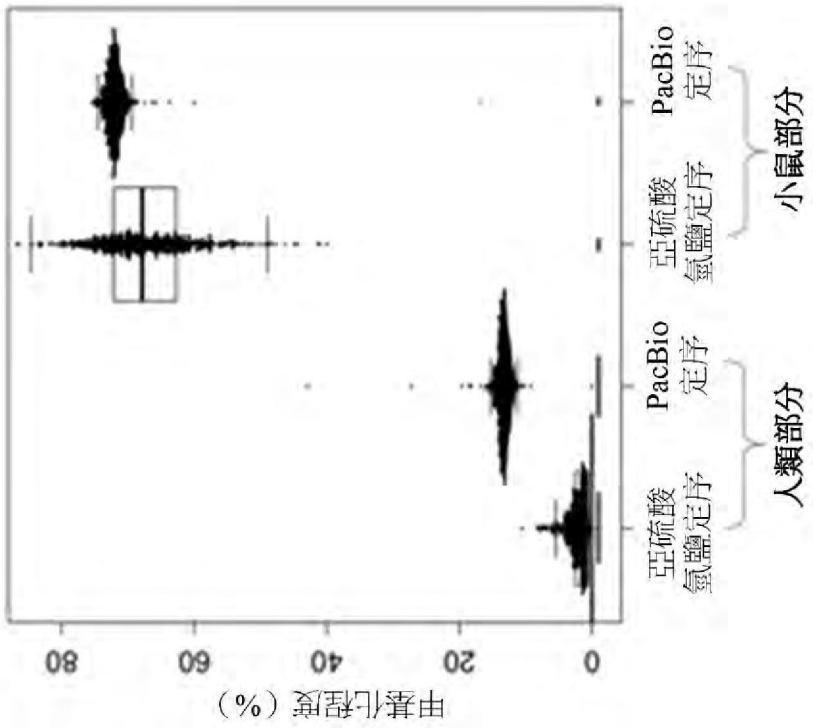
【圖119】


甲基化程度（%）
0
20
40
60
亞硫酸氫鹽定序
PacBio定序
亞硫酸氫鹽定序
PacBio定序
僅人類DNA
僅小鼠DNA

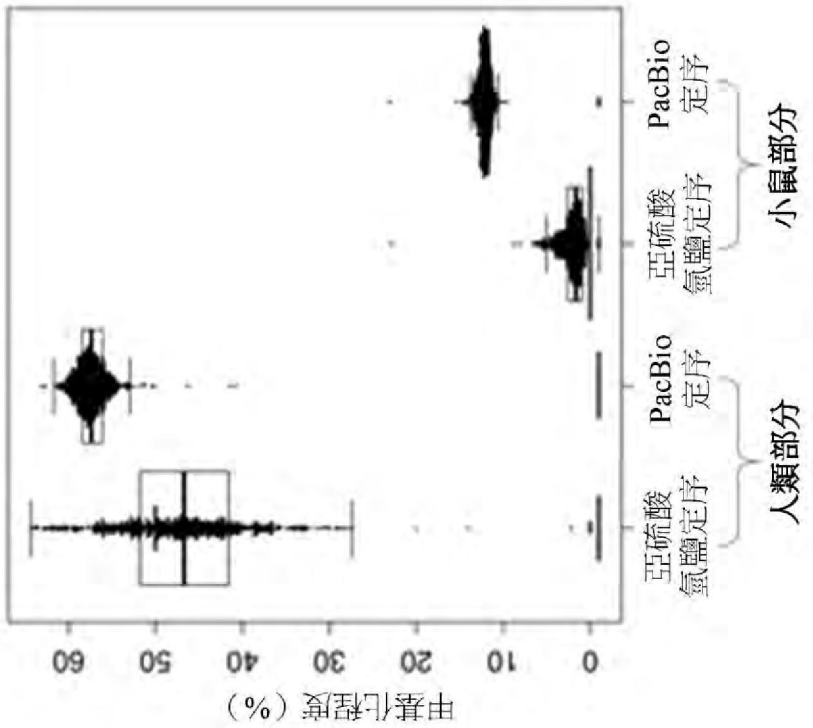
【圖120B】


甲基化程度（%）
0
10
20
30
40
50
60
亞硫酸氫鹽定序
PacBio定序
亞硫酸氫鹽定序
PacBio定序
僅人類DNA
僅小鼠DNA

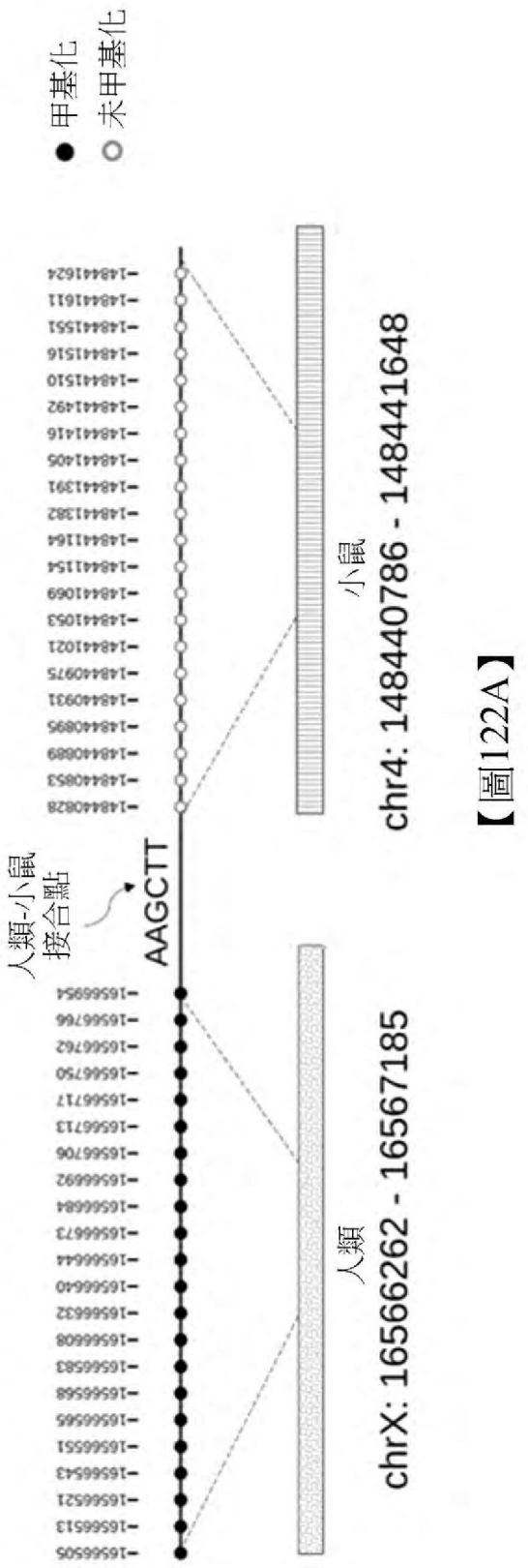
【圖120A】


甲基化程度（%）
0
20
40
60
80
亞硫酸氫鹽定序
PacBio定序
亞硫酸氫鹽定序
PacBio定序
人類部分
小鼠部分

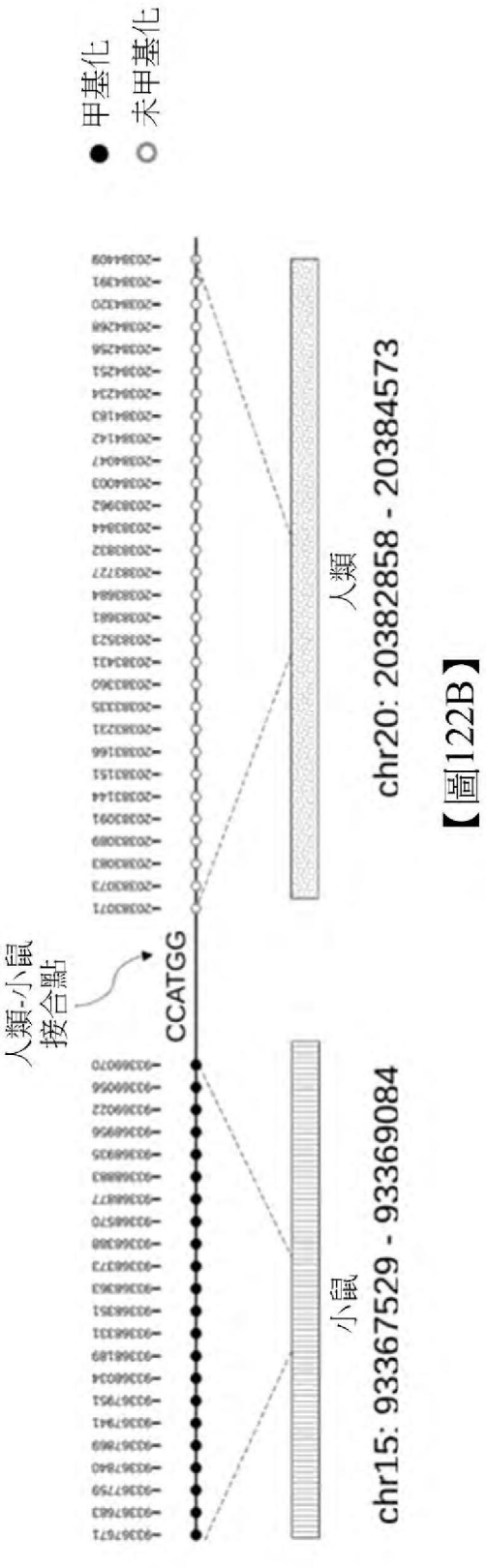
【圖121B】


甲基化程度（%）
0
10
20
30
40
50
60
亞硫酸氫鹽定序
PacBio定序
亞硫酸氫鹽定序
PacBio定序
人類部分
小鼠部分

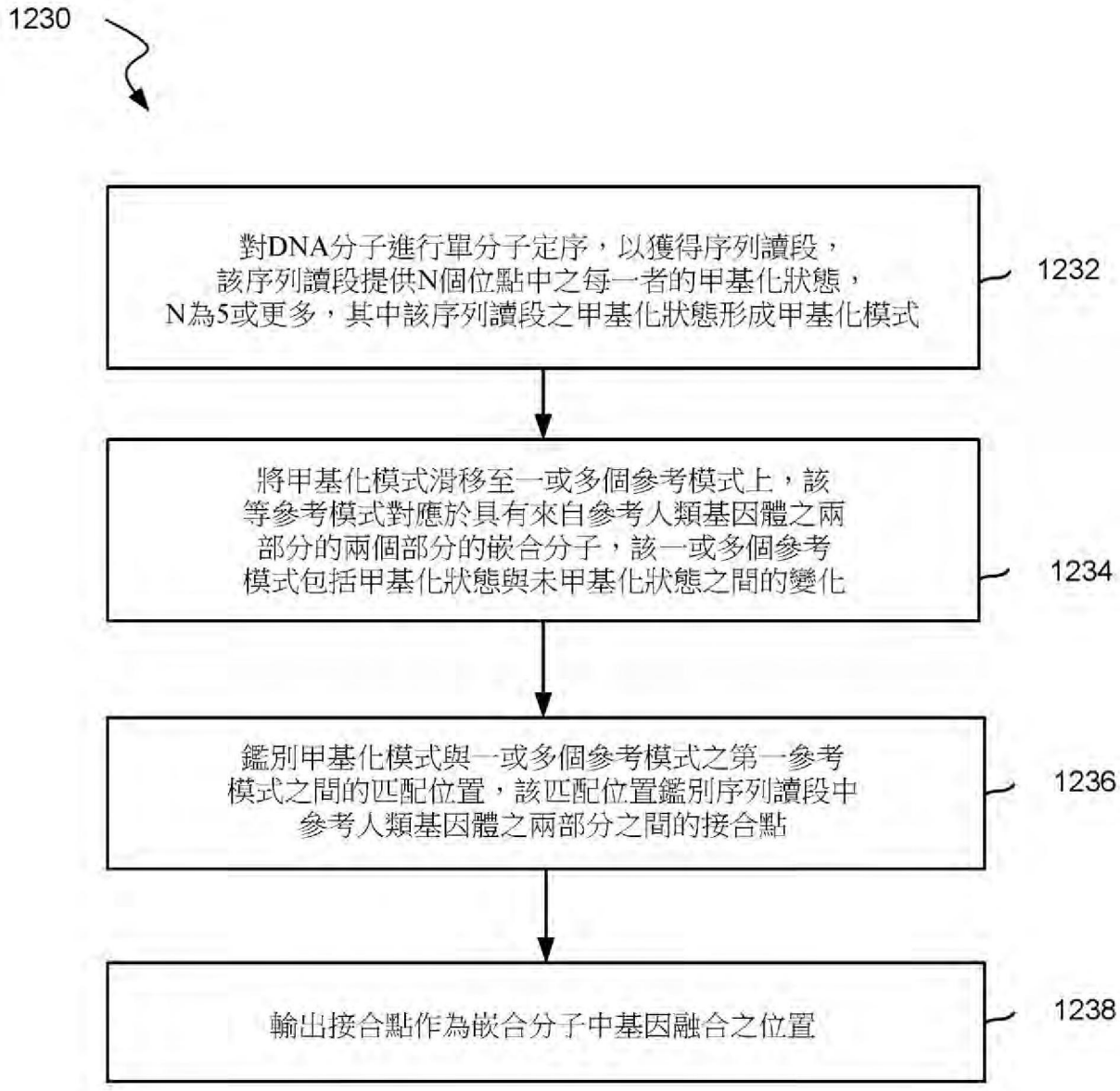
【圖121A】


甲基化
未甲基化
人類-小鼠
接合點
AAGCTT
人類
小鼠
chrX: 16566262 - 16567185
chr4: 148440786 - 148441648

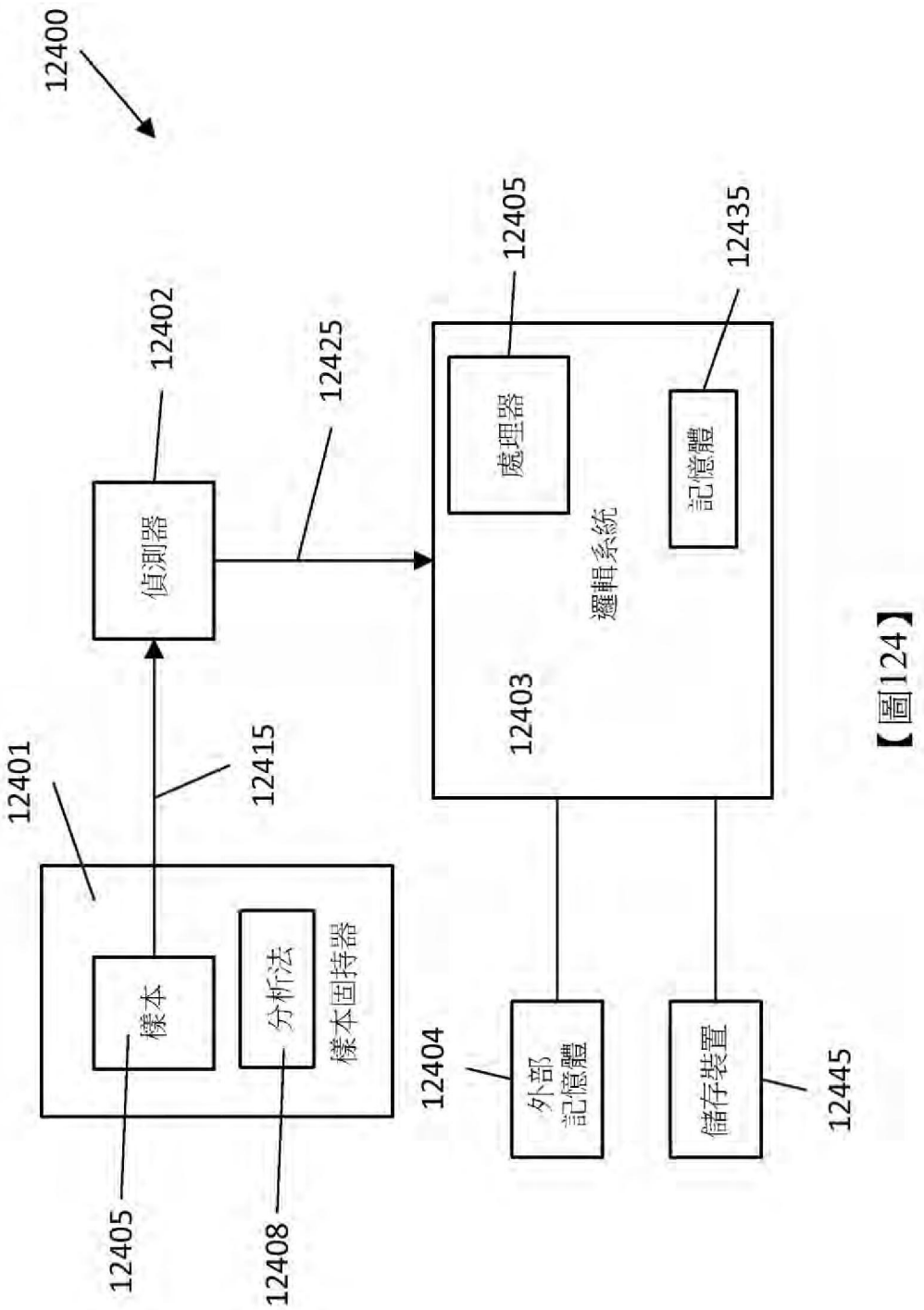
【圖122A】


甲基化
未甲基化
人類-小鼠
接合點
CCATGG
人類
小鼠
chr15: 93367529 - 93369084
chr20: 20382858 - 20384573

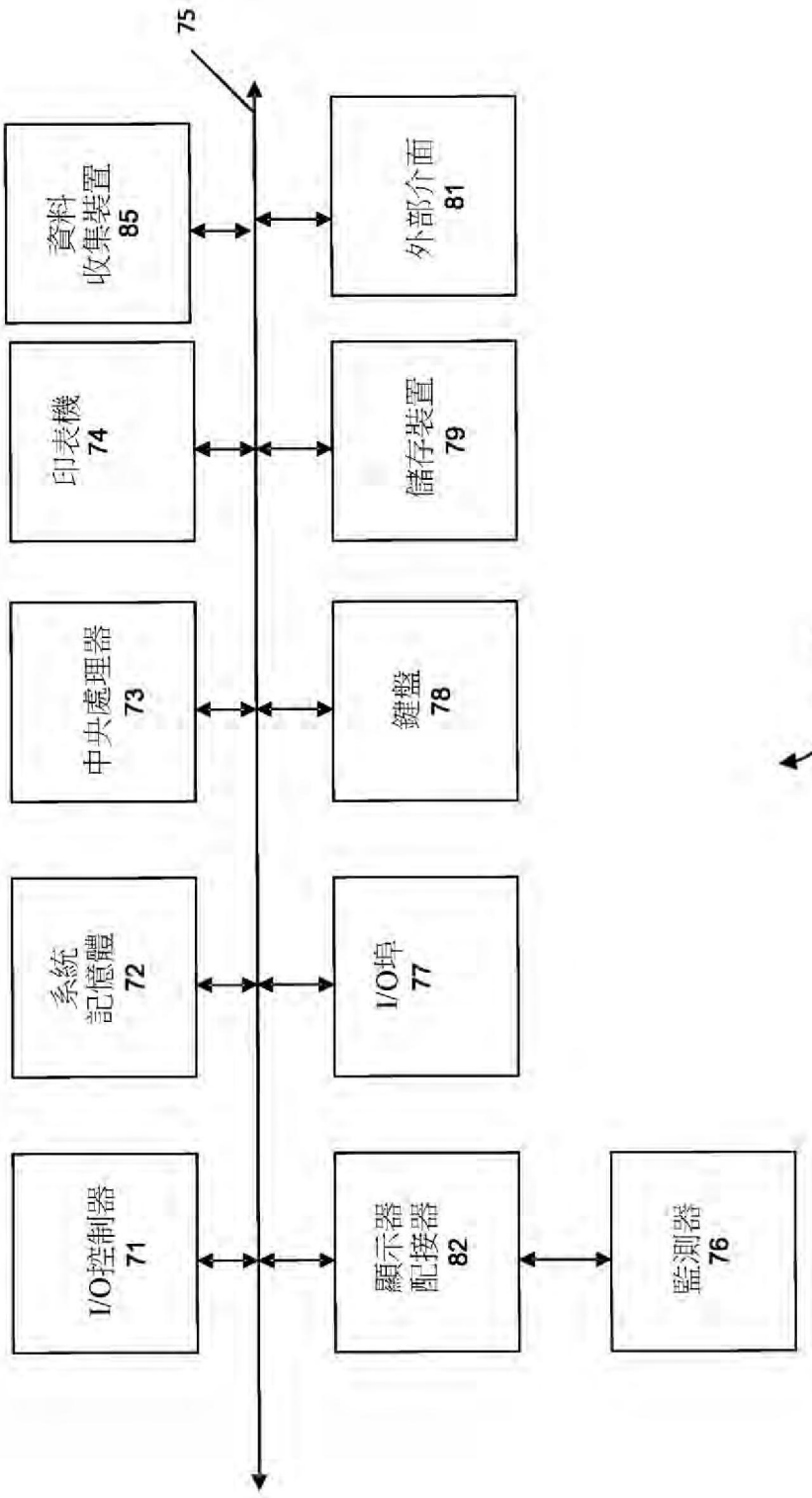
【圖122B】


1230
對DNA分子進行單分子定序，以獲得序列讀段，
該序列讀段提供N個位點中之每一者的甲基化狀態，
N為5或更多，其中該序列讀段之甲基化狀態形成甲基化模式
1232
將甲基化模式滑移至一或多個參考模式上，該
等參考模式對應於具有來自參考人類基因體之兩
部分的兩個部分的嵌合分子，該一或多個參考
模式包括甲基化狀態與未甲基化狀態之間的變化
1234
鑑別甲基化模式與一或多個參考模式之第一參考
模式之間的匹配位置，該匹配位置鑑別序列讀段中
參考人類基因體之兩部分之間的接合點
1236
輸出接合點作為嵌合分子中基因融合之位置
1238

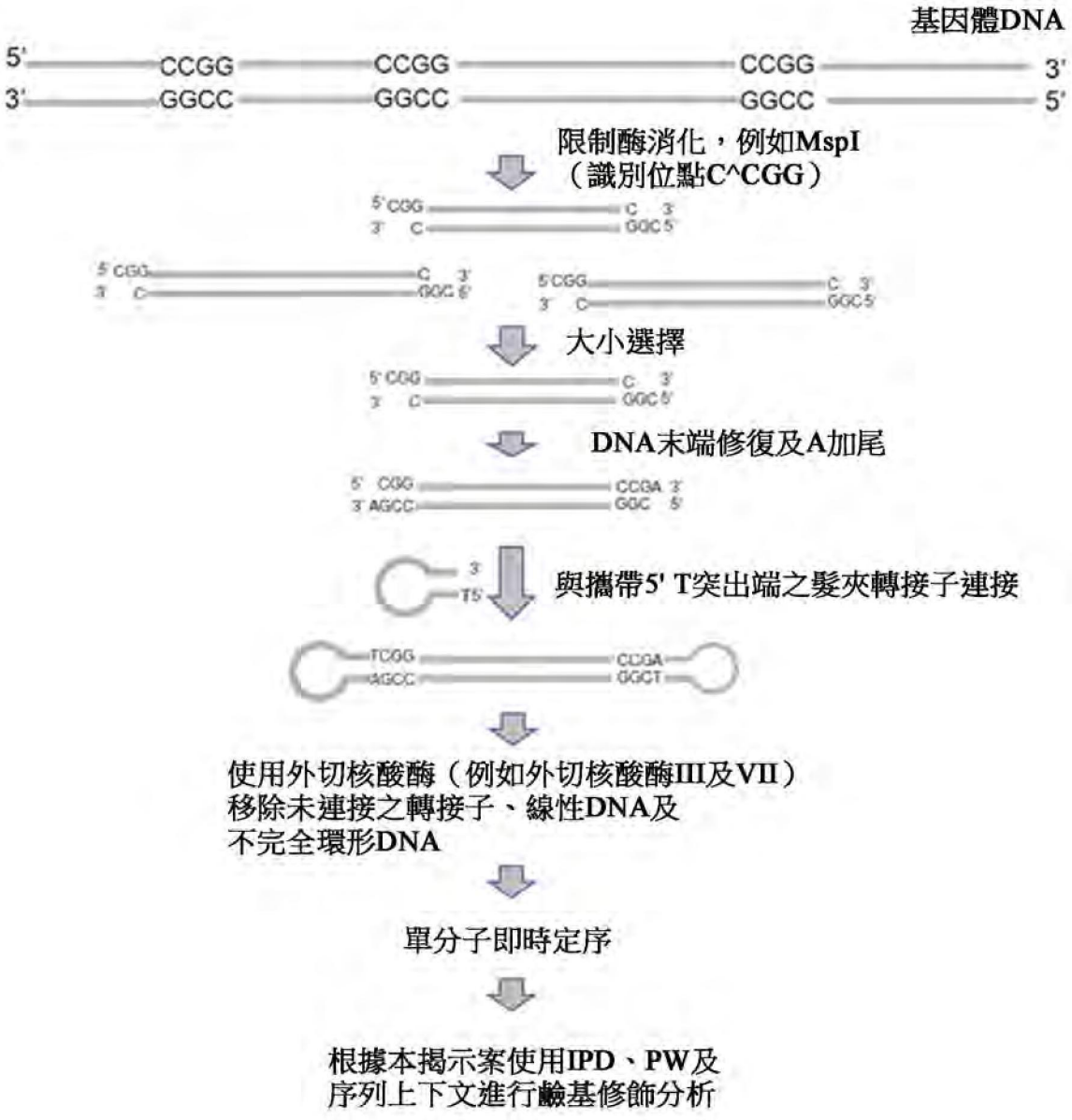
【圖123】


12400
12402
偵測器
12425
12405
處理器
12435
記憶體
邏輯系統
12403
12401
12415
樣本
分析法
樣本固持器
12404
外部
記憶體
儲存裝置
12445
12405
12408

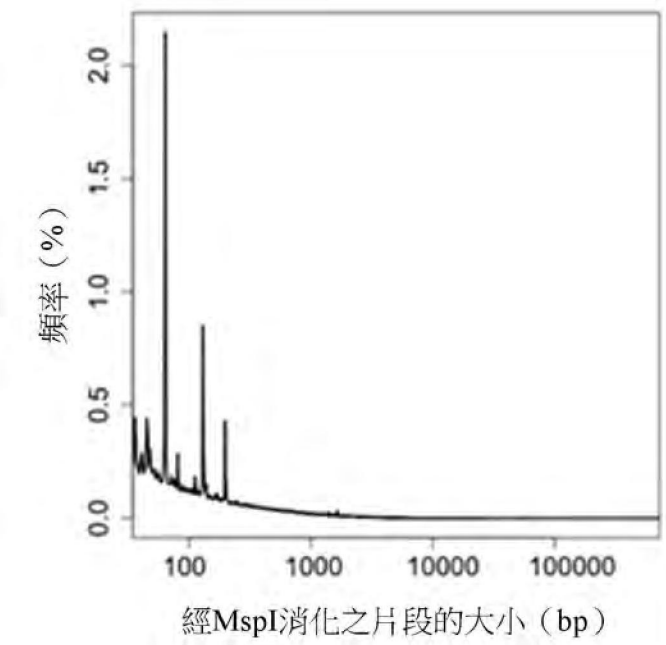
【圖124】


75
資料收集裝置 85
外部介面 81
印表機 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
I/O控制器 71
顯示器配接器 82
監測器 76
10

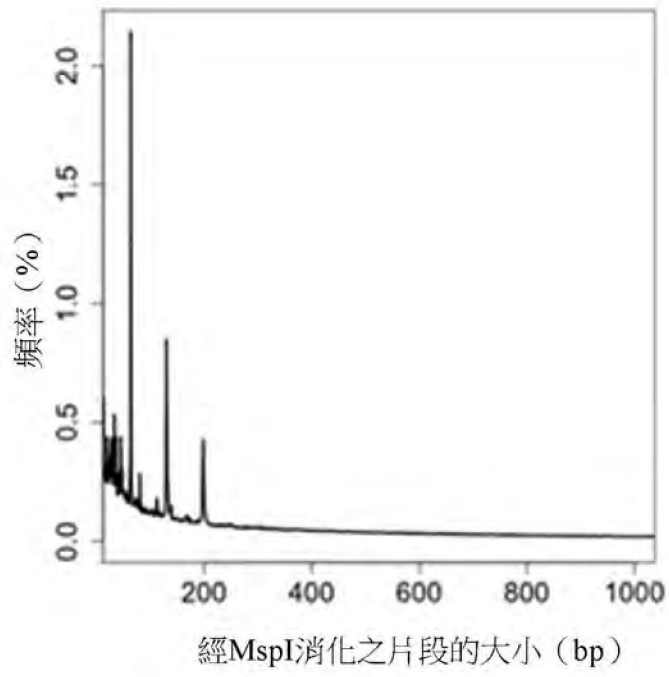
【圖125】


基因體DNA
5' CCGG CCGG CCGG 3'
3' GGCC GGCC GGCC 5'
限制酶消化，例如MspI
（識別位點C^CGG）
大小選擇
DNA末端修復及A加尾
與攜帶5' T突出端之髮夾轉接子連接
使用外切核酸酶（例如外切核酸酶III及VII）
移除未連接之轉接子、線性DNA及
不完全環形DNA
單分子即時定序
根據本揭示案使用IPD、PW及
序列上下文進行鹼基修飾分析

【圖126】


頻率（%）
0.0
0.5
1.0
1.5
2.0
100
1000
10000
100000
經MspI消化之片段的大小（bp）

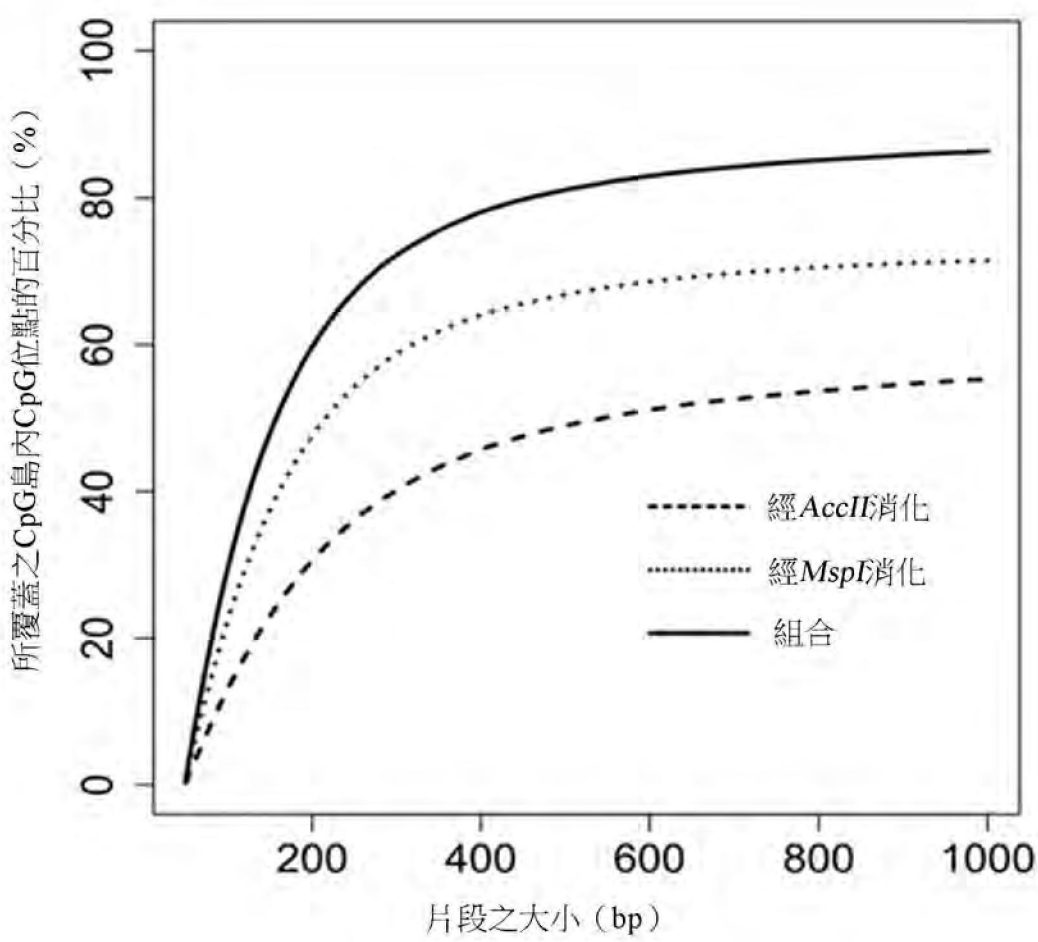
【圖127A】


頻率（%）
0.0
0.5
1.0
1.5
2.0
200
400
600
800
1000
經MspI消化之片段的大小（bp）

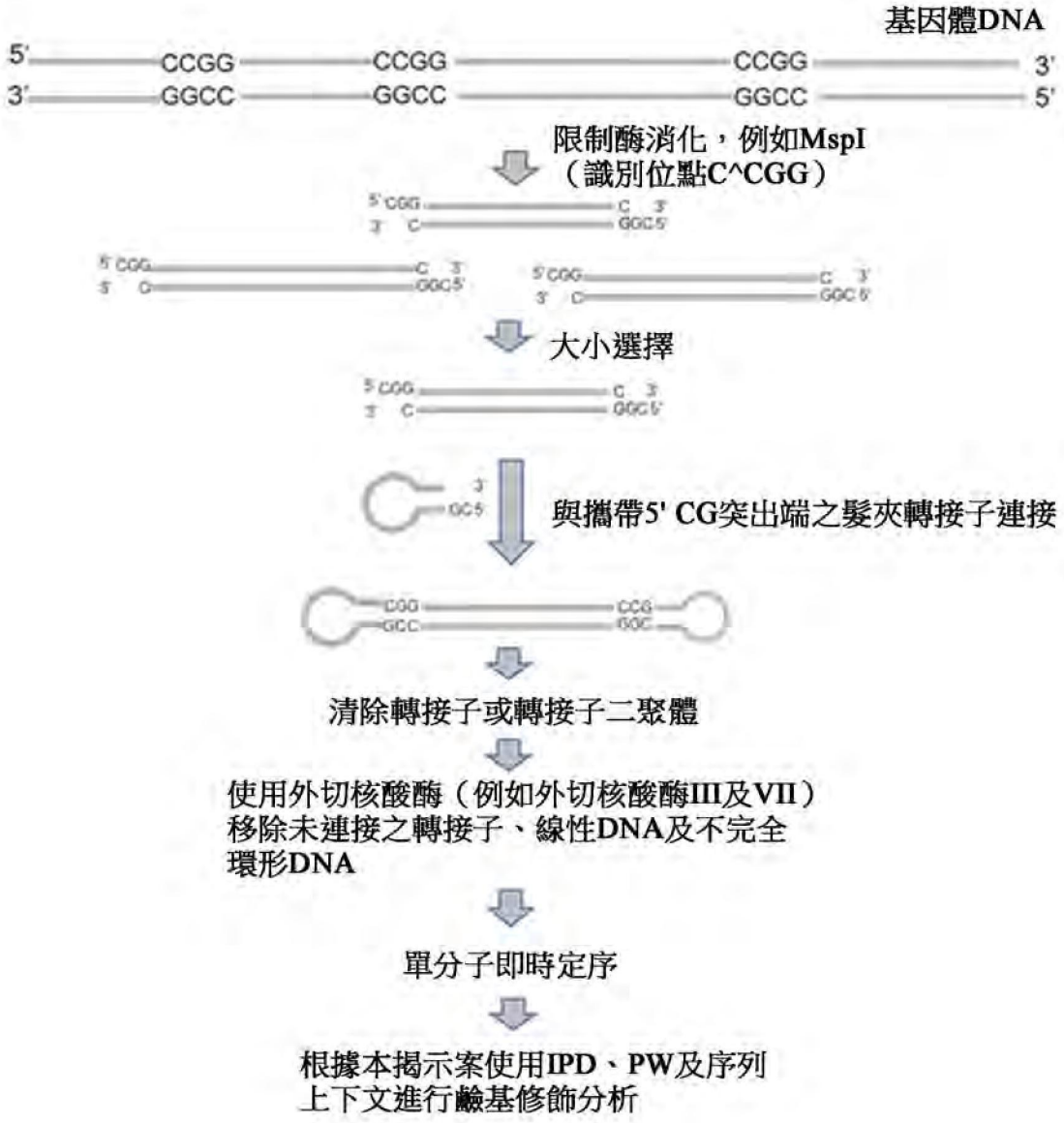
【圖127B】

大小範圍（bp）	分子數量	在大小範圍內之分子相對於總片段的百分比（%）	在大小範圍內與CpG島重疊的分子數量	在大小範圍內之分子與CpG島重疊的百分比（%）	經定序之CpG位點的數量	落入CpG島內之CpG位點的數量	藉由大小選擇範向且落入CpG島內之CpG位點的百分比（%）
50-200	526,543	23.03	104,059	19.76	2,358,020	885,041	37.53
200-400	269,562	11.79	23,927	8.88	1,781,556	353,087	19.82
400-600	177,776	7.77	7,369	4.15	1,468,561	107,130	7.29
600-800	133,927	5.86	3,673	2.74	1,326,544	48,851	3.68
800-1000	104,976	4.59	2,168	2.07	1,193,233	25,821	2.16
1000-2000	311,596	13.63	4,596	1.47	4,610,504	58,288	1.26
2000-3000	149,468	6.54	1,771	1.18	3,036,951	25,106	0.83
3000-4000	86,760	3.79	809	0.93	2,165,171	10,785	0.50
5000-6000	36,931	1.62	266	0.72	1,242,712	3,412	0.27
6000-7000	25,027	1.09	202	0.81	947,874	3,354	0.35
7000-8000	17,597	0.77	86	0.49	736,830	791	0.11
8000-9000	12,658	0.55	76	0.60	583,680	993	0.17
9000-10000	9,184	0.40	48	0.52	461,935	591	0.13
10000-15000	20,790	0.91	97	0.47	1,255,731	2,003	0.16
15000-20000	5,111	0.22	16	0.31	414,400	163	0.04
20000-25000	1,441	0.06	6	0.42	147,731	34	0.02

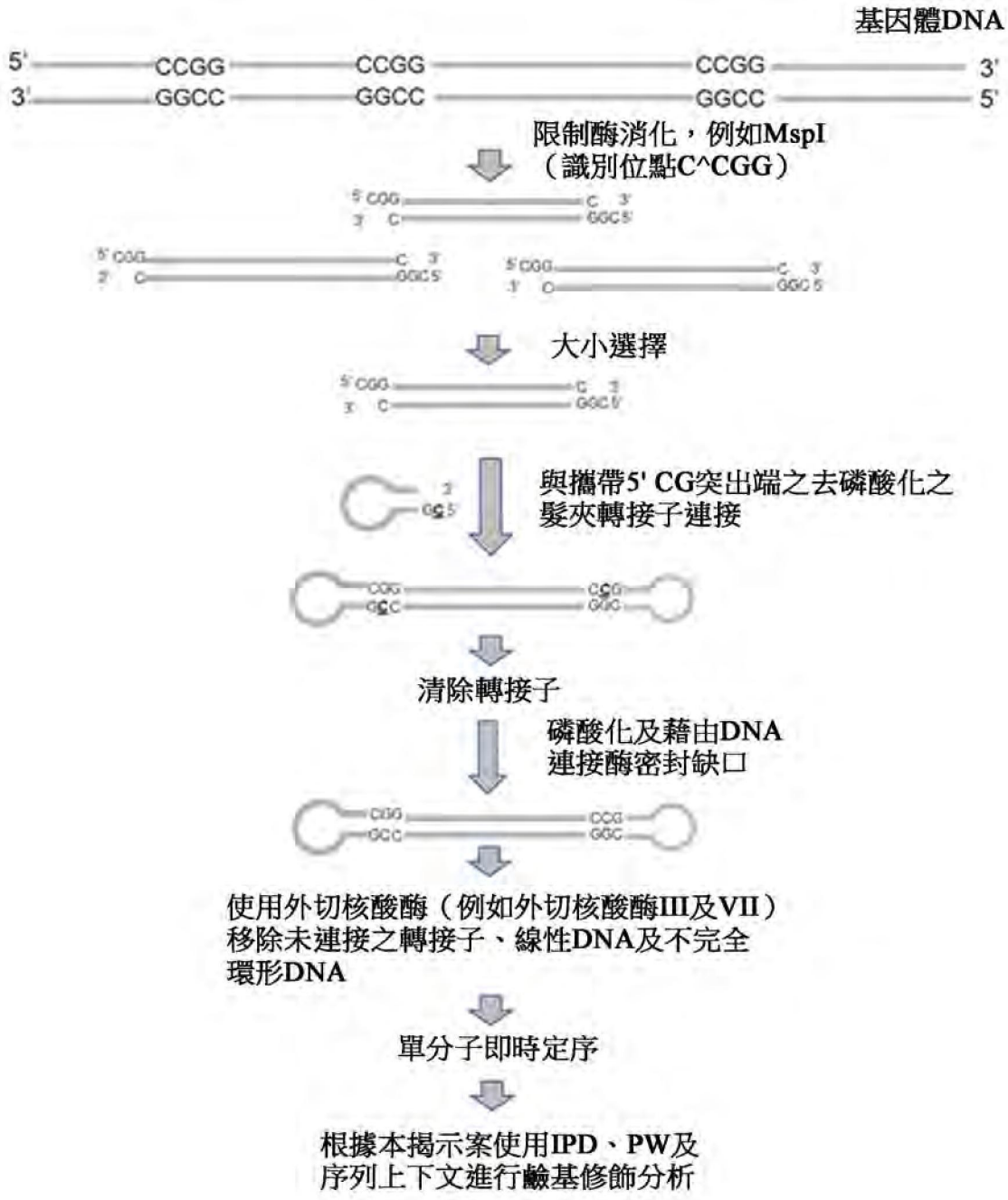
【圖128】


所覆蓋之CpG島內CpG位點的百分比（%）
0
20
40
60
80
100
200
400
600
800
1000
片段之大小（bp）
經AccII消化
經MspI消化
組合

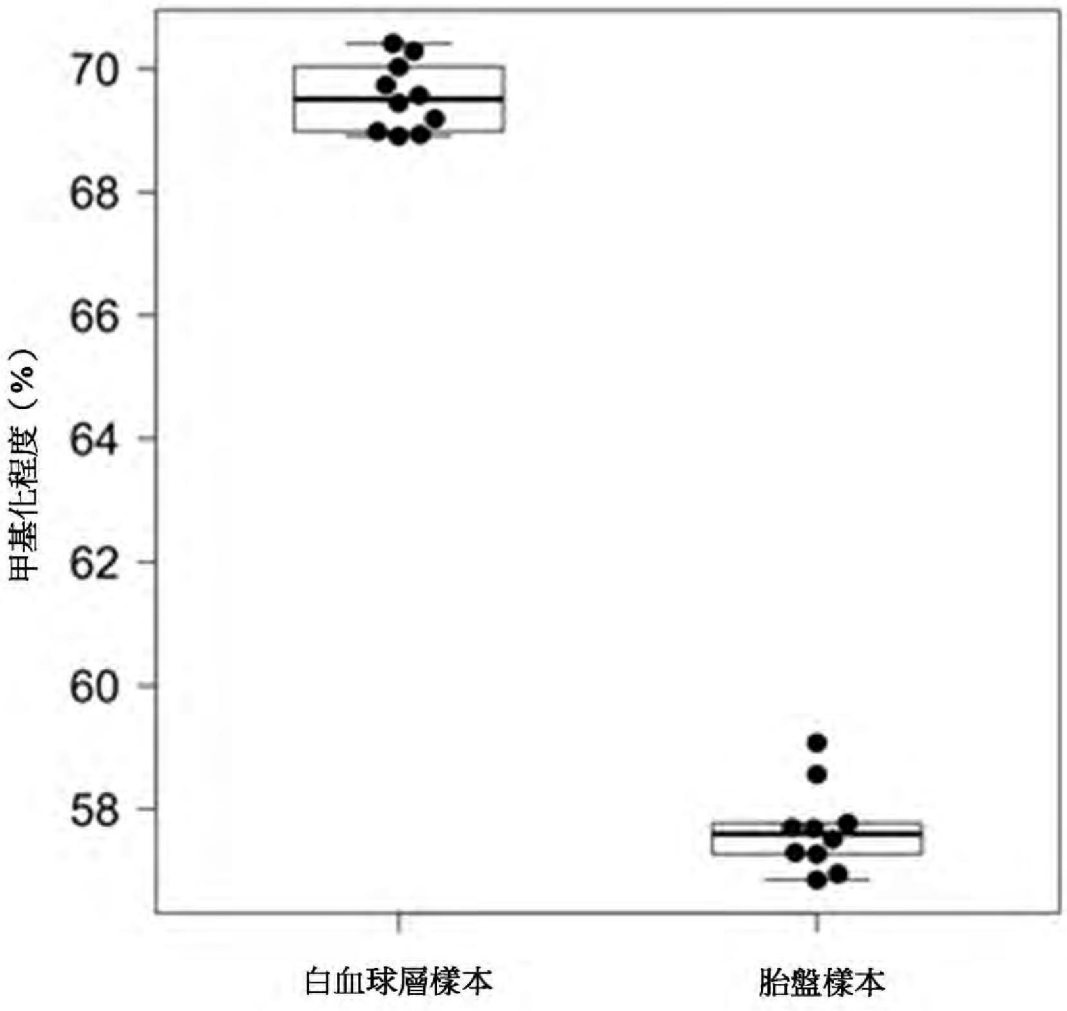
【圖129】


基因體DNA
5'
CCGG
CCGG
CCGG
3'
3'
GGCC
GGCC
GGCC
5'
限制酶消化，例如MspI
（識別位點C^CGG）
大小選擇
與攜帶5' CG突出端之髮夾轉接子連接
清除轉接子或轉接子二聚體
使用外切核酸酶（例如外切核酸酶III及VII）
移除未連接之轉接子、線性DNA及不完全
環形DNA
單分子即時定序
根據本揭示案使用IPD、PW及序列
上下文進行鹼基修飾分析

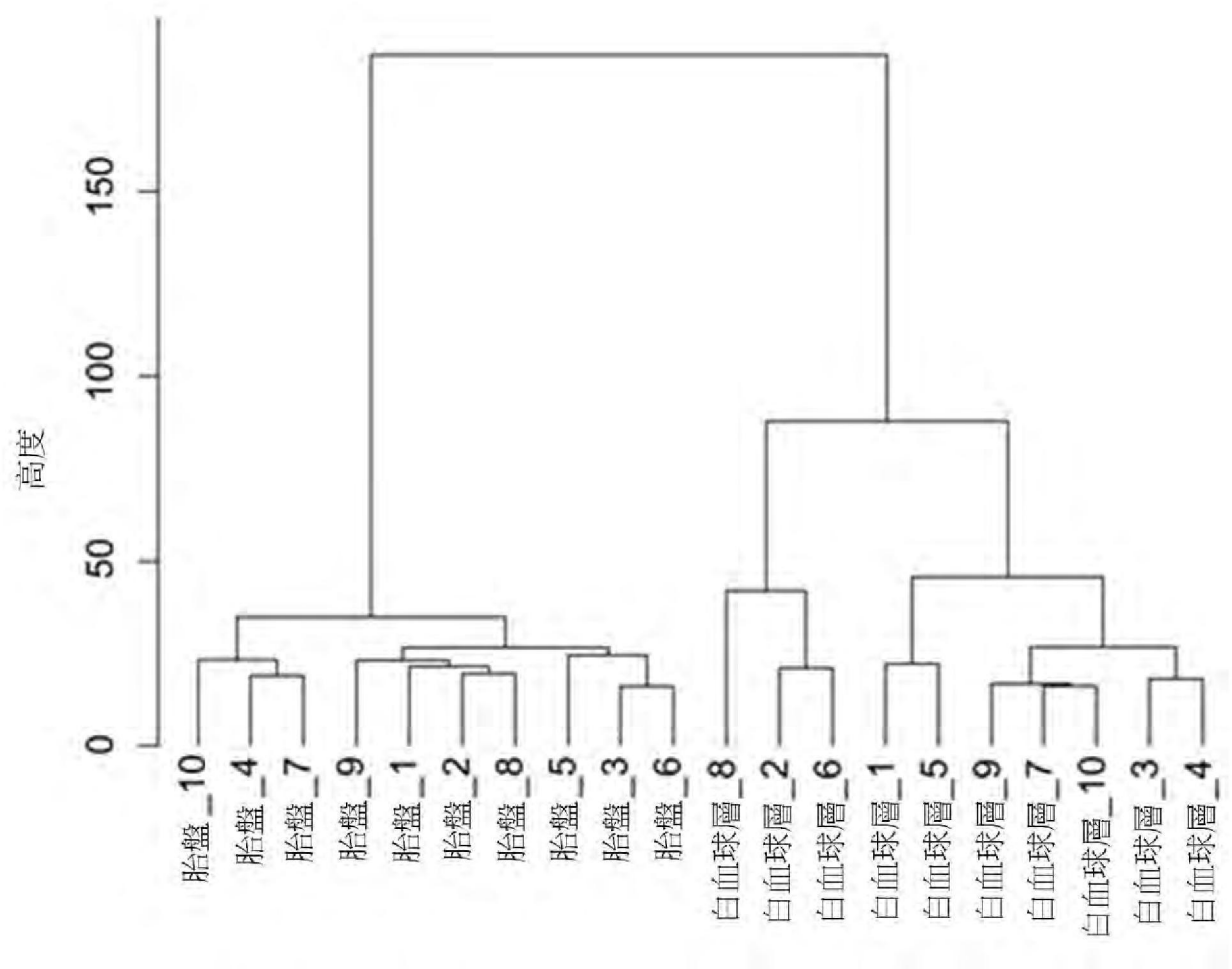
【圖130】


基因體DNA
5' CCGG CCGG CCGG 3'
3' GGCC GGCC GGCC 5'
限制酶消化，例如MspI
（識別位點C^CGG）
大小選擇
與攜帶5' CG突出端之去磷酸化之
髮夾轉接子連接
清除轉接子
磷酸化及藉由DNA
連接酶密封缺口
使用外切核酸酶（例如外切核酸酶III及VII）
移除未連接之轉接子、線性DNA及不完全
環形DNA
單分子即時定序
根據本揭示案使用IPD、PW及
序列上下文進行鹼基修飾分析

【圖131】



【圖132】


高度
0
50
100
150
胎盤_10
胎盤_4
胎盤_7
胎盤_9
胎盤_1
胎盤_2
胎盤_8
胎盤_5
胎盤_3
胎盤_6
白血球層_8
白血球層_2
白血球層_6
白血球層_1
白血球層_5
白血球層_9
白血球層_7
白血球層_10
白血球層_3
白血球層_4

【圖133】