I834642

(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 I834642 B**

(45)公告日： 中華民國 113 (2024) 年 03 月 11 日

(21)申請案號：108108527 (22)申請日： 中華民國 108 (2019) 年 03 月 13 日

(51)Int. Cl.： ***C12Q1/6886 (2018.01)***

(30)優先權：2018/03/13 美國 62/642,480

(71)申請人：美商格瑞爾有限責任公司 (美國) GRAIL, LLC (US)
美國

(72)發明人：古羅斯 山繆 S GROSS, SAMUEL S. (US)；達維多夫 康斯坦丁 DAVYDOV, KONSTANTIN (US)

(74)代理人：陳長文

(56)參考文獻：

WO 2017/106481A1

期刊 Wu et al., "Redefining CpG islands using hidden Markov models", Biostatistics, 2010, 11(3), pp 499-514.

審查人員：方冠岳

申請專利範圍項數：25 項 圖式數：11 共 95 頁

(54)名稱

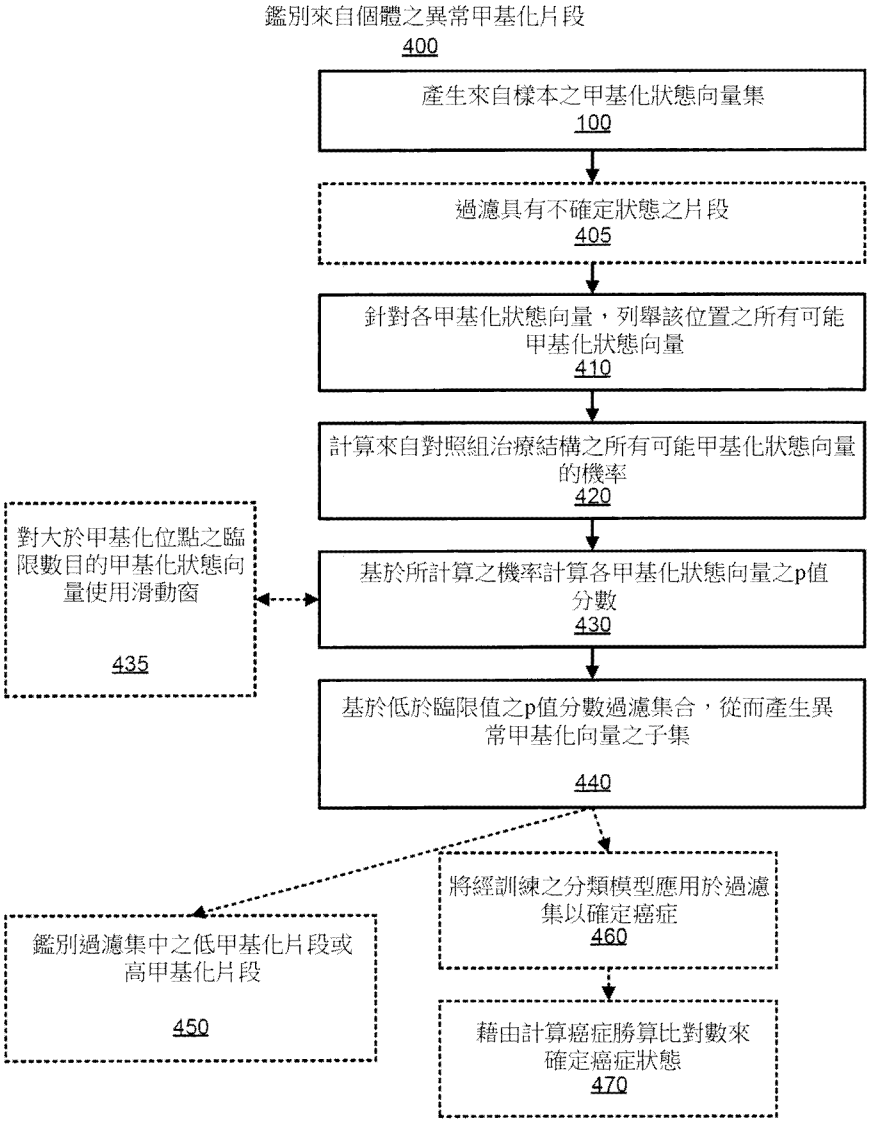
異常片段偵測及分類

(57)摘要

分析系統建立一種資料結構，該資料結構對健康對照組之甲基化向量字串進行計數。在給定來自個體之樣本片段的情況下，該分析系統列舉甲基化狀態向量之可能性，且利用馬可夫鏈機率(Markov chain probability)計算出所有可能性之機率。該分析系統藉由將小於或等於與測試甲基化狀態向量匹配之可能性之所計算機率的計算機率求和來產生該個體之測試甲基化狀態向量的 p 值分數。若該 p 值分數低於臨限分數，則該分析系統確定該測試甲基化狀態向量相較於健康對照組為異常甲基化的。在多個此類樣本片段存在下，該分析系統可以基於各 p 值分數來過濾樣本片段。該分析系統可以對過濾集運作分類模型以預測該個體是否患有癌症。

An analytics system creates a data structure counting strings of methylation vectors from a healthy control group. The analytics system enumerates possibilities of methylation state vectors given a sample fragment from a subject, and calculates probabilities for all possibilities with a Markov chain probability. The analytics system generates a p-value score for the subject's test methylation state vector by summing the calculated probabilities that are less than or equal to the calculated probability of the possibility matching the test methylation state vector. The analytics system determines the test methylation state vector to be anomalously methylated compared to the healthy control group if the p-value score is below a threshold score. With a number of such sample fragments, the analytics system can filter the sample fragments based on each p-value score. The analytics system can run a classification model on the filtered set to predict whether the subject has cancer.

指定代表圖：



【圖4】

符號簡單說明：

100・・・產生來自樣本之甲基化狀態向量集

400・・・鑑別來自個體之異常甲基化片段/自非癌症訓練組及癌症訓練組獲得異常片段

405・・・過濾具有不確定狀態之片段

410・・・針對各甲基化狀態向量，列舉該位置之所有可能甲基化狀態向量

420・・・計算來自對照組治療結構之所有可能甲基化狀態向量的機率

430・・・基於所計算之機率計算各甲基化狀態向量之 p 值分數

435・・・對大於甲基化位點之臨限數目的甲基化狀態向量使用滑動窗

440・・・基於低於臨限值之 p 值分數過濾集合，從而產生異常甲基化向量之子集

450・・・鑑別過濾集內之低甲基化片段或高甲基化片段

460・・・將經訓練之分類模型應用於過濾集以確定癌症

470・・・藉由計算癌症勝算比對數來確定癌症狀態

公告本

I834642

【發明摘要】

【中文發明名稱】

異常片段偵測及分類

【英文發明名稱】

ANOMALOUS FRAGMENT DETECTION AND CLASSIFICATION

【中文】

分析系統建立一種資料結構，該資料結構對健康對照組之甲基化向量字串進行計數。在給定來自個體之樣本片段的情況下，該分析系統列舉甲基化狀態向量之可能性，且利用馬可夫鏈機率(Markov chain probability)計算出所有可能性之機率。該分析系統藉由將小於或等於與測試甲基化狀態向量匹配之可能性之所計算機率的計算機率求和來產生該個體之測試甲基化狀態向量的p值分數。若該p值分數低於臨限分數，則該分析系統確定該測試甲基化狀態向量相較於健康對照組為異常甲基化的。在多個此類樣本片段存在下，該分析系統可以基於各p值分數來過濾樣本片段。該分析系統可以對過濾集運作分類模型以預測該個體是否患有癌症。

【英文】

An analytics system creates a data structure counting strings of methylation vectors from a healthy control group. The analytics system enumerates possibilities of methylation state vectors given a sample fragment from a subject, and calculates probabilities for all possibilities with a Markov chain probability. The analytics system generates a p-value score for the subject's test methylation state vector by summing

the calculated probabilities that are less than or equal to the calculated probability of the possibility matching the test methylation state vector. The analytics system determines the test methylation state vector to be anomalously methylated compared to the healthy control group if the p-value score is below a threshold score. With a number of such sample fragments, the analytics system can filter the sample fragments based on each p-value score. The analytics system can run a classification model on the filtered set to predict whether the subject has cancer.

【指定代表圖】

圖4

【代表圖之符號簡單說明】

100　產生來自樣本之甲基化狀態向量集

400　鑑別來自個體之異常甲基化片段/自非癌症訓練組及癌症訓練組獲得異常片段

405　過濾具有不確定狀態之片段

410　針對各甲基化狀態向量，列舉該位置之所有可能甲基化狀態向量

420　計算來自對照組治療結構之所有可能甲基化狀態向量的機率

430　基於所計算之機率計算各甲基化狀態向量之p值分數

435　對大於甲基化位點之臨限數目的甲基化狀態向量使用滑動窗

440　基於低於臨限值之p值分數過濾集合，從而產生異常甲基化向量之子集

450　鑑別過濾集內之低甲基化片段或高甲基化片段

460 將經訓練之分類模型應用於過濾集以確定癌症

470 藉由計算癌症勝算比對數來確定癌症狀態

【發明說明書】

【中文發明名稱】

異常片段偵測及分類

【英文發明名稱】

ANOMALOUS FRAGMENT DETECTION AND CLASSIFICATION

【技術領域】

【先前技術】

【0001】 DNA甲基化在調控基因表現方面起重要作用。異常DNA甲基化已牽涉到許多疾病過程，包括癌症。利用甲基化測序(例如全基因組亞硫酸氫鹽測序(WGBS))進行的DNA甲基化圖譜分析愈來愈公認為用於偵測、診斷及/或監測癌症之有價值的診斷工具。舉例而言，差異甲基化區域特異性模式及/或對偶基因特異性甲基化模式可適用作分子標記物以便利用循環的游離DNA進行非侵入性診斷。然而，此項技術中仍需要改良的方法來分析游離DNA之甲基化測序資料以便偵測、診斷及/或監測疾病，諸如癌症。

【發明內容】

【0002】 個體癌症之早期偵測因其允許較早治療且因此允許較大的生存機會而具有重要作用。對游離DNA (cfDNA)片段的測序及對片段中之胞嘧啶及鳥嘌呤(稱為CpG位點)之不同二核苷酸之甲基化狀態的分析能夠洞察個體是否患有癌症。為此，本說明書包括用於分析cfDNA片段之CpG位點之甲基化狀態的方法。具體言之，本發明提供一種鑑別具有或可能具有異常甲基化模式之cfDNA片段的方法。片段高頻率存在於無癌症個體中不大可能產生高區分度特點供癌症狀態分類用。因此，相對於來自健

康樣本(例如無癌症個體)的cfDNA片段，鑑別出具有異常甲基化模式的cfDNA片段對於選擇可作為以低雜訊偵測癌症特異性甲基化模式之指標的cfDNA片段具有重要作用。在低雜訊區域當中，可選擇在區分癌症患者與健康個體或患有其他健康病狀之個體方面具有最大資訊性之基因組區域所衍生的cfDNA片段。癌症患者與健康個體之間的區分可利用分類器執行，該分類器已用獲自患有癌症之個體的甲基化測序資料及/或未患癌症之個體的甲基化測序資料訓練。另外提供驗證資料，其證實利用本文所述方法鑑別出之異常甲基化cfDNA片段的分析可用於以高靈敏性及特異性偵測癌症。

【0003】 在一個實施例中，包括複數個cfDNA片段的測試樣本係自對照組個體獲得。處理測試樣本中之複數個cfDNA片段以將未甲基化胞嘧啶轉化成尿嘧啶，對cfDNA片段測序且與參考基因組進行比較以鑑別多個CpG位點中之每一者的甲基化狀態。分析系統建立一種資料結構，其針對參考基因組中之所鑑別出的各CpG位點，在該CpG位點開始對來自對照組的片段數目進行計數，該資料結構具有甲基化對未甲基化之CpG位點之一些數字的特定甲基化字串。

【0004】 該分析系統針對各測序片段產生甲基化狀態向量，其中該甲基化狀態向量包含位於片段中的CpG位點以及其甲基化狀態，例如甲基化、未甲基化或不確定。對於每一個片段而言，該分析系統利用機率性分析及對照組資料結構鑑別所觀測之指定片段(或其一部分)在該片段之CpG位點觀測到有甲基化狀態之意外性。在一個特定實施例中，此機率性分析係列舉參考基因組內與指定片段(或其一部分)具有相同長度(位點中)及位置之甲基化狀態向量的替代可能性，且利用得自資料結構的計數來確定各

此類可能性的機率。該分析系統可以利用馬可夫鏈機率分析(以及馬可夫鏈的指定最大階數)對各此類甲基化狀態向量可能性之機率進行模型化。計算出甲基化狀態向量之各種可能性的機率之後，該分析系統藉由將小於與測試甲基化狀態向量匹配之可能性之機率的甲基化狀態向量可能性之彼等機率求和來產生片段之p值分數。該分析系統對照臨限值比較所產生的p值，以鑑別出相對於對照組發生異常甲基化的cfDNA片段(在本文中亦稱為具有異常甲基化模式的片段)。

【0005】 除上述分析系統之外，分類器有助於基於機率對患有癌症或不患有癌症的個體進行分類。分類器經獲自患有癌症之個體的甲基化測序資料及/或不患有癌症之個體的甲基化測序資料訓練。在進行測序及針對各經測序之cfDNA片段產生甲基化狀態向量之後，利用cfDNA片段訓練分類器，該等cfDNA片段經鑑別相較於健康對照組為低甲基化的或高甲基化的。如本文所用，「低甲基化」cfDNA片段可定義為具有至少5個CpG位點、其中至少80% CpG位點未甲基化的片段。類似地，「高甲基化」cfDNA片段可定義為具有至少5個CpG位點、其中至少80% CpG位點發生甲基化的片段。接著，分類器跑遍基因組中之每一個CpG位點且計算低甲基化分數及高甲基化分數。兩種分數計算方式類似。對於低甲基化分數而言，分類器計算含有當前CpG位點之視為低甲基化之癌症片段相對於含有當前CpG位點之視為低甲基化之所有片段(癌症及非癌症)的比率。各CpG位點的高甲基化分數類似地藉由求得視為高甲基化之癌症片段相對於視為高甲基化之所有片段的比率來計算。

【0006】 現分類器自訓練組獲取個體以及其複數個cfDNA片段且對該等片段進行測序以產生甲基化狀態向量。分類器利用該個體之各甲基化

狀態向量計算高甲基化總分及低甲基化總分。各總分係基於各個CpG位點的低甲基化分數及高甲基化分數計算。分類器接著依據個體甲基化狀態向量的低甲基化總分對其進行排序且亦依據其高甲基化總分進行排序。分類器利用該兩種排序、藉由自排序中選擇分數來產生該個體的特徵向量。接著訓練分類器以將對應於非癌症訓練組的特徵向量與對應於癌症訓練組的特徵向量區分開來。在一個實施例中，分類器使用L2正則化內核邏輯回歸與高斯(Gaussian)徑向基底函數內核(RBF內核)。

【0007】 相應地，在一個態樣中，本發明提供一種用於偵測無細胞去氧核糖核酸(cfDNA)樣本片段中之異常甲基化模式的方法，該方法包含：存取資料結構，該資料結構包含參考基因組內之CpG位點之字串計數及其來自訓練片段集之各別甲基化狀態；針對樣本片段產生樣本狀態向量，該樣本片段包含參考基因組內的樣本基因組位置以及該樣本片段中之複數個CpG位點中之每一者的甲基化狀態，各甲基化狀態經測定為甲基化的或未甲基化的；列舉長度與樣本狀態向量相同之樣本基因組位置之甲基化狀態的複數種可能性；針對各種可能性，藉由存取資料結構中所儲存的計數來計算機率；鑑別出與樣本狀態向量匹配的可能性且相應地將所計算之機率鑑別為樣本機率；基於樣本機率，產生樣本狀態向量之樣本片段相對於訓練片段集的分數；以及基於所產生的分數確定樣本片段是否具有異常甲基化模式。

【0008】 在一些實施例中，各CpG位點字串包含參考基因組內之複數個基因組位置之CpG位點中之每一者的甲基化狀態，其中各甲基化狀態經測定為甲基化的或未甲基化的。

【0009】 在一些實施例中，該方法進一步包含：利用訓練片段集構

建資料結構且包含：針對該訓練片段集內之各訓練片段，產生訓練狀態向量，該訓練狀態向量包含參考基因組內之已知基因組位置及訓練片段中之複數個CpG位點中之每一者的甲基化狀態，各甲基化狀態經測定為甲基化的或未甲基化的；確定複數個字串，其中各字串為訓練狀態向量的一部分；量化來自訓練狀態向量之各字串的計數；以及將各字串之複數個計數儲存於資料結構中。

【0010】 在一些實施例中，基於所產生之分數確定樣本片段是否具有異常甲基化模式的步驟進一步包含確定針對樣本片段所產生之分數是否低於臨限分數，其中臨限分數指示樣本片段具有異常甲基化模式之信賴度。在一些實施例中，臨限分數為0.1或更小。

【0011】 在一些實施例中，訓練片段集包含來自一或多個健康個體的訓練片段，其中該一或多個健康個體缺乏特定醫學病症且其中相對於來自一或多個健康個體的訓練片段集確定樣本片段為異常甲基化的。

【0012】 在一些實施例中，針對樣本片段產生分數包含：鑑別出小於樣本機率的甲基化狀態可能性之所計算機率；以及針對樣本片段、藉由將所有鑑別出之機率與樣本機率求和來產生分數。在一些實施例中，藉由存取資料結構中所儲存之關於每種可能性的計數來計算機率的步驟包含：對於複數個條件元中之每一者(其中各條件元為考慮存在可能性之CpG位點子集的條件性機率)，利用資料結構中所儲存之複數個計數、藉由包含以下之步驟計算某一階數之馬可夫鏈機率：鑑別與該條件元匹配之字串數的第一計數；鑑別與該條件元之先前甲基化狀態直至全數目長度匹配之字串數的第二計數；以及藉由將第一計數除以第二計數來計算馬可夫鏈機率。在一些實施例中，階數選自由以下組成之群：1、2、3、4、5、6、

7、8、9、10、11、12、13、14及15。在一些實施例中，利用資料結構中所存儲之複數個計數計算某一階數之馬可夫鏈機率的步驟進一步包含執行平滑算法。

【0013】 第一窗口及第二窗口為樣本片段之兩個不同部分；其中鑑別與樣本狀態向量匹配的可能性及相應地將所計算之機率鑑別為樣本機率包含利用與第一窗口匹配的第一樣本機率鑑別第一可能性及利用與第二窗口匹配的第二樣本機率鑑別第二可能性；且其中所產生的分數係基於第一樣本機率及第二樣本機率之一。

【0014】 在一些實施例中，該方法進一步包含基於針對各樣本片段所產生的分數來過濾複數個樣本片段，從而產生具有異常甲基化模式之樣本片段的子集。

【0015】 在一些實施例中，該方法進一步包含當樣本片段包含至少臨限數目個CpG位點、其中超過臨限百分比之CpG位點發生甲基化時，將該樣本片段鑑別為高甲基化的。在一些實施例中，臨限數目個CpG位點為5個或更多個CpG位點，且其中甲基化CpG位點之臨限百分比為80%或更大。在一些實施例中，該方法進一步包含當樣本片段包含至少臨限數目個CpG位點、其中超過臨限百分比之CpG位點未甲基化時，將該樣本片段鑑別為低甲基化的。在一些實施例中，臨限數目個CpG位點為5個或更多個CpG位點，且其中未甲基化CpG位點之臨限百分比為80%或更大。

【0016】 在一些實施例中，該方法進一步包含：將樣本狀態向量應用於分類器，該分類器經來自一或多個患有癌症之個體的癌症訓練片段集及來自一或多個未患癌症之個體的非癌症訓練片段集訓練，其中分類器可以用於確定樣本片段是否來自患有癌症之個體。在一些實施例中，將樣本

狀態向量應用於分類器產生癌症機率及非癌症機率中的至少一者。在一些實施例中，該方法進一步包含基於癌症機率及非癌症機率中之至少一者產生癌症狀態分數。

【0017】 在另一態樣中，本發明提供一種確定測試個體是否患有癌症的方法，該方法包含：存取模型，該模型係由訓練程序利用來自一或多個患有癌症之訓練個體的癌症片段集及來自一或多個未患癌症之訓練個體的非癌症片段集所獲得，其中癌症片段集與非癌症片段集均包含複數個訓練片段，其中該訓練程序包含：對於各訓練片段而言，確定該訓練片段是否為低甲基化的或高甲基化的，其中低甲基化及高甲基化訓練片段中之每一者包含至少臨限數目個CpG位點，其中至少臨限百分比之CpG位點分別為未甲基化的或甲基化的，對於參考基因組中之複數個CpG位點中之每一者而言：對與CpG位點重疊之低甲基化訓練片段的計數及與CpG位點重疊之高甲基化訓練片段的計數進行量化；以及基於低甲基化訓練片段及高甲基化訓練片段之計數產生低甲基化分數及高甲基化分數；對於各訓練片段而言，基於訓練片段中之CpG位點的低甲基化分數產生低甲基化總分且基於訓練片段中之CpG位點的高甲基化分數產生高甲基化總分；對於各訓練個體而言：基於低甲基化總分對複數個訓練片段排序且基於高甲基化總分對複數個訓練片段排序；以及基於訓練片段之排序來產生特徵向量；獲得一或多個未患癌症之訓練個體的訓練特徵向量及一或多個患有癌症之訓練個體的訓練特徵向量；以及用一或多個未患癌症之訓練個體的特徵向量及一或多個患有癌症之訓練個體的特徵向量訓練該模型；以及將該模型應用於與測試個體對應的測試特徵向量以確定測試個體是否患有癌症。

【0018】 在一些實施例中，臨限數目為五或更大。在一些實施例

中，臨限百分比為80%或更大。

【0019】 在一些實施例中，針對參考基因組中的各CpG位點量化與該CpG位點重疊之低甲基化訓練片段的計數及與該CpG位點重疊之高甲基化訓練片段的計數進一步包含：對來自一或多個患有癌症之訓練個體之與該CpG位點重疊之癌症低甲基化訓練片段的計數及來自一或多個未患癌症之訓練個體之與該CpG位點重疊之非癌症低甲基化訓練片段的計數進行量化；以及對來自一或多個患有癌症之訓練個體之與該CpG位點重疊之癌症高甲基化訓練片段的計數及來自一或多個未患癌症之訓練個體之與該CpG位點重疊之非癌症高甲基化訓練片段的計數進行量化。

【0020】 在一些實施例中，針對參考基因組中的各CpG位點，基於低甲基化訓練片段及高甲基化訓練片段之計數產生低甲基化分數及高甲基化分數進一步包含：為了產生低甲基化分數，計算癌症低甲基化訓練片段計數相對於癌症低甲基化訓練片段計數與非癌症低甲基化訓練片段計數之低甲基化總和的低甲基化比率；以及為了產生高甲基化分數，計算癌症高甲基化訓練片段計數相對於癌症高甲基化訓練片段計數與非癌症高甲基化訓練片段計數之高甲基化總和的高甲基化比率。在一些實施例中，低甲基化及高甲基化比率進一步利用平滑算法計算。

【0021】 在一些實施例中，針對參考基因組中之各CpG位點，基於低甲基化訓練片段及高甲基化訓練片段之計數產生低甲基化分數及高甲基化分數進一步包含：為了產生低甲基化分數，計算癌症低甲基化訓練片段計數相對於非癌症低甲基化訓練片段計數的低甲基化比率對數；以及為了產生高甲基化分數，計算癌症高甲基化訓練片段計數相對於非癌症高甲基化訓練片段計數的高甲基化比率對數。在一些實施例中，低甲基化及高甲

基化比率進一步利用平滑算法計算。在一些實施例中，針對各訓練片段、基於該訓練片段中之CpG位點的低甲基化分數產生低甲基化總分及基於該訓練片段中之CpG位點的高甲基化分數產生高甲基化總分進一步包含將該訓練片段中之CpG位點的低甲基化最高分鑑別為低甲基化總分以及將該訓練片段中之CpG位點的高甲基化最高分鑑別為高甲基化總分。

【0022】 在一些實施例中，基於訓練片段之排序、針對各訓練個體產生訓練特徵向量進一步包含鑑別來自排序中的複數個低甲基化總分及來自排序中的複數個高甲基化總分以及產生包含複數個低甲基化總分及複數個高甲基化分數的訓練特徵向量。

【0023】 在一些實施例中，用來自一或多個未患癌症之訓練個體的訓練特徵向量及來自一或多個患有癌症之訓練個體的訓練特徵向量訓練模型係藉由非線性分類器訓練。

【0024】 在一些實施例中，對於各訓練個體而言，藉由該訓練個體之訓練片段的平均長度將訓練特徵向量標準化。在一些實施例中，該方法進一步包含獲得與測試個體對應之測試特徵向量的步驟，其中獲得測試特徵向量的步驟包含：自測試個體獲得測試片段集之序列讀段；針對各測試片段，確定該測試片段是否為低甲基化的或高甲基化的，其中低甲基化及高甲基化測試片段中之每一者包含至少臨限數目個CpG位點，其中至少臨限百分比之CpG位點分別為未甲基化的或甲基化的，針對參考基因組中之複數個CpG位點中之每一者：對與CpG位點重疊之低甲基化測試片段的計數及與CpG位點重疊之高甲基化測試片段的計數進行量化；以及基於低甲基化測試片段及高甲基化測試片段之計數來產生低甲基化分數及高甲基化分數；針對各測試片段，基於測試片段中之CpG位點的低甲基化分數產生

低甲基化總分且基於測試片段中之CpG位點的高甲基化分數產生高甲基化總分；針對測試個體，基於低甲基化總分對複數個測試片段排序且基於高甲基化總分對複數個測試片段排序；以及基於測試片段之排序來產生測試特徵向量。

【0025】 在一些實施例中，將該模型應用於測試個體之測試特徵向量以確定測試個體是否患有癌症包含：基於該模型產生該測試個體之癌症機率；及將該癌症機率與臨限機率進行比較以確定該測試個體是否患有癌症。

【0026】 在一些實施例中，診斷模型包含內核邏輯回歸分類器。

【0027】 在又另一態樣中，本發明提供一種用於確定測試個體是否患有癌症的方法，該方法包含：存取模型，該模型係由訓練程序利用來自一或多個患有癌症之訓練個體的癌症訓練片段集及來自一或多個未患癌症之訓練個體的非癌症訓練片段集所獲得，其中癌症訓練片段集與非癌症訓練片段集包含複數個訓練片段，其中訓練程序包含：針對各訓練片段，確定該訓練片段是否為低甲基化的或高甲基化的，其中低甲基化及高甲基化訓練片段中之每一者包含至少臨限數目個CpG位點，其中至少臨限百分比之CpG位點分別為未甲基化的或甲基化的，針對各訓練個體、基於低甲基化訓練片段及高甲基化訓練片段產生訓練特徵向量，以及用來自一或多個未患癌症之訓練個體的訓練特徵向量及來自一或多個患有癌症之訓練個體的特徵向量訓練該模型；以及將該模型應用於與測試個體對應的測試特徵向量以確定測試個體是否患有癌症。

【0028】 在一些實施例中，針對各訓練個體來產生訓練特徵向量包含：針對參考基因組中之複數個CpG位點中之每一者：對與CpG位點重疊

之低甲基化訓練片段的計數及與CpG位點重疊之高甲基化訓練片段的計數進行量化；以及基於低甲基化訓練片段及高甲基化訓練片段之計數來產生低甲基化分數及高甲基化分數；針對訓練個體之各訓練片段，基於訓練片段中之CpG位點的低甲基化分數產生低甲基化總分且基於訓練片段中之CpG位點的高甲基化分數產生高甲基化總分；以及基於低甲基化總分對訓練個體之複數個訓練片段排序且基於高甲基化總分對該訓練個體之複數個訓練片段排序，其中訓練個體之訓練特徵向量係依據基於低甲基化總分的排序及基於高甲基化總分的排序。

【0029】 在一些實施例中，該方法進一步包含獲得與測試個體對應之測試特徵向量的步驟，其中獲得測試特徵向量的步驟包含：自測試個體獲得測試片段集之序列讀段；針對各測試片段，確定該測試片段是否為低甲基化的或高甲基化的，其中低甲基化及高甲基化測試片段中之每一者包含至少臨限數目個CpG位點，其中至少臨限百分比之CpG位點分別為未甲基化的或甲基化的，針對參考基因組中之複數個CpG位點中之每一者：對與CpG位點重疊之低甲基化測試片段的計數及與CpG位點重疊之高甲基化測試片段的計數進行量化；以及基於低甲基化測試片段及高甲基化測試片段之計數來產生低甲基化分數及高甲基化分數；針對各測試片段，基於測試片段中之CpG位點的低甲基化分數產生低甲基化總分且基於測試片段中之CpG位點的高甲基化分數產生高甲基化總分；針對測試個體，基於低甲基化總分對複數個測試片段排序且基於高甲基化總分對複數個測試片段排序；以及基於測試片段之排序來產生測試特徵向量。在一些實施例中，將該模型應用於測試個體之測試特徵向量以確定測試個體是否患有癌症包含：基於該模型產生該測試個體之癌症機率；及將該癌症機率與臨限機率

進行比較以確定該測試個體是否患有癌症。在一些實施例中，診斷模型包含內核邏輯回歸分類器。

【0030】 在一個態樣中，本發明提供一種用於預測來自懷疑患有癌症之測試個體之測試片段是否具有異常甲基化模式的方法，該方法包含：存取資料結構，該資料結構包含參考基因組內之CpG位點字串之計數及得自訓練片段集之其各別甲基化狀態；針對測試片段產生測試狀態向量，其中測試狀態向量包含參考基因組內之測試基因組位置及測試片段中之複數個CpG位點中之每一者的甲基化狀態，其中各甲基化狀態經測定為以下中之一者：甲基化、未甲基化及不確定；基於資料結構中所儲存之計數，針對測試狀態向量計算測試機率；對長度與測試狀態向量相同之測試基因組位置之可能甲基化狀態向量的子集取樣；針對所取樣之可能甲基化狀態向量中之每一者，至少部分地基於資料結構中所儲存之計數來計算與所取樣之可能甲基化狀態向量對應的機率；計算所取樣之可能甲基化狀態向量的比例，該等可能甲基化狀態向量對應的所計算機率小於或等於測試機率；基於所計算的比例，針對測試片段產生估算分數；以及基於估算分數確定測試片段是否可能具有異常甲基化模式。

【0031】 在一些實施例中，該方法進一步包含：藉由比較估算分數與臨限分數來過濾測試片段，該臨限分數經選擇以使得與低於臨限分數之估算分數有關的測試片段更可能包括異常甲基化模式。在一些實施例中，該方法進一步包含：回應於確定測試片段可能具有異常甲基化模式，計算測試狀態向量之測試片段相對於訓練片段集的詳盡分數，其中該詳盡分數係基於測試機率及複數個可能甲基化狀態向量的機率；以及基於詳盡分數來確定測試片段是否具有異常甲基化模式。

【0032】 在一些實施例中，該方法進一步包含：將分類器應用於測試狀態向量，該分類器經來自患有癌症之一或多個訓練個體的第一訓練片段集及來自未患癌症之一或多個訓練個體的第二訓練片段集訓練，其中分類器可以用於確定測試個體是否患有癌症。

【0033】 在另一態樣中，本發明提供一種非暫時性電腦可讀儲存媒體，其儲存用於偵測游離去氧核糖核酸(cfDNA)樣本片段之異常甲基化模式的執行指令，該等指令當由硬體處理器執行時，促使硬體處理器執行包含以下之步驟：存取資料結構，該資料結構包含參考基因組內之CpG位點字串之計數及得自訓練片段集之其各別甲基化狀態；針對樣本片段產生包含參考基因組內之樣本基因組位置及樣本片段中之複數個CpG位點中之每一者之甲基化狀態的樣本狀態向量，各甲基化狀態經測定為甲基化的或未甲基化的；列舉長度與樣本狀態向量相同之樣本基因組位置之甲基化狀態的複數種可能性；針對每種可能性，藉由存取資料結構中所存儲之計數來計算機率；鑑別與樣本狀態向量匹配的可能性且相應地將所計算的機率鑑別為樣本機率；基於樣本機率，相對於訓練片段集產生樣本狀態向量之樣本片段的分數；以及基於所產生的分數確定樣本片段是否具有異常甲基化模式。

【0034】 在一些實施例中，如技術方案44之非暫時性電腦可讀儲存媒體，其中CpG位點字串中之每一者在參考基因組內之複數個基因組位置包含各CpG位點之甲基化狀態，其中各甲基化狀態經測定為甲基化的或未甲基化的。

【0035】 在一些實施例中，該等步驟進一步包含：利用訓練片段集構建資料結構且包含：針對訓練片段集內之各訓練片段，產生訓練狀態向

量，該訓練狀態向量包含參考基因組內之已知基因組位置及訓練片段中之複數個CpG位點中之每一者的甲基化狀態，各甲基化狀態經測定為甲基化的或未甲基化的；確定複數個字串，其中各字串為訓練狀態向量的一部分；量化來自訓練狀態向量之各字串的計數；以及將各字串之複數個計數儲存於資料結構中。

【0036】 在一些實施例中，基於所產生之分數確定樣本片段是否具有異常甲基化模式的步驟進一步包含確定針對樣本片段所產生之分數是否低於臨限分數，其中臨限分數指示樣本片段具有異常甲基化模式之信賴度。在一些實施例中，臨限分數為0.1或更小。

【0037】 在一些實施例中，訓練片段集包含來自一或多個健康個體的訓練片段，其中該一或多個健康個體缺乏特定醫學病症且其中相對於來自一或多個健康個體的訓練片段集確定樣本片段為異常甲基化的。

【0038】 在一些實施例中，針對樣本片段產生分數包含：鑑別出小於樣本機率的甲基化狀態可能性之所計算機率；以及針對樣本片段、藉由將所有鑑別出之機率與樣本機率求和來產生分數。

【0039】 在一些實施例中，藉由存取資料結構中所儲存之關於每種可能性的計數來計算機率的步驟包含：針對複數個條件元中之每一者(其中各條件元為考慮存在可能性之CpG位點子集的條件性機率)，利用資料結構中所儲存之複數個計數、藉由包含以下之步驟計算某一階數之馬可夫鏈機率：鑑別與該條件元匹配之字串數的第一計數；鑑別與該條件元之先前甲基化狀態直至全數目長度匹配之字串數的第二計數；以及藉由將第一計數除以第二計數來計算馬可夫鏈機率。在一些實施例中，階數選自由以下組成之群：1、2、3、4、5、6、7、8、9、10、11、12、13、14及

15。在一些實施例中，利用資料結構中所儲存之複數個計數計算某一階數之馬可夫鏈機率的步驟進一步包含執行平滑算法。

【0040】 第一窗口及第二窗口為樣本片段之兩個不同部分；其中鑑別與樣本狀態向量匹配的可能性及相應地將所計算之機率鑑別為樣本機率包含利用與第一窗口匹配的第一樣本機率鑑別第一可能性及利用與第二窗口匹配的第二樣本機率鑑別第二可能性；且其中所產生的分數係基於第一樣本機率及第二樣本機率之一。

【0041】 在一些實施例中，該等步驟進一步包含基於針對各樣本片段所產生的分數來過濾複數個樣本片段，從而產生具有異常甲基化模式之樣本片段的子集。

【0042】 在一些實施例中，該等步驟進一步包含當樣本片段包含至少臨限數目個CpG位點、其中超過臨限百分比之CpG位點發生甲基化時，將該樣本片段鑑別為高甲基化的。在一些實施例中，臨限數目個CpG位點為5個或更多個CpG位點，且其中甲基化CpG位點之臨限百分比為80%或更大。

【0043】 在一些實施例中，該等步驟進一步包含當樣本片段包含至少臨限數目個CpG位點、其中超過臨限百分比之CpG位點未甲基化時，將該樣本片段鑑別為低甲基化的。在一些實施例中，臨限數目個CpG位點為5個或更多個CpG位點，且其中未甲基化CpG位點之臨限百分比為80%或更大。

【0044】 在一些實施例中，該等步驟進一步包含：將樣本狀態向量應用於分類器，該分類器經來自一或多個患有癌症之個體的癌症訓練片段集及來自一或多個未患癌症之個體的非癌症訓練片段集訓練，其中該分類

器可以用於確定樣本片段是否來自患有癌症之個體。在一些實施例中，將樣本狀態向量應用於分類器產生癌症機率及非癌症機率中的至少一者。在一些實施例中，該等步驟進一步包含基於癌症機率及非癌症機率中之至少一者產生癌症狀態分數。

【0045】 在又另一態樣中，本發明提供一種非暫時性電腦可讀儲存媒體，其儲存用於確定測試個體是否患有癌症的執行指令，該等指令當由硬體處理器執行時，促使硬體處理器執行包含以下的步驟：存取模型，該模型係由訓練程序利用來自一或多個患有癌症之訓練個體的癌症片段集及來自一或多個未患癌症之訓練個體的非癌症片段集所獲得，其中癌症片段集與非癌症片段集均包含複數個訓練片段，其中該訓練程序包含：針對各訓練片段，確定該訓練片段是否為低甲基化的或高甲基化的，其中低甲基化訓練片段及高甲基化訓練片段中之每一者包含至少臨限數目個CpG位點，其中至少臨限百分比之CpG位點分別為未甲基化或甲基化的，針對參考基因組中之複數個CpG位點中之每一者：對與CpG位點重疊之低甲基化訓練片段的計數及與CpG位點重疊之高甲基化訓練片段的計數進行量化；以及基於低甲基化訓練片段及高甲基化訓練片段之計數來產生低甲基化分數及高甲基化分數；針對各訓練片段，基於訓練片段中之CpG位點的低甲基化分數產生低甲基化總分及基於訓練片段中之CpG位點的高甲基化分數產生高甲基化總分；針對各訓練個體：基於低甲基化總分對複數個訓練片段排序且基於高甲基化總分對複數個訓練片段排序；以及基於訓練片段之排序來產生特徵向量；獲得一或多個未患癌症之訓練個體的訓練特徵向量及一或多個患有癌症之訓練個體的訓練特徵向量；用一或多個未患癌症之訓練個體的訓練特徵向量及一或多個患有癌症之訓練個體的訓練特徵向量

訓練該模型；以及將該模型應用於與該測試個體對應的測試特徵向量以確定該測試個體是否患有癌症。在一些實施例中，臨限數目為五或更大。在一些實施例中，臨限百分比為80%或更大。

【0046】 在一些實施例中，針對參考基因組中的各CpG位點量化與該CpG位點重疊之低甲基化訓練片段的計數及與該CpG位點重疊之高甲基化訓練片段的計數進一步包含：對來自一或多個患有癌症之訓練個體之與該CpG位點重疊之癌症低甲基化訓練片段的計數及來自一或多個未患癌症之訓練個體之與該CpG位點重疊之非癌症低甲基化訓練片段的計數進行量化；以及對來自一或多個患有癌症之訓練個體之與該CpG位點重疊之癌症高甲基化訓練片段的計數及來自一或多個未患癌症之訓練個體之與該CpG位點重疊之非癌症高甲基化訓練片段的計數進行量化。在一些實施例中，針對參考基因組中的各CpG位點，基於低甲基化訓練片段及高甲基化訓練片段之計數產生低甲基化分數及高甲基化分數進一步包含：為了產生低甲基化分數，計算癌症低甲基化訓練片段計數相對於癌症低甲基化訓練片段計數與非癌症低甲基化訓練片段計數之低甲基化總和的低甲基化比率；以及為了產生高甲基化分數，計算癌症高甲基化訓練片段計數相對於癌症高甲基化訓練片段計數與非癌症高甲基化訓練片段計數之高甲基化總和的高甲基化比率。

【0047】 在一些實施例中，低甲基化及高甲基化比率進一步利用平滑算法計算。在一些實施例中，基於低甲基化訓練片段及高甲基化訓練片段之計數產生低甲基化分數及高甲基化分數進一步包含：為了產生低甲基化分數，計算癌症低甲基化訓練片段計數相對於非癌症低甲基化訓練片段計數的低甲基化比率對數；以及為了產生高甲基化分數，計算癌症高甲基

化訓練片段計數相對於非癌症高甲基化訓練片段計數的高甲基化比率對數。在一些實施例中，低甲基化及高甲基化比率進一步利用平滑算法計算。在一些實施例中，針對各訓練片段、基於該訓練片段中之CpG位點的低甲基化分數產生低甲基化總分及基於該訓練片段中之CpG位點的高甲基化分數產生高甲基化總分進一步包含將該訓練片段中之CpG位點的低甲基化最高分鑑別為低甲基化總分以及將該訓練片段中之CpG位點的高甲基化最高分鑑別為高甲基化總分。

【0048】 在一些實施例中，基於訓練片段之排序、針對各訓練個體產生訓練特徵向量進一步包含鑑別來自排序中的複數個低甲基化總分及來自排序中的複數個高甲基化總分以及產生包含複數個低甲基化總分及複數個高甲基化分數的訓練特徵向量。

【0049】 在一些實施例中，用來自一或多個未患癌症之訓練個體的訓練特徵向量及來自一或多個患有癌症之訓練個體的訓練特徵向量訓練模型係藉由非線性分類器訓練。

【0050】 在一些實施例中，該等步驟進一步包含針對各訓練個體，藉由該訓練個體之訓練片段的平均長度將訓練特徵向量標準化。

【0051】 在一些實施例中，該等步驟進一步包含：獲得與測試個體對應之測試特徵向量，其中獲得測試特徵向量的步驟包含：自測試個體獲得測試片段集之序列讀段；針對各測試片段，確定該測試片段是否為低甲基化的或高甲基化的，其中低甲基化及高甲基化測試片段中之每一者包含至少臨限數目個CpG位點，其中至少臨限百分比之CpG位點分別為未甲基化的或甲基化的，針對參考基因組中之複數個CpG位點中之每一者：對與CpG位點重疊之低甲基化測試片段的計數及與CpG位點重疊之高甲基化測

試片段的計數進行量化；以及基於低甲基化測試片段及高甲基化測試片段之計數來產生低甲基化分數及高甲基化分數；針對各測試片段，基於測試片段中之CpG位點的低甲基化分數產生低甲基化總分且基於測試片段中之CpG位點的高甲基化分數產生高甲基化總分；針對測試個體，基於低甲基化總分對複數個測試片段排序且基於高甲基化總分對複數個測試片段排序；以及基於測試片段之排序來產生測試特徵向量。

【0052】 在一些實施例中，將該模型應用於測試個體之測試特徵向量以確定測試個體是否患有癌症包含：基於該模型產生該測試個體之癌症機率；及將該癌症機率與臨限機率進行比較以確定該測試個體是否患有癌症。

【0053】 在又另一態樣中，本發明提供一種非暫時性電腦可讀儲存媒體，其儲存用於確定測試個體是否患有癌症之執行指令，該等指令當由硬體處理器執行時，促使硬體處理器執行包含以下之步驟：存取模型，該模型係由訓練程序利用來自一或多個患有癌症之訓練個體的癌症訓練片段集及來自一或多個未患癌症之訓練個體的非癌症訓練片段集所獲得，癌症訓練片段集與非癌症訓練片段集均包含複數個訓練片段，其中該訓練程序包含：針對各訓練片段，確定該訓練片段是否為低甲基化的或高甲基化的，其中低甲基化及高甲基化訓練片段中之每一者包含至少臨限數目個CpG位點，其中至少臨限百分比之CpG位點分別為未甲基化的或甲基化的，針對各訓練個體，基於低甲基化訓練片段及高甲基化訓練片段產生訓練特徵向量，以及用來自一或多個未患癌症之訓練個體的訓練特徵向量及來自一或多個患有癌症之訓練個體的特徵向量訓練該模型；以及將該模型應用於與測試個體對應的測試特徵向量以確定該測試個體是否患有癌症。

【0054】 在一些實施例中，針對各訓練個體來產生訓練特徵向量包含：針對參考基因組中之複數個CpG位點中之每一者：對與CpG位點重疊之低甲基化訓練片段的計數及與CpG位點重疊之高甲基化訓練片段的計數進行量化；以及基於低甲基化訓練片段及高甲基化訓練片段之計數來產生低甲基化分數及高甲基化分數；針對訓練個體之各訓練片段，基於訓練片段中之CpG位點的低甲基化分數產生低甲基化總分且基於訓練片段中之CpG位點的高甲基化分數產生高甲基化總分；以及基於低甲基化總分對訓練個體之複數個訓練片段排序且基於高甲基化總分對該訓練個體之複數個訓練片段排序，其中訓練個體之訓練特徵向量係依據基於低甲基化總分的排序及基於高甲基化總分的排序。

【0055】 在一些實施例中，該等步驟進一步包含：獲得與測試個體對應之測試特徵向量，其中獲得測試特徵向量的步驟包含：自測試個體獲得測試片段集之序列讀段；針對各測試片段，確定該測試片段是否為低甲基化的或高甲基化的，其中低甲基化及高甲基化測試片段中之每一者包含至少臨限數目個CpG位點，其中至少臨限百分比之CpG位點分別為未甲基化的或甲基化的，針對參考基因組中之複數個CpG位點中之每一者：對與CpG位點重疊之低甲基化測試片段的計數及與CpG位點重疊之高甲基化測試片段的計數進行量化；以及基於低甲基化測試片段及高甲基化測試片段之計數來產生低甲基化分數及高甲基化分數；針對各測試片段，基於測試片段中之CpG位點的低甲基化分數產生低甲基化總分且基於測試片段中之CpG位點的高甲基化分數產生高甲基化總分；針對測試個體，基於低甲基化總分對複數個測試片段排序且基於高甲基化總分對複數個測試片段排序；以及基於測試片段之排序來產生測試特徵向量。在一些實施例中，將

該模型應用於測試個體之測試特徵向量以確定測試個體是否患有癌症包含：基於該模型產生該測試個體之癌症機率；及將該癌症機率與臨限機率進行比較以確定該測試個體患有癌症。

【0056】 在一個態樣中，本發明提供一種非暫時性電腦可讀儲存媒體，其儲存用於確定來自懷疑患有癌症之測試個體之測試片段是否具有異常甲基化模式的執行指令，該等指令當由硬體處理器執行時，促使硬體處理器執行包含以下之步驟：存取資料結構，該資料結構包含參考基因組內之CpG位點字串之計數及其來自訓練片段集之各別甲基化狀態；針對測試片段產生測試狀態向量，其中該測試狀態向量包含參考基因組內之測試基因組位置及該測試片段中之複數個CpG位點中之每一者的甲基化狀態，其中各甲基化狀態經測定為以下中之一者：甲基化、未甲基化及不確定；基於資料結構中所存儲之計數來計算測試狀態向量之測試機率；對長度與測試狀態向量相同之測試基因組位置之可能甲基化狀態向量的子集取樣；針對所取樣之可能甲基化狀態向量中之每一者，至少部分地基於資料結構中所存儲之計數來計算與所取樣之可能甲基化狀態向量對應的機率；計算與小於或等於測試機率之所計算機率對應之所取樣之可能甲基化狀態向量的比例；基於所計算的比例，產生測試片段之估算分數；以及基於所估算的分數確定該測試片段是否可能具有異常甲基化模式。

【0057】 在一些實施例中，該等步驟進一步包含：藉由比較估算分數與臨限分數來過濾測試片段，該臨限分數經選擇以使得與低於臨限分數之估算分數有關的測試片段更可能包括異常甲基化模式。在一些實施例中，該等步驟進一步包含：回應於確定測試片段可能具有異常甲基化模式，計算測試狀態向量之測試片段相對於訓練片段集的詳盡分數，其中該

詳盡分數係基於測試機率及複數個可能甲基化狀態向量的機率；以及基於詳盡分數來確定測試片段是否具有異常甲基化模式。在一些實施例中，該等步驟進一步包含：將分類器應用於測試狀態向量，該分類器經來自患有癌症之一或多個訓練個體的第一訓練片段集及來自未患癌症之一或多個訓練個體的第二訓練片段集訓練，其中分類器可以用於確定測試個體是否患有癌症。

【0058】 在另一態樣中，本發明提供一種儲存可執行指令的非暫時性電腦可讀儲存媒體，該等指令當由硬體處理器執行時，促使該處理器執行分類器以診斷癌症，其中該分類器係由包含以下之方法產生：a.自一或多個患有癌症之個體獲得癌症片段集之序列讀段且自一或多個未患癌症之個體獲得非癌症片段集之序列讀段，其中癌症片段集與非癌症片段集均包含複數個樣本片段；b.針對各片段，確定該片段是否為低甲基化的或高甲基化的，其中低甲基化及高甲基化片段包含至少臨限數目個CpG位點，其中至少臨限百分比之CpG位點分別為未甲基化的或甲基化的；c.針對參考基因組中之複數個CpG位點中之每一者：i.對與CpG位點重疊之低甲基化片段的計數及與CpG位點重疊之高甲基化片段的計數進行量化；以及ii.基於低甲基化片段及高甲基化片段之計數來產生低甲基化分數及高甲基化分數；d.針對各個體：i.基於低甲基化總分對複數個片段排序及基於高甲基化總分對複數個片段排序；以及ii.基於片段排序來產生特徵向量；e.基於來自一或多個患有癌症之個體之所產生特徵向量及來自一或多個未患癌症之個體之所產生特徵向量來訓練診斷模型，該診斷模型經組態以接收來自測試個體之測試特徵向量集且基於來自測試個體之測試特徵向量集輸出癌症之可能性；以及f.將代表該診斷模型之參數集儲存於非暫時性電腦可讀

儲存媒體上。

【0059】在一些實施例中，診斷模型包含具有複數個層之神經網路，該等層包括用於接收來自一或多個患有癌症之個體及一或多個未患癌症之個體之特徵向量的輸入層及基於特徵向量指示癌症之可能性的輸出層。在一些實施例中，該診斷模型進一步包含藉由重複地反向傳播誤差項來更新神經網路，該等誤差項係藉由將來自複數個訓練實例之訓練實例應用於診斷模型且計算損失函數而獲得，其中該複數個層係基於所計算的損失函數來更新。在一些實施例中，診斷模型包含內核邏輯回歸分類器。在一些實施例中，確定片段是否為低甲基化或高甲基化包含：a.存取資料結構，該資料結構包含參考基因組內之CpG位點字串的計數及其來自訓練片段集的各別甲基化狀態；b.針對該片段產生狀態向量，該狀態向量包含參考基因組內之基因組位置及片段中之複數個CpG位點中之每一者的甲基化狀態，各甲基化狀態經測定為甲基化的或未甲基化的；c.列舉長度與該狀態向量相同之基因組位置之複數個可能甲基化狀態；d.針對各種可能甲基化狀態，基於資料結構中所儲存之字串計數計算相應機率；e.鑑別與該狀態向量匹配的可能甲基化狀態及與經鑑別之可能甲基化狀態對應的所計算機率；f.基於經鑑別之計算機率，產生狀態向量之片段相對於訓練片段集的分數；以及g.基於所產生的分數確定該片段是否為低甲基化及高甲基化之一。在一些實施例中，將診斷模型應用於測試個體之測試特徵向量，該診斷模型經組態以輸出測試個體之癌症機率且將所輸出之癌症機率與臨限機率進行比較以確定測試個體是否患有癌症。

【0060】在另一態樣中，本發明提供一種儲存可執行指令的非暫時性電腦可讀儲存媒體，該等可執行指令當由硬體處理器執行時，促使該處

理器執行分類器以診斷癌症，其中該分類器係藉由包含以下之方法產生：a.自一或多個患有癌症之個體獲得癌症片段集之序列讀段且自一或多個未患癌症之個體獲得非癌症片段集之序列讀段，其中癌症片段集與非癌症片段集均包含複數個樣本片段；b.針對各片段，確定該片段是否具有異常甲基化模式，藉此獲得異常甲基化片段集；c.針對各異常甲基化片段，確定該異常甲基化片段是否為低甲基化的或高甲基化的，其中低甲基化及高甲基化片段包含至少臨限數目個CpG位點，其中至少臨限百分比之CpG位點分別為未甲基化的或甲基化的；d.針對參考基因組中之複數個CpG位點中之每一者：i.對與CpG位點重疊之低甲基化片段的計數及與CpG位點重疊之高甲基化片段的計數進行量化；及ii.基於低甲基化片段及高甲基化片段之計數來產生低甲基化分數及高甲基化分數；e.針對各個體：i.基於低甲基化總分將複數個片段排序且基於高甲基化總分將複數個片段排序；及ii.基於片段排序來產生特徵向量；f.基於來自一或多個患有癌症之個體之所產生特徵向量及來自一或多個未患癌症之個體之所產生特徵向量，訓練診斷模型，該診斷模型經組態以接收來自測試個體之測試特徵向量集且基於來自該測試個體之測試特徵向量集輸出癌症可能性；以及g.將代表該診斷模型的參數集儲存於非暫時性電腦可讀儲存媒體上。

【0061】 在一些實施例中，診斷模型包含具有複數個層之神經網路，該等層包括用於接收來自一或多個患有癌症之個體及一或多個未患癌症之個體之特徵向量的輸入層及基於特徵向量指示癌症之可能性的輸出層。在一些實施例中，該診斷模型進一步包含藉由重複地反向傳播誤差項來更新神經網路，該等誤差項係藉由將來自複數個訓練實例之訓練實例應用於診斷模型且計算損失函數而獲得，其中該複數個層係基於所計算的損

失函數來更新。在一些實施例中，診斷模型包含內核邏輯回歸分類器。在一些實施例中，確定片段是否發生異常甲基化包含：a.存取資料結構，該資料結構包含參考基因組內之CpG位點字串之計數及其來自訓練片段集之各別甲基化狀態；b.針對該片段產生狀態向量，該狀態向量包含參考基因組內之基因組位置及該片段中之複數個CpG位點中之每一者的甲基化狀態，各甲基化狀態經測定為甲基化的或未甲基化的；c.列舉長度與該狀態向量相同之基因組位置的複數個可能甲基化狀態；d.針對各種可能甲基化狀態，基於資料結構中所儲存之字串計數計算相應機率；e.鑑別與該狀態向量匹配的可能甲基化狀態及與經鑑別之可能甲基化狀態對應的所計算機率；f.基於經鑑別之計算機率產生該狀態向量之片段相對於訓練片段集的分數；及g.基於所產生的分數確定該片段是否發生異常甲基化。在一些實施例中，將診斷模型應用於測試個體之測試特徵向量，該診斷模型經組態以輸出測試個體之癌症機率且將所輸出之癌症機率與臨限機率進行比較以確定測試個體是否患有癌症。

【圖式簡單說明】

【0062】 圖1A為流程圖，其描述根據一個實施例對游離(cf) DNA之片段進行測序以獲得甲基化狀態向量的方法。

【0063】 圖1B為根據一個實施例對游離(cf) DNA之片段進行測序以獲得甲基化狀態向量之圖1A方法的說明。

【0064】 圖1C及1D展示利用對照組驗證測序一致性之資料的三個圖。

【0065】 圖2為流程圖，其描述根據一個實施例建立對照組之資料結構的方法。

【0066】 圖3為流程圖，其描述根據一個實施例驗證圖2之對照組資料結構的另一步驟。

【0067】 圖4為流程圖，其描述根據一個實施例鑑別個體之異常甲基化片段的方法。

【0068】 圖5為根據一個實施例的實例p值分數計算說明。

【0069】 圖6為流程圖，其描述根據一個實施例、基於片段之甲基化狀態訓練分類器的方法。

【0070】 圖7A-7C圖示針對不同癌症測定之跨越不同癌症階段的癌症勝算比對數。

【0071】 圖8A為根據一個實施例對核酸樣本進行測序之裝置的流程圖。圖8B提供根據一個實施例分析cfDNA甲基化狀態的分析系統。

【0072】 圖9提供自VIII.A所述之實驗獲得之訓練集及測試集內之所有癌症(左)及高信號癌症類型(右)的ROC曲線。上圖描繪特異性之完整範圍；下圖聚焦於90-100%特異性以更清晰地描繪95%、98%及99%特異性時的靈敏度，如所示。

【0073】 圖10展示訓練集與測試集效能之間的一致。使用WGBS分析報導訓練(x軸)及測試(y軸)中之各腫瘤類型在98%特異性時的靈敏度。指示高信號癌症及樣本大小。灰色陰影表示擬合線之95%信賴區間。

【0074】 圖11A及B在訓練集(圖11A)及測試集(圖11B)中展示當藉由WGBS分析時，各腫瘤類型(x軸)在98%特異性(y軸)時的靈敏度。誤差條表示95%信賴區間。

【0075】 圖式描繪本發明之各種實施例僅出於說明之目的。熟習此項技術者自以下論述將容易認識到可以使用本文所說明之結構及方法的替

代實施例而此等替代實施例不悖離本文所述之原理。

【實施方式】

相關申請案之交叉參考

【0076】本申請案主張2018年3月13日申請之美國臨時申請案第62/642,480號的權益及優先權，該美國臨時申請案以全文引用之方式併入本文中用於所有目的。

I.概述

【0077】根據本發明，處理來自測試個體之cfDNA片段以將未甲基化胞嘧啶轉化成尿嘧啶，測序且將序列讀段與參考基因組進行比較以鑑別該片段內之一或多個CpG位點處的甲基化狀態。相較於健康個體，鑑別異常甲基化cfDNA片段可以洞察個體的癌症狀態。如此項技術中所熟知，DNA甲基化異常(相較於健康對照組)會引起可能導致癌症的不同效應。在鑑別異常甲基化cfDNA片段中出現各種挑戰。首先，確定一或多個cfDNA片段異常甲基化僅相較於假設片段正常甲基化之對照個體組保持權重。另外，在一組對照個體中，甲基化狀態可以變化，此在確定個體之cfDNA異常甲基化時可能難以顧及。另外值得注意的是，CpG位點之胞嘧啶甲基化有原因地影響後續CpG位點之甲基化。囊封此相依性本身為一種挑戰。

【0078】當胞嘧啶鹼基之嘧啶環上的氫原子轉化成甲基時，去氧核糖核酸(DNA)典型地發生甲基化，形成5-甲基胞嘧啶。特定言之，甲基化容易發生於胞嘧啶及鳥嘌呤之二核苷酸處，在本文中稱為「CpG位點」。在其他情況下，甲基化可以發生於不為CpG位點之一部分的胞嘧啶處，或發生於不為胞嘧啶的另一核苷酸處；然而，此等發生較罕見。在本發明

中，為清楚起見，參考CpG位點論述甲基化。可以進一步鑑別cfDNA片段異常甲基化為高甲基化或低甲基化的，此兩者均可以指示癌症狀態。

【0079】 熟習此項技術者將瞭解本文所述之原理同樣適用於偵測非CpG情形下之甲基化，包括非胞嘧啶甲基化。在此類實施例中，用於偵測甲基化之濕式實驗室分析可不同於本文所述之彼等分析。另外，甲基化狀態向量可以含有通常作為向量位點的元件，在該等位點處，甲基化已發生或尚未發生(即使彼等位點並非特異性地為CpG位點)。經由該取代，本文所述方法的其餘部分為相同的，並且本文所述之本發明概念因此適用於甲基化之彼等其他形式。

【0080】 術語「游離核酸」、「游離DNA」或「cfDNA」係指在個體身體之體液(例如血流)中循環且來源於一或多個健康細胞及/或一或多個癌細胞的核酸片段或DNA片段。另外，cfDNA可以來自其他來源，例如病毒、胎兒等。

【0081】 術語「循環腫瘤DNA」或「ctDNA」係指來源於腫瘤細胞或其他類型之癌細胞的核酸片段，其可以作為生物過程之結果(諸如死亡細胞之細胞凋亡或壞死)而釋放至個體身體之體液(諸如血液、汗液、尿液或唾液)中，或由活的腫瘤細胞主動釋放。

【0082】 術語「個體」係指人類個體。術語「健康個體」係指經推定未患有癌症或疾病之個體。術語「個體」係指已知患有或潛在地患有癌症或疾病之個體。

【0083】 術語「序列讀段」係指來自自個體獲得之樣本的核苷酸序列讀段。序列讀段可以經由此項技術中已知之各種方法獲得。

【0084】 術語「讀取區段」或「讀段」係指自個體獲得之任何核苷

酸序列，包括序列讀段，及/或來源於初始序列讀段的核苷酸序列，該初始序列讀段來自自個體獲得的樣本。

II.樣本處理

【0085】 圖1A為流程圖，其描述根據一個實施例對游離(cf) DNA之片段進行測序以獲得甲基化狀態向量的方法100。為了分析DNA甲基化，分析系統首先自個體獲得110包含複數個cfDNA片段之樣本。一般而言，樣本可來自健康個體、已知患有或懷疑患有癌症之個體，或先前資訊未知之個體。測試樣本可為選自由血液、血漿、血清、尿液、糞便及唾液樣本組成之群的樣本。或者，測試樣本可以包含選自由以下組成之群的樣本：全血、血液分離物、組織切片、胸膜液、心包液、腦脊髓液及腹膜液。

【0086】 處理來自樣本的cfDNA片段以將未甲基化胞嘧啶轉化成尿嘧啶。在一個實施例中，方法利用亞硫酸氫鹽處理cfDNA片段，從而將未甲基化胞嘧啶轉化成尿嘧啶而不使甲基化胞嘧啶轉化。舉例而言，亞硫酸氫鹽轉化係使用市售套組，諸如EZ DNA Methylation™ - Gold、EZ DNA Methylation™ - Direct或EZ DNA Methylation™ - Lightning套組(獲自Zymo Research Corp (Irvine, CA))。在另一個實施例中，未甲基化胞嘧啶轉化成尿嘧啶係利用酶反應完成。舉例而言，該轉化可以利用將未甲基化胞嘧啶轉化成尿嘧啶的市售套組，諸如APOBEC-Seq (NEBiolabs, Ipswich, MA)。

【0087】 利用經轉化之cfDNA片段製備130測序文庫。視情況，可以使用複數個雜交探針，針對癌症狀態資訊豐富的cfDNA片段或基因組區域富集135測序文庫。雜交探針為能夠與標靶cfDNA片段雜交或與來源於一或多個靶區域之cfDNA片段雜交的短寡核苷酸，且針對彼等片段或區域

進行富集以便隨後測序及分析。雜交探針可以用於對所關注之指定CpG位點集進行靶向高深度分析。測序文庫或其一部分一經製備，即可測序以獲得複數個序列讀段。序列讀段可呈電腦可讀的數位形式以供電腦軟體處理及解譯。

【0088】 分析系統利用序列讀段，基於與參考基因組的比對來確定150一或多個CpG位點中之每一者的位置及甲基化狀態。分析系統針對各片段產生160甲基化狀態向量，該甲基化狀態向量指定片段在參考基因組中之位置(例如如根據各片段中之第一CpG位點之位置或另一種類似度量標準所指定)、片段中之CpG位點數目，及各CpG位點的甲基化狀態，該甲基化狀態是否為甲基化的(例如標示為M)、未甲基化的(例如標示為U)或不確定(例如標示為I)。觀測到之狀態為甲基化及未甲基化狀態；然而，未觀測到之狀態為不確定的。甲基化狀態向量可以儲存於暫時性或持久性電腦存儲器中供日後使用及處理。另外，分析系統可以移除來自單一個體之雙重複讀段或雙重複甲基化狀態向量。在另一實施例中，分析系統可以確定某一片段含有一或多個具有不確定甲基化狀態之CpG位點。不確定甲基化狀態可能來源於測序誤差及/或DNA片段互補股之甲基化狀態之間的不一致。分析系統不僅可以決定排除此類片段或選擇性地包括此類片段，而且可以構建模型，解釋此類不確定的甲基化狀態。下文將結合圖4描述一種此類模型。

【0089】 圖1B為根據一個實施例對cfDNA片段測序以獲得甲基化狀態向量之圖1A方法100的說明。作為一個實例，分析系統獲取cfDNA片段112。在這個實例中，cfDNA片段112含有三個CpG位點。如所示，cfDNA片段112之第一及第三CpG位點發生甲基化114。在處理步驟120期

間，使cfDNA片段112轉化以產生經轉化的cfDNA片段122。在處理120期間，未甲基化之第二CpG位點中的胞嘧啶被轉化成尿嘧啶。然而，第一及第三CpG位點不轉化。

【0090】 轉化之後，製備測序文庫130且測序140，產生序列讀段142。分析系統將序列讀段142與參考基因組144比對150。參考基因組144提供關於片段cfDNA來源於人類基因組中之何位置的訊文。在此簡化實例中，分析系統對序列讀段進行比對150，以使得三個CpG位點與CpG位點23、24及25相關聯(為了方便說明，使用任意參考標識符)。分析系統從而產生關於cfDNA片段112上之所有CpG位點之甲基化狀態及人類基因組中之與CpG位點對應之位置的資訊。如所示，序列讀段142上之甲基化CpG位點作為胞嘧啶讀取。在此實例中，胞嘧啶在序列讀段142中僅在第一及第三CpG位點出現，讓人推斷出第一及第三CpG位點在原始cfDNA片段中發生甲基化。然而，第二CpG位點作為胸腺嘧啶讀取(在測序過程期間，U轉化成T)，且因此可以推斷出第二CpG位點在原始cfDNA片段中未甲基化。利用此兩塊資訊：甲基化狀態及位置，分析系統針對cfDNA片段112產生160甲基化狀態向量152。在此實例中，所得甲基化狀態向量152為$< M_{23}, U_{24}, M_{25} >$，其中M對應於甲基化CpG位點，U對應於未甲基化CpG位點，且下標數字對應於各CpG位點在參考基因組中之位置。

【0091】 圖1C及1D展示利用對照組驗證測序一致性之資料的三個圖。第一個圖170展示自處於癌症之不同階段 - 階段I、階段II、階段III、階段IV及無癌症之個體之測試樣本獲得之cfDNA片段上之未甲基化胞嘧啶轉化成尿嘧啶(步驟120)的轉化準確性。如所示，cfDNA片段上之未甲基化胞嘧啶轉化成尿嘧啶存在均勻的一致性。總體轉化準確度為

99.47%，精確度為±0.024%。第二個圖180展示癌症之不同階段的平均覆蓋率。僅使用與基因組對應值得信賴的彼等DNA片段，計算所有群組之平均覆蓋率，其為DNA片段之基因組覆蓋率之平均約34倍。第三個圖190展示在癌症之不同階段，每個樣本之cfDNA濃度。

III.對照資料結構

III.A.建立

【0092】 圖2為流程圖，其描述根據一個實施例產生健康對照組之資料結構的方法200。為建立健康對照組資料結構，分析系統接收來自複數個個體的複數個DNA片段(例如cfDNA)。鑑別各片段的甲基化狀態向量，例如經由方法100來鑑別。

【0093】 利用各片段的甲基化狀態向量，分析系統將甲基化狀態向量再劃分210成CpG位點字串。在一個實施例中，分析系統再劃分210甲基化狀態向量，使得所得字串皆小於指定長度。舉例而言，長度11之甲基化狀態向量可以再分為長度小於或等於3之字串，得到長度3之9個字串、長度2之10個字串，及長度1之11個字串。在另一實例中，長度7之甲基化狀態向量再分為長度小於或等於4之字串得到長度4之4個字串、長度3之5個字串、長度2之6個字串，及長度1之7個字串。若甲基化狀態向量比指定的字串長度短或長度與指定的字串長度相同，則甲基化狀態向量可以轉化成含有向量中之所有CpG位點的單一字串。

分析系統對字串計數220，此係藉由針對向量中之各可能CpG位點及甲基化狀態可能性，對存在於對照組中的字串數計數來達成，該字串中具有指定CpG位點作為第一CpG位點且具有甲基化狀態之該可能性。舉例而言，在指定的CpG位點且將字串長度視為3，存在2^3或8種可能的字串組

態。在該指定的CpG位點，針對8種可能字串組態中之每一者，分析系統計算220各甲基化狀態向量可能性在對照組中出現多少次。繼續以此為例，對於參考基因組中之各起始CpG位點x而言，此可以包括計算以下數量：$< M_x, M_{x+1}, M_{x+2} >$、$< M_x, M_{x+1}, U_{x+2} >$、...、$< U_x, U_{x+1}, U_{x+2} >$。分析系統建立230資料結構，該資料結構儲存各起始CpG位點及字串可能性之所記錄計數。

【0094】 設置字串長度上限有若干好處。首先，視字串最大長度而定，分析系統所建立之資料結構的尺寸可以在尺寸上顯著增加。舉例而言，最大字串長度為4意謂每個CpG位點具有至少2^4個數字對長度4之字串計數。將最大字串長度增加至5意謂每個CpG位點具有額外2^4或16個數字來計數，相較於先前字串長度，使計數的數字(及所必需之電腦記憶體)倍增。就合理的計算及儲存而言，減小字串尺寸有助於保持資料結構建立及效能(例如用於日後存取，如下文所述)。其次，限制最大字串長度的統計學考量係避免過度擬合之下游模型利用字串計數。若CpG位點之長字串在生物學上對結果(例如預測作為癌症存在前兆的異常)不產生有力影響，則基於CpG位點之大字串計算機率可能存在問題，原因在於其需要的大量資料無法獲得，且因此對於模型欲適當發揮效能而言亦過於稀少。舉例而言，計算侷限於先前100個CpG位點上之異常/癌症的機率將需要對資料結構中長度100之字串進行計數，理想情況下，一些字串計數與先前100種甲基化狀態恰好匹配。若長度100之字串僅有稀少的計數可供利用，則確定測試樣本中長度100之指定字串是否異常的資料不足。

III.A. 資料結構驗證

【0095】 資料結構已建立後，分析系統可以試圖驗證240資料結構

且/或利用資料結構驗證任何下游模型。一種驗證類型係校驗對照組資料結構內的一致性。舉例而言，若對照組內之個體、樣本及/或片段存在任何離群值，則分析系統可以執行各種計算來確定是否將任何片段自彼等類別之一中排除。在一個代表性實例中，健康對照組可以含有未經診斷、但癌變之樣本，因此該樣本含有異常甲基化片段。此第一種驗證類型確保自健康對照組中移除潛在癌變樣本，以免影響對照組純度。

【0096】 第二種驗證類型係利用來自資料結構本身(亦即，來自健康對照組)的計數校驗用於計算p值的機率模型。p值計算方法結合圖5描述如下。分析系統產生驗證組之甲基化狀態向量的p值後，分析系統利用p值構建累積密度函數(CDF)。有了CDF，分析系統可以對CDF執行各種計算以驗證對照組的資料結構。一項測試係利用如下事實：CDF在理想情況下應等於或低於恆等函數，使得CDF(x)≤x。反之，高於恆等函數揭露用於對照組資料結構之機率模型內存在一些不足。舉例而言，若1/100片段之p值分數為1/1000，意謂CDF(1/1000) = 1/100>1/1000，則第二種驗證類型失敗，表示機率模型存在問題。

【0097】 第三種驗證類型係利用自構建資料結構所用之彼等驗證樣本分離出的健康驗證樣本集，其測試資料結構是否正確地構建且模型是否發揮作用。用於實施此驗證類型之實例方法結合圖3描述如下。第三種驗證類型可以量化健康對照組如何良好地歸納於健康樣本之分佈。若第三種驗證類型失敗，則健康對照組不能良好地歸納於健康分佈。

【0098】 第四種驗證類型係測試來自非健康驗證組之樣本。分析系統計算p值且構建CDF用於非健康驗證組。在非健康驗證組存在下，分析系統預期至少一些樣本經歷CDF(x)>x，或換言之，與使用健康對照組及

健康驗證組的第二種驗證類型及第三種驗證類型中所預期的相反。若第四種驗證類型失敗，則此表示該模型不適當地鑑別其經設計用於鑑別的異常。

【0099】 圖3為流程圖，其描述根據一個實施例驗證圖2之對照組資料結構的另一步驟240。在驗證資料結構之此步驟240中，分析系統係使用個體、樣本及/或片段之組成與對照組假設類似的驗證組。舉例而言，若分析系統選擇未患癌症之健康個體作為對照組，則分析系統亦使用未患癌症之健康個體作為驗證組。

【0100】 分析系統獲取驗證組且產生100甲基化狀態向量集，如圖1中所述。分析系統針對來自驗證組之各甲基化狀態向量執行p值計算。p值計算方法將結合圖4及5進一步描述。針對甲基化狀態向量之各種可能性，分析系統計算320對照組資料結構之機率。計算出甲基化狀態向量之可能性的機率後，分析系統基於所計算之機率計算330甲基化狀態向量之p值分數。p值分數表示發現該特定甲基化狀態向量及對照組中具有甚至更低之機率之其他可能甲基化狀態向量的預期。因此，低p值分數通常對應於相對而言未預期的甲基化狀態向量(相較於對照組中之其他甲基化狀態向量)，其中高p值分數通常對應於預期相對大於對照組中所發現之其他甲基化狀態向量的甲基化狀態向量。分析系統產生驗證組之甲基化狀態向量之p值分數後，分析系統利用來自驗證組的p值分數構建340累積密度函數(CDF)。CFD可以用於如此章節中之以上別處所述的驗證測試。

IV. 鑑別具有異常甲基化模式的片段

IV.A. 通用方法

【0101】 圖4為流程圖，其描述根據一個實施例鑑別個體之異常甲

基化片段的方法400。方法400之實例可視地說明於圖5中，且在圖4之說明下方進一步描述。在方法400中，分析系統產生100來自個體之cfDNA片段的甲基化狀態向量。分析系統如下處理各甲基化狀態向量。

【0102】 在一些實施例中，分析系統過濾405在一或多個CpG位點具有不確定狀態的片段。在此類實施例中，分析系統實施預測模型以鑑別不可能具有異常甲基化模式之片段過濾。針對樣本片段，預測模型計算相較於健康對照組資料結構發生樣本片段甲基化狀態向量的樣本機率。預測模型對涵蓋樣本片段甲基化狀態向量中之CpG位點的可能甲基化狀態向量子集隨機取樣。預測模型計算與所取樣之可能甲基化狀態向量中之每一者對應的機率。片段甲基化狀態向量及所取樣之可能甲基化狀態向量的機率計算可以根據馬可夫鏈模型計算，如下文在章節IV.B.*實例P值分數計算*(*Example P-Value Score Calculation*)中所述。預測模型計算與小於或等於樣本機率之機率對應之所取樣之可能甲基化狀態向量的比例。預測模型基於所計算之比例，針對片段產生估算p值分數。預測模型可以過濾與高於臨限值之p值分數對應的片段且保留與低於臨限值之p值分數對應的片段。

【0103】 在其他實施例中，預測模型可以計算信賴機率，預測模型利用該信賴機率確定何時繼續或何時終止取樣。信賴機率描述片段真實p值分數(真實p值分數之計算進一步描述於章節IV.B.*實例P值分數計算*)低於臨限值的可能程度，該臨限值係基於所取樣之可能甲基化狀態向量的估算p值分數及機率。預測模型可以對額外的一或多個可能甲基化狀態向量進行取樣，同時以迭代方式計算估算p值分數及信賴機率。接著，當信賴機率高於信賴臨限值時，預測模型可以終止取樣。

【0104】 對於指定的甲基化狀態向量而言，分析系統列舉410甲基化狀態向量中具有相同起始CpG位點及相同長度(亦即，CpG位點集)之甲基化狀態向量的所有可能性。由於所觀測到之各甲基化狀態可以是甲基化的或未甲基化的，因此各CpG位點處僅存在兩種可能狀態，且因此，甲基化狀態向量之不同可能性的計數依賴於冪2，使得長度n之甲基化狀態向量與甲基化狀態向量之2^n種可能性相關。在甲基化狀態向量包括一或多個CpG位點之不確定狀態的情況下，分析系統可以列舉410僅考慮CpG位點具有所觀測到之狀態之甲基化狀態向量的可能性。

【0105】 分析系統藉由存取健康對照組資料結構，根據所鑑別之起始CpG位點/甲基化狀態向量長度來計算420觀測到甲基化狀態向量之各種可能性的機率。在一個實施例中，計算觀測到指定可能性之機率係利用馬可夫鏈機率對聯合機率計算進行模型化，其將更詳細地結合下圖5描述。在其他實施例中，使用除馬可夫鏈機率外之計算方法來確定觀測到甲基化狀態向量之各種可能性的機率。

【0106】 分析系統利用各種可能性之所計算機率來計算430甲基化狀態向量之p值分數。在一個實施例中，此包括鑑別對應於與所討論之甲基化狀態向量匹配之可能性的所計算機率。特定言之，此為CpG位點集與甲基化狀態向量相同的可能性，或類似地，起始CpG位點及長度與甲基化狀態向量相同的可能性。分析系統對機率小於或等於所鑑別之機率之任何可能性的所計算機率求和，以產生p值分數。

【0107】 此p值表示觀測到片段之甲基化狀態向量或健康對照組中可能性甚至更小之其他甲基化狀態向量的機率。因此，低p值分數通常對應於健康個體中罕見的甲基化狀態向量，且相對於健康對照組，引起所標

記之片段發生異常甲基化。相對而言，預期健康個體中存在與甲基化狀態向量通常相關的高p值分數。舉例而言，若健康對照組為非癌症群組，則低p值表示片段相對於非癌症群組已發生異常甲基化，且因此可以指示測試個體存在癌症。

【0108】 如上所述，分析系統計算複數個甲基化狀態向量中之每一者的p值分數，該等甲基化狀態向量各代表測試樣本中之cfDNA片段。為了鑑別哪些片段發生異常甲基化，分析系統可以基於其p值分數過濾440甲基化狀態向量集。在一個實施例中，藉由比較p值分數與臨限值且保持僅彼等片段低於臨限值來進行過濾。此臨限值p值評分階數可為約0.1、0.01、0.001、0.0001或類似者。

【0109】 根據方法400之實例結果，分析系統針對訓練中之未患癌症之參與者產生中值(範圍)為2,800 (1,500-12,000)個具有異常甲基化模式之片段，且針對訓練中之患有癌症之參與者產生中值(範圍)為3,000 (1,200-220,000)個具有異常甲基化模式之片段。此等經過濾之具有異常甲基化模式之片段集可以用於下游分析，如下文在章節IV.D.*經過濾之異常片段集之實例使用情況(Example Use Cases for Filtered Sets of Anomalous Fragments)*中所述。

IV.B. 實例P值分數計算

【0110】 圖5為根據一個實施例的實例p值分數計算說明500。為了計算在測試甲基化狀態向量505指定情況下的p值分數，分析系統獲取該測試甲基化狀態向量505且列舉410甲基化狀態向量之可能性。在此說明性實例中，測試甲基化狀態向量505為< M23, M24, M25, U26 >。由於測試甲基化狀態向量505之長度為4，因此涵蓋CpG位點23-26之甲基化狀態

向量存在2^4種可能性。在一通用實例中，甲基化狀態向量之可能性的數目為2^n，其中n為測試甲基化狀態向量之長度或替代地為滑動窗之長度(下文進一步描述)。

【0111】 分析系統計算420所列舉甲基化狀態向量之可能性的機率515。由於甲基化條件性地依賴於鄰近CpG位點之甲基化狀態，因此計算觀測到所指定甲基化狀態向量可能性之機率的一種方式為使用馬可夫鏈模型。一般而言，甲基化狀態向量，諸如$<S_1, S_2, \cdots, S_n>$，其中S表示甲基化狀態是否為甲基化的(標示為M)、未甲基化的(標示為U)或不確定(標示為I)，具有可以利用如下機率鏈規則擴增的聯合機率：

$$P(< S_1, S_2, \ldots, S_n >)$$

$$= P(S_n | S_1, \ldots, S_{n-1}) * P(S_{n-1} | S_1, \ldots, S_{n-2}) * \ldots * P(S_2 | S_1) * P(S_1) \quad (1)$$

【0112】 馬可夫鏈模型可以用於更高效地計算各種可能性之條件性機率。在一個實施例中，分析系統選擇馬可夫鏈階數k進行條件性機率計算，該馬可夫鏈階數k對應於向量(或窗口)中欲考慮多少個先前CpG位點，使得條件性機率具有如下模型：$P(S_n | S_1, \cdots, S_{n-1}) \sim P(S_n | S_{n-k-2}, \cdots, S_{n-1})$。

【0113】 為了計算甲基化狀態向量之可能性的各種馬可夫模型化機率，分析系統存取對照組資料結構，特定言之，CpG位點及狀態之各種字串之計數。為了計算$P(M_n | S_{n-k-2}, \cdots, S_{n-1})$，分析系統獲取來自資料結構之與$< S_{n-k-2}, \cdots, S_{n-1}, M_n >$匹配之字串數目之所儲存計數除以來自資料結構之與$< S_{n-k-2}, \cdots, S_{n-1}, M_n >$及$< S_{n-k-2}, \cdots, S_{n-1}, U_n >$匹配之字串數目之所儲存計數之總和的比率。因此，$P(M_n | S_{n-k-2}, \cdots, S_{n-1})$為具有以下形式之計算比率：

$$\frac{< S_{n-k-2}, \ldots, S_{n-1}, M_n > \text{之數目}}{< S_{n-k-2}, \ldots, S_{n-1}, M_n > \text{之數目} + < S_{n-k-2}, \ldots, S_{n-1}, U_n > \text{之數目}} \quad (2)$$

計算另外可以藉由應用事前分佈來執行計數平滑。在一個實施例中，事前分佈為如拉普拉斯平滑(Laplace smoothing)中的均一先驗。作為其實例，將一種常數添加至上述方程式之分子中且將另一種常數(例如為分子中之常數的兩倍)添加至上述方程式之分母中。在其他實施例中，使用算法技術，諸如克萊瑟-奈伊平滑(Knesser-Ney smoothing)。

【0114】 在說明中，將上述公式式應用於覆蓋位點23至26之測試甲基化狀態向量505。一旦所計算之機率515完成，則分析系統計算430 p值分數525，該p值分數為小於或等於與測試甲基化狀態向量505匹配之甲基化狀態向量可能性之機率的機率總和。

【0115】 在具有不確定狀態之實施例中，分析系統可以計算出作為片段甲基化狀態向量中具有不確定狀態之CpG位點之總和的p值分數。分析系統鑑別出甲基化狀態向量之排除不確定狀態之所有甲基化狀態共有的所有可能性。分析系統可以向甲基化狀態向量賦予作為所鑑別出之可能性之機率總和的機率。作為一個實例，由於觀測到CpG位點1及3之甲基化狀態且為CpG位點1及3之片段甲基化狀態所共有，因此分析系統將< M_1, I_2, U_3 >之甲基化狀態向量之機率作為< M_1, M_2, U_3 >與< M_1, U_2, U_3 >之甲基化狀態向量之可能性之機率總和來計算。對具有不確定狀態之CpG位點求和的此方法係利用多達2^i種可能性之機率的計算，其中i表示甲基化狀態向量中之不確定狀態的數目。在其他實施例中，可以實施動態程式化算法以計算具有一或多種不確定狀態之甲基化狀態向量之機率。有利地，動態程式化算法係在線性計算時間中運作。

【0116】 在一個實施例中，可以藉由快取至少一些計算來進一步減

小計算機率及/或p值分數的計算負擔。舉例而言，分析系統可以對甲基化狀態向量(或其窗口)中之可能性之機率的暫時性或持久性記憶體計算進行快取。若其他片段具有相同CpG位點，則快取可能性機率允許高效計算p值分數而無需重新計算潛在可能性機率。同樣，分析系統可以計算與來自向量(或其窗口)之CpG位點集有關之甲基化狀態向量之各種可能性的p值分數。分析系統可以快取p值分數用於確定包括相同CpG位點之其他片段的p值分數。一般而言，具有相同CpG位點之甲基化狀態向量中之可能性的p值分數可以用於確定來自相同CpG位點集之可能性中之不同者的p值分數。

IV.C. 滑動窗

【0117】 在一個實施例中，分析系統利用435滑動窗來確定甲基化狀態向量之可能性且計算p值。分析系統僅針對連續CpG位點之窗口列舉可能性及計算p值，而非針對整個甲基化狀態向量列舉可能性及計算p值，其中窗口(CpG位點)長度比至少一些片段短(否則，窗口將無的放失)。窗口長度可為靜態的、依使用者決定的、動態的，或以其他方式選擇。

【0118】 計算大於窗口之甲基化狀態向量的p值時，自向量中之第一個CpG位點開始，窗口鑑別向量在窗口內的連續CpG位點集。分析系統針對包括第一個CpG位點之窗口計算p值分數。分析系統接著將窗口「滑動」至向量中之第二個CpG位點，且針對第二窗口計算另一個p值分數。因此，對於窗口尺寸l及甲基化向量長度m而言，各甲基化狀態向量將產生$m-l+1$個p值分數。針對向量之各部分完成p值計算之後，將得自所有滑動窗的最低p值分數視為甲基化狀態向量的總體p值分數。在另一個實施例中，分析系統將甲基化狀態向量之p值分數加總，產生總體p值分數。

【0119】 使用滑動窗有助於減少甲基化狀態向量之所列舉可能性的數目及其原本需要進行的相應機率計算。實例機率計算展示於圖5中，但甲基化狀態向量之可能性數目相對於甲基化狀態向量之尺寸通常以2之倍數呈指數級增加。以現實為例，片段有可能具有多達54個CpG位點。分析系統不是計算2^54 (約1.8×10^16)種可能性之機率來產生單一p分數，而是可以改用(例如)尺寸5之窗口，從而針對該片段之甲基化狀態向量之50個窗口中之每一者進行50次p值計算。50次計算中之每一者列舉甲基化狀態向量之2^5 (32)種可能性，其總共產生50×2^5 (1.6×10^3)次機率計算。由此使得待進行的計算大量減少，而對異常片段之準確鑑別不產生意味深長的衝擊。當利用驗證組的甲基化狀態向量驗證240對照組時，亦可應用此額外步驟。

IV.D. 經過濾之異常片段集的實例使用情況

【0120】 分析系統可以對異常片段集進行任何多種及/或可能的額外分析。一種額外分析係鑑別來自過濾集的450個低甲基化片段或高甲基化片段。低甲基化或高甲基化的片段可以定義為特定長度之CpG位點(例如超過3、4、5、6、7、8、9、10個等)的片段，其中分別存在高百分比的甲基化CpG位點(例如超過80%、85%、90%或95%，或50%-100%範圍內的任何其他百分比)或高百分比的未甲基化CpG位點(例如超過80%、85%、90%或95%，或50%-100%範圍內的任何其他百分比)。下述圖6說明基於異常甲基化片段集來鑑別基因組中之此等低甲基化或高甲基化部分的實例方法。

【0121】 一種替代分析係對異常片段集應用460經訓練的分類模型。經訓練之分類模型可以訓練成鑑別可以利用甲基化狀態向量鑑別出之

任何所關注病狀。在一個實施例中，經訓練之分類模型為基於自患有癌症之個體組獲得之cfDNA片段之甲基化狀態及視情況基於自未患癌症之健康個體組獲得之cfDNA片段之甲基化狀態訓練的二進位分類器，且接著基於異常甲基化狀態向量用於對測試個體患癌或不患癌之機率進行分類。在其他實施例中，可以利用已知患有特定癌症(例如乳癌、肺癌、前列腺癌等)之個體組訓練不同分類器以預測測試個體是否患有彼等特定癌症。

【0122】 在一個實施例中，基於得自方法450之關於高/低甲基化區域的資訊訓練分類器，如下文關於圖6所述。

【0123】 另一種額外分析係計算來自個體之異常片段大體上指示癌症或特定癌症類型的勝算比對數。勝算比對數可以藉由將癌變機率相對於非癌變機率(亦即，一減去癌變機率)之比率取對數來計算。兩種機率均如應用的460分類模型所確定。

【0124】 圖7A-7C展示多個個體之跨越不同階段之多種癌症的圖，其標繪出根據上文關於圖4所述之方法鑑別之異常片段的勝算比對數。此資料係經由測試超過1700個臨床上可評估之個體而獲得，其中過濾掉逾1400個個體，包括近600個未患癌症的個體及剛好逾800個患有癌症的個體。圖7A中之第一個圖700展示所有癌症個案，其跨越三個不同層級 - 非癌症；階段I/II/II；及階段IV。階段IV之癌症勝算比對數顯著大於階段I/II/II及非癌症之勝算比對數。圖7A中之第二個圖710展示跨越癌症之所有階段及非癌症的乳癌個案，在癌症之漸進階段中，勝算比對數增加出現類似的進展。圖7B中之第三個圖720展示乳癌亞型。值得注意的是，亞型HER2+及TNBC較多散開，而HR+/HER2-濃度更接近約1。圖7C中之第四個圖730展示跨越癌症之所有階段及非癌症、在肺癌之漸進階段中穩定進

展之肺癌個案。第五個圖740展示跨越癌症之所有階段及非癌症的結腸直腸癌個案，再次展示結腸直腸癌之漸進階段中的穩定進展。圖7C中之第六個圖750展示跨越癌症之所有階段及非癌症的前列腺癌個案。此實例不同於先前說明的大部分，僅階段IV顯著不同於其他階段I/II/II及非癌症。

V. 高/低甲基化區域及分類器

【0125】 圖6為流程圖，其描述根據一個實施例、基於cfDNA片段之甲基化狀態訓練分類器的方法600。分析系統可以用於執行方法600。該方法存取兩個訓練樣本組 - 非癌症組及癌症組並且獲得400非癌症甲基化狀態向量集及癌症甲基化狀態向量集，每組包含樣本之異常片段。異常片段可以根據例如圖4之方法400鑑別。

【0126】 分析系統針對各甲基化狀態向量確定610甲基化狀態向量是否為低甲基化的或高甲基化的。在此，若至少一些數目個CpG位點具有特定狀態(分別為甲基化的或未甲基化的)且/或有臨限百分比之位點呈特定狀態(再次分別為甲基化的或未甲基化的)，則賦予高甲基化或低甲基化標記。如上文所定義，若cfDNA片段具有至少五個未甲基化或甲基化的CpG位點且(邏輯AND)高於80%的片段CpG位點為未甲基化的或甲基化的，則鑑別該等片段分別為低甲基化的或高甲基化的。

【0127】 在一個替代實施例中，分析系統考慮甲基化狀態向量之一部分且確定該部分是否為低甲基化的或高甲基化的，且可以區分低甲基化或高甲基化的該部分。此替代方案解決尺寸較大、但含有至少一個緻密低甲基化或高甲基化區域的缺失甲基化狀態向量。界定低甲基化及高甲基化之此方法可適用於圖4之步驟450中。

【0128】 分析系統針對基因組中之每個CpG位點產生620低甲基化

分數及高甲基化分數。為了產生所指定CpG位點之分數，分類器在該CpG位點獲取四種計數 - (1)與CpG位點重疊之經標記之低甲基化癌症集內之(甲基化狀態)向量的計數；(2)與CpG位點重疊之經標記之高甲基化癌症集內之向量的計數；(3)與CpG位點重疊之經標記之低甲基化非癌症集內之向量的計數；及(4)與CpG位點重疊之經標記之高甲基化非癌症集內之向量的計數。另外，該方法可以將各組之此等計數標準化以考慮非癌症組與癌症組之間在群組尺寸上之偏差。

【0129】 在一個實施例中，所指定CpG位點之低甲基化分數定義為(1)相對於(3)之比率的對數。類似地，高甲基化分數作為(2)相對於(4)之比率的對數計算。另外，此等比率可以利用如上文所論述的另一種平滑技術計算。

【0130】 在另一個實施例中，低甲基化分數定義為(1)相對於(1)與(3)總和之比率。高甲基化分數定義為(2)相對於(2)與(4)總和之比率。類似於上述實施例，可以對該等比率實施平滑技術。

【0131】 分析系統針對各異常甲基化狀態向量產生630低甲基化總分及高甲基化總分。高甲基化總分及低甲基化總分係基於甲基化狀態向量中之CpG位點的高甲基化分數及低甲基化分數來確定。在一個實施例中，高甲基化總分及低甲基化總分分別作為各狀態向量中之位點之最大高甲基化分數及低甲基化分數來指定。然而，在替代實施例中，總分可以基於使用各向量中之位點之高/低甲基化分數所得到的平均值、中值或其他計算。在一個實施例中，分析系統將低甲基化總分及高甲基化總分中之較大者賦予異常甲基化狀態向量。

【0132】 分析系統接著根據該個體之所有甲基化狀態向量的低甲基

化總分及高甲基化總分對該等向量進行排序640，每位個體得到兩個秩。該方法自低甲基化秩中選擇低甲基化總分且自高甲基化秩中選擇高甲基化總分。利用所選分數，分類器產生650各個體之單一特徵向量。在一個實施例中，選自任一個秩的分數係根據固定的階數選擇，每個訓練組中之各個體之所產生之各特徵向量的固定階數相同。作為一個實例，在一個實施例中，分類器自各秩中選擇第一、第二、第四、第八、第十六、第三十二、第六十四個高甲基化總分且類似地選擇各種低甲基化總分並且將彼等分數寫入該個體之特徵向量(特徵向量中存在總共14個特徵)。在其他實施例中，為了調整樣本測序深度，分析系統依線性比例將秩調整至相對樣本深度。舉例而言，若相對樣本深度為x，則內插分數係在x*初始秩取得(亦即，x=1.1，吾等獲取在秩1.1、2.2、…、x*2i所計算的分數)。分析系統接著可以基於用於進一步分類之調整秩來定義特徵向量。

【0133】 分析系統訓練660二進位分類器以區分癌症與非癌症訓練組之間的特徵向量。分析系統可以將訓練樣本歸入一或多個訓練樣本集以對二進位分類器進行迭代分批訓練。輸入所有訓練樣本集(包括其訓練特徵向量)且調整分類參數之後，對二進位分類器進行充分訓練，根據測試樣本的特徵向量、在一些誤差邊際內標記測試樣本。舉例而言，在一個實施例中，分類器確定樣本特徵向量來自患有癌症之個體的似然度或機率分數(例如0至100)。在一些實施例中，對機率分數與臨限機率進行比較以確定個體是否患有癌症。在其他實施例中，機率分數大於或等於60表示個體患有癌症。在仍其他實施例中，機率分數大於或等於65、大於或等於70、大於或等於75、大於或等於80、大於或等於85、大於或等於90或大於或等於95表示個體患有癌症。一般而言，可以使用多種分類技術中之任

一者。此等技術有許多，包括潛在使用內核方法、機器學習算法，諸如多層神經網路等。

【0134】 在一個實施例中，分類器為非線性分類器。在一個特定實施例中，分類器為利用L2正則化內核邏輯回歸與高斯徑向基底函數(RBF)內核的非線性分類器。特定言之，利用各向同性徑向基底函數(冪指數2)作為內核以及比例參數γ及L2正則化參數λ，訓練正則化內核邏輯回歸分類器(KLR)。在指定的訓練資料內，利用內部交叉驗證使γ及λ最佳化以便留出損失對數，且在最大值開始且每個步驟將參數減半，在相乘步驟中利用網格搜尋來最佳化。

VI. 實例測序儀及分析系統

【0135】 圖8A為根據一個實施例對核酸樣本進行測序之裝置的流程圖。此說明性流程圖包括諸如測序儀820及分析系統800之裝置。測序儀820及分析系統800可以串聯工作以執行圖1A之方法100、圖2之200、圖3之240、圖4之400、圖6之600及本文所述之其他方法中的一或多個步驟。

【0136】 在各種實施例中，測序儀820接收經富集之核酸樣本810。如圖8A中所示，測序儀820可以包括能夠讓使用者與特定任務(例如起始測序或終止測序)交互作用之圖形使用者介面825，以及一個以上裝載站830用於裝載包括經富集之片段樣本的測序筒柱及/或用於裝載執行測序分析所需的緩衝液。因此，測序儀820之使用者向測序儀820之裝載站830提供必需的試劑及測序筒柱後，使用者可以藉由與測序儀820之圖形使用者介面825交互作用來起始測序。一旦起始，測序儀820即執行測序且輸出來自核酸樣本810之富集片段的序列讀段。

【0137】 在一些實施例中，測序儀820與分析系統800通信耦接。分

析系統800包括一些數目個用於處理序列讀段的計算裝置用於多種應用，諸如評估一或多個CpG位點之甲基化狀態、變異體識別或品質控制。測序儀820可以向分析系統800提供呈BAM檔案格式的序列讀段。分析系統800可以經由無線、有線通信技術或無線與有線通信技術之組合與測序儀820通信耦接。一般而言，分析系統800經組態而具有處理器及儲存電腦指令之非暫時性電腦可讀儲存媒體，該等電腦指令當由處理器執行時，促使處理器處理序列讀段或執行本文所揭示之任一種方法或程序的一或多個步驟。

【0138】 在一些實施例中，可以利用此項技術中已知可確定比對位置資訊的方法(例如圖1A中之方法100之步驟140的一部分)將序列讀段與參考基因組比對。比對位置可以大體描述參考基因組中之區域的開始位置及終止位置，其對應於所指定之序列讀段的開始核苷酸鹼基及終止核苷酸鹼基。對應於甲基化測序，可以根據與參考基因組的比對來概括比對位置資訊以指出序列讀段中所包括的第一個CpG位點及最後一個CpG位點。比對位置資訊可以進一步指出所指定之序列讀段中之所有CpG位點的甲基化狀態及位置。可以使參考基因組中之區域與基因或基因區段關聯；因此，分析系統800可以用與序列讀段比對的一或多個基因標記序列讀段。在一個實施例中，片段長度(或尺寸)係由開始及終止位置決定。

【0139】 在各種實施例中，例如，當使用成對末端測序法時，序列讀段包含標示為R_1及R_2的讀段對。舉例而言，第一個讀段R_1可以自雙股DNA (dsDNA)分子的第一末端測序，而第二讀段R_2可以自雙股DNA (dsDNA)之第二末端測序。因此，可以將第一讀段R_1與第二讀段R_2之核苷酸鹼基對與參考基因組之核苷酸鹼基進行一致性比對(例如依

相反取向)。來源於讀段對R_1及R_2的比對位置資訊可以包括參考基因組中的起始位置，該起始位置與第一讀段(例如R_1)之末端對應；及參考基因組中的終止位置，該終止位置與第二讀段(例如R_2)的末端對應。換言之，參考基因組中之起始位置及終止位置表示參考基因組內之與核酸片段所對應的可能位置。可以產生且輸出具有SAM (序列比對圖譜)格式或BAM (二進位)格式的輸出檔案用於進一步分析。

【0140】 現參看圖8B，圖8B為根據一個實施例之用於處理DNA樣本之分析系統800的方塊圖。分析系統實施用於分析DNA樣本的一或多個計算裝置。分析系統800包括序列處理器840、序列資料庫845、模型資料庫855、模型850、參數資料庫865及評分引擎860。在一些實施例中，分析系統800執行圖1A之方法100、圖2之200、圖3之240、圖4之400、圖6之600及本文所述之其他方法中的一或多個步驟。

【0141】 序列處理器840針對來自樣本之片段產生甲基化狀態向量。在片段上之各CpG位點，序列處理器840經由圖1A之方法100產生各片段之甲基化狀態向量，該甲基化狀態向量指明片段在參考基因組中之位置、片段中之CpG位點數目，及片段中之各CpG位點是否甲基化、未甲基化或不確定的甲基化狀態。序列處理器840可以將片段之甲基化狀態向量儲存於序列資料庫845中。可以組織序列資料庫845中之資料以使得來自樣本之甲基化狀態向量彼此相關聯。

【0142】 另外，可以將多個不同模型850儲存於模型資料庫855中或擷取以供測試樣本使用。在一個實例中，模型為一種經訓練之癌症分類器，其利用來源於異常片段的特徵向量確定測試樣本之癌症預測。癌症分類器之訓練及使用將結合章節V.*高/低甲基化區域及分類器(Hyper/Hypo*

Methylated Regions and a Classifier)進一步論述。分析系統800可以訓練一或多個模型850且將多個經訓練的參數儲存於參數資料庫865中。分析系統800將模型850以及函數儲存於模型資料庫855中。

在推理期間，評分引擎860利用一或多個模型850返回輸出。評分引擎860存取模型資料庫855中之模型850以及參數資料庫865中之訓練參數。根據各模型，評分引擎接收模型之適當輸入且基於所接收之輸入、各模型之參數及函數計算輸出，從而使輸入與輸出關聯。在一些使用個案中，評分引擎860進一步計算度量值，從而與來自模型之所計算輸出之信賴度相關聯。在其他使用個案中，評分引擎860計算模型中使用的其他中間值。

VII. 應用

【0143】 在一些實施例中，本發明之方法、分析系統及/或分類器可以用於偵測癌症之存在、監測癌症進展或復發、監測治療反應或有效性、確定最少殘留病變(MRD)之存在或監測最小殘餘疾病(MRD)，或其任何組合。舉例而言，如本文所述，分類器可以用於產生樣本特徵向量來自患有癌症之個體的似然度或機率分數(例如0至100)。在一些實施例中，對機率分數與臨限機率進行比較以確定個體是否患有癌症。在其他實施例中，可以在不同時間點(例如治療之前或之後)評估似然度或機率分數以監測疾病進展或監測治療有效性(例如治療功效)。在仍其他實施例中，似然度或機率分數可以用於產生或影響臨床決策(例如癌症診斷、治療選擇、治療有效性評估等)。舉例而言，在一個實施例中，若似然度或機率分數超過臨限值，則醫師可以開具適當療法處方。

VII.A. 癌症之早期偵測

【0144】 在一些實施例中，本發明之方法及/或分類器係用於偵測懷疑患有癌症之個體中存在或不存在癌症。舉例而言，分類器(如本文所述)可以用於確定樣本特徵向量來自患有癌症之個體的似然度或機率分數。

【0145】 在一個實施例中，機率分數大於或等於60可以表示個體患有癌症。在仍其他實施例中，機率分數大於或等於65、大於或等於70、大於或等於75、大於或等於80、大於或等於85、大於或等於90或大於或等於95表示個體患有癌症。在其他實施例中，機率分數可以指示疾病嚴重程度。舉例而言，相較於低於80的分數(例如分數70)，機率分數80可以指示癌症之較嚴重形式或晚期。類似地，機率分數隨時間(例如在第二個後續時間點)增加可以指示疾病進展或機率分數隨時間(例如在第二個後續時間點)減小可以指示治療成功。

【0146】 在另一個實施例中，可以藉由對癌變機率相對於非癌變機率(亦即，一減去癌變機率)之比率取對數來計算測試個體之癌症勝算比對數，如本文所述。根據此實施例，癌症勝算比對數大於1可以指示個體患有癌症。在仍其他實施例中，癌症機勝算比對數大於1.2、大於1.3、大於1.4、大於1.5、大於1.7、大於2、大於2.5、大於3、大於3.5或大於4指示個體患有癌症。在其他實施例中，癌症勝算比對數可以指示疾病之嚴重程度。舉例而言，相較於低於2之分數(例如1分)，癌症勝算比對數大於2可以指示癌症之較嚴重形式或晚期。類似地，癌症勝算比對數隨時間(例如在第二個後續時間點)增加可以指示疾病進展，或癌症勝算比對數隨時間(例如第二個後續時間點)減小可以指示治療成功。

【0147】 根據本發明之態樣，可以訓練本發明之方法及系統以偵測多種癌症適應症或對其進行分類。舉例而言，本發明之方法、系統及分類

器可以用於偵測一或多種、兩種或更多種、三種或更多種、五種或更多種、十種或更多種、十五種或更多種、或二十種或更多種不同癌症類型之存在。

【0148】 可以使用本發明之方法、系統及分類器偵測之癌症實例包括癌瘤、淋巴瘤、母細胞瘤、肉瘤及白血病或淋巴惡性疾病。更特定而言，此類癌症之實例包括(但不限於)鱗狀細胞癌(例如上皮鱗狀細胞癌)；皮膚癌；黑色素瘤；肺癌，包括小細胞肺癌、非小細胞肺癌(「NSCLC」)、肺腺癌及肺鱗狀癌；腹膜癌；胃或胃臟癌，包括胃腸癌；胰臟癌(例如胰管腺癌)；子宮頸癌；卵巢癌(例如高嚴重度漿液性卵巢癌瘤)、肝癌(例如肝細胞癌(HCC))、肝瘤、肝癌瘤、膀胱癌(例如尿道上皮膀胱癌)、睪丸(生殖細胞腫瘤)癌症、乳癌(例如HER2陽性、HER2陰性及三陰性乳癌)、腦癌(例如星形細胞瘤、神經膠質瘤(例如神經膠母細胞瘤))、結腸癌、直腸癌、結腸直腸癌、子宮內膜或子宮癌、唾液腺癌瘤、腎臟或腎癌(例如腎細胞癌、腎母細胞瘤或威爾姆斯氏腫瘤(Wilms' tumor)、前列腺癌、外陰癌、甲狀腺癌、肛門癌瘤、陰莖癌瘤、頭頸癌、食道癌及鼻咽癌(NPC)。癌症之其他實例包括(但不限於)視網膜母細胞瘤、泡膜細胞瘤、卵巢男胚瘤、血液惡性疾病(包括(但不限於)非霍奇金氏淋巴瘤(non-Hodgkin's lymphoma，NHL)、多發性骨髓瘤及急性血液惡性疾病)、子宮內膜異位、纖維肉瘤、絨膜癌、喉癌、卡波西氏肉瘤(Kaposi's sarcoma)、神經鞘瘤、寡樹突神經膠質瘤、神經母細胞瘤、橫紋肌肉瘤、骨原性肉瘤、平滑肌肉瘤及泌尿道癌瘤。

【0149】 在一些實施例中，癌症為以下中之一或多者：肛門直腸癌、膀胱癌、乳癌、子宮頸癌、結腸直腸癌、食道癌、胃癌、頭頸癌、肝

膽癌、白血病、肺癌、淋巴瘤、黑色素瘤、多發性骨髓瘤、卵巢癌、胰臟癌、前列腺癌、腎癌、甲狀腺癌、子宮癌或其任何組合。

【0150】 在一些實施例中，一或多種癌症可為「高信號」癌症(定義為5年癌症特異性死亡率大於50%的癌症)，諸如肛門直腸癌、結腸直腸癌、食道癌、頭頸癌、肝膽癌、肺癌、卵巢癌及胰臟癌，以及淋巴瘤及多發性骨髓瘤。高信號癌症具有較強侵襲性傾向且在獲自患者的測試樣本中典型地具有高於平均值的游離核酸濃度。

VII.B.癌症及治療監測

【0151】 在一些實施例中，可以在不同時間點(例如治療之前或之後)評估似然度或機率分數以監測疾病進展或監測治療有效性(例如治療功效)。舉例而言，本發明包括的方法包括：在第一個時間點自癌症患者獲得第一樣本(例如第一血漿cfDNA樣本)；確定其第一似然度或機率分數(如本文所述)；在第二時間點獲得第二測試樣本(例如第二血漿cfDNA樣本)；及確定其第二似然度或機率分數(如本文所述)。

【0152】 在某些實施例中，第一時間點係在癌症治療之前(例如在切除手術或治療性干預之前)，且第二時間點係在癌症治療之後(例如在切除手術或治療性干預之後)，且該方法用於監測治療有效性。舉例而言，若第二似然度或機率分數相較於第一似然度或機率分數減小，則認為治療已成功。然而，若第二似然度或機率分數相較於第一似然度或機率分數增加，則認為治療尚未成功。在其他實施例中，第一與第二時間點係在癌症治療之前(例如在切除手術或治療性干預之前)。在仍其他實施例中，第一時間點與第二時間點均在癌症治療之後(例如在切除手術或治療性干預之前)且該方法用於監測治療有效性或治療有效性降低。在仍其他實施例

中，在第一及第二時間點自癌症患者獲得cfDNA樣本且加以分析，例如監測癌症進展、確定癌症是否緩解(例如治療後)、監測或偵測殘餘疾病或疾病復發，或監測治療(例如治療性)功效。

【0153】 熟習此項技術者將容易瞭解，可在任何所需的一組時間點自癌症患者獲得測試樣本且根據本發明之方法加以分析以監測患者之癌症狀態。在一些實施例中，第一及第二時間點相隔一定時長，該時長範圍為約15分鐘至長達約30年，諸如約30分鐘，諸如約1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、 、17、18、19、20、21、22、23或約24小時，諸如約1、2、3、4、5、10、15、20、25或約30天，或諸如約1、2、3、4、5、6、7、8、9、10、11或12個月，或諸如約1、1.5、2、2.5、3、3.5、4、4.5、5、5.5、6、6.5、7、7.5、8、8.5、9、9.5、10、10.5、11、11.5、12、12.5、13、13.5、14、14.5、15、15.5、16、16.5、17、17.5、18、18.5、19、19.5、20、20.5、21、21.5、22、22.5、23、23.5、24、24.5、25、25.5、26、26.5、27、27.5、28、28.5、29、29.5或約30年。在其他實施例中，測試樣本可以自患者獲得，每3個月至少一次、每6個月至少一次、一年至少一次、每2年至少一次、每3年至少一次、每4年至少一次，或每5年至少一次。

VII.C.治療

【0154】 在仍另一個實施例中，似然度或機率分數可以用於產生或影響臨床決策(例如癌症診斷、治療選擇、治療有效性評估等)。舉例而言，在一個實施例中，若似然度或機率分數超過臨限值，則醫師可以開具適當治療處方(例如切除手術、輻射療法、化學療法及/或免疫療法)。

【0155】 分類器(如本文所述)可以用於確定樣本特徵向量來自患有

癌症之個體的似然度或機率分數。在一個實施例中，當該似然度或臨限值超過一種臨限值時，開具適當治療處方(例如切除手術或治療)。舉例而言，在一個實施例中，若似然度或機率分數大於或等於60，則開具一或多種適當治療處方。在另一實施例中，若似然度或機率分數大於或等於65、大於或等於70、大於或等於75、大於或等於80、大於或等於85、大於或等於90、或大於或等於95，則開具一或多種適當治療處方。在其他實施例中，癌症勝算比對數可以指示癌症治療有效性。舉例而言，癌症勝算比對數隨時間(例如在治療後的第二個時間點)增加可以指示治療無效。類似地，癌症勝算比對數隨時間(例如在治療後的第二個時間點)減小可以指示治療成功。在另一個實施例中，若癌症勝算比對數大於1、大於1.5、大於2、大於2.5、大於3、大於3.5或大於4，則開具一或多種適當治療處方。

【0156】 在一些實施例中，療法為一或多種選自由以下組成之群的癌症治療劑：化學治療劑、標靶癌症治療劑、分化治療劑、激素治療劑及免疫治療劑。舉例而言，療法可為一或多種選自由以下組成之群的化學治療劑：烷基化劑、抗代謝物、蒽環黴素(anthracyclines)、抗腫瘤抗生素、細胞骨架瓦解劑(紫杉烷(taxans))、拓樸異構酶抑制劑、有絲分裂抑制劑、皮質類固醇、激酶抑制劑、核苷酸類似物、基於鉑之藥劑及其任何組合。在一些實施例中，治療為一或多種選自由以下組成之群的標靶癌症治療劑：信號轉導抑制劑(例如酪胺酸激酶及生長因子受體抑制劑)、組蛋白脫乙醯基酶(HDAC)抑制劑、視黃酸受體促效劑、蛋白酶體抑制劑、血管生成抑制劑及單株抗體結合物。在一些實施例中，療法為一或多種分化治療劑，包括類視黃素，諸如維甲酸(tretinoin)、亞利崔托寧(alitretinoin)及貝瑟羅汀(bexarotene)。在一些實施例中，療法為一或多種選自由以下

組成之群的激素治療劑：抗雌激素、芳香酶抑制劑、孕激素、雌激素、抗雄激素及GnRH促效劑或類似物。在一個實施例中，療法為一或多種選自包含以下之群的免疫治療劑：單株抗體療法，諸如利妥昔單抗(rituximab)(RITUXAN)及阿侖單抗(alemtuzumab)(CAMPATH)、非特異性免疫療法及佐劑，諸如BCG、介白素-2 (IL-2)及干擾素-α；免疫調節藥物，例如沙立度胺(thalidomide)及來那度胺(lenalidomide)(REVLIMID)。基於諸如以下之特徵選擇適當癌症治療劑係在熟練醫師或腫瘤學家之能力範圍內：腫瘤類型、癌症階段、先前對癌症療法或治療劑的暴露，及癌症之其他特徵。

VIII. 實例

VIII.A.利用偵測異常甲基化片段之方法診斷癌症

【0157】 研究設計及樣本：CCGA (NCT02889978)為一種結合縱向隨訪的前瞻性、多中心、個案對照、觀測型研究。自142個地點、自約15,000位參與者收集經去除身分資訊的生物樣品。將樣本分成訓練集(1,785)及測試集(1,015)；選擇樣本以確保跨地點之癌症類型及非癌症在各群組中的預定分佈，並且使癌症及非癌症樣本的頻率與性別、年齡匹配。

【0158】 全基因組亞硫酸氫鹽測序：自血漿中分離出Cf DNA，且利用全基因組亞硫酸氫鹽測序(WGBS；30x深度)進行cfDNA分析。使用經改良之QIAamp循環核酸套組(Qiagen；Germantown, MD)自每個患者兩管血漿(至多10 ml組合體積)中提取CfDNA。使用EZ-96 DNA甲基化套組(Zymo Research, D5003)使至多75 ng血漿cfDNA經受亞硫酸氫鹽轉化。利用經轉化之cfDNA、使用Accel-NGS Methyl-Seq DNA文庫製備套

組(Swift BioSciences；Ann Arbor, MI)製備雙索引化測序文庫且所構建之文庫係使用Illumina平台之KAPA文庫定量套組(Kapa Biosystems；Wilmington, MA)定量。將四個文庫與10% PhiX v3文庫(Illumina，FC-110-3001)彙集且在Illumina NovaSeq 6000 S2流動池上聚類，隨後進行150 bp成對末端測序(30x)。

【0159】 **cfDNA分析及癌症相對於非癌症之分類：**對於各樣本而言，將WGBS片段集簡化成具有異常甲基化模式之片段的小子集。另外，選擇高甲基化或低甲基化cfDNA片段。針對具有異常甲基化模式及高或高甲基化而選擇的cfDNA片段在本文中稱為「極端甲基化狀態之異常片段」或「UFXM」。未患癌症之個體中高頻率存在或具有不穩定甲基化的片段不大可能產生有高區分度的特徵供癌症狀態分類用。吾等因此利用來自CCGA研究之108位未患癌症之非吸菸參與者(年齡：58±14歲，79 [73%]女性)之獨立參考集(亦即，參考基因組)產生典型片段的統計學模型及資料結構。此等樣本係用於訓練馬可夫鏈模型(3階)，從而估算片段內之CpG甲基化狀態之所指定序列的似然度，如上文在IV.B中進一步所述。此模型已證明被校準於正常片段範圍內(p值>0.001)且用於拒斥馬可夫模型p值因>=0.001而異常度不足的片段。

【0160】 如上文所述，另一資料簡化步驟僅選擇覆蓋至少5個CpG且平均甲基化>0.9 (高甲基化)或<0.1 (低甲基化)的片段。此程序對於訓練中之未患癌症之參與者產生2,800 (1,500-12,000)個UFXM片段之中值(範圍)，且對於訓練中之患有癌症之參與者產生3,000 (1,200-220,000)個UFXM片段之中值(範圍)。由於此資料簡化程序僅使用參考集資料，因此此階段僅需應用於各樣本一次。

【0161】在基因組內之所選基因座，分開構建高甲基化UFXM與低甲基化UFXM之癌症狀態資訊豐富的近似比率對數分數。首先，產生該基因座處之各樣本的二進位特徵：若UFXM片段與該樣本內之該基因座不重疊，則二進位特徵為0；若UFXM片段與該基因座存在重疊，則二進位特徵為1。接著對來自患有癌症(C_c)及未患癌症(C_nc)之參與者之樣本的正值(1s)數目計數。比率對數分數接著如下構建：log(C_c+1)-log(C_nc+1)，將正則項添加至計數中，且將與每組內之樣本總數目相關的標準化項捨棄，原因在於其為恆定的(log[N_nc+2]-log[N_c+2])。構建基因組內之所有CpG位點之位置處的分數，從而向約25M個基因座賦予分數：一個分數賦予UFXM高甲基化片段且一個分數賦予UFXM低甲基化片段。

【0162】在基因座特異性比率對數分數給定的情況下，藉由在片段內之基因座之所有比率對數分數中取最大值且與高甲基化或低甲基化之甲基化類別匹配來對樣本中之UFXM片段評分。由此使樣本內之每個UFXM片段產生一個分數。

【0163】藉由獲取各樣本內之極端排序片段子集之分數(高甲基化片段與低甲基化片段分開獲取)而將樣本內的此片段層面分數簡化成每個樣本之小特徵集。以此方式，利用小型適用特徵集捕捉各樣本中資訊最豐富之片段的資訊。在低cfDNA腫瘤分離物樣本中，預期具有豐富異常資訊的片段僅佔少數。

【0164】在各類片段中，針對各類高甲基化及低甲基化UFXM內的片段選擇秩1、2、4…64 (2^i，i屬於0:6)的最大分數，從而產生14個特徵(7與7)。為了調整樣本測序深度，排序程序係作為與分數映射的函數秩處

理，且吾等在觀測到之分數之間內插以獲得對應於調整秩的分數。秩依線性比例調整至相對樣本深度：若相對樣本深度為x，則在x*初始秩獲取所內插之分數(亦即，x=1.1，吾等獲取在秩1.1、2.2、…、$x*2^i$所計算的分數)。接著將用於進一步分類的經調整之14個極端秩分數之集合賦予每個樣本。

【0165】 在特徵向量給定的情況下，在利用特徵預測癌症/非癌症狀態時，使用內核邏輯回歸分類器捕捉潛在非線性。特定言之，使用各向同性徑向基底函數(冪指數2)作為內核，結合比例參數γ及L2正則化參數λ(藉由除以m^2來調整，其中m為樣本數目，因此λ自然地隨著訓練資料之量按比例調整)，訓練正則化內核邏輯回歸分類器(KLR)。在指定的訓練資料內，利用內部交叉驗證使γ及λ最佳化以便留出損失對數，且在最大值開始且每個步驟將參數減半，在7個相乘步驟中、在範圍$1,1e^{-2}$ (γ)、$1e^{3}$-$1e^{1}$ (λ)內利用網格搜尋來最佳化。內部交叉驗證折回期間的中值最佳參數對於γ為0.125且對於λ為125。

【0166】 **驗證經訓練之癌症分類器：**為了評估此極端秩分數分類器程序對CCGA子研究資料集之效能，將交叉驗證應用於訓練集，將樣本分成10個折回。留出每一個折回且用剩餘9/10資料訓練ERS分類器(在彼等折回內利用內部交叉驗證最佳化γ及λ)。特徵化中所用的比率對數分數僅自訓練折回存取資料。將得自每一次留出折回的輸出分數彙集且用於構建接收者操作特徵(ROC)曲線用於實施。為評估CCGA測試集，使用整個訓練資料集構建分數及單一KLR分類器，其接著應用於測試資料集。

【0167】 利用分類器估算靈敏度及特異性；針對分析特異性干擾生物學信號(例如CH、血液科條件、與年齡有關的變異)校正各分類器或抑

制該等信號。利用非癌症個案估算校正干擾信號之後的特異性。靈敏度與特異性之間的關係藉由圖9中所提供的接收者操作特徵(ROC)曲線描繪，其證明分析潛在地具有高特異性。訓練資料集與測試資料集的曲線下面積(AUC)值相似。相較於所有癌症(0.73及0.71)，特定癌症組之AUC值顯著較高(0.88及0.87)。展示分析具有高特異性且稱為「高信號癌症」(定義為5年癌症特異性死亡率大於50%的癌症)的癌症組包括若干實體癌(肛門直腸癌、結腸直腸癌、食道癌、頭頸癌、肝膽癌、肺癌、卵巢癌及胰臟癌)以及淋巴瘤及多發性骨髓瘤。

【0168】 進一步研究截止值98%特異性時的靈敏度，從而(1)估算出年齡≥50歲個人的隱性癌症率(SEER)為每年約1.3%，及(2)考慮非癌症參與者之持續隨訪。靈敏度估算值在跨越癌症類型之訓練集與測試集之間大體一致(圖10)。結果進一步描繪於圖11A-B及表1中。

表1

	訓練/交叉驗證		測試	
靈敏度(95% CI)	**n**	**WGBS**	**n**	**WGBS**
肛門直腸	7	86% (42-100)	2	100% (16-100)
膀胱	10	40% (12-74)	1	0% (0-98)
乳房	339	22% (18-27)	170	14% (9-20)
子宮頸	13	46% (19-75)	8	25%(3-65)
結腸直腸	45	78% (63-89)	39	62% (45-77)
食道	24	67% (45-84)	7	43% (10-82)
胃	11	36% (11-69)	13	46% (1975)
頭頸	19	74% (49-91)	12	50% (21-79)
肝膽	13	92% (64-100)	14	79% (49-95)
白血病	10	40% (12-74)	13	23% (5-54)
肺	118	63% (53-71)	46	70% (54-82)
淋巴瘤	22	64% (41-83)	18	67% (41-87)
黑色素瘤	10	10% (0-45)	8	25% (3-65)
多發性骨髓瘤	11	64% (31-89)	8	62% (24-91)
卵巢	17	82% (57-96)	7	71% (29-96)
胰臟	26	77% (56-91)	22	77% (55-92)
前列腺	69	7% (2-16)	55	0% (0-6)

腎	26	23% (9-44)	13	15% (2-45)
甲狀腺	13	0 (0-25)	5	0% (0-52)
子宮	27	11% (2-29)	9	22% (3-60)
多原發癌	6	50% (12-88)	0	-
不明原發癌/其他	19	74% (49-91)	15	53% (27-79)
CI：信賴區間。WGS：全基因組測序。WGBS：全基因組亞硫酸氫鹽測序。 資料包括階段I-IV。				

【0169】 在跨越所有癌症類型之訓練集中，98%特異性時之總體靈敏度為39.5% (36-43%)；此與測試集一致(34.2% [30-39%])。如所預期，靈敏度隨著癌症階段而增加。在98%特異性之訓練集中，高信號癌症在98%特異性時之靈敏度為70.2% (65-75%)，此與測試集一致(66.9% [59-74%])。

【0170】 結果表明cfDNA測序及對其甲基化狀態的分析可以高特異性偵測癌症。此支持用於早期跨階段偵測(潛在地偵測較大比例之癌症，包括一些高死亡率之癌症)之可行性。

VIII. 其他考量

【0171】 應理解，本發明之圖式及說明書已經簡化以說明對於清楚理解本發明相關之元件，同時出於清晰之目的消除在典型系統中發現之許多其他元件。一般熟習此項技術者可以認識到，在實施本發明時需要及/或必需其他元件及/或步驟。然而，由於此類元件及步驟在此項技術中已熟知，且由於其無助於對本發明之更好理解，因此本文中不提供對此類元件及步驟之論述。本文中之揭示內容係關於對熟習此項技術者已知之此類元件及方法的所有此類變更及修飾。

【0172】 上述說明之一些部分依據算法及運算符號表示資訊來描述。熟習資料處理技術者常用此等演算法說明及表示來將其實質性工作有效地傳達給熟習此項技術的其他者。此等運算雖然在函數上、計算上或邏

輯上描述，但應理解為由電腦程式或等效電路、微碼或類似者來實施。所述運算及其相關模組可以軟體、韌體、硬體或其任何組合實施。

【0173】 如本文所用，對「一個實施例」或「一實施例」的任何提及意謂結合實施例所述的特定零件、特點、結構或特徵包括於至少一個實施例中。片語「在一個實施例中」出現於說明書中之多處不一定皆指相同實施例，藉此提供所述實施例之多種可能性在一起發揮作用的構架。

【0174】 如本文所用，術語「包含(comprises/comprising)」、「包括(includes/including)」、「具有(has/having)」或其任何其他變化形式意欲涵蓋非排他性包括。舉例而言，包含元件清單之製程、方法、物件或裝置不必僅限於彼等元件，而是可以包括未明確列舉或此類製程、方法、物件或裝置所固有的其他元件。此外，除非明確相反地陳述，否則「或」係指包括性的或，而非指排他性的或。舉例而言，以下中之任一者滿足條件A或條件B：A為真(或存在)且B為假(或不存在)；A為假(或不存在)且B為真(或存在)；且A與B均為真(或存在)。

【0175】 另外，「一(a)」或「一(an)」之使用係用於描述本文實施例之元件及組件。此舉僅為方便起見及得到本發明的普遍意義。除非明顯另有所指，否則此描述應理解為包括一者或至少一者，且單數亦包括複數。

【0176】 雖然已說明且描述特定實施例及應用，但應理解，所揭示之實施例不限於本文中所揭示之確切構造及組件。在不脫離隨附申請專利範圍中所界定之精神及範疇的情況下，可對本文中所揭示之方法及裝置之配置、操作及細節作出熟習此項技術者顯而易見的各種修飾、變化及變更。

【符號說明】

【0177】

23 CpG位點

24 CpG位點

25 CpG位點

100 產生樣本中之游離(cf) DNA片段之甲基化狀態向量/產生對照組之甲基化狀態向量集/產生驗證組之甲基化狀態向量集/產生來自樣本之甲基化狀態向量集

110 自個體獲得樣本且分離出cfDNA片段

112 片段cfDNA

114 甲基組

120 處理經分離之cfDNA片段，使未甲基化胞嘧啶轉化成尿嘧啶

122 經轉化之cfDNA

130 製備測序文庫

135 使用可以靶向特定區域之雜交探針富集測序文庫

140 測序以獲得序列讀段

142 序列讀段

144 參考基因組

150 藉由將序列讀段與參考基因組比對來確定複數個甲基化位點之甲基化狀態

152 甲基化狀態載體

160 產生甲基化狀態向量，該甲基化狀態向量指明cfDNA片段中之甲基化位點之位置及各甲基化位點之甲基化狀態

170 轉化準確度

180 一致片段長度

190 每個樣本之cfDNA濃度

200 產生對照組之資料結構

210 將各甲基化狀態向量再劃分成甲基化位點字串

220 對各位置與甲基化狀態組合之字串計數

230 建立資料結構，該資料結構儲存來自對照組之所有可能字串的計數

240 驗證資料結構一致性

310 針對各甲基化狀態向量，列舉該位置之所有可能甲基化狀態向量

320 計算來自對照組資料結構之所有可能甲基化狀態向量的機率

330 基於所計算之機率計算各甲基化狀態向量之p值分數

340 構建驗證組之所有p值的累積密度函數(CDF)

350 驗證對照組資料結構內之p值一致性

400 鑑別來自個體之異常甲基化片段/自非癌症訓練組及癌症訓練組獲得異常片段

405 過濾具有不確定狀態之片段

410 針對各甲基化狀態向量，列舉該位置之所有可能甲基化狀態向量

420 計算來自對照組治療結構之所有可能甲基化狀態向量的機率

430 基於所計算之機率計算各甲基化狀態向量之p值分數

435 對大於甲基化位點之臨限數目的甲基化狀態向量使用滑動窗

440 基於低於臨限值之p值分數過濾集合，從而產生異常甲基化向量之子集

450 鑑別過濾集內之低甲基化片段或高甲基化片段

460 將經訓練之分類模型應用於過濾集以確定癌症

470 藉由計算癌症勝算比對數來確定癌症狀態

500 利用馬可夫鏈模型計算P值

505 測試甲基化狀態向量

515 可能甲基化狀態向量之機率

525 測試甲基化狀態向量之P值

600 訓練分類器

610 針對來自兩個訓練組之各甲基化狀態向量，確定是否為高甲基化的或低甲基化的

620 針對基因組中之各甲基化位點，產生低甲基化分數及高甲基化分數

630 針對各甲基化狀態向量，產生低甲基化總分及高甲基化總分

640 針對每位個體，依據低甲基化總分及高甲基化總分將甲基化狀態向量排序

650 針對每位個體，基於排序來產生特徵向量

660 進行訓練以將來自非癌症訓練組與來自癌症訓練組的特徵向量區分開來

700 所有癌症勝算比對數

710 乳癌勝算比對數

720 乳癌亞型勝算比對數

730 肺癌勝算比對數

740 結腸直腸癌勝算比對數

750 前列腺癌勝算比對數

800 分析系統

810 核酸樣本

820 測序儀

825 圖形使用者介面

830 裝載站

840 序列處理器

845 序列資料庫

850 模型

855 模型資料庫

860 評分引擎

865 參數資料庫

【發明申請專利範圍】

【第1項】

一種自游離去氧核糖核酸(cfDNA)樣本片段中偵測測試個體之癌症之方法，該cfDNA樣本片段係收集自該測試個體及隨後測序，該方法包含：

存取資料結構，該資料結構包含參考基因組內之CpG位點字串的計數及其來自訓練片段集的各別甲基化狀態；

針對樣本片段產生樣本狀態向量，該樣本狀態向量包含該參考基因組內之樣本基因組位置及該樣本片段中之複數個CpG位點中之每一者的甲基化狀態，各甲基化狀態經測定為甲基化的或未甲基化的；

列舉長度與該樣本狀態向量相同之該樣本基因組位置之甲基化狀態的複數種可能性；

針對該等可能性中之每一者，藉由存取該資料結構中所儲存之該等計數來計算機率；

鑑別與該樣本狀態向量匹配的該可能性且相應地將所計算之該機率鑑別為樣本機率；

基於該樣本機率，相對於該訓練片段集，針對該樣本片段產生該樣本狀態向量的分數，其藉由：

鑑別出一或多個小於該樣本機率的甲基化狀態可能性之所計算機率，及

針對該樣本片段，藉由將該一或多個鑑別出之所計算機率與該樣本機率求和來產生分數；

基於所產生之該分數，確定該樣本片段是否具有異常甲基化模式；及

回應於確定該樣本片段具有異常甲基化模式，將癌症分類器應用於該樣本狀態向量，以確定對產生該樣本片段之測試對象的癌症預測，其中根據該癌症預測開具癌症治療處方。

【第2項】

如請求項1之方法，其中該等CpG位點字串中之每一者包含該參考基因組內之複數個基因組位置之該等CpG位點中之每一者的甲基化狀態，其中該等甲基化狀態中之每一者經測定為甲基化的或未甲基化的。

【第3項】

如請求項1之方法，其進一步包含：

利用該訓練片段集構建該資料結構且包含：

針對該訓練片段集內之各訓練片段，產生訓練狀態向量，該訓練狀態向量包含該參考基因組內之已知基因組位置及該訓練片段中之該複數個CpG位點中之每一者的甲基化狀態，各甲基化狀態經測定為甲基化的或未甲基化的；

確定複數個字串，其中各字串為該訓練狀態向量的一部分，

量化該等訓練狀態向量中之各字串的計數；及

將各字串之複數個計數儲存於該資料結構中。

【第4項】

如請求項1之方法，其中基於所產生之該分數確定該樣本片段是否具有異常甲基化模式進一步包含確定針對該樣本片段所產生之該分數是否低於臨限分數，其中該臨限分數指示該樣本片段具有異常甲基化模式之信賴度。

【第5項】

如請求項1之方法，其中該訓練片段集包含來自一或多個健康個體的訓練片段，其中該一或多個健康個體缺乏特定醫學病症且其中相對於來自該一或多個健康個體的該訓練片段集確定該樣本片段為異常甲基化的。

【第6項】

如請求項1之方法，其中藉由存取該資料結構中所儲存之該等計數來計算該等可能性中之每一者之機率包含：

針對複數個條件元中之每一者，其中各條件元為考慮CpG位點子集具有該可能性的條件性機率，利用該資料結構中所儲存之該等複數個計數計算某一階數之馬可夫鏈機率，其藉由：

鑑別與該條件元匹配之字串數目的第一計數；

鑑別與該條件元之先前甲基化狀態直至全數目長度匹配之字串數目的第二計數；及

藉由將該第一計數除以該第二計數來計算該馬可夫鏈機率。

【第7項】

如請求項6之方法，其中利用該資料結構中所儲存之該複數個計數計算某一階數之馬可夫鏈機率進一步包含執行平滑算法。

【第8項】

如請求項1之方法，其中將該樣本狀態向量分割成包含第一窗口及第二窗口的複數個窗口，其中該第一窗口及該第二窗口為該樣本片段之兩個不同部分；其中鑑別與該樣本狀態向量匹配的該可能性及相應地將所計算之該機率鑑別為該樣本機率包含利用與該第一窗口匹配的第一樣本機率鑑別第一可能性及利用與該第二窗口匹配的第二樣本機率鑑別第二可能性；且其中所產生的該分數係基於該第一樣本機率及該第二機率機率之一。

【第9項】

如請求項1之方法，其進一步包含基於針對各樣本片段所產生的該分數來過濾複數個樣本片段，從而產生具有異常甲基化模式之樣本片段的子集。

【第10項】

如請求項1之方法，其進一步包含：

當該樣本片段包含至少臨限數目個CpG位點、其中超過臨限百分比之CpG位點發生甲基化時，將該樣本片段鑑別為高甲基化的，其中該臨限數目個CpG位點為5個或更多個CpG位點，且其中甲基化CpG位點之該臨限百分比為80%或更大；或

當該樣本片段包含至少臨限數目個CpG位點、其中超過臨限百分比之CpG位點未甲基化時，將該樣本片段鑑別為低甲基化的，其中該臨限數目個CpG位點為5個或更多個CpG位點，且其中未甲基化CpG位點之該臨限百分比為80%或更大。

【第11項】

如請求項1之方法，其中該癌症分類器經來自一或多個患有癌症之個體的癌症訓練片段集及來自一或多個未患癌症之個體的非癌症訓練片段集訓練。

【第12項】

如請求項1之方法，其中將該樣本狀態向量應用於該癌症分類器產生癌症機率及非癌症機率中的至少一者，其中該癌症預測包含基於該癌症機率及該非癌症機率中之至少一者的癌症狀態分數。

【第13項】

一種非暫時性電腦可讀儲存媒體，其儲存用於自游離去氧核糖核酸(cfDNA)樣本片段中偵測測試個體之癌症的指令，該等指令當由處理器執行時，促使處理器執行包含以下之操作：

存取資料結構，該資料結構包含參考基因組內之CpG位點字串的計數及其來自訓練片段集的各別甲基化狀態；

針對樣本片段產生樣本狀態向量，該樣本狀態向量包含該參考基因組內之樣本基因組位置及該樣本片段中之複數個CpG位點中之每一者的甲基化狀態，各甲基化狀態經測定為甲基化的或未甲基化的；

列舉長度與該樣本狀態向量相同之該樣本基因組位置之甲基化狀態的複數種可能性；

針對該等可能性中之每一者，藉由存取該資料結構中所儲存之該等計數來計算機率；

鑑別與該樣本狀態向量匹配的該可能性且相應地將所計算之該機率鑑別為樣本機率；

基於該樣本機率，相對於該訓練片段集，針對該樣本片段產生該樣本狀態向量的分數，其藉由：

鑑別出一或多個小於該樣本機率的甲基化狀態可能性之所計算機率，及

針對該樣本片段，藉由將一或多個鑑別出之所計算機率與該樣本機率求和來產生分數；

基於所產生之該分數，確定該樣本片段是否具有異常甲基化模式；及

回應於確定該樣本片段具有異常甲基化模式，將癌症分類器應用於

該樣本狀態向量，以確定對產生該樣本片段之測試對象的癌症預測，其中根據該癌症預測開具癌症治療處方。

【第14項】

如請求項13之非暫時性電腦可讀儲存媒體，其中該等CpG位點字串中之每一者包含該參考基因組內之複數個基因組位置之該等CpG位點中之每一者的甲基化狀態，其中該等甲基化狀態中之每一者經測定為甲基化的或未甲基化的。

【第15項】

如請求項13之非暫時性電腦可讀儲存媒體，其進一步包含：

利用該訓練片段集構建該資料結構且包含：

針對該訓練片段集內之各訓練片段，產生訓練狀態向量，該訓練狀態向量包含該參考基因組內之已知基因組位置及該訓練片段中之該複數個CpG位點中之每一者的甲基化狀態，各甲基化狀態經測定為甲基化的或未甲基化的；

確定複數個字串，其中各字串為該訓練狀態向量的一部分，

量化該等訓練狀態向量中之各字串的計數；及

將各字串之複數個計數儲存於該資料結構中。

【第16項】

如請求項13之非暫時性電腦可讀儲存媒體，其中基於所產生之該分數確定該樣本片段是否具有異常甲基化模式進一步包含確定針對該樣本片段所產生之該分數是否低於臨限分數，其中該臨限分數指示該樣本片段具有異常甲基化模式之信賴度。

【第17項】

如請求項13之非暫時性電腦可讀儲存媒體，其中該訓練片段集包含來自一或多個健康個體的訓練片段，其中該一或多個健康個體缺乏特定醫學病症且其中相對於來自該一或多個健康個體的該訓練片段集確定該樣本片段為異常甲基化的。

【第18項】

如請求項13之非暫時性電腦可讀儲存媒體，其中藉由存取該資料結構中所儲存之該等計數來計算該等可能性中之每一者之機率包含：

針對複數個條件元中之每一者，其中各條件元為考慮CpG位點子集具有該可能性的條件性機率，利用該資料結構中所儲存之該等複數個計數計算某一階數之馬可夫鏈機率，其藉由：

鑑別與該條件元匹配之字串數目的第一計數；

鑑別與該條件元之先前甲基化狀態直至全數目長度匹配之字串數目的第二計數；及

藉由將該第一計數除以該第二計數來計算該馬可夫鏈機率。

【第19項】

如請求項18之非暫時性電腦可讀儲存媒體，其中利用該資料結構中所儲存之該複數個計數計算某一階數之馬可夫鏈機率進一步包含執行平滑算法。

【第20項】

如請求項13之非暫時性電腦可讀儲存媒體，其中將該樣本狀態向量分割成包含第一窗口及第二窗口的複數個窗口，其中該第一窗口及該第二窗口為該樣本片段之兩個不同部分；其中鑑別與該樣本狀態向量匹配的該可能性及相應地將所計算之該機率鑑別為該樣本機率包含利用與該第一窗

口匹配的第一樣本機率鑑別第一可能性及利用與該第二窗口匹配的第二樣本機率鑑別第二可能性；且其中所產生的該分數係基於該第一樣本機率及該第二機率機率之一。

【第21項】

如請求項13之非暫時性電腦可讀儲存媒體，該操作進一步包含基於針對各樣本片段所產生的該分數來過濾複數個樣本片段，從而產生具有異常甲基化模式之樣本片段的子集。

【第22項】

如請求項13之非暫時性電腦可讀儲存媒體，該操作進一步包含：

當該樣本片段包含至少臨限數目個CpG位點、其中超過臨限百分比之CpG位點發生甲基化時，將該樣本片段鑑別為高甲基化的，其中該臨限數目個CpG位點為5個或更多個CpG位點，且其中甲基化CpG位點之該臨限百分比為80%或更大；或

當該樣本片段包含至少臨限數目個CpG位點、其中超過臨限百分比之CpG位點未甲基化時，將該樣本片段鑑別為低甲基化的，其中該臨限數目個CpG位點為5個或更多個CpG位點，且其中未甲基化CpG位點之該臨限百分比為80%或更大。

【第23項】

如請求項13之非暫時性電腦可讀儲存媒體，其中該癌症分類器經來自一或多個患有癌症之個體的癌症訓練片段集及來自一或多個未患癌症之個體的非癌症訓練片段集訓練。

【第24項】

如請求項13之非暫時性電腦可讀儲存媒體，其中將該樣本狀態向量

應用於該癌症分類器產生癌症機率及非癌症機率中的至少一者，其中該癌症預測包含基於該癌症機率及該非癌症機率中之至少一者的癌症狀態分數。

【第25項】

一種自游離去氧核糖核酸(cfDNA)樣本片段中偵測測試個體之癌症之系統，該系統包含：

一電腦處理器；及

一非暫時性電腦可讀儲存媒體，其儲存用於自游離去氧核糖核酸(cfDNA)樣本片段中偵測測試個體之癌症的指令，該等指令當由處理器執行時，促使處理器執行包含以下之操作：

存取資料結構，該資料結構包含參考基因組內之CpG位點字串的計數及其來自訓練片段集的各別甲基化狀態；

針對樣本片段產生樣本狀態向量，該樣本狀態向量包含該參考基因組內之樣本基因組位置及該樣本片段中之複數個CpG位點中之每一者的甲基化狀態，各甲基化狀態經測定為甲基化的或未甲基化的；

列舉長度與該樣本狀態向量相同之該樣本基因組位置之甲基化狀態的複數種可能性；

針對該等可能性中之每一者，藉由存取該資料結構中所儲存之該等計數來計算機率；

鑑別與該樣本狀態向量匹配的該可能性且相應地將所計算之該機率鑑別為樣本機率；

基於該樣本機率，相對於該訓練片段集，針對該樣本片段產生該樣本狀態向量的分數，其藉由：

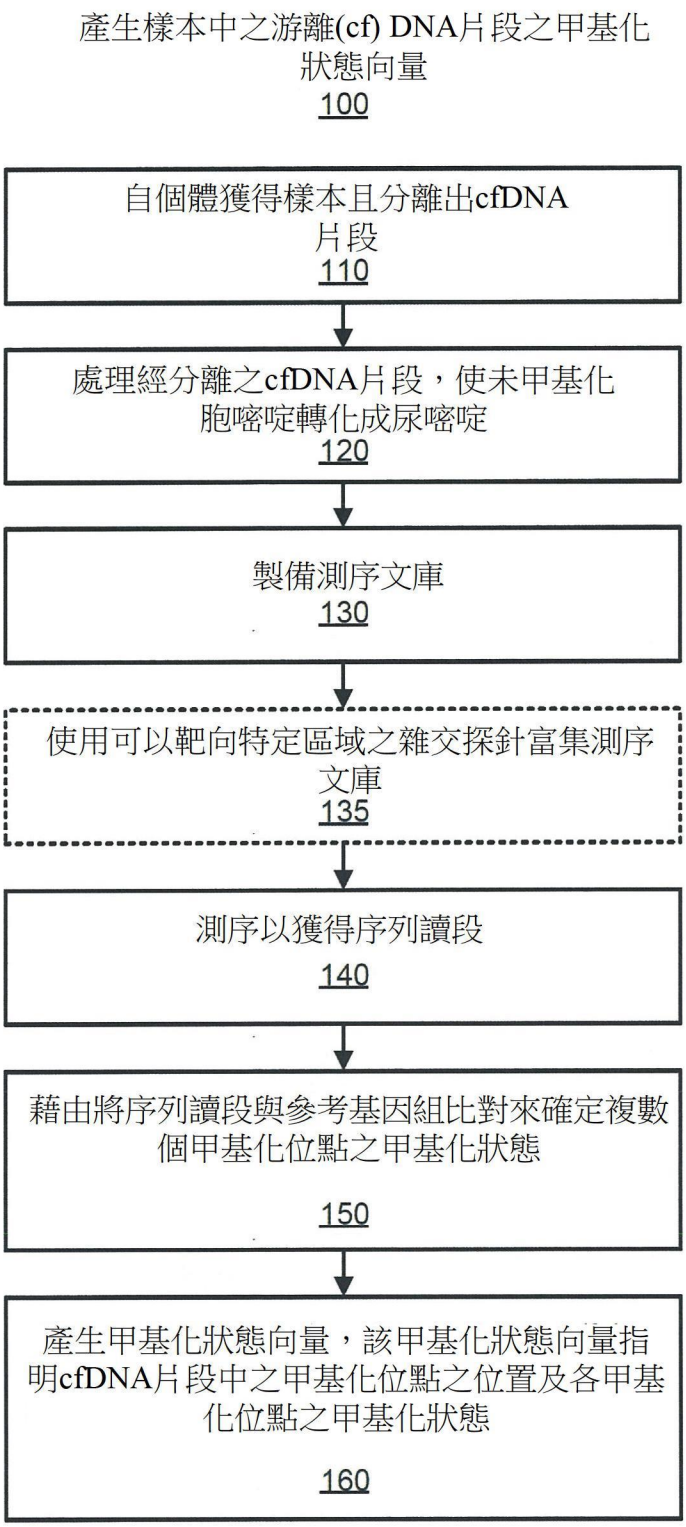
鑑別出一或多個小於該樣本機率的甲基化狀態可能性之所計算機率，及

針對該樣本片段，藉由將一或多個鑑別出之所計算機率與該樣本機率求和來產生分數；

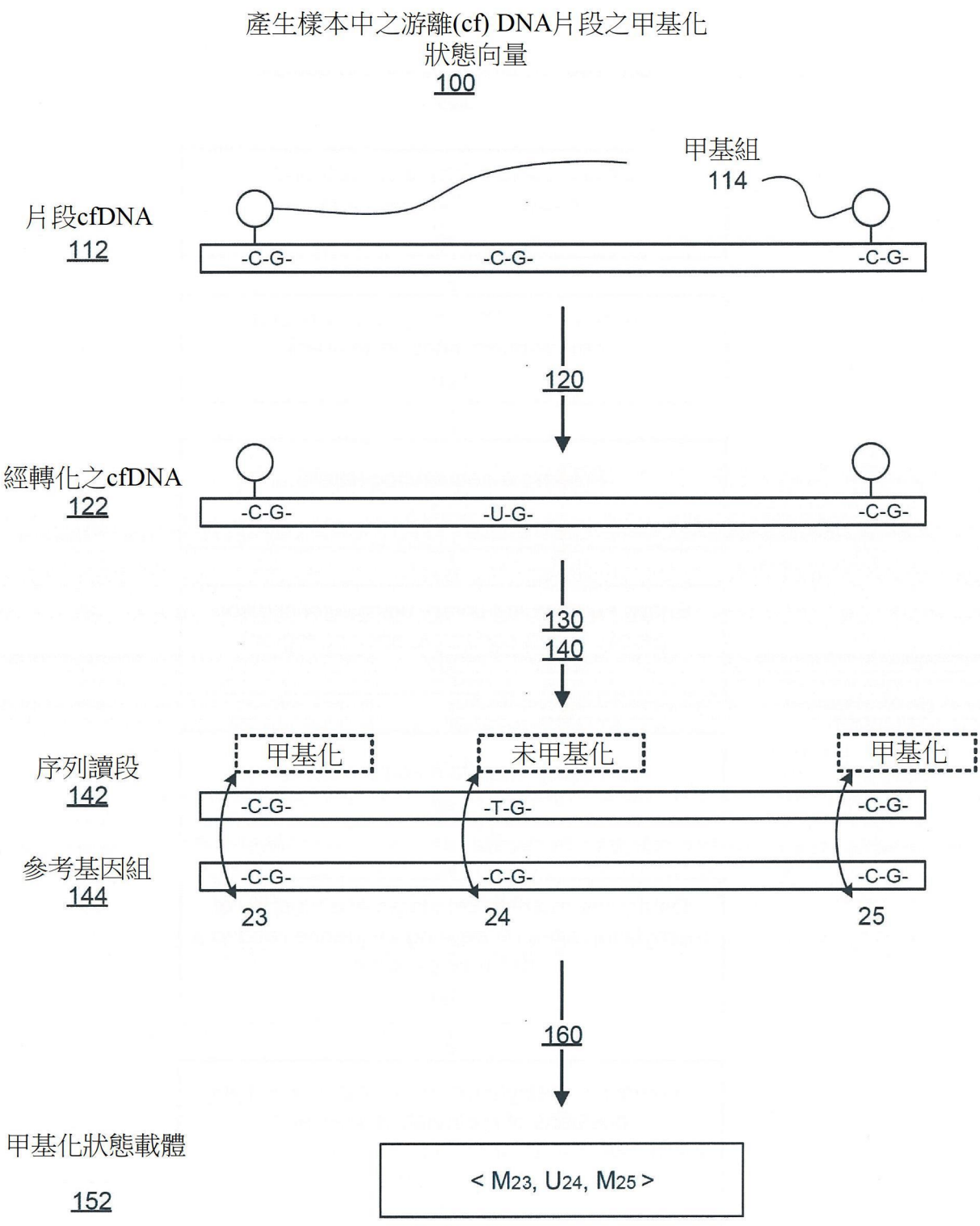
基於所產生之該分數，確定該樣本片段是否具有異常甲基化模式；及

回應於確定該樣本片段具有異常甲基化模式，將癌症分類器應用於該樣本狀態向量，以確定對產生該樣本片段之測試對象的癌症預測，其中根據該癌症預測開具癌症治療處方。

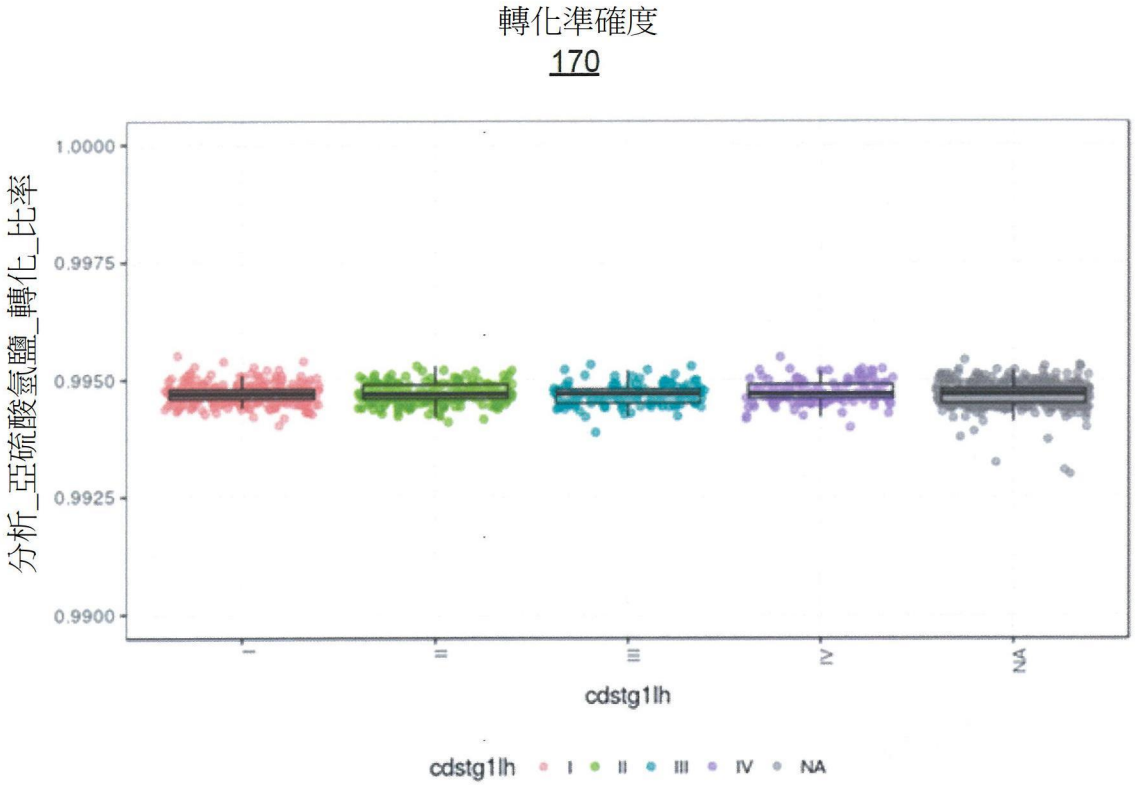
【發明圖式】

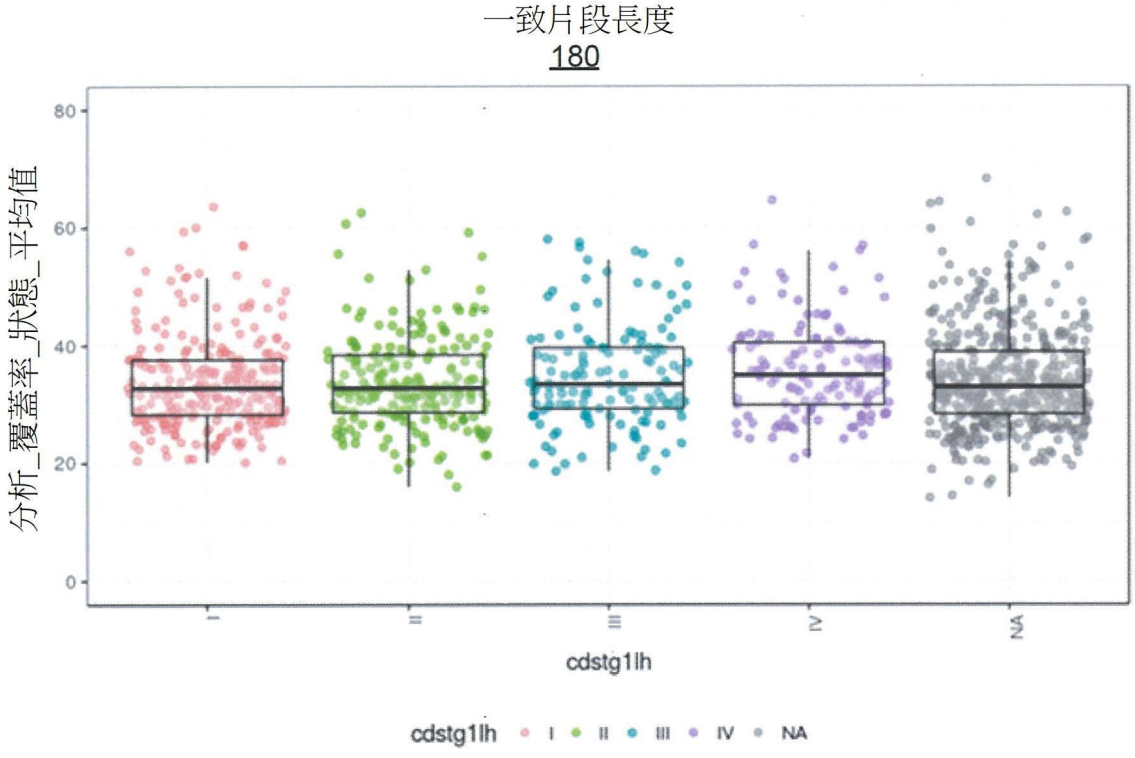

產生樣本中之游離(cf) DNA片段之甲基化狀態向量
100
自個體獲得樣本且分離出cfDNA片段
110
處理經分離之cfDNA片段，使未甲基化胞嘧啶轉化成尿嘧啶
120
製備測序文庫
130
使用可以靶向特定區域之雜交探針富集測序文庫
135
測序以獲得序列讀段
140
藉由將序列讀段與參考基因組比對來確定複數個甲基化位點之甲基化狀態
150
產生甲基化狀態向量，該甲基化狀態向量指明cfDNA片段中之甲基化位點之位置及各甲基化位點之甲基化狀態
160

【圖1A】


產生樣本中之游離(cf) DNA片段之甲基化
狀態向量
100
甲基組
114
片段cfDNA
112
-C-G-
-C-G-
-C-G-
120
經轉化之cfDNA
122
-C-G-
-U-G-
-C-G-
130
140
序列讀段
142
甲基化
未甲基化
甲基化
-C-G-
-T-G-
-C-G-
參考基因組
144
-C-G-
-C-G-
-C-G-
23
24
25
160
甲基化狀態載體
152
< M23, U24, M25 >

【圖1B】

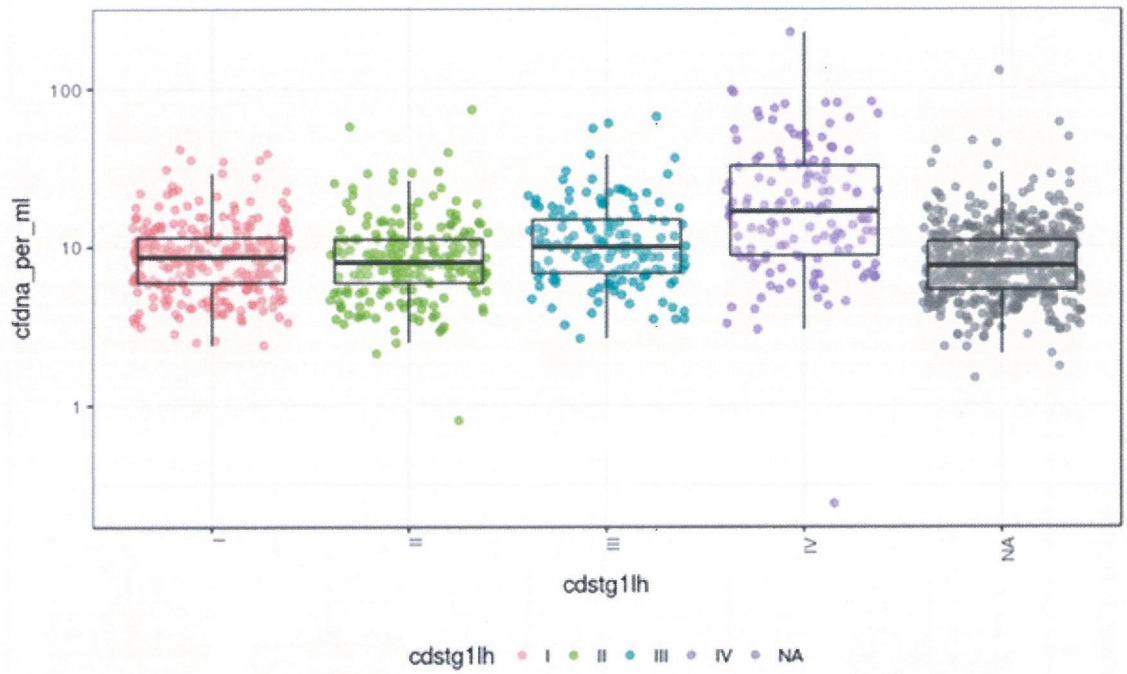

轉化準確度
170
分析_亞硫酸氫鹽_轉化_比率
1.0000
0.9975
0.9950
0.9925
0.9900
I
II
III
IV
NA
cdstg1lh
cdstg1lh I II III IV NA


一致片段長度
180
分析_覆蓋率_狀態_平均值
80
60
40
20
0
I
II
III
IV
NA
cdstg1lh
cdstg1lh I II III IV NA

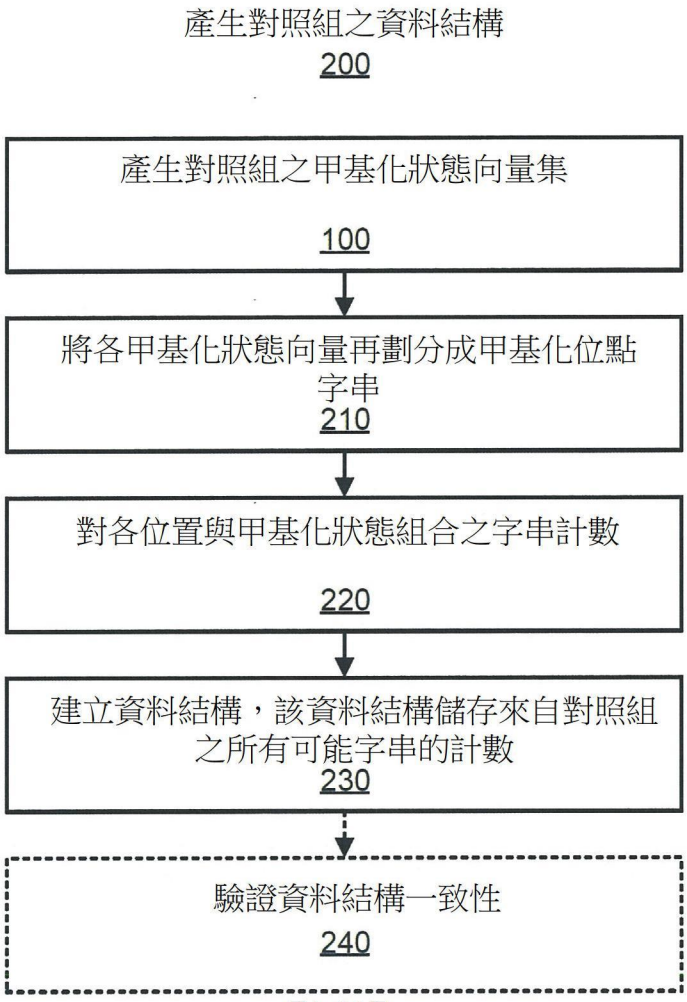
【圖1C】

每個樣本之cfDNA濃度

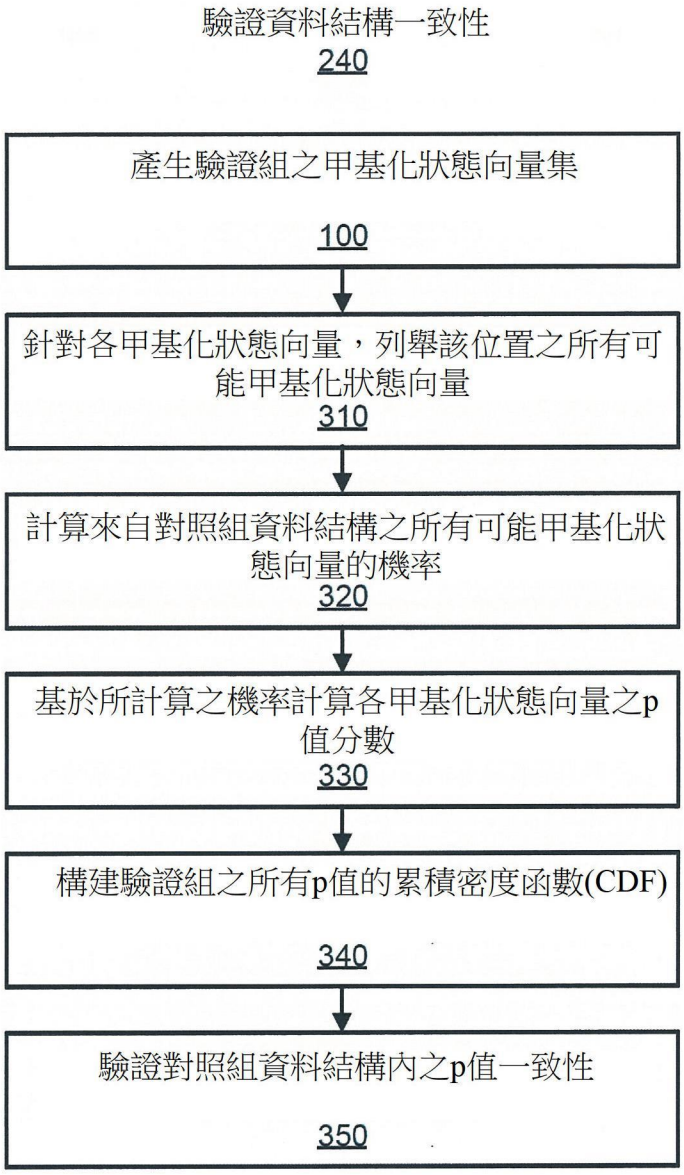
190



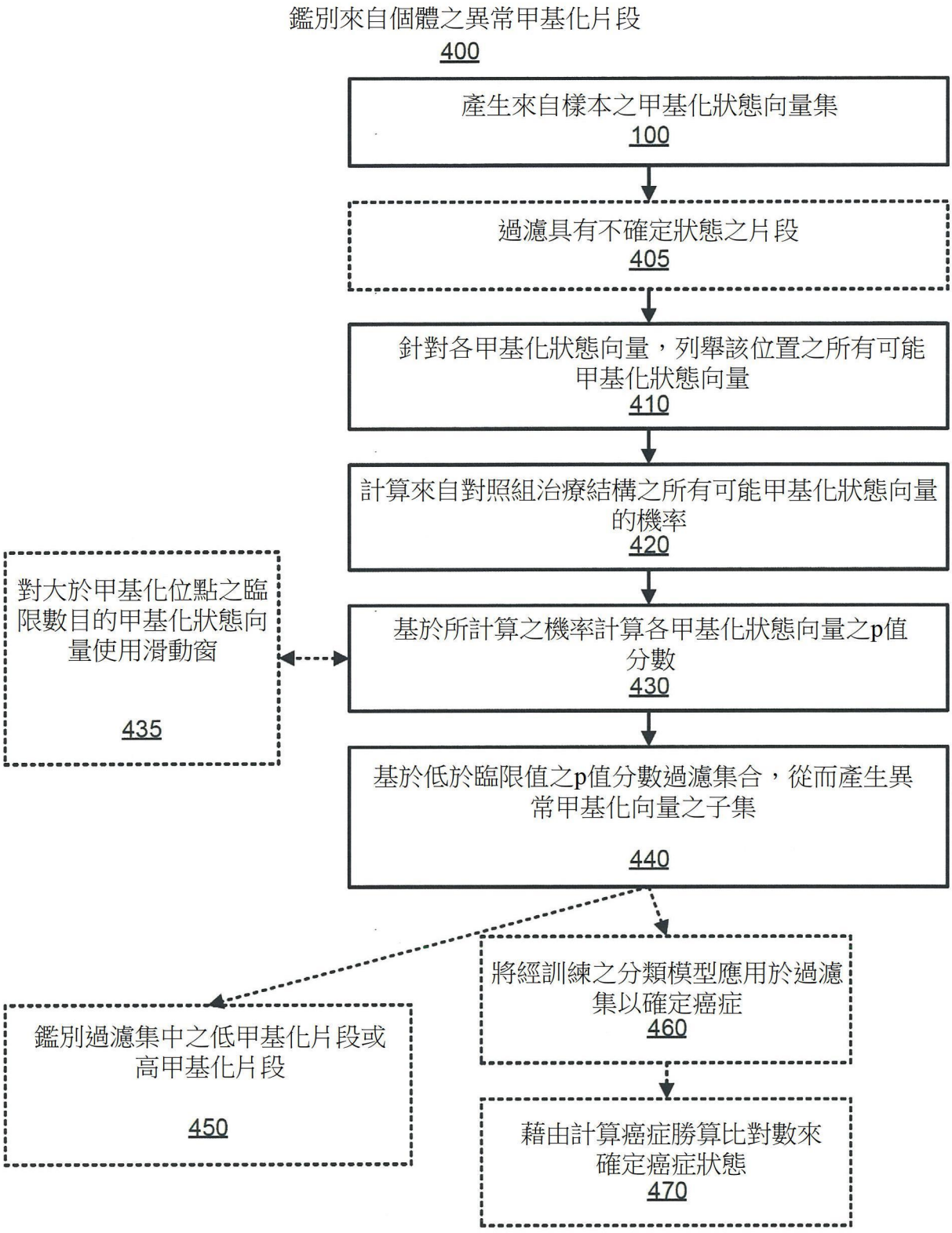
【圖1D】


產生對照組之資料結構
200
產生對照組之甲基化狀態向量集
100
將各甲基化狀態向量再劃分成甲基化位點字串
210
對各位置與甲基化狀態組合之字串計數
220
建立資料結構，該資料結構儲存來自對照組之所有可能字串的計數
230
驗證資料結構一致性
240

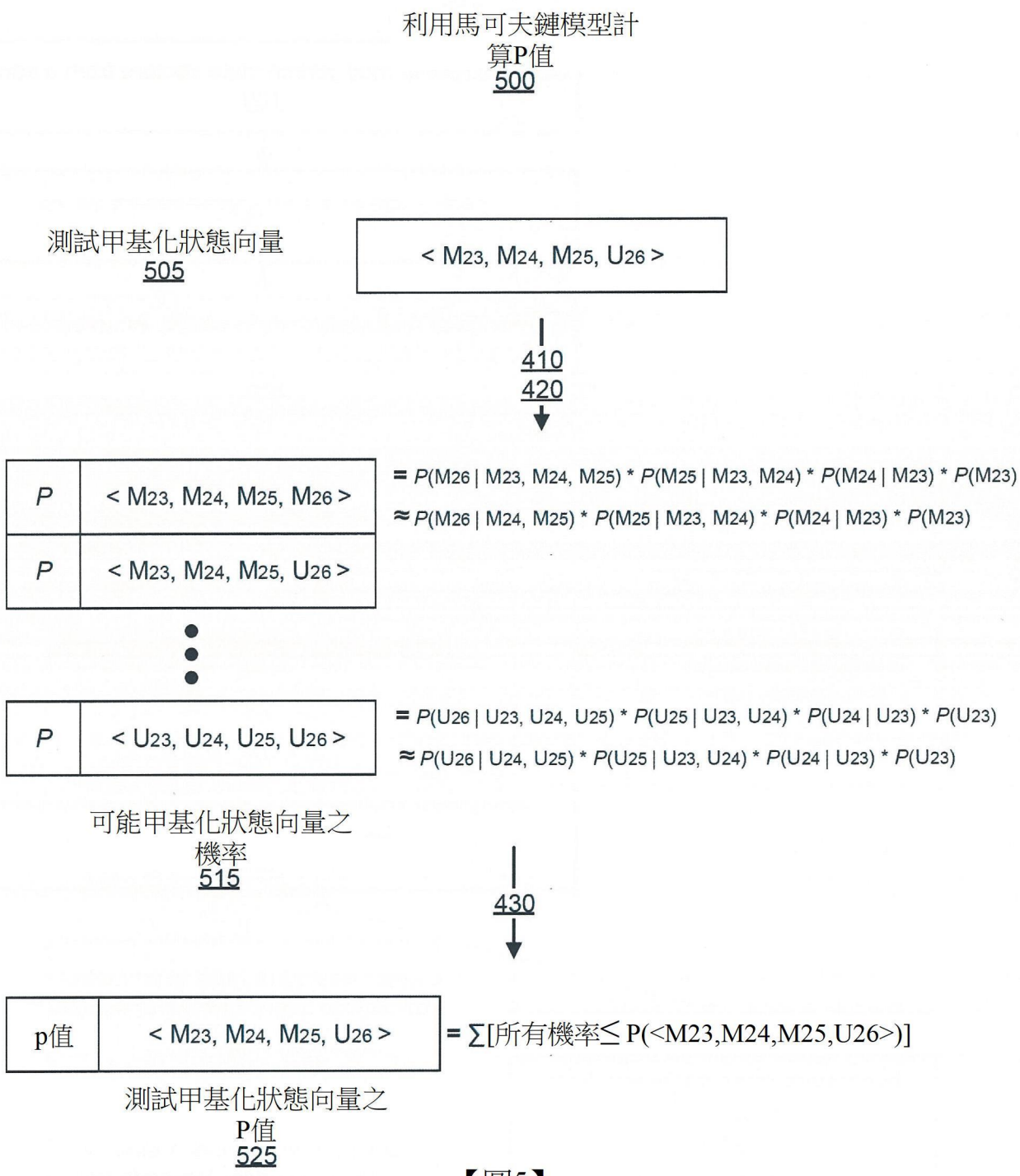
【圖2】


驗證資料結構一致性
240
產生驗證組之甲基化狀態向量集
100
針對各甲基化狀態向量，列舉該位置之所有可能甲基化狀態向量
310
計算來自對照組資料結構之所有可能甲基化狀態向量的機率
320
基於所計算之機率計算各甲基化狀態向量之p值分數
330
構建驗證組之所有p值的累積密度函數(CDF)
340
驗證對照組資料結構內之p值一致性
350

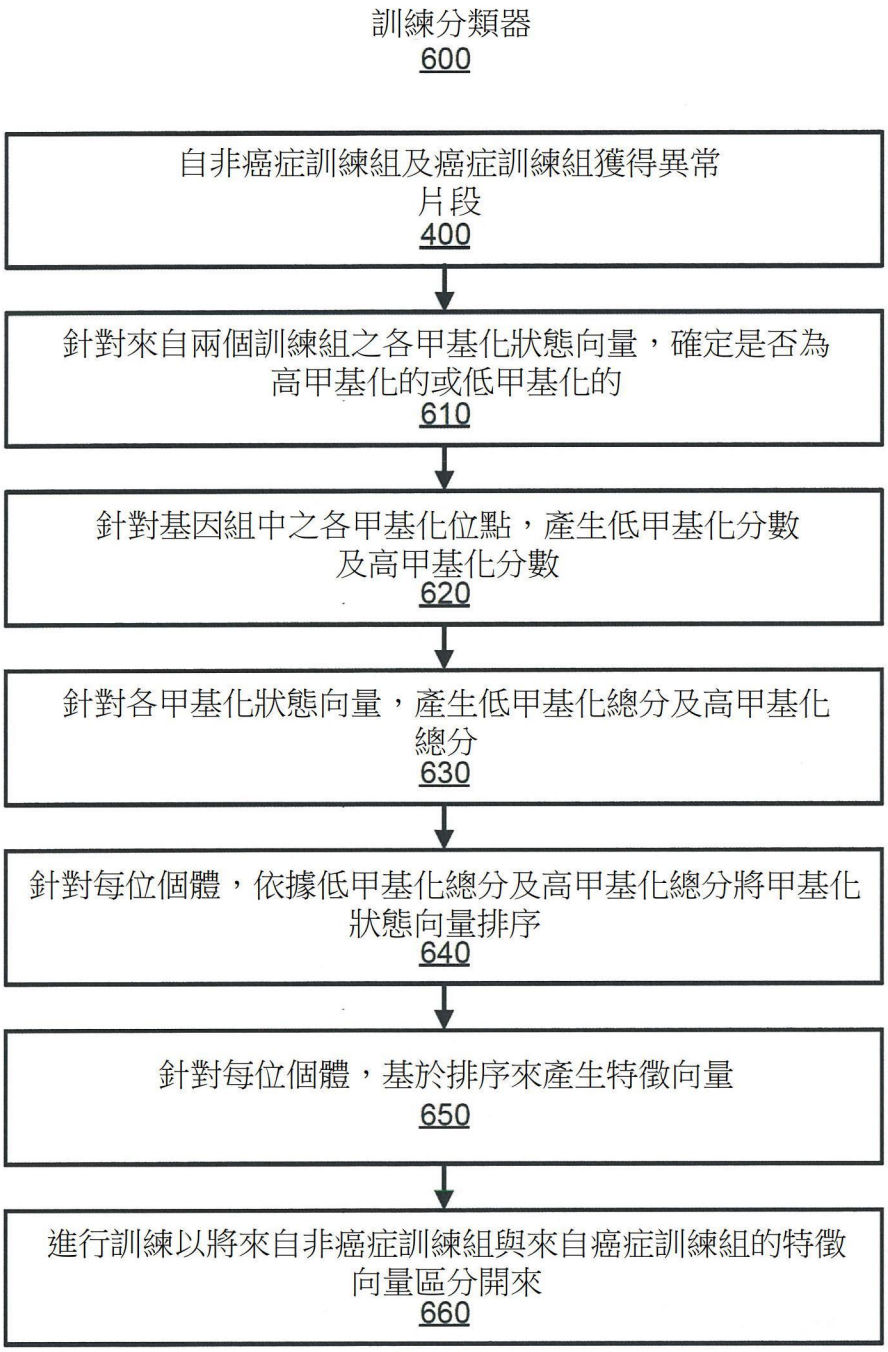
【圖3】


鑑別來自個體之異常甲基化片段
400
產生來自樣本之甲基化狀態向量集
100
過濾具有不確定狀態之片段
405
針對各甲基化狀態向量，列舉該位置之所有可能甲基化狀態向量
410
計算來自對照組治療結構之所有可能甲基化狀態向量的機率
420
對大於甲基化位點之臨限數目的甲基化狀態向量使用滑動窗
435
基於所計算之機率計算各甲基化狀態向量之p值分數
430
基於低於臨限值之p值分數過濾集合，從而產生異常甲基化向量之子集
440
鑑別過濾集中之低甲基化片段或高甲基化片段
450
將經訓練之分類模型應用於過濾集以確定癌症
460
藉由計算癌症勝算比對數來確定癌症狀態
470

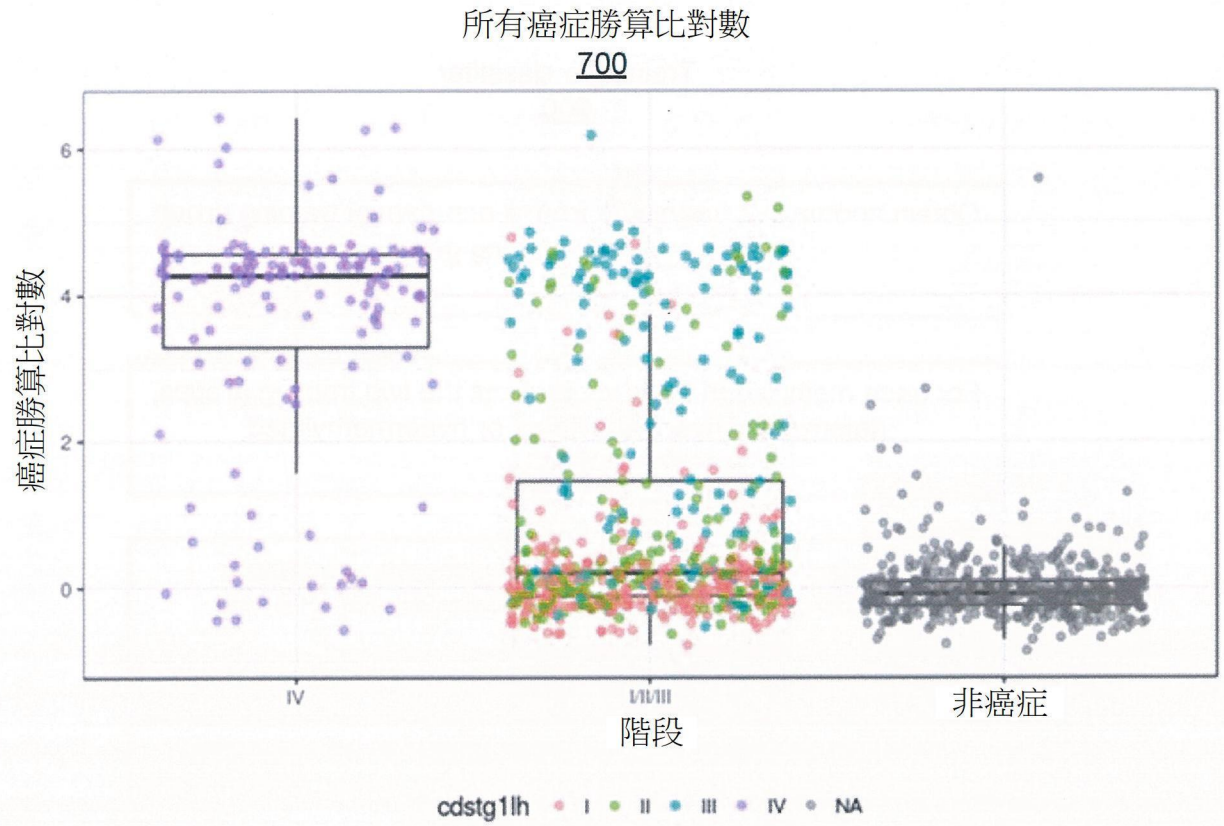
【圖4】

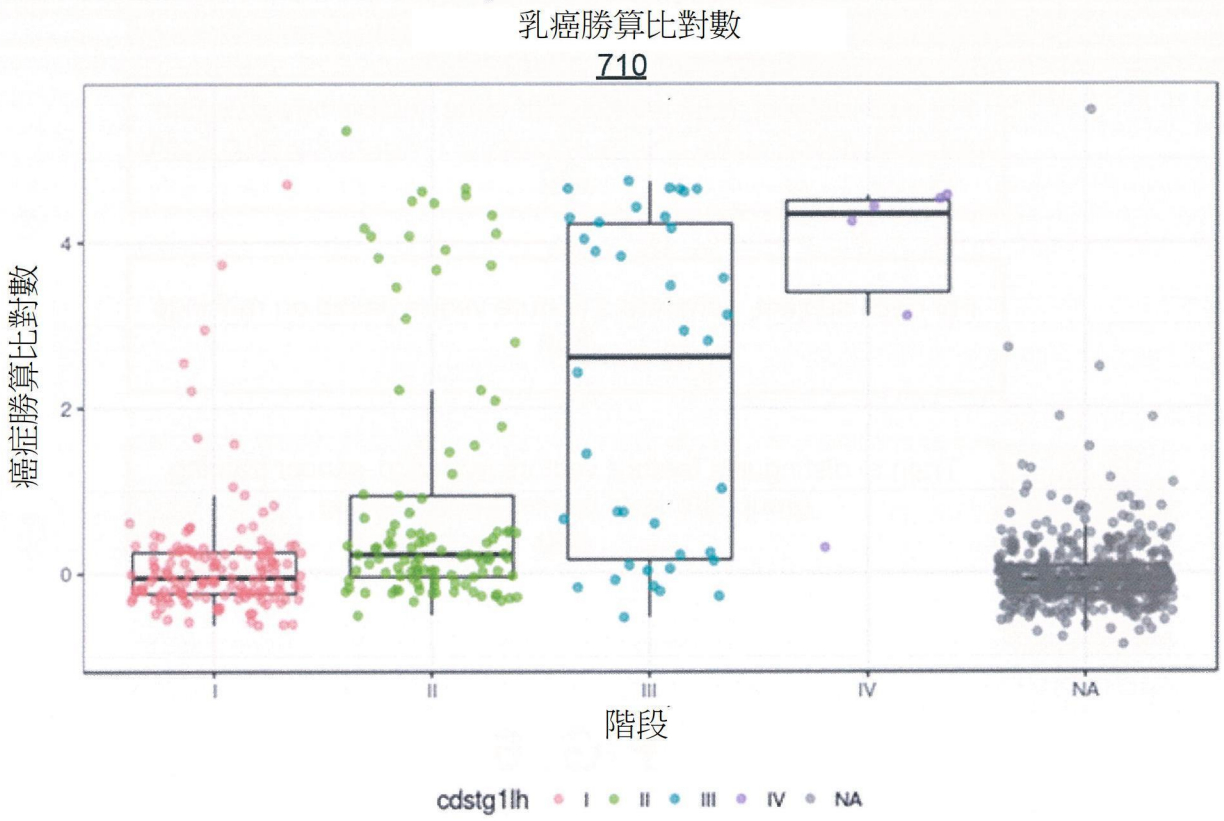

利用馬可夫鏈模型計算P值
500
測試甲基化狀態向量
505
< M23, M24, M25, U26 >
410
420
P
< M23, M24, M25, M26 >
= P(M26 | M23, M24, M25) * P(M25 | M23, M24) * P(M24 | M23) * P(M23)
≈ P(M26 | M24, M25) * P(M25 | M23, M24) * P(M24 | M23) * P(M23)
P
< M23, M24, M25, U26 >
P
< U23, U24, U25, U26 >
= P(U26 | U23, U24, U25) * P(U25 | U23, U24) * P(U24 | U23) * P(U23)
≈ P(U26 | U24, U25) * P(U25 | U23, U24) * P(U24 | U23) * P(U23)
可能甲基化狀態向量之機率
515
430
p值
< M23, M24, M25, U26 >
= ∑[所有機率≤ P(<M23,M24,M25,U26>)]
測試甲基化狀態向量之P值
525

【圖5】

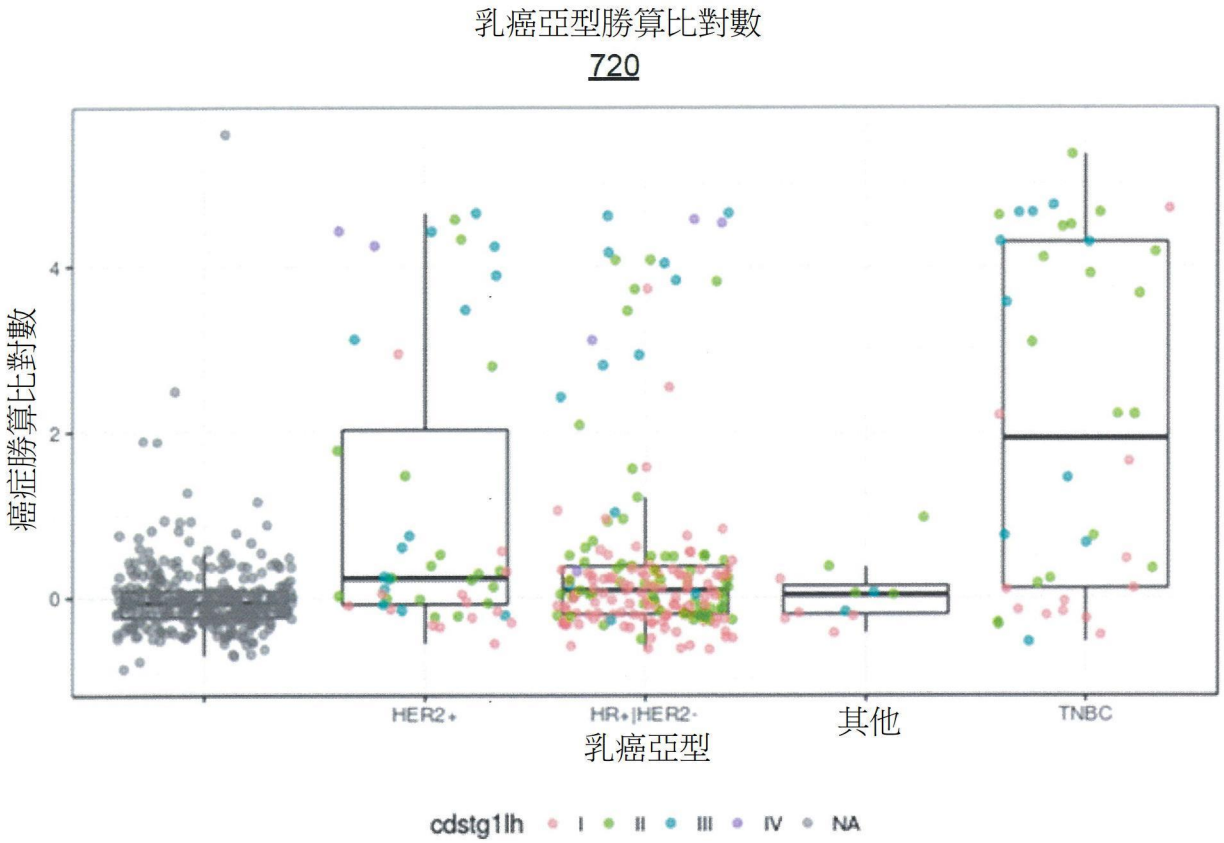

訓練分類器
600
自非癌症訓練組及癌症訓練組獲得異常片段
400
針對來自兩個訓練組之各甲基化狀態向量，確定是否為高甲基化的或低甲基化的
610
針對基因組中之各甲基化位點，產生低甲基化分數及高甲基化分數
620
針對各甲基化狀態向量，產生低甲基化總分及高甲基化總分
630
針對每位個體，依據低甲基化總分及高甲基化總分將甲基化狀態向量排序
640
針對每位個體，基於排序來產生特徵向量
650
進行訓練以將來自非癌症訓練組與來自癌症訓練組的特徵向量區分開來
660

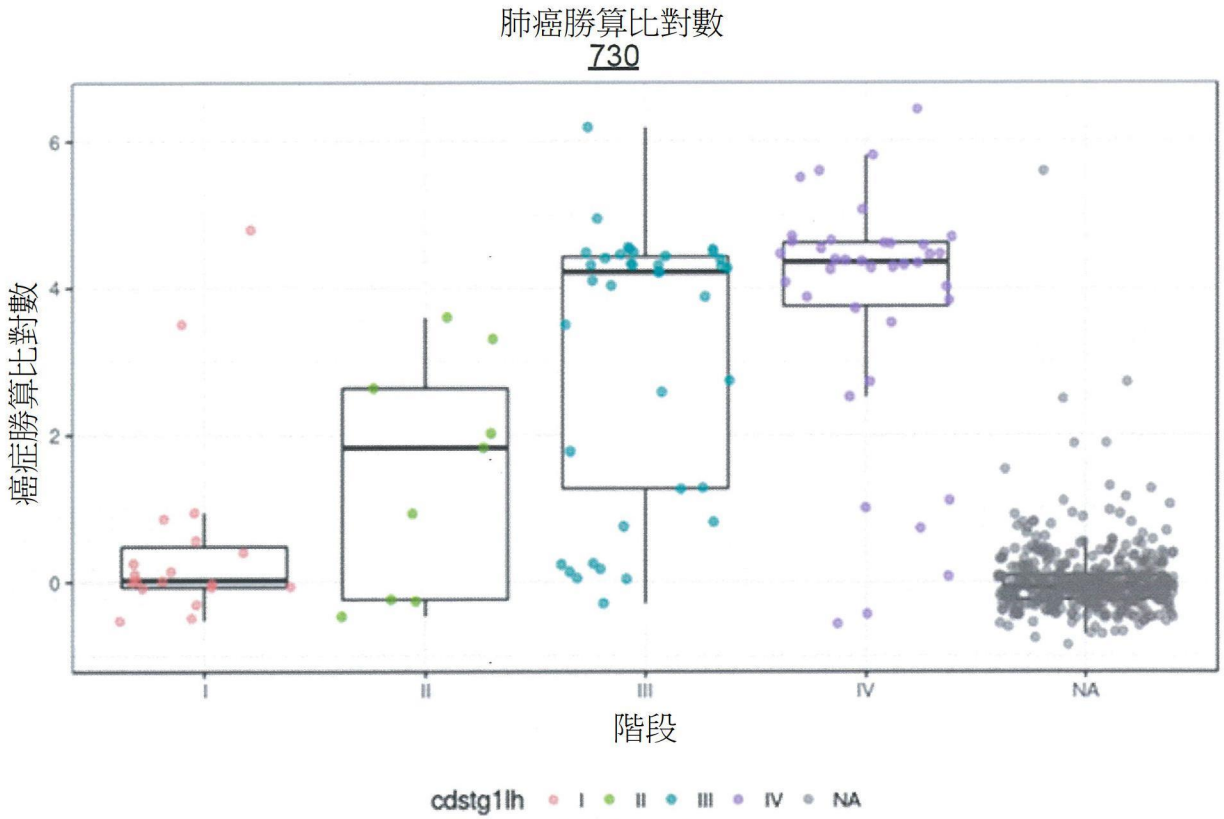
【圖6】


所有癌症勝算比對數
700
癌症勝算比對數
6
4
2
0
IV
I/II/III
非癌症
階段
cdstg1lh I II III IV NA

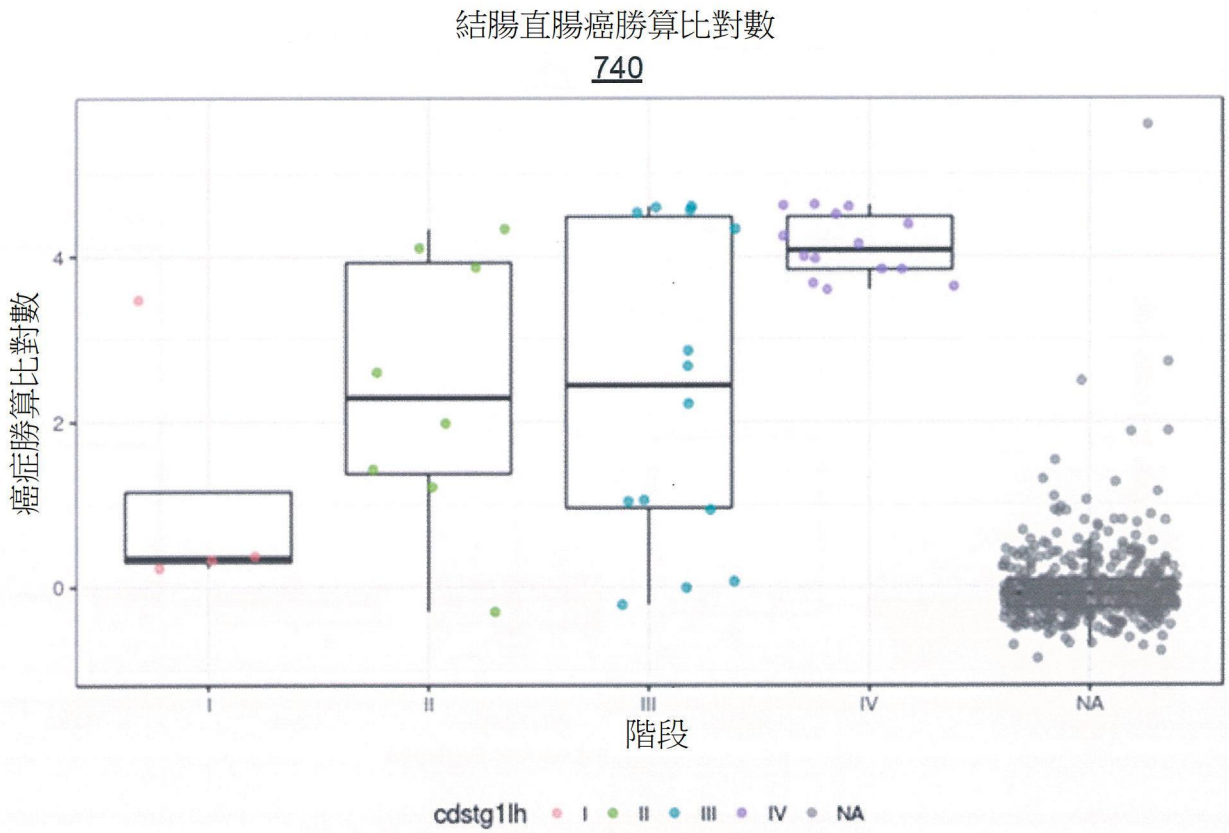

乳癌勝算比對數
710
癌症勝算比對數
4
2
0
I
II
III
IV
NA
階段
cdstg1lh I II III IV NA

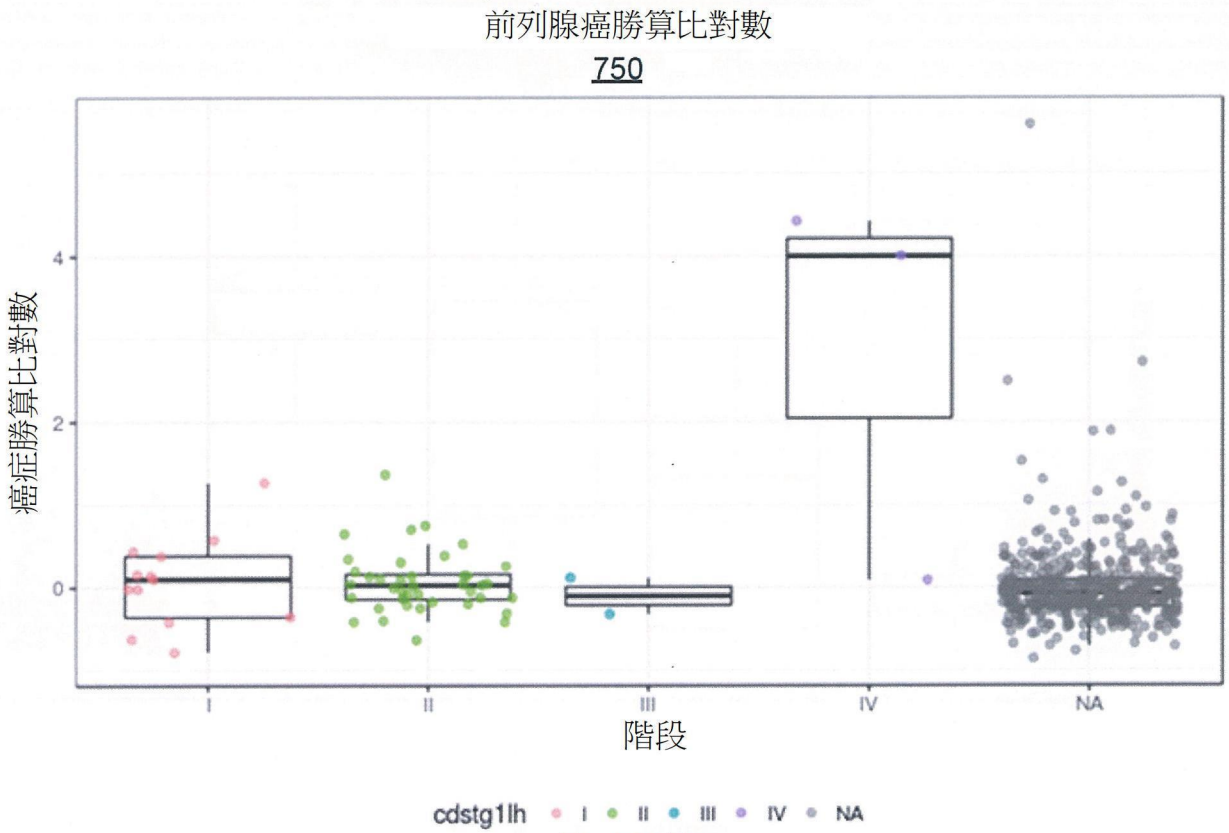
【圖7A】


乳癌亞型勝算比對數
720
癌症勝算比對數
4
2
0
HER2+
HR+|HER2-
其他
TNBC
乳癌亞型
cdstg1lh
I
II
III
IV
NA

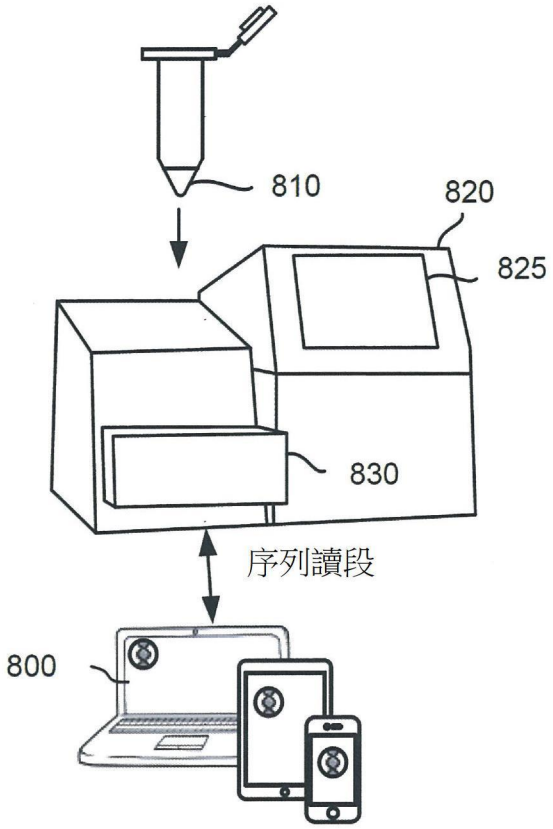

肺癌勝算比對數
730
癌症勝算比對數
6
4
2
0
I
II
III
IV
NA
階段
cdstg1lh
I
II
III
IV
NA

【圖7B】

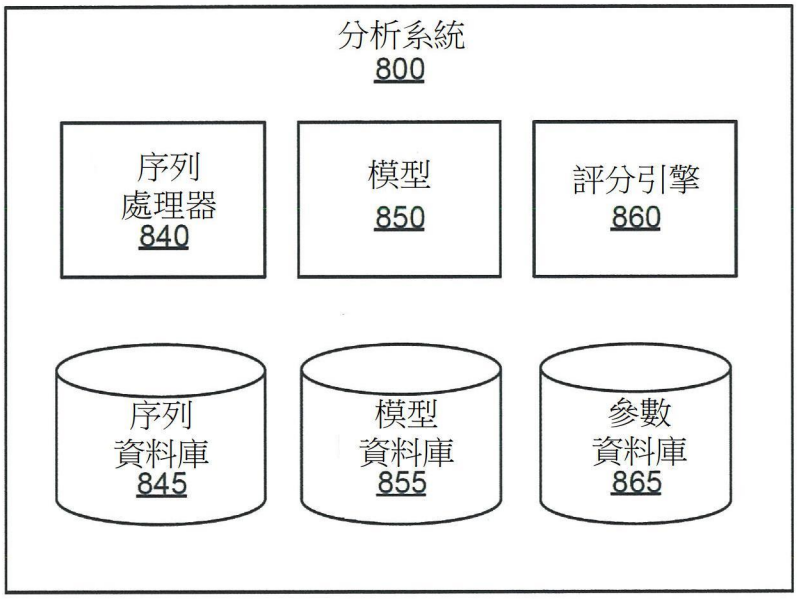

結腸直腸癌勝算比對數
740
癌症勝算比對數
4
2
0
I
II
III
IV
NA
階段
cdstg1lh
I
II
III
IV
NA


前列腺癌勝算比對數
750
癌症勝算比對數
4
2
0
I
II
III
IV
NA
階段
cdstg1lh
I
II
III
IV
NA

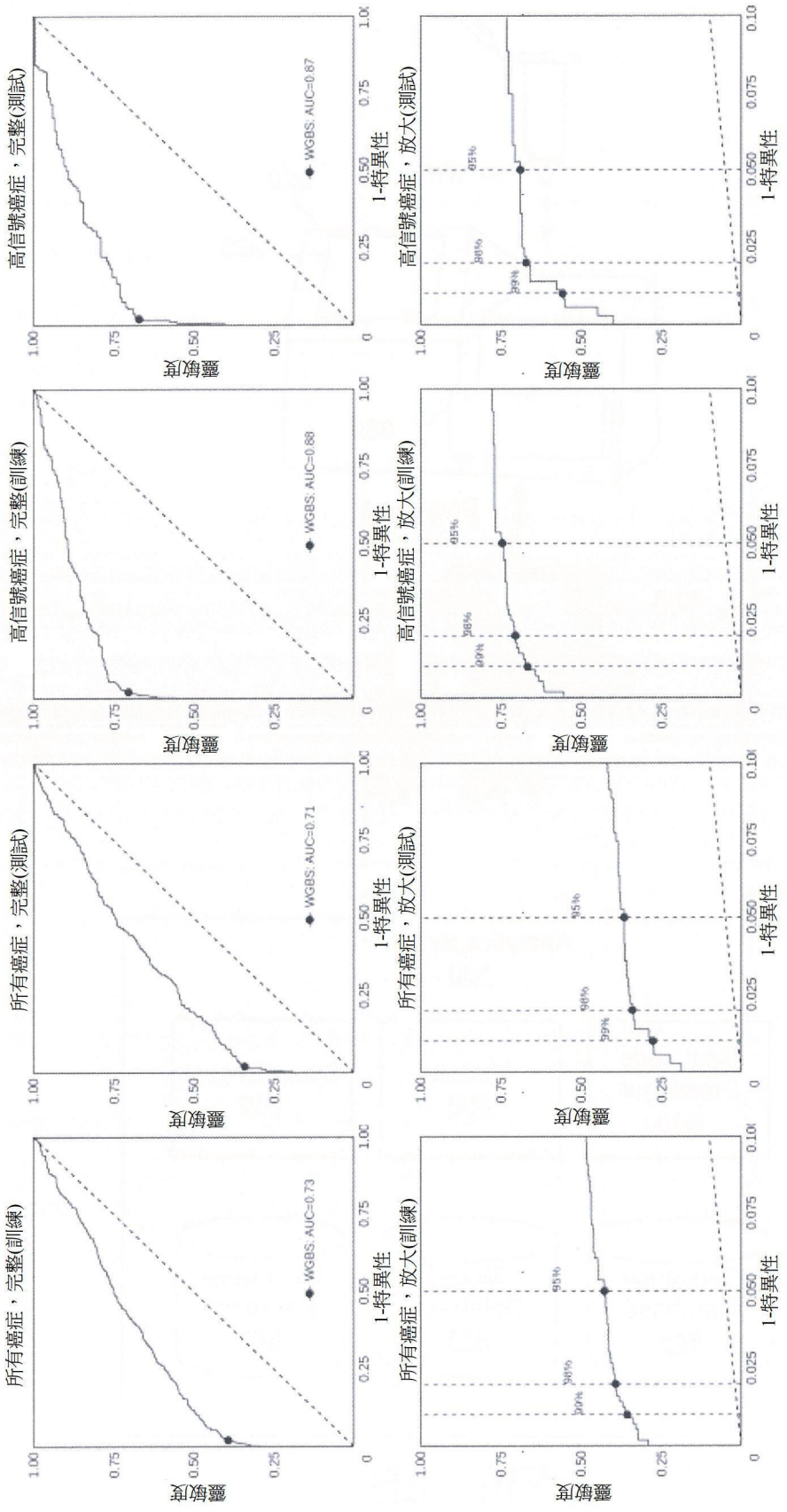
【圖7C】


810
820
825
830
序列讀段
800

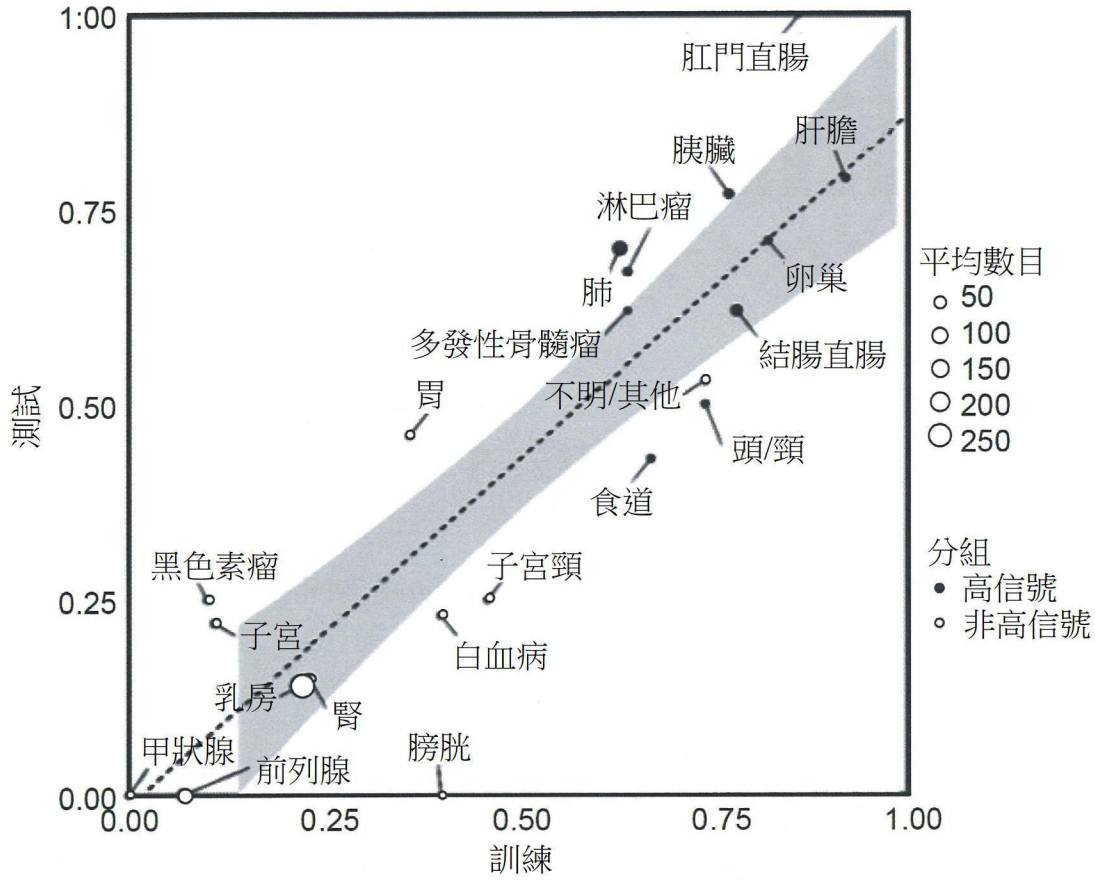
【圖8A】


分析系統
800
序列
處理器
840
模型
850
評分引擎
860
序列
資料庫
845
模型
資料庫
855
參數
資料庫
865

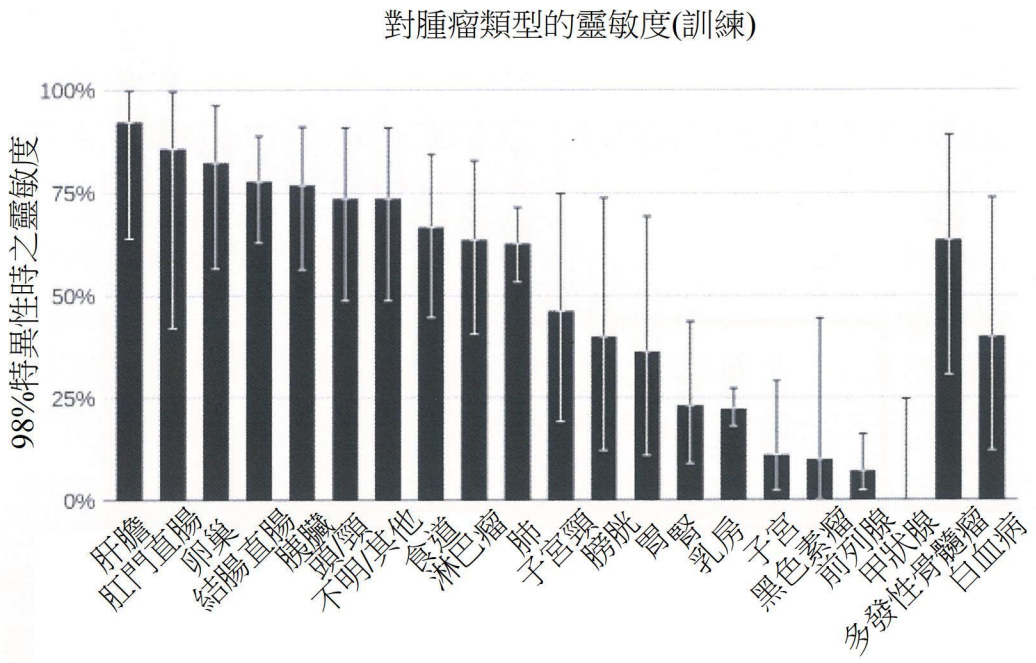
【圖8B】


所有癌症，完整(訓練)
所有癌症，完整(測試)
高信號癌症，完整(訓練)
高信號癌症，完整(測試)
所有癌症，放大(訓練)
所有癌症，放大(測試)
高信號癌症，放大(訓練)
高信號癌症，放大(測試)
靈敏度
1-特異性
WGBS: AUC=0.73
WGBS: AUC=0.71
WGBS: AUC=0.88
WGBS: AUC=0.87
99%
98%
95%

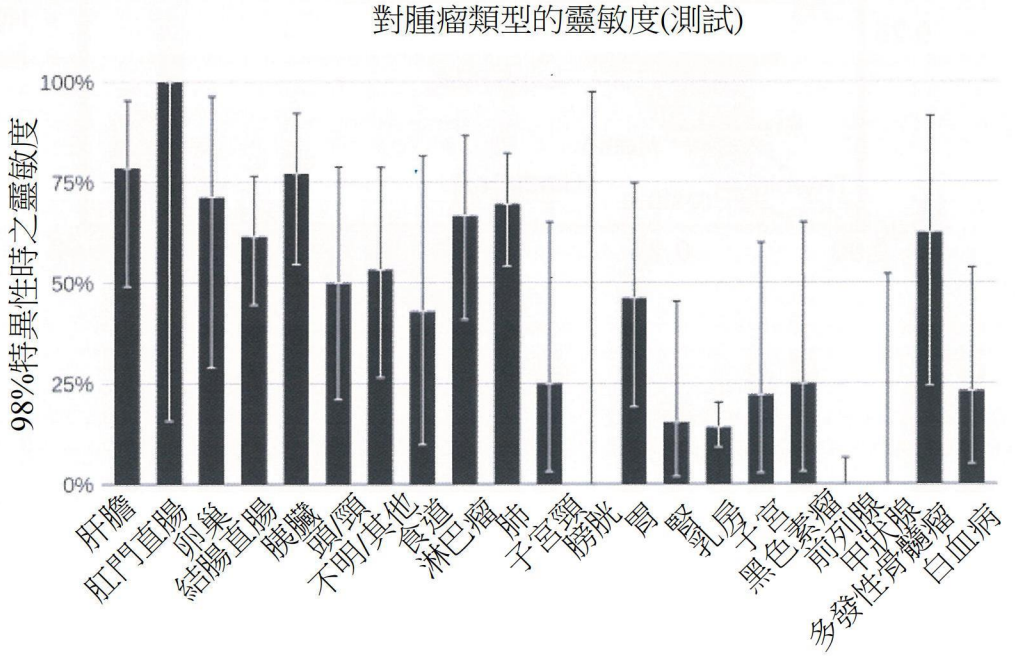
【圖9】


1:00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
驗證
訓練
肛門直腸
肝膽
胰臟
淋巴瘤
卵巢
肺
多發性骨髓瘤
結腸直腸
胃
不明/其他
頭/頸
食道
黑色素瘤
子宮頸
子宮
白血病
乳房
腎
甲狀腺
前列腺
膀胱
平均數目
50
100
150
200
250
分組
高信號
非高信號

【圖10】


對腫瘤類型的靈敏度(訓練)
98%特異性時之靈敏度
100%
75%
50%
25%
0%
肝膽
肛門直腸
卵巢
結腸直腸
胰臟
頭/頸
不明/其他
食道
淋巴瘤
肺
子宮頸
膀胱
胃
腎
乳房
子宮
黑色素瘤
前列腺
甲狀腺
多發性骨髓瘤
白血病

【圖11A】


對腫瘤類型的靈敏度(測試)
98%特異性時之靈敏度
100%
75%
50%
25%
0%
肝膽
肛門直腸
卵巢
結腸直腸
胰臟
頭/頸
不明/其他
食道
淋巴瘤
肺
子宮頸
膀胱
胃
腎
乳房
子宮
黑色素瘤
前列腺
甲狀腺
多發性骨髓瘤
白血病

【圖11B】