



(21)申請案號：109111392

(22)申請日：中華民國 109 (2020) 年 04 月 01 日

(51)Int. Cl. :

C12Q1/6883 (2018.01)

C12Q1/6886 (2018.01)

G01N33/48 (2006.01)

(30)優先權：2019/04/02美國62/828,224

2020/01/15美國62/961,517

(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：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)；林偉棋 LAM, WAI KEI (HK)；吉璐 JI, LU (CN)

(74)代理人：陳長文

(56)參考文獻：

WO2018/081130A1

審查人員：陳婷妤

申請專利範圍項數：16 項圖式數：22共 127 頁

(54)名稱

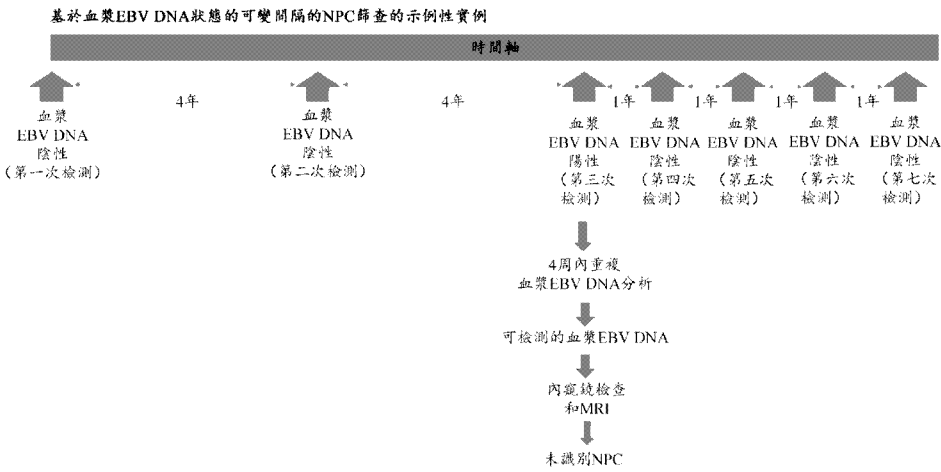
與病毒相關的癌症的風險分級

(57)摘要

本文提供了基於對來自受試者的生物樣品的無細胞核酸分子的分析來對所述受試者發展病原體相關疾病的風險進行分級的方法和系統。在各種實例中，基於風險分析來確定篩查頻率。本文還提供了用於分析無細胞核酸分子中的病原體基因組的變體模式的方法和系統。

Provided herein are methods and systems for stratifying risk for a subject to develop a pathogen-associated disorder based on analysis of cell-free nucleic acid molecules from a biological sample of the subject. In various examples, screening frequency is determined based on the risk analysis. Also provided herein are methods and systems for analyzing variant patterns of a pathogen genome in cell-free nucleic acid molecules.

指定代表圖：



【圖 2】



I858044

【發明摘要】

【中文發明名稱】

與病毒相關的癌症的風險分級

【英文發明名稱】

STRATIFICATION OF RISK OF VIRUS ASSOCIATED CANCERS

【中文】

本文提供了基於對來自受試者的生物樣品的無細胞核酸分子的分析來對所述受試者發展病原體相關疾病的風險進行分級的方法和系統。在各種實例中，基於風險分析來確定篩查頻率。本文還提供了用於分析無細胞核酸分子中的病原體基因組的變體模式的方法和系統。

【英文】

Provided herein are methods and systems for stratifying risk for a subject to develop a pathogen-associated disorder based on analysis of cell-free nucleic acid molecules from a biological sample of the subject. In various examples, screening frequency is determined based on the risk analysis. Also provided herein are methods and systems for analyzing variant patterns of a pathogen genome in cell-free nucleic acid molecules.

【指定代表圖】

圖2

【代表圖之符號簡單說明】

無

【發明說明書】

【中文發明名稱】

與病毒相關的癌症的風險分級

【英文發明名稱】

STRATIFICATION OF RISK OF VIRUS ASSOCIATED CANCERS

【技術領域】

【先前技術】

【0001】許多疾病和病況可能與病原體(如病毒)的感染有關。鼻咽癌(NPC)是中國南部和東南亞地區最流行的癌症之一，且NPC的發病機理可能與EB病毒(EBV)感染密切相關。在NPC高發區，幾乎所有的NPC腫瘤都會攜帶EBV基因組。基於EBV與NPC之間的密切關係，血漿EBV DNA已被開發為NPC的生物標記物。使用實時聚合酶鏈反應(PCR)分析，顯示血漿EBV DNA的檢測對檢測NPC具有95%的靈敏性和93%的特異性(Lo等人，*Cancer Res.* 1999;59:1188-91)。根據來自生物樣品中病原體的無細胞核酸分子的分析，開發用於對這些病原體相關疾病的風險進行分級的無創或微創診斷分析可能具有重大的臨床益處。

【發明內容】

【0002】在一些方面，本文提供了一種在受試者中篩查病原體相關疾病的方法，包括：接收在第一時間點進行的第一次測定的數據，所述第一次測定包括確定來自受試者的生物樣品中病原體的無細胞核酸分子的特性，其中來自所述病原體的無細胞核酸分子的特性包括數量、甲基化狀態、變體模式、片段大小或與生物樣品中來自受試者的無細胞核酸分子相比的相對豐度，並且其中所述特性指示受試者發展所述病原體相關疾病的

風險；以及基於所述特性，確定進行第二次測定以在受試者中篩查所述病原體相關疾病的第二時間點，其中所述第一時間點和所述第二時間點之間的間隔與風險反相關。

【0003】 在一些方面，本文提供了一種在受試者中預測病原體相關疾病的方法，包括：接收第一次測定的數據，所述第一次測定包括確定來自受試者的生物樣品中病原體的無細胞核酸分子的特性，其中來自所述病原體的無細胞核酸分子的特性包括數量、甲基化狀態、變體模式、片段大小或與所述生物樣品中來自受試者的無細胞核酸分子相比的相對豐度；以及基於來自所述病原體的無細胞核酸分子的所述特性和受試者的年齡、受試者的吸菸習慣、受試者的病原體相關疾病的家族史、受試者的基因型因素、受試者的種族或受試者的飲食史中的一種或多種因素生成指示受試者發展所述病原體相關疾病的風險的報告。

【0004】 在一些情況下，第一次測定的結果不會導致對所述受試者進行病原體相關疾病的醫學治療。在一些情況下，所述醫學治療包括治療劑治療、放射療法或外科手術治療。在一些情況下，在通過假陽性率低於1%的臨床診斷檢查確定第二時間點之前，所述受試者被診斷為不具有病原體相關疾病。在一些情況下，所述臨床診斷檢查包括身體檢查、侵入性活檢、內窺鏡檢查、核磁共振成像、正電子放射斷層掃描、計算機斷層掃描或x-射線成像。在一些情況下，所述臨床診斷檢查包括侵入性活檢，所述侵入性活檢包括組織學分析、細胞學分析或細胞核酸分析。在一些情況下，間隔為至少約2個月、4個月、6個月、8個月、10個月或12個月。在一些情況下，間隔為約12個月。

【0005】 在一些情況下，所述方法還包括進行第一次測定。在一些

情況下，進行第一次測定包括：(i)從受試者獲得第一生物樣品；以及(ii)測量來自第一生物樣品中的所述病原體的無細胞核酸分子的第一數量。在一些情況下，測量第一數量包括測量來自所述第一生物樣品中的所述病原體的無細胞核酸分子的拷貝數。在一些情況下，所述測量包括聚合酶鏈反應(PCR)。在一些情況下，所述測量包括定量聚合酶鏈反應(qPCR)。在一些情況下，所述第一數量包括測量來自第一生物樣品中的所述病原體的無細胞核酸分子的第一百分比。在一些情況下，所述第一次測定還包括：(iii)如果所述第一數量高於閾值，則從受試者獲得第二生物樣品，並測量來自第二生物樣品中的所述病原體的無細胞核酸分子的第二數量。在一些情況下，第二生物樣品在第一生物樣品之後約4周獲得。在一些情況下，與所述第二數量低於所述閾值時的間隔相比，如果所述第一數量和所述第二拷貝數均高於所述閾值，則所述第一時間點和所述第二時間點之間的時間隔更短。在一些情況下，與所述第一數量高於所述閾值時的間隔相比，如果所述第一數量低於所述閾值，則所述第一時間點和所述第二時間點之間的時間隔更長。在一些情況下，如果所述第一數量和所述第二數量均高於所述閾值，則所述第一時間點和所述第二時間點之間的時間隔為約1年。在一些情況下，如果所述第二數量低於所述閾值，則所述第一時間點和所述第二時間點之間的時間隔為約2年。在一些情況下，如果所述第一數量低於所述閾值，則所述第一時間點和所述第二時間點之間的時間隔為約4年。在一些情況下，第一次測定包括：確定來自生物樣品中所述病原體的無細胞核酸分子的甲基化狀態。在一些情況下，確定甲基化狀態包括用甲基化敏感的限制性酶或亞硫酸氫鹽處理所述生物樣品中的無細胞核酸分子。在一些情況下，確定甲基化狀態包括對受試者的生物樣品中的無細胞核酸進行甲

基化-感知測序。在一些情況下，所述甲基化-感知測序包括亞硫酸氫鹽將未甲基化的胞嘧啶轉化為尿嘧啶。在一些情況下，所述甲基化-感知測序包括用甲基化敏感的限制性酶處理。在一些情況下，第一次測定包括：確定來自生物樣品中的所述病原體的無細胞核酸分子的片段大小分布。在一些情況下，確定所述片段大小分布包括對所述生物樣品中的無細胞核酸分子進行測序，並基於映射到所述病原體的參考基因組的序列讀數，確定來自所述生物樣品中的所述病原體的無細胞核酸分子的片段大小。

【0006】 在一些情況下，第一次測定包括：確定來自所述生物樣品中所述病原體的無細胞核酸分子的變體模式。在一些情況下，確定所述變體模式包括對所述生物樣品中的無細胞核酸分子進行測序，以及基於映射到所述病原體的參考基因組的序列讀數，確定來自所述生物樣品中的所述病原體的無細胞核酸分子的變體模式。在一些情況下，來自所述病原體的無細胞核酸分子的變體模式包括單核苷酸變體。在一些情況下，識別所述變體模式包括：確定映射到所述病原體的所述參考基因組的序列讀數與所述病原體的疾病相關的參考基因組之間的相似度水平。在一些情況下，所述病原體的疾病相關的參考基因組包括在患病組織中識別出的所述病原體的基因組。在一些情況下，確定所述相似度水平包括：將所述病原體的所述參考基因組分離到多個箱中；以及確定所述多個箱中的每個箱對所述病原體的所述疾病相關的參考基因組的相似性指數，其中所述相似性指數與相應箱內的一部分變體位點相關，在所述變體位點上，至少有一個映射到所述病原體的參考基因組的中的序列讀數具有與所述病原體的疾病相關的參考基因組相同的核苷酸變體。在一些情況下，所述病原體的疾病相關的參考基因組包括多個所述病原體的疾病相關的參考基因組，並且其中確定

所述相似度水平包括：確定多個箱中的每個箱與所述病原體的疾病相關的參考基因組中的每一個的各自的相似性指數，以及針對所述相應箱中的各自的相似性指數高於截止值的箱，基於所述多個疾病相關的參考基因組的比例確定所述多個箱中的每一個箱的箱得分。在一些情況下，多個箱中的每個箱具有約100、200、300、400、500、600、700、800、900或1000 bp的長度。在一些情況下，第一次測定包括確定來自所述生物樣品中所述病原體的無細胞核酸分子的甲基化狀態、片段大小分布或變體模式。

【0007】 在一些情況下，所述方法還包括使用應用於數據輸入的分類器，計算受試者發展病原體相關疾病的風險評分，所述數據輸入包含來自所述生物樣品中所述病原體的無細胞核酸分子的特性，其中所述分類器被配置為將函數應用於包含來自所述生物樣品中的所述病原體的無細胞核酸分子的特性的數據輸入，以生成包含風險評分的輸出，所述風險評分評估受試者發展所述疾病的風險。在一些情況下，所述分類器被標記的數據集訓練。

【0008】 在一些情況下，所述的方法還包括在第二時間點進行第二次測定。在一些情況下，所述第二次測定與第一次測定相同。在一些情況下，第二次測定包括受試者的無細胞核酸分子的測定、受試者的侵入性活檢、受試者的內窺鏡檢查或受試者的磁共振成像檢查。

【0009】 在一些方面，本文提供了一種分析來自受試者的生物樣品的核酸分子的方法，包括：在計算機系統中獲得來自受試者的生物樣品的無細胞核酸分子的序列讀數，其中所述生物樣品包含來自受試者且可能來自病原體的無細胞核酸分子；在計算機系統中，將無細胞核酸分子的序列讀數與所述病原體的參考基因組比對；以及在計算機系統中，識別來自所

述病原體的無細胞核酸分子的變體模式，其中所述變體模式表徵在所述病原體的參考基因組的多個變體位點的每個點處映射到所述病原體參考基因組的序列讀數的核苷酸變體，其中所述多個變體位點包括跨越所述病原體的所述參考基因組上的至少30個位點，並且其中所述變體模式指示所述受試者中病原體相關疾病的狀態或風險。

【0010】 在一些情況下，所述多個變體位點包括跨越所述病原體的所述參考基因組的至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400、至少500、至少600、至少700、至少800、至少900、至少1000、至少1100或至少1200個位點。在一些情況下，所述多個變體位點包括跨越所述病原體的所述參考基因組的至少600個位點。在一些情況下，所述多個變體位點包括跨越所述病原體的所述參考基因組的約660個位點。在一些情況下，所述多個變體位點包括跨越所述病原體的所述參考基因組的至少1000個位點。在一些情況下，所述多個變體位點包括跨越所述病原體的所述參考基因組的約1100個位點。在一些情況下，多個變體位點由映射到所述病原體的參考基因組的序列讀數具有與所述病原體的參考基因組不同的核苷酸變體的所有位點組成。在一些情況下，比對序列讀數被配置為允許映射到所述病原體的所述參考基因組的序列讀數與所述病原體的所述參考基因組之間的最大錯配為10、9、8、7、6、5、4、3、2或1個鹼基。在一些情況下，比對序列讀數被配置為允許映射到所述病原體的所述參考基因組的序列讀數與所述病原體的所述參考基因組之間的最大錯配為2個鹼基。在一些情況下，所述方法還包括：(d)基於映射到所述病原體的所述參考基因組的序列讀數的變體模式，診斷、預測或監測受試者與所述病原體相關疾病。在一些情況

下，來自所述病原體的無細胞核酸分子的變體模式包括單核苷酸變體。在一些情況下，識別變體模式包括：確定映射到所述病原體的所述參考基因組的序列讀數與所述病原體的疾病相關的參考基因組之間的相似度水平。在一些情況下，所述病原體的與疾病相關的參考基因組包括在患病組織中識別出的所述病原體的基因組。在一些情況下，確定相似度水平包括：將所述病原體的所述參考基因組分離到多個箱中；以及針對所述病原體的疾病相關的參考基因組，確定所述多個箱中的每個箱的相似性指數，其中所述相似性指數與相應箱內的一部分變體位點相關聯，在所述變體位點上，至少有一個映射到所述病原體的參考基因組的中的序列讀數具有與所述病原體的疾病相關的參考基因組相同的核苷酸變體。在一些情況下，所述病原體的疾病相關的參考基因組包括多個所述病原體的疾病相關的參考基因組，並且其中確定所述相似度水平包括：確定多個箱中的每個箱與所述病原體的疾病相關的參考基因組中的每一個的各自的相似性指數，以及針對所述相應箱中的各自的相似性指數高於截止值的箱，基於所述多個疾病相關的參考基因組的比例確定所述多個箱中的每個箱的箱得分。在一些情況下，截止值為約0.9。在一些情況下，多個箱中的每個箱具有約100、200、300、400、500、600、700、800、900或1000 bp的長度。在一些情況下，所述方法還包括：使用應用於數據輸入的分類器，計算受試者發展病原體相關疾病的風險評分，所述數據輸入包含來自所述病原體的無細胞核酸分子的變體模式，其中所述分類器被配置為將函數應用於包含來自病原體的無細胞核酸分子的變體模式的數據輸入，以生成包含風險評分的輸出，所述風險評分評估受試者發展所述疾病的風險。在一些情況下，所述分類器被標記的數據集訓練。在一些情況下，所述分類器包括使用樸素

貝葉斯模型、邏輯回歸、隨機森林、決策樹、梯度提升樹、神經網絡、深度學習、線性/內核支持向量機(SVM)、線性/非線性回歸或線性判別分析的數學模型。

【0011】 在一些情況下，所述病原體是病毒。在一些情況下，所述病原體是EB病毒(EBV)。在一些情況下，所述病原體相關疾病包括鼻咽癌、NK細胞淋巴瘤、伯基特淋巴瘤、移植後淋巴增生性障礙或何杰金氏淋巴瘤。在一些情況下，來自所述病原體的無細胞核酸分子的變體模式表徵了在多個變體位點的每個位點上映射到所述病原體的所述參考基因組的序列讀數的核苷酸變體，所述多個變體位點包括選自表6中所列出的相對於EBV參考基因組(AJ507799.2)的基因組位點的至少30、40、50、100、150、200、250、300、350、400、450、500、550或600個位點。在一些情況下，所述多個變體位點包括表6中所列出的相對於EBV參考基因組(AJ507799.2)的基因組位點。在一些情況下，來自所述病原體的無細胞核酸分子的變體模式表徵了在所述多個變體位點的每個位點上映射到所述病原體的所述參考基因組的序列讀數的核苷酸變體，所述多個變體位點隨機選自表6中所列出的相對於EBV參考基因組(AJ507799.2)的基因組位點。在一些情況下，來自所述病原體的無細胞核酸分子的變體模式表徵了在所述多個變體位點的每個位點上映射到所述病原體的所述參考基因組的序列讀數的核苷酸變體，所述多個變體位點包括隨機選自表6中所列出的相對於EBV參考基因組(AJ507799.2)的基因組位點的至少30、40、50、100、150、200、250、300、350、400、450、500、550或600個位點。

【0012】 在一些情況下，所述病原體是人乳頭瘤病毒(HPV)。在一些情況下，所述病原體相關疾病包括宮頸癌、口咽癌和頭頸癌。在一些情

況下，所述病毒是乙型肝炎病毒。在一些情況下，所述病原體相關疾病包括肝硬化或肝細胞癌(HCC)。在一些情況下，所述變體模式表明受試者的病原體相關疾病的狀態，其中所述病原體相關疾病的狀態包括受試者中所述病原體相關疾病的存在、受試者中腫瘤組織的數量、受試者中所述腫瘤組織的大小、受試者中腫瘤所處的階段、受試者中的腫瘤負荷或受試者中腫瘤轉移的存在。在一些情況下，所述生物樣品選自全血、血漿、血清、尿液、腦脊液、血沉棕黃層、陰道分泌物、陰道沖洗液、唾液、口腔沖洗液、鼻沖洗液、鼻刷樣品及其組合。

【0013】 在一些方面，本文提供了一種包含機器可執行代碼的非暫時性計算機可讀介質，所述代碼在由一個或多個計算機處理器執行時實現以上任何一種方法。

【0014】 在一些方面，本文提供了一種計算機產品，包括存儲用於控制計算機系統執行以上任何一種方法中的操作的多個指令的非暫時性計算機可讀介質。

【0015】 在一些方面，本文提供一種系統，所述系統包括：本文所述的計算機產品；以及一個或多個處理器，所述處理器用於執行存儲在計算機可讀介質上的指令。

【0016】 在一些方面，本文提供一種系統，其包括用於執行以上任何一種方法的裝置。

【0017】 在一些方面，本文提供一種被配置為執行以上任何一種方法的系統。

【0018】 在一些方面，本文提供了一種系統，所述系統包括分別執行以上任何一種方法的步驟的模塊。

【圖式簡單說明】

【0019】 本文所述的新穎性特徵在所附申請專利範圍中具體闡述。通過參考對在其中利用本文所述原理的示例說明性實施方案加以闡述的以下詳細描述和附圖，將會對本文所述的特徵和優點獲得更好的理解；在附圖中：

【0020】 圖1是對超過20,000名受試者的隊列中進行NPC篩查研究的設計示意圖。

【0021】 圖2示出了根據本公開內容的NPC篩查方案的示例性示意圖。

【0022】 圖3概述了基於來自NPC患者和非NPC受試者的樣本的EBV變體譜的系統樹分析。

【0023】 圖4概述了基於來自NPC患者和非NPC受試者的樣本的EBV變體譜的系統樹分析，所述EBV變體不包括29種報告的變體。

【0024】 圖5概述了基於來自NPC患者、非NPC受試者和前NPC受試者的樣本的EBV變體譜的系統樹分析。

【0025】 圖6概述了基於來自NPC患者、非NPC受試者和前NPC受試者的樣本的EBV變體譜的系統樹分析，所述EBV變體不包括29種報告的變體。

【0026】 圖7圖示了基於塊的變體模式分析的原理。

【0027】 圖8概述了13個NPC、16個非NPC和4個前NPC樣本的EBV DNA變體模式的基於塊的分析。

【0028】 圖9概述了13個NPC、16個非NPC和4個前NPC樣本的EBV DNA變體模式的基於塊的分析，所述EBV DNA變體不包括29種報告的變

體。

【0029】 圖10A示出了使用經訓練的分類器而計算出的NPC風險評分，該分類器基於使用基於塊的變體分析對所有EBV變體的分析。圖10B示出了使用經訓練的分類器而計算出的NPC風險評分，該分類器基於對29個報告的EBV變體的分析。圖10C示出了使用經訓練的分類器而計算出的NPC風險評分，該分類器基於使用基於塊的變體分析對所有EBV變體(但是不包括29種報告的變體)的分析。

【0030】 圖11概述了具有一過性陽性EBV DNA或持續性陽性EBV DNA的NPC患者和非NPC受試者的甲基化水平。

【0031】 圖12圖示了非癌症受試者的血漿DNA的大小變化的示意圖，所述受試者通過甲基化敏感性酶消化誘導且具有陽性血漿EBV DNA。填充和未填充的棒棒糖形狀分別代表甲基化和未甲基化的CpG位點。d橫線代表血漿EBV DNA分子。隨著酶消化，大小分布向左側移動。

【0032】 圖13是圖示了NPC患者的血漿DNA的大小變化的示意圖，所述患者具有通過甲基化敏感的酶消化誘導的陽性EBV DNA。填充和未填充的棒棒糖形狀分別代表甲基化和未甲基化的CpG位點。黃色(填充的)橫條代表血漿EBV DNA分子。隨著酶消化，大小分布向左側移動。

【0033】 圖14顯示了用甲基化敏感的限制性酶HpaII進行和不進行計算機內消化的血漿EBV DNA的大小分布。

【0034】 圖15示出了對於NPC患者和未患有NPC的受試者，使用和不使用甲基化敏感的限制性酶消化的血漿EBV DNA的累積大小分布圖。

【0035】 圖16A是表明了與NPC相關的整個EBV基因組的661個SNV位點的訓練集中的三個假設位點A、B和C的示意圖。制定測試樣本

的NPC風險評分，以通過血漿EBV DNA讀數(例如，具有可用的基因型信息)所覆蓋的這661個SNV位點的子集的基因型模式確定。從測試樣本的血漿測序數據來看，基因型信息僅可用於位點A和位點C，而不能用於位點B，因為該位點B沒有被任何測序的EBV DNA讀數覆蓋。圖16B是通過分析訓練集中所有的63個NPC樣本和88個非NPC樣本的在位點A和C處的基因型表明這兩個位點處的基因型權重的示意圖。構建邏輯回歸模型，以報告位點A和位點C上高危基因型的權重。圖16C是表明基於位點A和位點C的基因型，用其從訓練模型中得出的相應係數加權而推導出測試樣本的NPC風險評分的過程的示意圖。圖16D示出了來自訓練集中的NPC樣本和非NPC樣本的EBV基因組上的5678個SNV的分布(示出了在EBV基因組上的1000個核苷酸的滑動窗口中的變體總數)。

【0036】 圖17A和圖17B是概述使用留一法的訓練集中的NPC風險評分的圖。圖17A示出了訓練集中的NPC血漿樣本和非NPC血漿樣本的NPC風險評分。圖17B示出了通過NPC風險評分分析對NPC樣本和非NPC樣本進行區分的ROC曲線分析。

【0037】 圖18A和圖18B是概述測試集中的NPC風險評分的圖。圖18A示出了測試集中的NPC血漿樣本和非NPC血漿樣本的NPC風險評分。圖18B示出了通過NPC風險評分分析對NPC樣本和非NPC樣本進行區分的ROC曲線分析。

【0038】 圖19A和圖19B是通過分析EBER區域上的基因型模式來概述NPC風險分析的圖。圖19A示出了測試集中的NPC血漿樣本和非NPC血漿樣本的NPC風險評分，所述評分通過分析EBER區域上的基因型模式得出。圖19B示出了基於EBER區域上的NPC風險評分分析對NPC樣本和非

NPC樣本進行區分的ROC曲線分析。

【0039】 圖20A和圖20B是通過分析BALF2區域上的基因型模式來概述NPC風險的圖。圖20A示出了測試集中的NPC血漿樣本和非NPC血漿樣本的NPC風險評分，所述評分通過分析BALF2區域上的基因型模式得出。圖20B示出了基於BALF2區域上的NPC風險評分分析對NPC樣本和非NPC樣本進行區分的ROC曲線分析。

【0040】 圖21示出了計算機控制系統，其可以被編程或以其他方式配置為實現本文所提供的方法。

【0041】 圖22示出了本文公開的方法和系統的示意圖。

【實施方式】

交叉引用

【0042】 本申請要求於2020年1月15日提交的美國臨時申請號62/961,517及2019年4月2日提交的美國臨時申請號62/828,224的權益，這些臨時申請各自通過引用整體併入本文。

援引併入

【0043】 本說明書中提及的所有出版物、專利和專利申請均通過引用而併入於此，程度如同具體地和單獨地指出通過引用而併入每一單個出版物、專利或專利申請。

概述

【0044】 在一些方面，本文提供了用於篩查受試者中病原體相關疾病的方法和系統。該方法和系統可以基於來自受試者的生物樣品中的病原體的無細胞核酸分子的特性來評估受試者發展與病原體相關疾病的風險。其中，風險預測可以能够確定適當的篩查頻率。適當和及時的隨訪篩查不

僅可以節省受試者的費用，而且能够及早發現疾病。例如，EBV-NPC的階段分布向早期階段轉移可以顯著改善NPC患者的無進展生存期。

【0045】 受試者發展與病原體相關疾病的風險可指受試者傾向於發展與病原體相關疾病的可能性。在一些情況下，本文所述的風險是指在將來的時間點病原體相關疾病在受試者中發展為可被臨床檢測的狀態(“臨床可檢測的疾病”)的可能性。在一些情況下，在第一時間點通過測定來自受試者生物樣品中的病原體的無細胞核酸分子的篩查測定篩查受試者，並且當受試者在第一時間點被診斷為不具有臨床可檢測的病原體，來自受試者的生物樣品中的病原體的無細胞核酸分子的特性可以表明受試者在未來的時間點具有臨床上可檢測到的疾病的風險。

【0046】 臨床可檢測的疾病可以指可通過一種或多種完善的臨床診斷檢查來檢測的表現出病理症狀的疾病。在一些情況下，完善的臨床診斷檢查包括對病原體相關疾病的假陽性檢出率較低的醫學檢查/測定，所述假陽性檢出率例如低於30%、20%、10%、8%、7%、6%、5%、4%、3%、2.5%、2%、1%、0.8%、0.5%、0.25%、0.15%、0.1%、0.08%、0.05%、0.02%、0.01%、0.005%、0.002%、0.001%或甚至更低。完善的臨床診斷檢查包括還可以具有很高的檢測病原體相關疾病的敏感性的醫學檢查/測定，所述敏感性例如至少30%、40%、50%、60%、70%、80%、85%、90%、92%、94%、95%、96%、97%、98%、99%或99.5%或100%。在一些情況下，病原體相關疾病是與病原體相關的增生性疾病，如癌症，並且所述癌症可以通過一項或多項侵入性活檢，然後對活檢組織進行組織學檢查或其他檢查(例如組織分析、細胞檢查，如細胞DNA或蛋白質分析)、成像檢查，例如X射線、磁共振成像(MRI)、正電子發射斷層

掃描(PET)或計算機斷層掃描(CT) 或PET-CT、實驗室檢查(例如血液或尿液檢查)或身體檢查在臨床上以高置信度和低假陽性率診斷。病原體相關疾病的診斷可以由有資格的醫師根據上述或其他完善的臨床檢查結果進行。在一些情況下，第一次篩查測定的結果不會導致對病原體相關疾病的受試者進行藥物治療，因為通過完善的臨床診斷檢查診斷出該受試者不具有該疾病。

【0047】 在一些情況下，基於評估的風險，該方法包括確定受試者中病原體相關篩查測定的頻率。篩查測定的頻率可以與風險相關，並且兩次篩查測定的間隔，例如本文所述的篩查測定和隨後的隨訪篩查測定之間的間隔可以與風險成反相關。在一些情況下，該方法包括接收在第一時間點進行的第一次篩查測定的數據。第一次篩查測定可以包括確定來自受試者的生物樣品中的病原體的無細胞核酸分子的特性。例如，第一次篩查測定包括從受試者獲得生物樣品，並且該生物樣品包括來自受試者並且可能來自病原體的無細胞核酸分子，例如無細胞DNA。第一次篩查測定還可以包括確定來自生物樣品中的病原體的無細胞核酸分子的特性。可以在本文提供的方法和系統中使用的來自病原體的無細胞核酸分子的非限制性特性包括數量(例如，拷貝數或百分比)、甲基化狀態、片段大小、變體模式和與生物樣品中來自受試者的無細胞核酸分子相比的相對豐度。如本文所述，關於對受試者或受試者的生物樣品進行的檢查或測定的時間點可以指受試者進行檢查的時間點或從受試者獲得生物樣品的時間點，而不是對生物樣品進行實際測定的時間點。

【0048】 在一些方面，本文提供的方法包括：(a)接收在第一時間點進行的第一次測定的數據，所述第一次測定包括確定來自受試者的生物樣

品中的病原體的無細胞核酸分子的特性，其中來自所述病原體的無細胞核酸分子的特性包括數量(例如，拷貝數或百分比)、甲基化狀態、變體模式、片段大小或與所述生物樣品中來自受試者的無細胞核酸分子相比的相對豐度，並且其中所述特性指示所述受試者發展所述病原體相關疾病的風險；以及(b)基於所述特性，確定第二時間點，在所述第二時間點進行第二次測定以篩查受試者與所述病原體相關疾病，其中所述第一時間點和所述第二時間點之間的間隔與風險反相關。

【0049】 在一些情況下，如本文所述，受試者的生物樣品中的無細胞核酸分子的一個或多個特性使得能夠採用非侵入性方法來在受試者中來評估病原體相關疾病(例如癌症)的狀態或受試者未來發展病原體相關疾病的風險。不期望受到某種理論的束縛，可以用於該方法和系統中的無細胞核酸分子的一個或多個特性與該受試者發展病原體相關疾病的風險之間的關聯可能至少存在兩種可能的情况。在一種可能的情况下，經受與病原體相關疾病例如病原體相關的腫瘤的患病組織可能已經在初始篩查時(例如，第一次篩查測定)就已經存在。但是，患病組織(例如腫瘤)的大小可能太小，無法通過其他經典的醫學檢查方法，例如，檢測病原體相關疾病的假陽性率低於10%、5%、2%、1%、0.5%、0.1%或0.05%的方法，如內窺鏡檢查和磁共振成像(MRI)發現。隨著疾病的發展，例如，患病組織(例如腫瘤)的大小增長，然後可以在隨後的篩查(第二次篩查測定)中檢測到更晚期的患病組織，如腫大的組織(例如腫大的腫瘤)。另一個可能的情况是：病原體的核酸分子(例如EBV DNA)可以由處於初步患病狀態的細胞(例如惡化前細胞)釋放，並且這些細胞以後可以潛在地發展為患病細胞，例如癌細胞。無論關聯背後的确切的情况如何，本文所述的主題都可

以用於對受試者進行隨後可臨床檢測的NPC的風險分級。

【0050】 在一些情況下，根據健康經濟考慮因素(例如，篩查的成本)，受試者的偏好(例如，較頻繁的篩查間隔可能對某些受試者的生活方式造成更大的干擾)和其他臨床參數(例如個體的基因型(例如HLA狀態(Bei等人，*Nat Genet.* 2010;42:599-603; Hildesheim等人，*J Natl Cancer Inst.* 2002;94:1780-9.))、NPC家族史、飲食史、種族來源(例如，廣東人))來調整用於本文所述的特定篩查程序的實際時間間隔。

【0051】 在一些情況下，本文提供了方法包括：接收來自第一次測定的數據，所述第一次測定包括確定來自受試者的生物樣品中的病原體的無細胞核酸分子的特性，其中來自所述病原體的無細胞核酸分子的特性包括數量(例如，拷貝數或百分比)、甲基化狀態、變體模式、片段大小、片段末端的坐標、片段末端的序列基序或與所述生物樣品中來自所述受試者的無細胞核酸分子相比的相對豐度；以及基於來自所述病原體的無細胞核酸分子的特性和所述受試者的年齡、所述受試者的吸菸習慣、所述受試者的病原體相關疾病的家族史、受試者的基因型因素或所述受試者的飲食史中的一種或多種因素生成指示所述受試者發展所述病原體相關疾病的風險的報告。

【0052】 在各個方面，本文提供了用於分析來自受試者的生物樣品中的核酸分子的方法和系統。方法和系統的實例可以涉及分析來自生物樣品中的病原體的核酸分子的變體模式。在一些情況下，來自生物樣品中的病原體的核酸分子包括無細胞核酸分子。變體模式分析可涉及將被識別為源自病原體的生物樣品中的核酸分子的序列與病原體的一個或多個參考基因組進行比較，並隨後確定來自生物樣品中的病原體的核酸分子中的核苷

酸變體模式。

【0053】 在一些情況下，本文提供的方法和系統包括基於來自生物樣品中的病原體的核酸分子中的變體模式，確定受試者中病原體相關的病症的狀態或風險。例如，在血漿中檢測到的EBV基因組的遺傳變異可以用於預測未來NPC發展的風險。儘管先前已經報導存在於EBV相關腫瘤和對照樣本中的EBV菌株可能有所不同(Palser等人，*J Virol* 2015;89:5222-37)，但本研究中的腫瘤和對照樣本是從不同的地理位置收集。考慮到EBV變體的地理差異，因此很難得出結論——腫瘤樣本中已識別出的變體是地理關聯的還是疾病關聯的。

【0054】 在一些情況下，如本文所述的變體模式分析涉及來自生物樣品中的病原體的核酸分子與病原體的一個或多個參考基因組之間的全基因組比較。全基因組比較可能涉及整個病原體基因組的序列比對以及核苷酸變體模式的後續聚類分析。在一些情況下，全基因組比較涉及病原體參考基因組中大量位點的核苷酸變體的分析。這些位點可以包括整個病原體基因組中的所有位點。或者，病原體的參考基因組上的這些位點或變體位點可以包括至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400、至少500、至少600、至少700、至少800、至少900、至少1000、至少1100、至少1200、至少1300、至少1400、至少1500、至少1600、至少1700、至少1800、至少1900、至少2000、至少3000、至少4000或至少5000個通常可以發現核苷酸變體的位點。如本文所述的核苷酸變體可以包括單核苷酸變體(SNV)。如本文提供的用於變體模式分析的變體位點可以包括在病原體的基因組中識別的典型SNV。在一些情況下，變體位點可包括插入、缺失和融合。

【0055】 本文提供的全基因組變體模式分析可能優於單個單核苷酸多態性(SNP)的分析。在示例性情況下，雖然固定數目位點上的SNP可能與可導致受試者病理的病原體的特定菌株或亞型相關，但是基於對這些單個SNP進行的分析的風險評估可能受限於病原體的特定菌株或亞型，如果存在病原體的其他致病菌株或亞型，則基於這些單個SNP的分析進行的風險評估不足以提供對風險的準確評估。在另一示例性情況下，當生物樣品中的病原體核酸分子稀少時，例如，當分析生物樣品如血漿中的無細胞核酸分子時，本文提供的全基因組變體模式分析可能是有益處的。生物樣品中可用的病原體核酸分子可能沒有大量覆蓋病原體基因組。結果，涉及整個病原體基因組中大量變體位點的全基因組變體模式分析可以提供相對更全面的關於來自生物樣品中的病原體的無細胞核酸分子的基因型特徵的讀數，而涉及固定數目的單個多態性的分析受限於基因組的一個相對較小的區域或多個較小的區域，因此可以提供相對有限的來自生物樣品的病原體中的無細胞核酸分子的基因型特徵的讀數。

【0056】 在一些情況下，本文提供的變體模式分析包括基於塊的模式分析，其涉及將病原體的參考基因組分離到多個箱中，並分析相對於多個箱中的每箱的序列讀數。在一些情況下，該方法包括針對所述病原體的所述疾病相關的參考基因組，確定所述多個箱中的每一個的相似性指數。該相似性指數與相應箱內的一部分變體位點相關，在所述變體位點上，至少有一個映射到所述病原體的參考基因組的中的序列讀數具有與所述病原體的疾病相關的參考基因組相同的核苷酸變體。在一些情況下，病原體的疾病相關的參考基因組包括多個病原體的疾病相關的參考基因組，該方法包括確定多個箱中的每個箱與病原體的疾病相關的參考基因組中的每個組

的各自的相似性指數，以及基於多個疾病相關的參考基因組的比例確定多個箱中的每個箱的箱得分，相應箱的各自的相似性指數高於該得分的截止值。

無細胞核酸分子的測定

【0057】 來自受試者的生物樣品的無細胞核酸分子的篩查測定可以是任何合適的核酸測定。例如，可以採用測序方法來分析無細胞核酸分子的數量(例如，拷貝數或百分數)、甲基化狀態、片段大小或相對豐度。替代地或附加地，還可以使用基於擴增或雜交的方法，例如各種聚合酶鏈反應(PCR)方法或基於微陣列的方法。在一些情況下，使用例如免疫沉澱方法來分析核酸分子的甲基化狀態。

【0058】 在本公開內容的一些實例中，用於檢測無細胞病原體核酸分子(例如，無細胞EBV DNA)的篩查測定包括在不同時間點進行的一項以上的測試，並且無細胞病原體核酸分子在多次測試中的可檢測性可以指示受試者發展病原體相關疾病的風險。例如，該測定可以包括兩步測定，或者包括3、4、5、6、7、8、9、10或甚至更多步測試的測定方案。一些測試可以在同一時間點進行，而其他一些測試可以在不同的時間點進行，或者所有測試可以在不同的時間點進行。

【0059】 可以通過本文提供的方法和系統確定不同篩查測定的時間或篩查頻率。第一次篩查測定和第二次篩查測定之間的時間間隔可以是至少約2個月、4個月、6個月、8個月、10個月或12個月。在一些情況下，間隔為至少約12個月。第一次篩查測定和第二次篩查測定之間的時間間隔可以是大約1年、1.5年、2年、2.5年、3年、3.5年、4年、4.5年、5年、6年、7年、8年、9年、10年或更長時間。當受試者通過完善的臨床診斷方法正

常診斷為沒有病原體相關疾病(例如，沒有臨床上可檢測的病原體相關疾病)，即使首次篩查測定可以給出表明存在病原體相關疾病的陽性結果，間隔時間也可以很長。本文提供的方法和系統可以能夠預測受試者在未來例如6個月、12個月、2年、3年、5年或10年之內發展與病原體相關疾病的風險。基於評估的風險，可以確定適當的隨訪時間點。

【0060】 獲得樣品與進行測定之間的時間可以經優化以提高測定或方法的敏感性和/或特異性。在一些實施方案中，可以在進行測定之前立即獲得樣品(例如，在進行第一次測定之前獲得第一樣品，並且在進行第一次測定之後但在進行第二次測定之前獲得第二樣品)。在一些實施方案中，可以獲得樣品並在進行測定之前將樣品存儲一段時間(例如，數小時、數天或數周)。在一些實施方案中，可以在從受試者獲得樣品後1天、2天、3天、4天、5天、6天、1周、2周、3周、4周、5周、6周、7周、8周、3個月、4個月、5個月、6個月、1年或多於1年內對樣品進行測定。

【0061】 在進行測定(例如，第一次測定或第二次測定)與確定樣品是否包括指示疾病例如腫瘤的標誌物或標誌物集之間的時間可以變化。在一些情況下，時間可以被優化以提高測定或方法的敏感性和/或特異性。在一些實施方案中，確定樣品是否包含指示腫瘤的標誌物或標誌物集可以在進行檢測的至多0.1小時、0.5小時、1小時、2小時、4小時、8小時、12小時、24小時、2天、3天、4天、5天、6天、1周、2周、3周或1個月內發生。

【0062】 可以進行本文所述的生物樣品的測序分析，以分析來自病原體的無細胞核酸分子的一個或多個特性。本文提供的方法可以包括對生

物樣品中的核酸分子，例如無細胞核酸分子、細胞核酸分子或兩者進行測序。在一些情況下，本文提供的方法包括分析來自生物學樣品的核酸分子的測序結果，例如測序讀數。本文提供的方法和系統可以涉及或不涉及測序的活性步驟。該方法和系統可以包括或提供用於從測序儀接收和處理測序數據的裝置。該方法和系統還可以包括或提供用於向測序儀提供命令，例如基於對測序結果的分析的命令以調節測序過程的參數的裝置。

【0063】 商業上可獲得的測序裝置可以用於本公開內容中提供的方法，例如Illumina測序平臺和454/Roche平臺。可以使用本領域已知的任何方法對核酸進行測序。例如，測序可以包括下一代測序。在一些情況下，可以使用鏈終止測序、雜交測序、Illumina測序(例如，使用可逆終止劑染料)、ion torrent半導體測序、質譜法測序、大規模平行特徵測序(MPSS)、Maxam-Gilbert測序、納米孔測序、聚合酶克隆測序、焦磷酸測序、鳥槍法測序、單分子實時(SMRT)測序、SOLiD測序(使用四個螢光標記物的雙鹼基探針雜交)、通用測序或其任意組合進行核酸測序。

【0064】 在本文提供的方法中可以使用的一種測序方法可以涉及配對端測序，例如，使用Illumina的“配對端模塊”及其基因組分析儀。使用該模塊，在基因組分析儀完成第一次測序讀取後，配對端模塊可以指導原始模板的重新合成和第二輪簇的生成。通過在本文提供的方法中使用成對的末端讀數，可以從核酸分子的兩端獲得序列信息，並將兩端映射到參考基因組，例如病原體的基因組或宿主生物的基因組。在映射兩端之後，可以根據本文提供的方法的一些實施方案確定病原體集成概況。

【0065】 在配對末端測序期間，從核酸分子的第一端讀取的序列可以包括至少20、至少25、至少30、至少35、至少40、至少45、至少50、

至少55、至少60、至少65、至少70、至少75、至少80、至少85、至少90、至少95、至少100、至少105、至少110、至少105、至少120、至少125、至少130、至少135、至少140、至少145、至少150、至少155、至少160、至少165、至少170、至少175或至少180個連續核苷酸。從核酸分子的第一端讀取的序列可以包括至多24、至多28、至多32、至多38、至多42、至多48、至多52、至多58、至多62、至多68、至多72、至多78、至多82、至多88、至多92、至多98、至多102、至多108、至多122、至多128、至多132、至多138、至多142、至多148、至多152、至多158、至多162、至多168、至多172或至多180個連續核苷酸。從核酸分子的第一端讀取的序列可以包括約20、約25、約30、約35、約40、約45、約50、約55、約60、約65、約70、約75、約80、約85、約90、約95、約100、約105、約110、約105、約120、約125、約130、約135、約140、約145、約150、約155、約160、約165、約170、約175或約180個連續核苷酸。從核酸分子的第二端讀取的序列可以包括至少20、至少25、至少30、至少35、至少40、至少45、至少50、至少55、至少60、至少65、至少70、至少75、至少80、至少85、至少90、至少95、至少100、至少105、至少110、至少105、至少120、至少125、至少130、至少135、至少140、至少145、至少150、至少155、至少160、至少165、至少170、至少175或至少180個連續核苷酸。從核酸分子的第二端讀取的序列可以包括至多24、至多28、至多32、至多38、至多42、至多48、至多52、至多58、至多62、至多68、至多72、至多78、至多82、至多88、至多92、至多98、至多102、至多108、至多122、至多128、至多132、至多138、至多142、至多148、至多152、至多158、至多162、至多168、至多172

或至多180個連續核苷酸。從核酸分子的第二端讀取的序列可以包括約20、約25、約30、約35、約40、約45、約50、約55、約60、約65、約70、約75、約80、約85、約90、約95、約100、約105、約110、約115、約120、約125、約130、約135、約140、約145、約150、約155、約160、約165、約170、約175或約180個連續核苷酸。在一些情況下，從核酸分子的第一端讀取的序列可以包含至少75個連續核苷酸。在一些情況下，從核酸分子的第二端讀取的序列可以包含至少75個連續核苷酸。從核酸分子的第一端和第二端讀取的序列可以具有相同的長度或不同的長度。從生物樣品中的多個核酸分子讀取的序列可以具有相同的長度或不同的長度。

【0066】 本文提供的方法中的測序可以在各種測序深度下進行。測序深度可以指基因座被由基因座比對的序列讀數覆蓋的次數。該基因座可以小至核苷酸，大至染色體臂，或大至整個基因組。本文提供的方法中的測序深度可以是50x、100x，等等，其中“x”之前的數字是指基因座被序列讀數覆蓋的次數。測序深度還可以應用於多個基因座或整個基因組，在這種情況下，x可以指分別對基因座或單倍體基因組或整個基因組進行測序的平均次數。在一些情況下，在本文所述的方法中進行超深測序，所述超深測序可以指進行至少100x的測序深度。

【0067】 核酸內特定核苷酸在測序過程期間可被讀取的次數或平均次數(例如，測序深度)可以比被測序的核酸的長度大數倍。在一些情況下，當測序深度相比核酸的長度足夠大(例如，至少5倍)時，該測序可被稱為“深度測序”。在一些實例中，測序深度可以比被測序的核酸的長度平均大至少約5倍、至少約10倍、至少約20倍、至少約30倍、至少約40倍、

至少約50倍、至少約60倍、至少約70倍、至少約80倍、至少約90倍、至少約100倍。在一些情況下，樣品可以富含特定的分析物(例如，核酸片段或癌症特異性核酸片段)。

【0068】 本文提供的方法中產生的序列讀數(或測序讀數)可以指從任何部分或全部的核酸分子測序的一串核苷酸。例如，序列讀數可以是存在於生物樣品中的與核酸片段互補的短串核苷酸序列(例如20-150)、與核酸片段的末端互補的核苷酸序列串或與整個核苷酸互補的核苷酸序列串。可以以多種方式，例如，使用測序技術獲得序列讀數。

數量/檢測能力

【0069】 可以在所述方法和系統中使用的無細胞核酸分子的特性之一是來自病原體的無細胞核酸分子的數量(例如，拷貝數或百分比)。本公開內容的一些方面涉及基於對來自受試者的生物樣品中的病原體的無細胞核酸分子的數量(例如，拷貝數或百分比)的評估，對受試者發展與病原體相關的病症的風險進行分級。

【0070】 生物樣品中核酸分子的拷貝數可以與核酸分子的可檢測性相關。給定特定的測定方法，核酸模板的可檢測性可以與模板分子的拷貝數相關，例如，低於測定方法的檢測下限的拷貝數可能是不可檢測的，而拷貝數是等於或高於測定方法的檢測下限可以稱為“可檢測的”。例如，定量聚合酶鏈反應(qPCR)方法通常可以具有檢測極限，在該極限下不能將模板分子的信號與背景噪聲區分開。因此，在一些情況下，本文提供的方法和系統直接依賴於生物樣品中無細胞核酸分子的可檢測性，其可以與生物樣品中的拷貝數相關。在一些情況下，直接測定生物樣品中無細胞核酸分子的拷貝數。在其他情況下，通過檢測無細胞核酸分子本身來隱含測定

或推斷拷貝數。

【0071】 可以進行檢測測定，例如聚合酶鏈反應(PCR)或定量PCR(qPCR)，以評估生物樣品中病原體的無細胞核酸分子的存在或不存在或其拷貝數。可以將探針設計為靶向病原體特異性基因組區域，例如EBV特異性基因組DNA序列、人乳頭瘤病毒(HPV)特異性基因組DNA序列或乙肝病毒(HBV)特異性基因組DNA序列。

【0072】 儘管本文提供了多個實例和實施方案，但是涉及例如拷貝數和NPC的其他技術和實施方案可見於2011年11月30日提交的PCT AU/2011/001562，其全部內容通過引用並於本文。NPC與EBV感染密切相關聯。在中國南部，幾乎所有NPC患者的腫瘤組織中都可以發現EBV基因組。源自NPC組織的血漿EBV DNA已被開發為NPC的腫瘤標誌物(Lo等。癌症研究(Lo等人，*Cancer Res* 1999; 59:1188-1191)。具體地，實時qPCR測定可以用於靶向EBV基因組的BamHI-W片段的血漿EBV DNA分析。EBV基因組5中每一個大約有六至十二個BamHI-W片段重複，每個NPC腫瘤細胞中大約有50個EBV基因組(Longnecker等人，*Fields Virology*, 第5版第61章“Epstein-Barr virus”；Tierney等人，*J Virol.* 2011; 85: 12362-12375)。換言之，在每個NPC腫瘤細胞中可以有大約300-600(例如約500)個PCR靶標的拷貝。每個腫瘤細胞的該大量靶標可以解釋為什麼血漿EBV DNA在早期NPC的檢測中是高度敏感的標記物。NPC細胞可以將EBV DNA片段沉積到受試者的血液中。該腫瘤標誌物可以用於NPC的監測(Lo等人，*Cancer Res* 1999; 59: 5452-5455)和預測(Lo等人，*Cancer Res* 2000; 60: 6878-6881)。

【0073】 也可以以類似於本文所述的用於EBV的方式使用qPCR測定

法，以測定樣品中HPV、HBV或任何其他病毒DNA的量。這樣的分析對於篩查宮頸癌(CC)、頭頸部鱗狀細胞癌(HNSCC)、肝硬化或肝細胞癌(HCC)尤其有用。在一個實例中，qPCR測定法靶向HPV基因組的多態L1區域內的區域(例如200個核苷酸)。更具體地，本文考慮了qPCR引物的使用，所述qPCR引物與編碼L1區域中的一個或多個高變表面環的序列選擇性地雜交。

【0074】或者，可以使用測序技術對來自病原體的無細胞核酸分子進行檢測和定量。例如，可以對cfDNA片段進行測序並與HPV參考基因組進行比對和定量。或在其他實例中，對cfDNA片段的序列讀數與EBV或HBV的參考基因組進行比對和定量。

【0075】通過本文提供的測定法測量的來自病原體的無細胞核酸分子的可檢測性或拷貝數可以指示受試者發展與病原體相關疾病的風險。在一些實例中，來自病原體的無細胞核酸分子的拷貝數越高，則受試者傾向於發展病原體相關疾病的風險越高。在一些情況下，來自病原體的無細胞核酸分子在一個特定時間點或多個時間點上通過一種或多種測定法的可檢測性指示了受試者發展病原體相關疾病的風險。當與通過本文提供的測定法不可檢測到的分子相比，來自受試者的生物樣品中的病原體的無細胞核酸分子是可檢測的時，該受試者可能處於病原體相關疾病的更高的風險中。可以在如上所述的時間進行多步檢測測定。

【0076】在本公開內容的一些實例中，進行兩步測定法以檢測生物樣品中的無細胞病原體核酸分子。在一些情況下，進行兩步測定的第一次測定，然後根據第一時間點的測定結果，進行或不進行兩步分析的第二次測試。例如，如果第一次測試提供陽性結果，例如，在第一生物樣品中檢

測到無細胞病原體核酸分子，則可以進行兩步檢測測定的第二次測試；如果從第一次測試中獲得陰性結果，則可以不用進行第二次測試。在其他情況下，無論第一次測試如何，都進行第二次測試。在一些實例中，兩步檢測測定的兩個測試均具有陽性結果的情況被稱為永久陽性，而僅第一次或第二次測試具有陽性結果的情況被稱為一過性陽性。在一個示例性實例中，與“陰性”測定結果相比，“陽性”測定結果指示受試者發展病原體相關疾病例如EBV相關的NPC的更高風險，而“永久陽性”測定結果與“一過性陽性”檢測結果相比表明更高的風險。在一些示例性實例中，與獲得一過性陽性結果時相比，在第一時間點進行的兩步檢測測定中獲得永久陽性結果時，可以在第一時間點和第二時間點之間設置更長的時間間隔。例如，在EBV相關的NPC篩查中，如果從兩步檢測法的第一次測定中獲得了永久陽性結果，則建議應在第一次檢測測定的大約一年內進行隨訪的第二次篩查測定。相反，如果從兩步檢測法的第一次測定中獲得了一過性陽性結果，則可以在第一次檢測測定的大約兩年內進行隨訪的第二次篩查測定。如果獲得陰性結果，則可以在四年甚至更長的時間進行隨訪篩查測定。在一些情況下，指示較高風險的先前的陽性結果可以推翻間隔選擇，該間隔選擇將由指示較低風險的隨訪結果進行處理。例如，在第1年獲得永久陽性結果，然後不管在隨後的4年內所獲得的隨訪測定的結果如何，在隨後的4年內每年都會對受試者進行隨訪。示例性實例在圖2中給出並在實施例2中更詳細地進行描述。類似於檢測測定，基於來自病原體的無細胞核酸分子的其他特性的風險評估也可以遵循該示例性或類似的篩查方案。

【0077】可以在第一次測定後的數小時、數天或數周內進行該測定法的第二次測試。在一個實例中，可以在第一次測定後立即進行第二次測

定。在其他情況下，可以在第一次測定後1天、2天、3天、4天、5天、6天、1周、2周、3周、4周、5周、6周、7周、8周、3個月、4個月、5個月、6個月、1年或多於1年內進行第二次測定。在特定的實例中，可以在第一樣品的2周內進行第二次測定。通常，該測定法的第二次測試可以用於改善在患者中可以檢測病原體相關疾病(例如腫瘤)的特異性。可以通過實驗確定進行第一測試和第二測試之間的時間。在一些實施方案中，該方法可以包括2個或更多次測試，並且兩次測試都使用相同的樣品(例如，在進行第一次測定之前從受試者(例如患者)獲得單個樣品並保存一段時間直到進行第二次測定為止)。例如，可以同時從受試者獲得兩管血液。第一管可以用於第一次測試。僅當受試者的第一次測試結果為陽性時，才能使用第二管。可以使用本領域技術人員已知的任何方法(例如低溫地)保存樣品。在某些情況下，這種保存可能是有益的，例如，受試者可能收到陽性測試結果(例如，第一次測定指示癌症)，並且患者可能直到第二次測定前不願意等待，而寧願選擇尋求第二種意見。

甲基化狀態

【0078】 本公開內容的一些方面涉及基於對來自受試者的生物樣品中的病原體的無細胞核酸分子的甲基化狀態的評估，對受試者發展病原體相關的疾病的風險進行分級。

【0079】 無細胞病原體核酸分子的甲基化可以將樣品與具有病原體相關疾病的患者(例如EBV相關的NPC或HPV相關的宮頸癌)和無此疾病的受試者(例如非NPC受試者)區分開來。例如，與NPC相關的血漿EBV DNA的甲基化狀態可能不同於在非NPC受試者中檢測到的血漿EBV DNA的甲基化狀態，如美國專利申請16/046,795中所示，該專利申請通過引用

而整體併入本文。通過亞硫酸氫鹽測序進行分析時，來自NPC患者的血漿DNA與具有可檢測的EBV DNA的非NPC受試者之間可能存在差異化的甲基化區域。結果，對這些差異化的甲基化區域的甲基化狀態分析可以區分NPC和非NPC受試者。如本文所述，與NPC相關的EBV DNA甲基化狀態也可以預測NPC發展的風險，並且可以用於調節NPC篩查的間隔。例如，與不具有與NPC相關的EBV DNA甲基化模式的受試者相比，具有與NPC相關的EBV DNA甲基化模式的受試者可能被更頻繁地篩查。在一些情況下，還可以使用另一種類型的甲基化感知測序，例如使用Pacific Biosciences的單分子測序系統(Kelleher等人，*Methods Mol Biol.* 2018;1681:127-137；Powers等人，*BMC Genomics.* 2013;14:675)以及Oxford Nanopore(Simpson等人，*Nat Methods.* 2017;14:407-10)，以及在測序前使用甲基化敏感的限制性酶處理代替亞硫酸氫鹽測序。在另一種情況下，可以使用甲基化感知的且不基於測序的分子方法，例如甲基化特異性PCR(Herman等人，*Proc Natl Acad Sci U S A.* 1996;93:9821-6)、基於甲基化敏感酶(例如限制性酶)和亞硫酸氫鹽轉化並隨後進行質譜檢測的檢測系統(van den Boom 等人，*Methods Mol Biol.*(2009), 507(207):-27；Inoue等人，*Clin.Clin Chem.*2010;56:1627-35)，以及基於DNA分子甲基化狀態差異沉澱的方法(例如，使用抗甲基化的胞嘧啶抗體(Shen等人，*Nature.*2018;563:579-83；Zhou等人，*PLoS One.*2018;13:e0201586)或甲基化結合蛋白質(Zhang等人，*Nat Commun.*2013;4:1517)。

【0080】 在一些情況下，無細胞病原體核酸分子(例如血漿EBV DNA)的甲基化模式可用於檢測病原體相關疾病(例如病原體相關的癌症，例如NPC)或預測未來患有臨床可檢測的疾病的風險。如上所述，一種方

法是使用亞硫酸氫鹽處理核酸分子以將未甲基化的胞嘧啶轉化為尿嘧啶。甲基化的胞嘧啶不會被亞硫酸氫鹽改變，並保留為胞嘧啶。亞硫酸氫鹽處理的核酸分子的後續檢查(如測序)可以用於檢測生物樣品中核酸分子的甲基化狀態。

【0081】 在一個實例中，使用甲基化敏感的限制性酶分析來確定血漿甲基EBV DNA的甲基化水平的差異。甲基化敏感的限制性酶的一個非限制性實例是HpaII，其可以切割帶有未甲基化的“CCGG”基序的分子，但是使得分子不含“CCGG”或使得甲基化的“CCGG”的分子保持不變。替代地或另外地，可以使用其他甲基化敏感的限制性酶。在一個實例中，由於非癌症受試者中血漿EBV DNA的甲基化水平較低，因此非癌症受試者中血漿EBV DNA可能更易於被甲基化敏感的限制性酶切割。可以確定酶消化的敏感度，例如但不限於大規模平行測序、凝膠電泳、毛細管電泳、聚合酶鏈反應(PCR)和實時PCR。

【0082】 在使用測序(例如大規模平行測序)來分析甲基化敏感的限制性酶的消化程度的情況下，有酶消化和無酶消化的病原體無細胞核酸分子(例如血漿EBV DNA)的大小分布可以用來反映消化的程度。如圖12和圖13所示，大小分布曲線向左移動可以指示血漿EBV DNA的大小分布的縮短。曲線向左移動的次數越多，可以反映出酶的消化程度越高，並且暗示著DNA的甲基化水平越低。

【0083】 本文所述的無細胞病原體核酸分子的甲基化狀態可以包括單個甲基化位點的甲基化密度、病原體的基因組的連續區域上甲基化/未甲基化位點的分布、病原體的基因組的一個或多個特定區域內或病原體的整個基因組內每個甲基化位點的甲基化的模式或水平以及非CpG甲基化。

在一些情況下，甲基化狀態包括各個差異的甲基化位點的甲基化水平(或甲基化密度)，這些甲基化位點可以被例如患有病原體相關疾病(例如，EBV相關的NPC或HPV相關的宮頸癌)的患者和沒有患有該疾病的受試者(例如非NPC受試者)之間的樣品識別。對於給定的甲基化位點，甲基化密度可以指在包含這樣的甲基化位點的目標核酸分子總數中在給定的甲基化位點處甲基化的核酸分子的分數。例如，肝組織中的第一甲基化位點的甲基化密度可以指在總的肝臟DNA分子上在第一位點處甲基化的肝臟DNA分子的分數。在一些情況下，甲基化狀態包括各個甲基化位點之間甲基化/非甲基化狀態的一致性(coherence)(例如，模式或單倍型)。

【0084】 在一些情況下，本文所述的篩查測定(例如，第一次測定或第二次測定)可以包括通過任何可用的技術，例如但不限於進行甲基化-感知測序、甲基化敏感的擴增或甲基化敏感的沉澱來確定無細胞核酸分子的甲基化狀態。儘管在本文提供了實例和實施方案，但是涉及例如確定甲基化狀態的其他技術和實施方案可見於2013年9月20日提交的PCT AU/2013/001088，其全部內容通過引用並於本文。

片段大小

【0085】 本公開內容的一些方面涉及基於對來自受試者的生物樣品中的病原體的無細胞核酸分子的片段大小的評估，對受試者發展與病原體相關的疾病的風險進行分級。

【0086】 無細胞病原體核酸分子的片段大小分布和/或相對豐度可以將患有病原體相關疾病的患者(例如EBV相關的NPC或HPV相關的宮頸癌)的樣品和沒有患有此疾病的受試者(例如非NPC受試者)的樣品區分開來。例如，血漿EBV DNA分子的大小分布以及映射到EBV基因組和人類基因

組的循環DNA分子的比例對於將NPC患者與可檢測血漿EBV DNA的非NPC受試者區分開很有幫助，如在Lam等人*Proc Natl Acad Sci U S A*. 2018;115:E5115-E5124中使用大規模平行測定所示，其全部內容通過引用併入本文。根據本公開內容的一些實例，映射到EBV基因組和人類基因組的循環DNA分子的與NPC相關的大小分布和相對豐度也可以用於預測未來發展可在臨床上檢測到的NPC的風險。在一個實施方案中，與具有可檢測血漿EBV DNA但不具有這些NPC相關的特徵的受試者相比，在血漿DNA測序上具有這些與NPC相關的特徵但沒有可檢測的NPC的受試者的隨訪頻率更高。與使用如上所述的兩步測定相比，使用這種基於測序的分析對NPC的風險進行分級的一個潛在的實際優勢是可以省略從患者身上收集另一份血液樣品。

【0087】 在一些情況下，測定(例如，第一次測定或第二次測定)可以包括進行測定(例如下一代測序測定)以分析核酸片段大小，例如血漿EBV DNA的片段大小。在一些情況下，測序用於評估樣品中無細胞病毒核酸的大小。例如，每個經序的血漿DNA分子的大小可以從該序列的開始坐標和結束坐標得出，其中該坐標可以通過將序列讀數映射(比對)到病毒基因組來確定。在各個實例中，可以從兩個成對的末端讀數或覆蓋兩個端點的單個讀數來確定DNA分子的起始坐標和終點坐標，這可以在單分子測序中實現。在一些情況下，基於擴增或雜交的方法也可以用於片段大小分析。例如，可以將探針設計為靶向各種長度的基因組區域，當具有等於或大於比靶區域的長度時，擴增(例如PCR或qPCR)或雜交信號可以指示靶基因組區域上無細胞核酸片段的數目。因此可以推斷出片段的大小分布。片段大小測定和分析的方法可以包括在美國專利公開號

US20180208999A1中所述的方法，該專利通過引用全文併入本文。

【0088】 片段的大小分布可以以直方圖的形式顯示，橫軸上表示核酸片段大小。可以確定每種大小(例如，在1 bp的分辨率內)的核酸片段的數目，並可以在縱軸上繪製，例如，作為原始數目或頻率百分比。大小的分辨率可以大於1 bp(例如2、3、4或5 bp的分辨率)。以下對大小分布(也稱為尺寸分布)的分析顯示，在統計學上，來自NPC受試者的無細胞混合物中的病毒DNA片段在統計學上比沒有可觀察到病理學的受試者的病毒片段更長。在一個示例性實例中，在從血漿EBV DNA分析獲得的片段大小分布曲線中，在NPC患者血漿EBV DNA大小分布中可能存在特徵性的166 bp峰(核小體模式)，而非癌症血漿EBV DNA受試者沒有表現出典型的核小體模式。

【0089】 在一些情況下，計算來自病原體的無細胞核酸分子與來自受試者的無細胞核酸分子相比的相對豐度以評估風險。在一些情況下，分析在大小比值方面的相對豐度。在各個實例中，病原體片段與來自受試者的無細胞片段的大小比值是指來自病原體的無細胞核酸片段與來自受試者的無細胞核酸片段之間的數量比值。例如，EBV DNA片段在80和110個鹼基對之間的大小比值可以是：

$$\text{大小} \geq 110\text{bp 比值} = \frac{\text{80-110bp內EBV DNA片段的比例}}{\text{80-110bp內常染色體DNA片段的比例}}$$

【0090】 在各個情況下，設定截止值或閾值用於評估。例如，可以存在用於確定病原體片段和受試者常染色體片段之間的大小比值的大小閾值。或在一些情況下，設置大小閾值，以便將尺寸小於或大於閾值的多個片段視為指示受試者患有病原體相關疾病的風險。應當理解，大小閾值可以是任何值。大小閾值可以是至少約10 bp、20 bp、25 bp、30 bp、35

bp、40 bp、45 bp、50 bp、55 bp、60 bp、65 bp、70 bp、75 bp、80 bp、85 bp、90 bp、95 bp、100 bp、105 bp、110 bp、115 bp、120 bp、125 bp、130 bp、135 bp、140 bp、145 bp、150 bp、155 bp、160 bp、165 bp、170 bp、175 bp、180 bp、185 bp、190 bp、195 bp、200 bp、210 bp、220 bp、230 bp、240 bp、250 bp或大於250 bp。例如，大小閾值可以是150 bp。在另一個實例中，大小閾值可以是180bp。在一些實施方案中，可以使用大小上限閾值和大小下限閾值(例如，數值範圍)。在一些實施方案中，上限大小閾值和下限大小閾值可以用於選擇長度在上限閾值和下限閾值之間的核酸片段。在一些實施方案中，上限截止值和下限截止值可以用於選擇長度大於上限截止值且小於下限大小截止值的核酸片段。在一些情況下，使用大小比值的截止值來確定受試者是否具有發展病原體相關疾病(例如，NPC)的風險或該風險是多少。例如，與具有假陽性血漿EBV DNA結果的受試者相比，具有NPC的受試者在80至110 bp的大小範圍內具有較低的大小比值。在一些情況下，大小比值的截止值可以是約0.1、約0.5、約1、約2、約3、約4、約5、約6、約7、約8、約9、約10、約11、約12、約13、約14、約15、約16、約17、約18、約19、約20、約25、約50、約100或大於約100。在一些情況下，大小指數的截止值可以是約或至少10、約或至少2、約或至少1、約或至少0.5、約或至少0.333、約或至少0.25、約或至少0.2、約或至少0.167、約或至少0.143、約或至少0.125、約或至少0.111、約或至少0.1、約或至少0.091、約或至少0.083、約或至少0.077、約或至少0.071、約或至少0.067、約或至少0.063、約或至少0.059、約或至少0.056、約或至少0.053、約或至少0.05、約或至少0.04、約或至少0.02、約或至少0.001或小於約0.001。

【0091】 可以確定核酸片段的大小分布的各個統計值。例如，可以使用大小分布的平均值、衆數、中位數或均值。還可以使用其他統計值，例如，給定大小的累積頻率或不同大小的核酸片段的數量的各種比值。累積頻率可以對應於給定大小或更小或大於給定大小的DNA片段的比例(例如百分比)。統計值提供有關核酸片段大小分布的信息，以便與一個或多個截止值進行比較，以確定由病原體引起的病理學水平。可以使用健康受試者、已知具有一種或多種病理的受試者、對與病原體相關的病理呈假陽性的受試者以及本文提及的其他受試者的隊列來確定截止值。本領域技術人員將知道如何基於本文的描述確定這種截止值。

【0092】 在一些實例中，可以將病原體片段大小的第一統計值與來自人類基因組的大小的參考統計值進行比較。例如，可以確定第一統計值和參考統計值之間的時間隔值(例如，差異或比值)，例如，從病原體參考基因組中的其他區域確定或從人類核酸確定。間隔值也可以從其他值中來確定。例如，參考值可以從多個區域的統計值確定。可以將間隔值與大小閾值進行比較以獲得大小分類(例如，DNA片段是否比正常區域短、長還是與正常區域相同)。

【0093】 一些實例可以計算參數(間隔值)，該參數可以使用以下等式定義為參考病原體基因組與參考人類基因組之間短DNA片段比例的差值：

$$\Delta F = P|_{\leq 150\text{bp}}^{\text{測試}} - P|_{\leq 150\text{bp}}^{\text{參考}}$$

其中 $P|_{\leq 150\text{bp}}^{\text{測試}}$ 表示源自測試區域的大小≤150 bp的測序片段的比例，並且 $P|_{\leq 150\text{bp}}^{\text{參考}}$ 表示源自參考區域的大小≤150 bp的測序片段的比例。在其他實施方案中，可以使用其他大小閾值，例如但不限於100 bp、110 bp、120

bp、130 bp、140 bp、160 bp和166 bp。在其他實施方案中，大小閾值可以以鹼基或核苷酸或其他單位表達。

【0094】 可以使用對照受試者的平均值和SD值來計算基於大小的z得分。

$$\text{基於大小的z-得分} = \frac{\Delta F_{\text{樣品}} - \text{平均值}\Delta F_{\text{對照}}}{SD\Delta F_{\text{對照}}}$$

【0095】 在一些實施方案中，大於3的基於大小的z得分指示病原體的短片段的比例增加，而小於-3的基於大小的z得分指示病原體的短片段的比例減小。可以使用其他大小閾值。基於大小的方法的更多細節可見於美國專利號8,620,593和8,741,811，以及美國專利公開號2013/0237431，其中每一個均通過引用整體併入本文。

【0096】 爲了確定核酸片段的大小，本公開內容的至少一些實例可以使用可以分析染色體的起源和分子的長度的任何單分子分析平臺，例如，電泳、光學方法（例如，光學映射及其變體，en.wikipedia.org/wiki/Optical_mapping#cite_note-Nanocoding-3 和 Jo 等人，*Proc Natl Acad Sci USA*.(2007) 104:2673-2678)、基於熒光的方法、基於探針的方法、數字PCR(基於微流體或基於乳液的方法，例如 BEAMing(Dressman 等人，*Proc Natl Acad Sci USA*.(2003) 100:8817-8822)、RainDance(www.raindancetech.com/technology/pcr-genomics-research.asp))、滾環擴增、質譜、熔解分析(或熔解曲線分析)、分子篩等。以質譜的爲例，較長的分子將具有較大的質量(例如，大小值的實例)。

【0097】 在一個實例中，可以使用配對末端測序方案對核酸分子進行隨機測序。兩端的兩個讀數可以被映射(比對)到參考基因組，其可以被

重複屏蔽(例如，當與人類基因組比對時)。DNA分子的大小可以從兩個讀數所映射到的基因組位置之間的距離來確定。

變體模式分析

【0098】 本公開內容的一些方面涉及基於對來自受試者的生物樣品中的病原體的無細胞核酸分子的變體模式的評估，對受試者發展病原體相關的病症的風險進行分級。在生物樣品中檢測到的病原體基因組的遺傳變體可以用於預測未來發展病原體相關疾病的風險。

【0099】 與沒有患有病原體相關疾病的受試者的樣品相比，患有病原體相關疾病(例如，病原體相關的惡性腫瘤)的患病組織中的病原體核酸分子的變體模式可能不同。據報導，存在於EBV相關的腫瘤和對照樣品中的EBV菌株可能有所不同(Palser等人，*J Virol.* 2015;89:5222-37)。但是，在此之前的研究中，腫瘤和對照樣品是從不同地理位置收集的。考慮到EBV變體的潛在的地理差異，可能很難對腫瘤樣品中已識別出的變體是地理關聯的還是疾病關聯的得出結論。以前曾嘗試通過分析NPC腫瘤樣品來識別與NPC相關的EBV變體。在一項全基因相關研究(Hui等人，*Int J Cancer* 2019, doi.org/10.1002/ijc.32049)中，分析了來自相同的地理區域的沒有患有EBV相關疾病的個體的NPC腫瘤樣品和唾液樣品，存在29種經識別低於假髮現率(調整後的P為0.05)的多態性(單核苷酸多態性(SNP)或插入)。這些29個與NPC相關的EBV變體顯示存在於90%以上的NPC病例中，但僅占對照病例的40%-50%。

【0100】 與發展NPC的個體EBV多態性分析相反(Hui等人，*Int J Cancer* 2019, doi.org/10.1002/ijc.32049; Feng等人，*Chin J Cancer* 2015;34:61)，本公開內容的各方面提供了用於以全基因組方式分析變體

模式的病原體核酸分子的方法和系統。此外，本公開內容的各方面提供了通過分析無細胞的病原體核酸分子來分析病原體變體模式的方法和系統，而不是通過分析腫瘤和細胞系樣品來識別與疾病相關的EBV變體(Palser等人，*J Virol.* 2015;89:5222-37, Correia等人，*J Virol.*2018;92:e01132-18, Hui等人，*Int J Cancer* 2019, doi.org/10.1002/ijc.32049)，所述無細胞的病原體核酸分子例如位於血液(例如血漿或血清)、鼻沖洗液、鼻刷樣品或相比於侵入性腫瘤活檢物通過無創或微創手術獲得的其他體液中。在一個示例性實例中，血液中EBV DNA分子的低豐度和碎片化性質可能給分析帶來技術挑戰。以無創的方式分析無細胞病毒DNA分子的變體模式可以提高臨床應用，包括篩查、預測醫學、風險分級、監測和預後。在一個實例中，該分析可用於區分具有不同病毒相關病症的受試者，例如，在篩查的背景下具有可檢測血漿EBV DNA的NPC患者和非NPC受試者。在另一個實例中，它可以用於疾病或癌症風險預測。

【0101】可以使用不同的方法來獲取變體模式。非限制性測定方法可以包括大規模平行測序(MPS)、Sanger測序(例如Lorenzetti等人，*J Clin Microbiol.* 2012;50:609-18中所用的)以及基於微陣列的SNP分析(例如Wang等人，PNAS 2002; 99 : 15687-92中所述)、雜交分析和質譜分析。在一個示例性實例中，使用測序方法，例如具有捕獲富集的靶向測序、MPS或Sanger測序，並且參考病原體的參考基因組(例如，EBV參考基因組)以每個核苷酸為基礎分析序列讀數。該方法可以包括從受試者的生物樣品中獲得無細胞核酸分子的序列讀數。該方法可以進一步包括將序列讀數與病原體的參考基因組進行比對。該方法可以進一步包括通過分析病原體的參考基因組和映射到病原體的參考基因組的序列讀數之間的核苷

酸變異來分析整個病原體的參考基因組的核苷酸變體模式。本文提供的變體模式可以表徵在病原體參考基因組上多個變體位點的每個位點處映射到的病原體參考基因組的序列讀數的核苷酸變體。多個變體位點可以包括跨越病原體的參考基因組的至少30、至少40、至少50、至少60、至少70、至少80、至少90、至少100、至少200、至少300、至少400、至少500、至少600、至少700、至少800、至少900、至少1000、至少1100或至少1200個位點。在一些情況下，多個變體位點包括跨越病原體的參考基因組的至少1000個位點。在一些情況下，多個變體位點包括跨越所述病原體的所述參考基因組的約1100個位點。在一些情況下，多個變體位點包括跨越病原體的參考基因組的至少600個位點。在一些情況下，多個變體位點包括跨越病原體的所述參考基因組的約660個位點。在一些情況下，多個變體位點包括選自表6中所提出的相對於EBV參考基因組(AJ507799.2)的基因組位點的至少30、40、50、100、150、200、250、300、350、400、450、500、550或600個位點。在一些情況下，多個變體位點包含表6中所列出的相對於EBV參考基因組(AJ507799.2)的基因組位點。

【0102】 在一些情況下，來自病原體的無細胞核酸分子的變體模式表徵在多個變體位點的每個位點上映射到病原體的所述參考基因組的序列讀數的核苷酸變體，所述多個變體位點隨機選自表6中所列出的相對於EBV參考基因組(AJ507799.2)的基因組位點。在一些情況下，本文提供的方法包括從在表6中所列出的相對於EBV參考基因組(AJ507799.2)的基因組位點中隨機選擇多個變體位點的步驟。該方法還可以包括通過分析病原體的參考基因組與映射到病原體的參考基因組的序列讀數之間的核苷酸變

體來分析在隨機選擇的多個變體位點上的核苷酸變體模式。

【0103】 在一些情況下，來自病原體的無細胞核酸分子的變體模式表徵在多個變體位點的每個位點上映射到病原體的所述參考基因組的序列讀數的核苷酸變體，所述多個變體位點包括隨機選自從表6中所列出的相對於EBV參考基因組(AJ507799.2)的基因組位點的30、40、50、100、150、200、250、300、350、400、450、500、550或600個位點。

【0104】 在一些情況下，所述多個變體位點由映射到病原體的參考基因組的序列讀數具有與病原體的參考基因組不同的核苷酸變體的所有位點組成。

【0105】 在一些情況下，野生型病原體基因組用作參考基因組。例如，野生型EBV基因組(GenBank：AJ507799.2)可以用作參考EBV基因組。在其他情況下，其他病原體基因組用作參考基因組。在又一實例中，多個病原體基因組(例如，EBV基因組)被用作參考。在另一個實例中，將共有序列用作參考。可以通過組合不同病原體基因組序列的變體來建立共有序列，例如，在de Jesus等人，*J Gen Virol.* 2003;84:1443-50中所述的EBV基因組的共有序列。

【0106】 本文提供的方法和系統中利用的、例如用於分析拷貝數、甲基化狀態、片段大小、相對豐度或變體模式的序列比對可以通過任何適當的生物信息學算法、程序、工具包或程序包來進行。例如，可以使用短寡核苷酸分析包(SOAP)作為應用本文提供的方法和系統的比對工具。可以在本文提供的方法和系統中使用的短序列讀取分析工具的實例包括Arioc、BarraCUDA、BBMap、BFAST、BigBWA、BLASTN、BLAT、Bowtie、Bowtie2、BWA、BWA-PSSM、CASHX、Cloudburst、CUDA-

EC 、 CUSHAW 、 CUSHAW2 、 CUSHAW2-GPU 、 CUSHAW3 、 drFAST 、 ELAND 、 ERNE 、 GASSST 、 GEM 、 Genalice MAP 、 Geneious Assembler 、 GensearchNGS 、 GMAP和GSNAP 、 GNUMAP 、 HIVE-hexagon 、 Isaac 、 LAST 、 MAQ 、 mrFAST 、 mrsFAST 、 MOM 、 MOSAIK 、 MPscan 、 Novoalign & NovoalignCS 、 NextGENe 、 NextGenMap 、 Omixon Variant Toolkit 、 PALMapper 、 Partek Flow 、 PASS 、 PerM 、 PRIMEX 、 QPalma 、 RazerS 、 REAL 、 cREAL 、 RMAP 、 rNA 、 RTG Investigator 、 Segemehl 、 SeqMap 、 Shrec 、 SHRiMP 、 SLIDER 、 SOAP 、 SOAP2 、 SOAP3 、 SOAP3-dp 、 SOCS 、 SparkBWA 、 SSAHA 、 SSAHA2 、 Stampy 、 SToRM 、 Subread 、 Subjunc 、 Taipan 、 UGENE 、 VelociMapper 、 XpressAlign和ZOOM 。

【0107】 序列讀取中的多個連續核苷酸(“序列延伸”)可以用於比對參考基因組，以進行關於比對的調用(call)。例如，比對可以包括將序列讀數的至少4、至少6、至少8、至少10、至少12、至少14、至少16、至少18、至少20、至少22、至少24、至少25、至少26、至少28、至少30、至少32、至少34、至少35、至少36、至少38、至少40、至少42、至少44、至少45、至少46、至少48、至少50、至少52、至少54、至少55、至少56、至少58、至少60、至少62、至少64、至少65、至少66、至少67、至少68、至少69、至少70、至少71、至少72、至少73、至少74、至少75、至少76、至少78、至少80、至少82、至少84、至少85、至少86、至少88、至少90、至少92、至少94、至少95、至少96、至少98、至少100、至少102、至少104、至少106、至少108、至少110、至少112、至少114、至少116、至少118、至少120、至少122、至少124、至少126、至少

128、至少130、至少132、至少134、至少136、至少138、至少140、至少142、至少145、至少146、至少148或至少150個連續核苷酸與參考基因組例如病原體的參考基因組或宿主生物的參考基因組進行比對。在一些情況下，本文提到的比對可以包括將序列讀數的至多5、至多7、至多9、至多11、至多13、至多15、至多17、至多19、至多21、至多23、至多25、至多27、至多29、至多31、至多33、至多35、至多37、至多39、至多41、至多43、至多45、至多47、至多49、至多51、至多53、至多55、至多57、至多59、至多61、至多63、至多65、至多67、至多68、至多69、至多70、至多71、至多72、至多73、至多74、至多75、至多76、至多78、至多80、至多81、至多83、至多85、至多87、至多89、至多91、至多93、至多95、至多97、至多99、至多101、至多103、至多105、至多107、至多109、至多111、至多113、至多115、至多117、至多119、至多121、至多123、至多125、至多127、至多129、至多131、至多133、至多135、至多137、至多139、至多141、至多143、至多145、至多147、至多149或至多151個連續核苷酸與參考基因組例如病原體的參考基因組或宿主生物的參考基因組進行比對。在一些情況下，本文提到的比對可以包括將序列讀數的約20、約22、約24、約25、約26、約28、約30、約32、約34、約35、約36、約38、約40、約42、約44、約45、約46、約48、約50、約52、約54、約55、約56、約58、約60、約62、約64、約65、約66、約67、約68、約69、約70、約71、約72、約73、約74、約75、約76、約78、約80、約82、約84、約85、約86、約88、約90、約92、約94、約95、約96、約98、約100、約102、約104、約106、約108、約110、約112、約114、約116、約118、約120、約122、約124、

約126、約128、約130、約132、約134、約136、約138、約140、約142、約145、約146、約148、約150、約152、約154、約155、約156、約158、約160、約162、約164、約165、約166、約168、約170、約172、約174、約175、約176、約178、約180、約185、約190、約195或約200個連續核苷酸與參考基因組例如病原體的參考基因組或宿主生物的參考基因組進行比對。

【0108】 在一些情況下，當序列延伸在整個序列讀數上與參考基因組的特定區域例如人類參考基因組具有至少80%、至少85%、至少90%、至少95%、至少98%、至少99%或100%的序列同一性或互補性時進行比對調用。在一些情況下，當序列延伸在整個序列讀取上與參考基因組的特定區域例如人類參考基因組具有至少80%的序列同一性或互補性時進行比對調用。在一些情況下，當序列延伸與參考基因組的特定區域例如人類參考基因組的特定區域相同或互補且錯配不超過20、15、10、9、8、7、6、5、4、3、2或1個鹼基或零錯配時進行比對調用。在一些情況下，當序列延伸與參考基因組的特定區域例如人類參考基因組的特定區域相同或互補且錯配不超過2個鹼基時，進行比對調用。最大錯配數或百分比或最小相似度數或百分比可以根據在本文中提供的方法和系統的應用目的和背景作為選擇標準而變化。

【0109】 在一些情況下，序列讀數與病原體參考基因組的比對允許最大錯配不超過20、15、10、9、8、7、6、5、4、3、2或1個鹼基。映射的序列讀數與病原體的參考基因組之間的錯配可以指示存在於生物樣品中的病原體基因組序列中的核苷酸變體，在其他情況下，其還可以指示測序錯誤。不希望受某一理論的束縛，在一個生物樣品中的給定基因組位點識

別出一個以上的核苷酸變體可能是由於無細胞病原體核酸分子所來源的患病細胞的測序錯誤或異質性。在一些情況下，如果在給定的生物樣品中識別出超過1、2或3個核苷酸變體，則將基因組位點的核苷酸變體從分析中排除。

【0110】 在示例性實例中，使用具有捕獲富集的靶向測序分析具有可檢測血漿EBV DNA的NPC受試者和非NPC受試者的循環中的無細胞病毒DNA分子。捕獲探針可以設計成覆蓋整個EBV基因組。在其他情況下，僅可以分析一部分的EBV基因組，並且捕獲探針被設計為僅覆蓋一部分的EBV基因組。在同一分析中，還可以包括捕獲探針以靶向人類基因組中的目標基因組區域。例如，可以包括靶向人類共同單核苷酸多態性(SNP)位點和人類白細胞抗原(HLA)SNP的探針。在一個實施方案中，可以將更多的探針設計成與其他病毒基因組序列，例如HPV或HBV基因組雜交。

【0111】 在一些情況下，可以通過直接比較映射到參考基因組的序列讀數和參考基因組來分析病原體基因組的變體模式。可以以任何適當的方式例如用於聚類分析或系統樹分析進一步處理比較結果。用於這些分析的可用生物信息學工具可以包括MEGA4、MEGA5、CLUSTALW、Phylip、RAxML、BEAST、PhyML、TreeView、MAFFT、MrBayes、BIONJ、MLTreeMap、Newick Utilities、Phylo.io、Phylogeny.fr、REALPHY、SuperTree和The Phylogenetic Web Repeater (POWER)。聚類分析或系統樹分析將映射到病原體參考基因組的序列讀數與一個或多個病原體基因組進行比較，這些病原體基因組是從患病組織或健康受試者那裏獲得的，或被指示為能夠或無法引起病原體相關疾病，或被指示為在

引起病原體相關疾病方面有效或無效。

【0112】在示例性實例中，本文提供的方法和系統包括基於塊的變體模式分析。基於塊的變體模式分析可以包括將病原體的參考基因組分離成多個箱(“塊”)。將映射到病原體參考基因組的序列讀數與多個箱中的每個箱中與疾病相關的病原體基因組進行比較。在某些情況下，存在多個，例如至少2、3、4、5、6、7、8、9、10、12、14、16、18、20、22、24、26、28、30，40、50、60、70、80、90、100、120、140、160、180、200、300、400、500、600、700、800、900或1000個不同的病原體基因組，包括與疾病相關的病原體基因組，以及可選的已知或表明無法或不能有效導致病原體相關疾病的病原體基因組(與疾病無關的病原體基因組)可以進行比較用於基於塊的分析。在基於塊的分析中，在多個箱中的每個箱中，基於映射到病原體參考基因組的序列讀數與每個與疾病相關的病原體基因組或與疾病無關的病原體基因組之間共享的核苷酸變體來計算相似性指數。相似性指數可以取決於變體位點的比例，在該變體位點上映射到病原體參考基因組的至少一個序列讀數具有與疾病相關或與疾病無關的病原體基因組相同的核苷酸變體。基於針對序列讀數進行比較所針對的每個病原體基因組的相似性指數，可以基於例如由相似性指數所反映的相似度水平來計算箱得分。在一種情況下，箱得分可以取決於高於預定截止值的相似性指數的比例。可以為相似性指數設置一個截止值，例如約0.6、0.7、0.75、0.8、0.85、0.9或0.95。截止值以上的相似性指數可以表明序列讀數與其進行比較所針對的病原體基因組“相似”。基於上述分析，之後可以使用計算出的相似性指數或箱分數在整個病原體基因組或部分病原體基因組上進行更大範圍的模式分析。與上述類似的聚類分析或系

統樹分析可以遵循基於塊的分析，以預測發展病原體相關的疾病(如EBV相關的NPC)的風險。

風險評分

【0113】 本公開內容的一些方面涉及基於對來自受試者的生物樣品中的病原體的無細胞核酸分子的一個或多個特性的組合考慮的評估，對受試者發展病原體相關的疾病的風險進行分級。在一些情況下，產生指示受試者發展病原體相關疾病(例如EBV相關的鼻咽癌)風險的風險評分。

【0114】 在一些情況下，本公開內容涉及基於對來自受試者的生物樣品中的病原體的無細胞核酸分子的一個或多個特性的組合考慮和受試者的年齡、受試者的吸菸習慣、受試者的NPC家族史、受試者的基因型因素、受試者的飲食史或受試者的種族的一個或多個因素，對受試者發展病原體相關的疾病的風險進行分級。在沒有臨床可檢測的NPC的受試者中，血漿EBV DNA的檢出陽性率與受試者的年齡之間可能存在正相關。受試者的吸菸習慣可以使受試者發展NPC的風險更高。有NPC家族史的受試者可能具有更高的自身發展NPC的風險。諸如HLA狀態等基因型因素也可能與NPC的風險相關，如Bei等人, *Nat Genet.* 2010;42:599-603 和 Hildesheim等人, *J Natl Cancer Inst.* 2002;94:1780-9中所表明，其每一個都以整體併入本文。另外，飲食史可能與NPC的風險相關，例如，食用大量醃製魚的受試者可能具有相對較高的NPC風險。某些種族，例如廣東人，也可能與發展NPC的高風險相關。

【0115】 在一些情況下，所述方法和系統還包括生成指示受試者發展病原體相關疾病的風險的報告。這樣的報告可以具有數字風險評分值或分類風險評估。在一些情況下，該報告包括篩查頻率的建議或隨訪篩查測

定的未來時間點。可以將報告提供給受試者、為受試者服務的醫療保健機構或醫療保健專業人員或任何相關的第三方，例如醫療保險公司。在報告發布之前或之後，可以由執業醫生審查、評估或編輯該報告。在一些情況下，執業醫生會根據他/她的醫學意見或獨立的檢查對風險評估提供額外的意見或促成最終的風險評估。

【0116】 在一些情況下，本公開內容提供了通過使用分類器來將發展病原體相關疾病例如病原體相關增生性疾病例如EBV相關NPC的風險進行分級的方法。這樣的分類器可以將本文所述的一個或多個因素作為數據輸入，並提供包括風險評分的輸出，該風險評分可以指示受試者發展病原體相關疾病的風險。可以輸入到分類器中的一種或多種因素可以包括無細胞病原體核酸分子的一種或多種特性、來自受試者的生物樣品中的病原體的無細胞核酸分子的一種或多種特性以及受試者的年齡、吸菸習慣、受試者的NPC家族史、受試者的基因型因素、飲食史和受試者的種族中的一個或多個因素。作為分類器的輸出的風險評分可以指示受試者當前患上或將來或發展病原體相關疾病的風險。在一些情況下，風險評分指示該受試者當前患有病原體相關疾病的可能性。在一些情況下，風險評分指示受試者在未來的一段時間內(例如但不限於1年、2年、3年、4年、5年、10年或15年內)發展病原體相關疾病的可能性。在一些情況下，分類器提供的輸出包括推薦的篩查頻率或用於隨訪篩查測定的未來時間點。這樣的輸出可以是臨床推薦的形式，或者可以在如上所述的報告中提供給受試者、醫療保健機構或醫療保健專業人員或任何第三方，例如醫療保險公司。

【0117】 如本文所述，分類器可以指實現分類的任何算法。在本公開內容中，分類器可以是建立在用於預測未來發展病原體相關疾病的風險

的任何適當算法上的分類模型。適當的算法可以包括機器學習算法和其他數學/統計模型，例如但不限於支持向量機(SVM)、樸素貝葉斯、邏輯回歸、隨機森林、決策樹、梯度提升樹、神經網絡、深度學習、線性/內核SVM、線性/非線性回歸、線性判別分析等。在一些情況下，分類器使用包含多個輸入-輸出對的標記的數據集進行訓練。例如，數據集是由來自多個被診斷為不患有NPC或患有NPC的受試者的樣品分析結果生成的。在一些情況下，數據集可以包括具有來自這些受試者的血漿EBV DNA特性的一個或多個因素(例如，變體模式、甲基化狀態、可檢測性/拷貝數或片段大小)、年齡、家族史、吸菸習慣、種族或飲食史的輸入以及指示相應受試者是否患有NPC的相應輸出。在示例性實例中，可以使用包括大量輸入-輸出對，例如至少10、20、50、100、200、500、1000、2000、5000、10000或20000對標記的數據集訓練分類器。

【0118】在一個實例中，提供了一種分類模型，以通過使用變體模式的分析來預測具有可檢測血漿EBV DNA的受試者未來發展NPC的風險。分類模型可以是使用支持向量機(SVM)算法按以下方式構造的分類器：

給定一個包含 n 個樣本的訓練數據集：

$$(M_1, Y_1), \dots, (M_n, Y_n)$$

其中 Y_i 表示樣本 i 的NPC狀態。對於來自NPC患者的樣本， Y_i 為1；對於不具有NPC的受試者的樣本， Y_i 為-1。 M_i 是包含樣本 i 的病毒變體模式的 p 維向量。例如， M_i 可以是一系列的變體位點(例如，如表6所示的與NPC相關的29個變體位點或與NPC相關的661個變體位點)。或者， M_i 可以是相對存在於已知具有NPC的受試者中的參考EBV變體的一系列基於塊的變體

相似性得分(例如，500 bp的非重疊窗口)。

【0119】 通過尋找滿足以下條件的一組係數(帶有 p 維向量的 W)，可以識別出一個“超平面”，該超平面在訓練數據集中將非NPC組和NPC組盡可能準確地分開：

標準1：

$$W \cdot M_i - b \geq 1 \quad (\text{對於NPC組中的任何受試者})$$

和

標準2

$$W \cdot M_i - b \leq -1 \quad (\text{對於非NPC組中的任何受試者})$$

其中 W 是確定超平面的係數的 p 維向量； M 是具有 p 個變體(或基於塊的相似性評分)和 n 個樣本的矩陣($p \times n$ 維)； b 是截距。

【0120】 這兩個標準(即標準1和標準2)也可以寫成：

$$Y_i(W \cdot M_i - b) \geq 1 \quad (\text{標準3})$$

其中 Y_i 為-1(非NPC)或1(NPC)。

【0121】 標準1和標準2之間的邊距(D)為： $\frac{2}{\|W\|}$ ，

其中 $\|W\|$ 利用點到平面方程的距離來計算。

【0122】 通過根據準則3最小化 $\|W\|$ 而將 D 最大化。

【0123】 基於該原理，可以確定分類器的參數(W 和 b)。因此，可以使用用訓練的參數(W 和 b)實施訓練的分類器來計算測試樣本的NPC風險評分。

【0124】 在一個示例性實例中，NPC風險評分被計算為整個病毒基因組中一組固定的SNV位點的EBV基因型的加權總和(作為二元邏輯回歸模型中的解釋性變量)。在該實例中，通過分析來自訓練集中NPC樣本和

非NPC樣本的EBV SNV譜的差異來識別一組與NPC相關的SNV。可以例如使用Fisher精確檢驗來分析跨越EBV基因組的每個變體與NPC病例的關聯。然後可以獲得固定的一組顯著的SNV，例如，假髮現率(FDR)控制在5%。測試樣本的NPC風險評分可以通過該特定的一組顯著的SNV位點的EBV基因型來確定，該SNV位點是從訓練集中確定的，該訓練集包括來自已知NPC受試者和非NPC受試者的血漿DNA樣本的測序數據。在一些情況下，血漿EBV DNA分子的濃度可能較低，因此測序的EBV DNA讀數可能無法完全覆蓋整個EBV基因組。可以將分數制定為由血漿EBV DNA讀數(例如，具有可用的基因型信息)覆蓋的那些SNV位點上的基因型模式確定。為了獲得NPC風險評分，可以首先確定樣本中血漿EBV DNA讀數覆蓋的顯著的SNV位點的子集，然後可以在顯著的SNV位點的子集中確定每個位點上基因型的權重(效應大小)。可以構建如下的邏輯回歸模型來報告NPC上每個SNV位點的風險基因型的效應大小：

$$P = \frac{1}{1 + e^{-(\beta_0 + \sum_{k=1}^n \beta_k X_k)}}$$

可以改寫為：

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{k=1}^n \beta_k X_k，$$

其中n是顯著的SNV位點數目； β_0 和 β_k 是可以由最大似然估計值確定的係數；P是EBV陽性患者患有NPC的概率；變量 X_k 代表基因組位置k上的SNV位點。如果樣本中存在與EBV參考基因組相同的變體，則可將 X_k 編碼為-1。如果樣本中存在替代變體，則可以將 X_k 編碼為1。如果分析的變體位點未包含在樣本中，則可以將 X_k 編碼為0。由此可以例如使用python中的“邏輯回歸”函數來估計係數 β_0 和 β_k 。這可以通過分析訓練數據集中NPC樣本和非NPC樣本中每個位點的基因型模式來實現。因此，可以基於

測試樣本自身在SNV位點的基因型並由從訓練模型中推理的相應的係數 β_0 和 β_k 加權來推導測試樣本的NPC風險評分。

生物樣品

【0125】 在本文提供的方法中使用的生物樣品可以包括衍生自存活的或死亡的受試者的任何組織或材料。生物樣品可以是無細胞樣品。生物樣品通常包含核酸(例如，DNA或RNA)或其片段。樣品中的核酸可以是無細胞核酸。樣品可以是液體樣品或固體樣品(例如細胞或組織樣品)。生物樣品可以是體液，如血液、血漿、血清、尿液、口腔沖洗液、鼻腔沖洗液、鼻刷樣品、陰道分泌物、來自鞘膜積液(例如，睪丸的鞘膜積液)的液體、陰道沖洗液、胸膜液、腹水液、腦脊髓液、唾液、汗液、淚液、痰液、支氣管肺泡灌洗液、乳頭排出液、來自身體不同部位(例如，甲狀腺、乳房)的抽吸液等。還可使用糞便樣品。在各個實例中，已經富集無細胞DNA的生物樣品(例如，經由離心方案獲得的血漿樣品)中的大部分DNA可以是無細胞的(例如，大於50%、60%、70%、80%、90%、95%或99%的DNA可以是無細胞的)。生物樣品可以經處理以物理破壞組織或細胞結構(例如，離心和/或細胞裂解)，從而將細胞內組分釋放到溶液中，該溶液可進一步含有用於製備樣品以供分析的酶、緩衝液、鹽、洗滌劑等。

【0126】 本文提供的方法和系統可以用於分析生物樣品中的核酸分子。核酸分子可以是細胞核酸分子、無細胞核酸分子或兩者。在本文提供的方法中使用的無細胞核酸可以是生物樣品中的細胞外的核酸分子。無細胞核酸分子可以存在於各種體液，例如血液、唾液、精液和尿液中。由於各種組織中的細胞死亡可以產生無細胞的DNA分子，這些死亡可能是由健康狀況和/或疾病，例如病毒感染和腫瘤生長引起的。無細胞核酸分子

可以包括由於病原體集成事件而產生的序列。

【0127】 在本文提供的方法中使用的無細胞核酸分子，例如無細胞DNA可以存在於血漿、尿液、唾液或血清中。無細胞的DNA可以以短片段的形式自然存在。無細胞DNA片段化可以指當無細胞DNA分子產生或釋放時，高分子量DNA(例如細胞核中的DNA)被切割、斷裂或消化成短片段的過程。在一些情況下，本文提供的方法和系統可以用於分析細胞核酸分子，例如，當患者患有白血病、淋巴瘤或骨髓瘤時，來自腫瘤組織的細胞DNA或來自白細胞的細胞DNA。根據本公開內容的一些實例，可以對從腫瘤組織獲取的樣品進行測定和分析。

受試者

【0128】 本文提供的方法和系統可以用於分析來自受試者例如生物體例如宿主生物的樣品。受試者可以是任何人類患者，例如癌症患者，有患癌症風險的患者或具有家族或個人癌症史的患者。在一些情況下，受試者處於癌症治療的特定階段。在一些情況下，受試者可能患有或疑似患有癌症。在一些情況下，受試者是否患有癌症是未知的。

【0129】 在一些情況下，受試者接受或不接受病原體相關疾病的藥物治療取決於本文提供的篩查測定的結果。在一個實例中，儘管第一次篩查測定顯示陽性結果表明受試者發展病原體相關疾病的高風險，但是通過隨訪診斷檢查被診斷為受試者沒有病原體相關疾病(例如，EBV相關NPC)。在這種情況下，受試者不接受醫學治療，例如但不限於用治療劑(例如化療)、放療、外科手術或其任意組合的治療。在另一個實例中，受試者被篩查為具有發展病原體相關疾病(例如，HPV相關宮頸癌)的高風險，並進一步診斷為患有該疾病。因此，受試者可以接受對該疾病的醫學

治療，例如但不限於外科手術、化療、放療、靶向療法、免疫療法或其任意組合。

【0130】 本文提供的方法和系統可以適用的病原體相關疾病可以包括增生性疾病，例如癌症。該疾病可以與病原體例如病毒、細菌或真菌相關或由病原體例如病毒、細菌或真菌引起。可以與本文所述疾病相關的病毒可以包括EBV、卡波西肉瘤相關疱疹病毒(KSHV)、HPV(例如但不限於HPV16、HPV18、HPV31、HPV33、HPV34、HPV35、HPV39、HPV45、HPV51、HPV52、HPV56、HPV58、HPV59、HPV66、HPV68和HPV70)(Burd等人，*Clin Microbiol Rev* 2003:16:1-17)、梅克爾細胞多瘤病毒(MCPV)、HBV、HCV和人T淋巴營養病毒1(HTLV1)。適用的病原體相關癌症可包括可能與EBV相關的伯基特淋巴瘤、霍奇金淋巴瘤、免疫抑制相關性淋巴瘤、T細胞和NK細胞淋巴瘤、鼻咽癌或胃癌。適用的與病原體相關的癌症可以包括可能與KSHV相關的原發性滲出性淋巴瘤或卡波西肉瘤。適用的病原體相關癌症可以包括可能與HPV相關的宮頸癌、頭頸癌或生殖道癌。適用的病原體相關癌症可以包括可能與MCPV相關的梅克爾細胞瘤。適用的病原體相關癌症可以包括可能與HBV或丙型肝炎病毒(HCV)相關的HCC。適用的病原體相關癌症可以包括可能與HTLV1相關的成人T細胞白血病/淋巴瘤。

【0131】 受試者可以患有任何類型的癌症或腫瘤，或具有發展為任何類型的癌症或腫瘤的風險。在一個實例中，受試者可以患有鼻咽癌或鼻腔癌。在另一實例中，受試者可以患有口咽癌或口腔癌。癌症的非限制性實例可包括但不限於腎上腺癌、肛門癌、基底細胞癌、膽管癌、膀胱癌、血液癌症、骨癌、腦腫瘤、乳腺癌、支氣管癌、心血管系統癌症、宮頸

癌、結腸癌、結直腸癌、消化系統癌症、內分泌系統癌症、子宮內膜癌、食管癌、眼癌、膽囊癌、胃腸道腫瘤、肝細胞癌、腎癌、造血系統惡性腫瘤、喉癌、白血病、肝癌、肺癌、淋巴瘤、黑素瘤、間皮瘤、肌肉系統癌症、骨髓增生異常綜合症(MDS)、骨髓瘤、鼻腔癌、鼻咽癌、神經系統癌症、淋巴系統癌、口腔癌、口咽癌、骨肉瘤、卵巢癌、胰腺癌、陰莖癌、垂體瘤、前列腺癌、直腸癌、腎盂癌、生殖系統癌症、呼吸系統癌症、肉瘤、唾液腺癌、骨胳系統癌、皮膚癌、小腸癌、胃癌、睪丸癌、喉癌、胸腺癌、甲狀腺癌、腫瘤、泌尿系統癌症、子宮癌、陰道癌或外陰癌。淋巴瘤可以是任何類型的淋巴瘤，包括B細胞淋巴瘤(例如，彌漫性大B細胞淋巴瘤、濾泡性淋巴瘤、小淋巴細胞淋巴瘤、套細胞淋巴瘤、邊緣區B細胞淋巴瘤、伯基特淋巴瘤、淋巴漿細胞淋巴瘤、多毛細胞白血病或原發性中樞神經系統淋巴瘤)或T細胞淋巴瘤(例如，前體T淋巴母細胞淋巴瘤或外周T細胞淋巴瘤)。白血病可以是任何類型的白血病，包括急性白血病或慢性白血病。白血病的類型包括急性髓性白血病、慢性髓性白血病、急性淋巴細胞白血病、急性未分化白血病或慢性淋巴細胞白血病。在一些情況下，癌症患者未患有特定類型的癌症。例如，在一些情況下，患者可能患有非乳腺癌的癌症。

【0132】 癌症的實例可包括引起實體瘤的癌症以及不引起實體瘤的癌症。此外，本文提及的任何癌症可以是原發性癌症(例如，以其首先開始生長的身體部位命名的癌症)或者繼發性或轉移性癌症(例如，源自身體的另一部位的癌症)。

【0133】 通過本文所述的任何方法診斷的受試者可以是任何年齡，並且可以是成人、嬰兒或兒童。在一些情況下，受試者是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、51、52、53、54、55、56、57、58、59、60、61、62、63、64、65、66、67、68、69、70、71、72、73、74、75、76、77、78、79、80、81、82、83、84、85、86、87、88、89、90、91、92、93、94、95、96、97、98或99歲，或在年齡在一定範圍內(例如2歲與20歲之間、20歲與40歲之間或者40與90歲之間)。可以受益的特定類別的患者可以是40歲以上的患者。可以受益的另一特定類別的患者可以是小兒患者。此外，通過本文所述的任何方法或組合物診斷的受試者可以是男性或女性。

【0134】 在一些實施方案中，本公開內容的方法可以檢測受試者中的腫瘤或癌症，其中所述腫瘤或癌症具有疾病的地理模式。在一個實例中，受試者可能患有與EBV相關的癌症(例如，鼻咽癌)，其在華南地區(例如，中國香港特別行政區)流行。在另一個實例中，受試者可以患有HPV相關的癌症(例如口咽癌)，其可以在美國和西歐流行。在又一個實例中，受試者可能患有HTLV-1相關的癌症(例如，成人T細胞白血病/淋巴瘤)，其在日本南部、加勒比海、非洲中部、南美部分地區以及在美國東南部的某些地區的的移民群體中流行。

【0135】 本文公開的任何方法也可以在非人類受試者上進行，例如實驗室或農場動物，或衍生自本文公開的生物的細胞樣品。非人類受試者的非限制性實例包括狗、山羊、豚鼠、倉鼠、小鼠、豬、非人靈長類動物(例如大猩猩、猿、猩猩、狐猴或狒狒)、老鼠、綿羊、母牛或斑馬魚。

計算機系統

【0136】 本文公開的任何方法可以由一個或多個計算機系統執行和/或控制。在一些實例中，本文公開的方法的任何步驟可以由一個或多個計算機系統全部、單獨地或順序地執行和/或控制。本文提及的任何計算機系統都可以利用任何合適數目的子系統。在一些實施方案中，計算機系統包括單個計算機設備，其中子系統可以是計算機設備的組件。在其他實施方案中，計算機系統可以包括多個計算機設備，每個計算機設備是具有內部組件的子系統。計算機系統可以包括臺式計算機和膝上型計算機、平板計算機、移動電話和其他移動裝置。

【0137】 子系統可以經由系統總線相互連接。其他子系統包括打印機、鍵盤、存儲裝置和耦合到顯示适配器的監視器。耦合到I/O控制器的外圍設備和輸入/輸出(I/O)設備可以通過本領域已知的任何數目的連接機構(例如輸入/輸出(I/O)端口(例如，USB、FireWire®)。例如，可以使用I/O端口或外部接口(例如，以太網、Wi-Fi等)將計算機系統連接到廣域網，例如因特網、鼠標輸入設備或掃描儀。經由系統總線的互連允許中央處理器與每個子系統進行通信，並控制來自系統內存或存儲設備(例如，諸如硬盤或光盤之類的固定磁盤)的多個指令的執行)，以及子系統之間的信息交換。系統存儲器和/或存儲設備可以體現計算機可讀介質。另一個子系統是數據收集設備，例如照相機、麥克風、加速計等。本文提到的任何數據都可以從一個組件輸出到另一組件，並可以輸出給用戶。

【0138】 計算機系統可以包括多個相同的組件或子系統，例如，通過外部接口或內部接口連接在一起。在一些實施方案中，計算機系統、子系統或裝置可以通過網絡進行通信。在這樣的情況下，一台計算機可以被視為客戶端，另一台計算機可以被視為服務器，其中每台計算機都可以被

視為同一計算機系統的一部分。客戶端和服務器可以各自包含多個系統、子系統或組件。

【0139】本公開內容提供了被編程用於實現本公開內容的用於對病原體相關疾病進行分級的方法的計算機控制系統。圖21示出了計算機系統1101，該計算機系統1101被編程或以其他方式配置為分析無細胞核酸分子或其序列讀數、分析與該疾病風險相關的其他因素、評估風險或生成指示如本文所述風險的報告。計算機系統1101可以實現和/或調節本公開內容中提供的方法的各個方面，例如，控制來自生物樣品的核酸分子的測序，執行如本文所述的測序數據的生物信息學分析的各個步驟，集成數據收集、分析和結果報告以及數據管理。計算機系統1101可以在用戶的電子裝置上或者是相對於該電子裝置遠程定位的計算機系統。電子裝置可以是移動電子裝置。

【0140】計算機系統1101包括中央處理單元(CPU，本文也稱為“處理器”和“計算機處理器”)1105，其可以是單核或多核處理器，或者是用於並行處理的多個處理器。計算機系統1101還包括存儲器或存儲位置1110(例如，隨機存取存儲器、只讀存儲器、閃存)、電子存儲單元1115(例如，硬盤)、用於與一個或多個其他系統通信的通信接口1120(例如，網絡適配器)、外圍裝置1125，諸如高速緩存存儲器、其他存儲器、數據存儲和/或電子顯示適配器。存儲器1110、存儲單元1115、接口1120和外圍裝置1125通過諸如主板的通信總線(實線)與CPU 1105通信。存儲單元1115可以用於儲存數據的數據存儲單元(或數據存儲庫)。計算機系統1101可以借助於通信接口1120而可操作地耦合到計算機網絡(“網絡”)1130。網絡1130可以是互聯網、因特網和/或外聯網和/或與互聯網通

信的內聯網和/或外聯網。在一些情況下，網絡1130是電信網絡和/或數據網絡。網絡1130可以包括一個或多個計算機服務器，其能够分布式計算，諸如雲計算。在一些情況下，借助於計算機系統1101，網絡1130可以實現點對點網絡，該點對點網絡可以使耦合到計算機系統1101的裝置能够發揮客戶端或服務器的作用。

【0141】 CPU 1105可以執行一系列機器可讀指令，該指令可體現在程序或軟件中。指令可以儲存在存儲器位置，諸如存儲器1110中。指令可以指向CPU 1105，其可以在隨後編程或以其他方式配置CPU 1105以便實現本公開內容的方法。CPU 1105所執行的操作的實例可以包括提取、解碼、執行和回寫。

【0142】 CPU 1105可以是諸如集成電路等電路的一部分。在該電路中可以包含系統1101的一個或多個其他組件。在一些情況下，該電路是專用集成電路(ASIC)。

【0143】 存儲單元1115可以存儲文件，諸如驅動器、庫和保存的程序。存儲單元1115可以儲存用戶數據，例如，用戶偏好和用戶程序。在一些情況下，計算機系統1101可以包括在計算機系統1101外部的一個或多個其他數據存儲單元，諸如位於通過內聯網或互聯網與計算機系統1001通信的遠程服務器上。

【0144】 計算機系統1101可以通過網絡1130與一個或多個遠程計算機系統通信。例如，計算機系統1101可以與用戶的遠程計算機系統(例如，安裝有接收並顯示從計算機系統1101發送的樣本分析結果的應用程序的智能電話)進行通信。遠程計算機系統的實例包括個人計算機(例如，便攜式PC)、平板或平板計算機(例如，Apple® iPad、Samsung® Galaxy

Tab)、電話、智能電話(例如，Apple® iPhone、支持Android的裝置、Blackberry®)或個人數字助理。用戶可以經由網絡1130訪問計算機系統1101。

【0145】 本文所述的方法可通過存儲在計算機系統1101的電子存儲位置上(例如，存儲器1110或電子存儲單元1115上)的機器(例如，計算機處理器)可執行代碼實施。該機器可執行或機器可讀代碼能夠以軟件的形式提供。在使用期間，代碼可由處理器1105執行。在一些情況下，可以從存儲單元1115檢索代碼並將其儲存在存儲器1110上，以便於處理器1105訪問。在一些情況下，可以取消電子存儲單元1115，並且在存儲器1110上儲存機器可執行指令。

【0146】 代碼可以預編譯和配置用於與具有適於執行該代碼的處理器的機器一起使用，或者可以在運行時期間編譯。代碼可以以可選擇的編程語言提供以使代碼能夠以預編譯或編譯的方式執行。

【0147】 本文提供的系統和方法的方面，如計算機系統1101，可以在編程中體現。技術的各個方面可被認為是“產品”或“製品”，其形式通常為承載或體現於某種類型的機器可讀介質上的機器(或處理器)可執行代碼和/或關聯數據。機器可執行代碼可以儲存在電子存儲單元上，諸如存儲器(例如，只讀存儲器、隨機存取存儲器、閃存)上，或者硬盤上。“存儲”型介質可以包括計算機、處理器等的任何或所有有形存儲器或者其關聯的模塊，諸如可以在任何時刻為軟件編程提供非暫時性存儲的各種半導體存儲器、磁帶驅動器、磁盤驅動器等。整個軟件或部分軟件可以在任何時刻通過因特網或各種其他電信網絡進行通信。這樣的通信例如能夠將軟件從一台計算機或處理器加載到另一台計算機或處理器，例如，從管理服務器

或主機加載到應用服務器的計算機平臺。因此，可以承載軟件元素的另一類型的介質包括光波、電波和電磁波，諸如跨本地設備之間的物理接口、通過有線和光陸線網絡以及通過各種空中鏈路而使用的光波、電波和電磁波。攜載這樣的波的物理元件，諸如有線或無線鏈路、光學鏈路等，也可被認為是承載軟件的介質。如本文所使用，除非限制於非暫時性、有形“存儲”介質，否則諸如計算機或機器“可讀介質”的術語是指參與向處理器提供指令以供執行的任何介質。

【0148】 因此，諸如計算機可執行代碼等機器可讀介質可以採取許多形式，包括但不限於有形存儲介質、載波介質或物理傳輸介質。非易失性存儲介質例如包括光盤或磁盤，諸如任何計算機等中的任何存儲設備，例如其可以用於實現附圖中所示的文庫等。易失性存儲介質包括動態存儲器，諸如這樣的計算機平臺的主存儲器。有形傳輸介質包括同軸線纜；銅線和光纖，包括在計算機系統內包含總線的電線。載波傳輸介質可以採取電信號或電磁信號的形式，或者採取聲波或光波的形式，諸如在射頻(RF)和紅外(IR)數據通信期間生成的那些。因此，計算機可讀介質的常見形式例如包括：軟盤、柔性盤、硬盤、磁帶、任何其他磁介質、**CD-ROM**、**DVD**或**DVD-ROM**、任何其他光介質、穿孔卡片紙帶、任何其他具有孔洞圖案的物理存儲介質、**RAM**、**ROM**、**PROM**和**EPROM**、**FLASH-EPROM**、任何其他存儲器芯片或存儲器盒、傳輸數據或指令的載波、傳輸這樣的載波的線纜或鏈路，或者任何其他計算機可從中讀取編程代碼和/或數據的介質。這些形式的計算機可讀介質中的許多介質可以參與將一個或多個指令的一個或多個序列加載到處理器以供執行。

【0149】 計算機系統1101可以包括電子顯示器1135或與電子顯示器

1135通信，該電子顯示器1135包括用於提供例如樣本分析的結果，例如但不限於示出病原體集成概況、病原體集成斷點的基因組定位、病理學分類(例如，疾病或癌症的類型和癌症水平)以及基於病理學分類的治療建議或預防步驟的推薦的圖形的用戶界面(UI)1140。UI的實例包括但不限於：圖形用戶界面(GUI)和基於網絡的用戶界面。

【0150】可以通過一種或多種算法來實現本公開內容的方法和系統。算法可以通過在由中央處理器1105執行時的軟件而實現。該算法可以例如控制來自樣品的核酸分子的測序、測序數據的直接收集、分析測序數據、執行基於塊的變體模式分析、評估風險或生成指示該風險的報告。

【0151】在一些情況下，如圖22所示，可以從諸如人類受試者的受試者1201獲得樣本1202。樣本1202可以經受本文所述的一種或多種方法，例如執行測定。在一些情況下，測定可以包括雜交、擴增、測序、標記、表觀遺傳修飾鹼基或其任意組合。方法的一個或多個結果可以輸入到處理器1204中。諸如樣本識別、受試者識別、樣本類型、參考或其他信息的一個或多個輸入參數可以被輸入到處理器1204中。可以將來自測定的一個或多個度量輸入到處理器1204中，以使處理器可以產生結果，例如病理學的分類(例如，診斷)或對治療的推薦。處理器可以將結果、輸入參數、度量、參考或其任意組合發送到顯示器1205，例如視覺顯示器或圖形用戶界面。處理器1204可以(i)向服務器1207發送結果、輸入參數、度量或其任意組合，(ii)從服務器1207接收結果、輸入參數、度量或其任意組合，(iii)或其組合。

【0152】本公開內容的各方面可以使用硬件(例如，專用集成電路或現場可編程門陣列)和/或使用具有一般可編程處理器的計算機軟件以模塊

化或集成方式以控制邏輯的形式來實現。如本文所用的，處理器包括單核處理器、在同一集成芯片上的多核處理器或在單個電路板上或聯網的多個處理單元。基於本文所提供的公開內容和教導，本領域普通技術人員將知道並理解使用硬件以及硬件和軟件的組合來實現本文所述實施方案的其他方式和/或方法。

【0153】 可以使用諸如Java、C、C++、C#、Objective-C、Swift的任何適當的計算機語言，或例如利用常規的或面向對象的技術的諸如Perl或Python的腳本語言，將本申請中描述的任何軟件組件或功能實現為由處理器執行的軟件代碼。可以將軟件代碼作為一系列指令或命令存儲在計算機可讀介質上，以進行存儲和/或傳輸。合適的非暫時性計算機可讀介質可以包括隨機存取存儲器(RAM)、只讀存儲器(ROM)、諸如硬盤驅動器或軟盤之類的磁性介質或者諸如光盤(CD)之类的光學介質或DVD(數字通用磁盤)、閃存等。計算機可讀介質可以是這樣的存儲設備或傳輸設備的任意組合。

【0154】 也可以使用載波信號來編碼和發送這樣的程序，該載波信號適用於經由符合包括因特網的各種協議的有線、光學和/或無線網絡進行發送。這樣，可以使用用這樣的程序編碼的數據信號來創建計算機可讀介質。可以將用程序代碼編碼的計算機可讀介質與兼容設備打包在一起，或者與其他設備分開提供用程序代碼編碼的計算機可讀介質(例如，經由因特網下載)。任何此類計算機可讀介質可以駐留在單個計算機產品(例如，硬盤驅動器、CD或整個計算機系統)上或內部，並且可以存在於系統或網絡內的不同計算機產品上或內部。計算機系統可以包括用於向用戶提供本文提到的任何結果的監視器、打印機或其他合適的顯示器。

【0155】 可以利用包括一個或多個處理器的計算機系統完全或部分地執行本文所述的任何方法，所述計算機系統可以被配置為執行步驟。因此，實施方案可以針對被配置為執行本文所述的任何方法的步驟的計算機系統，其中不同的組件執行相應的步驟或相應的步驟組。儘管以編號的步驟呈現，但是本文方法的步驟可以同時或以不同順序執行。另外，這些步驟的一部分可以與其他方法的其他步驟的一部分一起使用。同樣，步驟的全部或部分可以是可選的。另外，任何方法的任何步驟都可以使用用於執行這些步驟的模塊、單元、電路或其他方法來執行。

其他實施方案

【0156】 本文使用的章節標題僅用於組織的目的，不應當解釋為限制所描述的主題。

【0157】 應當理解，本文描述的方法不限於本文描述的具體方法、方案、受試者和測序技術，而本身可以變化。還應當理解，本文所用的術語僅用於描述特定實施方案，並非意圖限制本文描述的方法和組合物的範圍，其僅受所附的申請專利範圍限制。雖然本文已經示出和描述了本公開內容的一些實施方案，但對於本領域技術人員顯而易見的是，這樣的實施方案只是以示例的方式提供。本領域技術人員現將在不偏離本公開內容的情況下想到許多變化、改變和替代。應當理解，本文中所述的本公開內容的實施方案的各種替代方案可用於實施本公開內容。以下申請專利範圍旨在限定本公開內容的範圍，並且由此覆蓋這些申請專利範圍及其等同項的範圍內的方法和結構。

【0158】 幾個方面參考用於說明的應用實例進行描述。除非另有說明，否則任何實施方案均可以與其他任何實施方案組合。應當理解，爲了

提供對本文所述特徵的完整理解，闡述了很多具體的細節、關係和方法。然而，技術人員將會容易地認識到：本文所述的特徵可以在沒有一個或多個這些具體細節的情況下或用其它方法來實施。本文所述特徵不受所述的動作或事件順序的限制，因為一些動作可以以不同順序發生和/或與其他動作或事件同時發生。此外，並非需要所有描述的行爲或事件來實施根據本文所述特徵的方法。

實施例

【0159】 提供以下實施例是爲了進一步說明本公開內容的一些實施方案，但並不意於限制本公開內容的範圍；通過其示例性將使人們理解，可以可替代地使用本領域技術人員已知的其他程序、方法或技術。

實施例1.四年內逾二萬名受試者隊列的NPC篩查

【0160】 該實施例描述了約4年內對二萬名受試者隊列進行的大規模篩查研究。圖1圖示了本研究的設計示意圖。在第一輪篩查中，使用血漿EBV DNA分析，對逾二萬名40至62歲之間的男性進行了NPC篩查。具有可檢測的血漿EBV DNA的受試者於中位4周後重新測試了第二組血液樣本。這種安排的目的是爲了區分NPC患者與那些沒有患有NPC、但具有可檢測的血漿EBV DNA的患者。在先前的一項研究中，沒有患有NPC的受試者中血漿EBV DNA的存在是一種典型的一過性現象。在三分之二的這些個體中，在中位兩周後將檢測不到受試者血漿EBV DNA。對血漿EBV DNA持續性陽性的受試者進行鼻內窺鏡檢查和鼻咽磁共振成像(MRI)檢查，以確定或排除NPC的存在。根據這種安排，識別了34例NPC。

【0161】 之後，在第一輪篩查後的中位4年，對該隊列進行另一輪(第二輪)NPC篩查。在第二輪NPC篩查中，與第一輪篩查一樣，約4周後

對檢測結果為陽性的受試者進行重新檢測。對超過4周連續兩次檢查均具有陽性結果的受試者將經由鼻內窺鏡和磁共振成像進行進一步檢查。第二輪篩查於2017年開始。截至2018年9月15日，共有8335名受試者完成了第二輪篩查。784例(9.4%)受試者的血漿EBV DNA呈陽性。在四周的重新檢測中，230名受試者(2.7%)仍具有可檢測的血漿EBV DNA。表1總結了兩輪NPC篩查的檢測結果。

表1 第一、二輪NPC篩查中血漿EBV DNA的狀況表

第一輪篩查中 血漿EBV DNA狀況	數量	第二輪篩查中血漿EBV DNA狀況		
		陰性	一過性陽性	持續性陽性
陰性	7907	7267 (92%)	479 (6%)	161 (2%)
一過性陽性	276	218 (79%)	30 (11%)	28 (10%)
持續性陽性	152	66 (43%)	48 (32%)	38 (25%)

【0162】 如表1所示，第二輪NPC篩查中具有可檢測的血漿EBV DNA的概率與第一輪篩查中血漿EBV DNA的狀態相關。第一輪篩查中，血漿EBV DNA呈陰性、一過性陽性和持續性陽性的受試者在第二輪篩查的初步分析中，有8%、21%和57%的概率具有可檢測的血漿EBV DNA。此外，4周後，三組患者血漿具有持續性陽性的EBV DNA的概率從2%逐漸增加到25%。

【0163】 經由本文所述篩查確定的NPC患者比未接受NPC篩查的歷史隊列患者具有更早得多的分期分布。早期疾病(I期和II期)的比例分別為70%和20%。這種分期分布的改變導致危險比為0.1的患者的無進展生存期得到顯著改善。表2總結了第一輪和第二輪篩查中NPC病例的分期分布。在第二輪篩查8335名受試者後，識別了13例新的NPC病例。在第一輪和第二輪篩查中，患有早期疾病的患者比例分別為71%和69%。早期疾

病患者的比例沒有顯著性差異(P=0.93，卡方檢驗)。

表2 兩輪篩查中NPC病例的分期分布表

分期	第一輪篩查	第二輪篩查
I	16 (47%)	4 (31%)
II	8 (24%)	5 (38%)
III	8 (24%)	4 (31%)
IV	2 (6%)	0 (0%)

【0164】 如表3所總結，與在第一輪中具有不可檢測的血漿EBV DNA的受試者相比，第一輪篩查中具有一過性地和持續性可檢測的血漿EBV DNA的受試者在第一輪篩查後4年進行的第二輪篩查中檢測到患有NPC的風險更高。這兩組的相對風險值分別為7.2和19.7。

表3 按第一輪EBV DNA狀態分類的第二輪篩查NPC病例數

第一輪篩查中血漿EBV DNA的狀態	數量	第二輪檢測到的NPC數量 (具有相同的血漿EBV DNA狀態的受試者的百分比)	與第一輪中具有不可檢測的血漿EBV DNA的受試者相比NPC的相對風險
陰性	7907	8 (0.10%)	1
一過性陽性	276	2 (0.72%)	7.2
持續性陽性	152	3 (1.97%)	19.7

【0165】 這些結果表明，血漿EBV DNA分析不僅有助於篩查患有NPC的當前狀態，而且有助於預測未來患有臨床上可觀察到的NPC的風險。該發現的一個實際應用是，基於較早情況的篩查受試者的血漿EBV DNA狀態，定制重複篩查的間隔。例如，與具有不可檢測的血漿EBV DNA的受試者相比，具有在基線時可檢測的血漿EBV DNA但未識別出NPC的受試者可在較短的間隔後重新篩查。同樣作為說明，對於具有不可

檢測的、一過性可檢測和持續性可檢測的血漿EBV DNA的、受試者，重複篩查的間隔分別為4年、2年和1年。

實施例2.基於血漿EBV DNA可檢測性的NPC篩查

【0166】 本實施例描述了一種基於受試者血漿中EBV DNA的可檢測性為受試者設計的NPC篩查方案。圖2示出如本文所述方案的示意圖。

【0167】 根據該方案，在早期篩查中具有不可檢測的血漿EBV DNA的受試者在4年後重新篩查，因為在接下來的4年中具有不可檢測的EBV DNA的受試者患有NPC的風險相對較低。如果隨後的篩查對血漿EBV DNA呈陰性，則隨後的篩查間隔為4年。然而，當受試者在一次篩查中有可檢測到的EBV DNA，但沒有檢測到NPC時，下一次篩查安排在一年後。當血漿EBV DNA在4年內保持陰性時，篩查間隔恢復到4年。具體篩查項目的實際時間間隔還根據健康經濟考慮因素(如篩查成本)、受試者偏好(如更頻繁的篩查間隔可能對某些受試者的生活方式造成更大的干擾)和其他臨床參數(如個體基因型，NPC家族史、飲食史、種族(如廣東人))。

實施例3.無細胞EBV DNA分子的變體模式分析

【0168】 在本實施例中，使用捕獲富集的靶向測序法分析NPC受試者、具有可檢測的血漿EBV DNA的非NPC受試者和前NPC受試者循環中的無細胞病毒DNA分子(詳見下一節)。捕獲探針被設計成用於覆蓋整個EBV基因組。在同一分析中，還包括靶向~3000個人類常見單核苷酸多態性(SNP)位點和人類白細胞抗原(HLA)SNP的探針。

【0169】 在該實施例中，對13名NPC患者和16名具有可檢測的血漿EBV DNA的非NPC受試者的血漿EBV DNA進行分析。這13名NPC患者

均有症狀，並從威爾士親王醫院臨床腫瘤科或耳鼻咽喉科招募。16名非NPC受試者均來自如實施例1所述的20000多名受試者的NPC篩查隊列。

【0170】在該分析中，使用通過專門設計的捕獲探針進行捕獲富集的靶向測序。對於要分析的每個血漿樣本，使用QIAamp循環核酸試劑盒從4ml血漿中提取DNA。對於每一例，所有提取的DNA都被用於使用TruSeq納米DNA文庫製備試劑盒(Illumina)製備序列文庫。與特定分子識別(UMI)序列(xGen雙索引UMI適配器，集成DNA技術)相結合的雙索引系統進行條形碼編碼。使用TruSeq納米試劑盒(Illumina)對連接適配器的樣本進行了八個周期的PCR擴增。然後使用myBait定制捕獲面板系統(Arbor Biosciences)，借由定制設計的探針捕獲擴增產物，該探針覆蓋上述病毒和人類基因組區域。靶捕獲後，用14個PCR周期對捕獲的產物進行富集，產生DNA文庫。DNA文庫在NextSeq平臺(Illumina)上測序。對於每次測序運行，使用配對末端模式對10個具有唯一樣本條形碼的樣本進行測序。每個DNA片段將從兩端的每一端分別測序71個核苷酸。測序後，將序列讀數映射到一個人工組合的參考序列，該參考序列由整個人類基因組(hg19)、整個EBV基因組(GenBank:AJ507799.2)、整個HBV基因組和整個HPV基因組組成。使用SOAP2(Bioinformatics 2009；25:1966-7)進行比對，允許在插入大小不超過600 bp的正確方向上每次讀取最多2個錯配。映射到組合基因組序列中特定位置的測序讀數將用於下游分析。具有相同的特定分子標識符的所有重複片段都將被過濾。

【0171】根據比對結果，確定了測序讀數與EBV參考基因組(GenBank: AJ507799.2)之間的核苷酸差異，包括但不限於單核苷酸變體(SNV)。在來自13名NPC受試者、16名具有可檢測的血漿EBV DNA的非

NPC受試者和4名前NPC受試者的44份樣本中，識別了1116個SNV(四分位間距(IQR)：902 - 1216)。在這些血漿樣本中，在EBV基因組的某些核苷酸位置觀察到兩個不同的等位基因。這種觀察可能是由於測序錯誤或腫瘤異質性的存在。在血漿EBV DNA中只有26個位置(IQR: 20 - 35)的中位數具有一個以上的等位基因。

【0172】在如圖3所示的系統樹分析中，NPC受試者被聚集在一起，並與非NPC受試者分離。這些結果表明NPC患者和非NPC者之間存在不同的EBV變體譜。因此，血漿EBV DNA的EBV變體譜分析可用於在篩查背景下區分NPC患者和非NPC受試者。三名非NPC受試者(AC106、AP080和FF159)進行了兩次連續採集的樣本分析，這些樣本間隔4周收集。來自同一個體的兩個樣本聚集在一起，表明它們具有非常相似的變體。

【0173】對同一組13名NPC患者和16名具有可檢測的血漿EBV DNA非NPC者進行的系統樹分析也基於EBV變體，但不包括Hui等人在(Hui 等人，*Int J Cancer* 2019, doi.org/10.1002/ijc.32049)研究中報告的29個變體。如圖4所示，NPC受試者也聚集在一起，並與非NPC受試者分開。

【0174】在第一輪篩查中血漿EBV DNA呈持續性陽性(如實施例1所述)、但在內窺鏡檢查和MRI上未檢測到NPC的四名受試者隨後被診斷為患有NPC。所有患者(BB096、DN054、FK015和HB121)在第一輪篩查後3年被診斷為NPC。在耳鼻喉科門診隨訪期間，所有患者在第一輪篩查後的第1年都額外採集了一份血漿樣本。對於這四個受試者中的每一個受試者，對第一輪篩查和一年後收集的兩個樣本進行EBV變體的分析。如圖5所示，將來自前NPC受試者的樣本與NPC樣本聚集在一起，表明與NPC相關的EBV變體在癌症實際發生之前就存在。這表明，具有與NPC相關的

EBV變體的個體未來發展NPC的風險更高。對同一組13名NPC患者、非NPC者與前NPC患者的系統樹分析也基於EBV變體，但不包括Hui等人在(Hui等人，*Int J Cancer* 2019, doi.org/10.1002/ijc.32049)研究中報告的29個變體。如圖6所示，來自前NPC受試者的樣本仍與NPC樣本聚集在一處，進一步表明EBV變體分析將能夠預測未來患有NPC的風險。

實施例4.基於塊的變體模式分析

【0175】 本實施例描述一種示例性的基於塊的變體模式分析方法的工作原理及其在實施例3中所述樣本中的EBV變體模式分析中的應用。

【0176】 圖7示出了基於塊的變體模式分析的原理。基於塊的分析用於評估從不同樣本的血漿EBV DNA測序中獲得的EBV DNA變體模式與參考基因組的相似性，此處的NPC測序數據可在公共文庫(Kwok等人，*J Virol* 2014;88:10662-72, Li等人，*Nat Comm* 2017;8:14121)中獲得。在基於塊的分析中，將EBV基因組分為大小為500bp的箱(共344個箱)，比較每個箱與參考集中的24個NPC樣本的變體模式的相似性。例如，如果在一個特定的箱內有8個變體位點，則分析測試樣本的該箱內這些位點上的等位基因，並與24個參考樣本的相同位點上的等位基因進行比較。根據與參考樣本具有完全相同的等位基因的比例得出相似性指數。例如，如果測試樣本在8個變體位點中有7個具有與一個參考樣本完全相同的等位基因，則該箱與該參考樣本的相似性指數將為7/8。與24個參考樣本相比，該測試樣本的箱子會有24個相似性指數。基於該箱的24個相似性指數，計算箱得分，該得分表示變體模式與參考樣本的總體相似性。例如，如果相似性指數的截止值設置為0.9，則箱得分視為指數高於截止值的箱的比例。因此，如果24個相似性指數中只有兩個高於0.9，則箱得分為2/24。箱得分

越高，測試樣本的變體模式與參考樣本集越相似。

【0177】圖8顯示了13個NPC、16個非NPC和4個前NPC樣本的EBV DNA變體模式的基於塊的分析。對4名前NPC受試者的每一名受試者，分析來自兩個時間點的樣本，因此共得到8個受試者樣本。在這些樣本中，得出EBV基因組的344個箱的箱得分。基於這些樣本的箱得分，進行無監督聚類分析。NPC樣本(黑色)聚集在一起，非NPC樣本(以點標記)聚集在一起。前NPC受試者的EBV變體譜與NPC受試者EBV變體譜聚集在一起。值得注意的是，這4名前NPC受試者的變體譜是通過對其基線樣本的分析獲得的，這些樣本是在NPC發展前幾年收集的。

【0178】圖9顯示了對於同一組13名NPC、16名非NPC和4名前NPC受試者進行的基於EBV變體(但不包括Hui等人在(Hui等人, *Int J Cancer* 2019, doi.org/10.1002/ijc.32049)研究中報告的29個變體)的基於塊的分析。同樣，觀察到NPC樣本(黑色)的聚類。此外，將前NPC受試者的EBV變體譜與NPC受試者的EBV變體譜聚集在一起。對前NPC樣本和NPC樣本的聚類分析表明，變體分析可以預測NPC的未來發展。總之，實施例3和實施例4中的數據顯示，那些在招募時沒有患有NPC但後來發展為癌症的受試者在基線血液樣本中具有與其他NPC患者相似的EBV變體模式。

實施例5.使用數學模型的NPC風險預測

【0179】該實施例描述了分類模型的構建，以使用變體模式的分析來預測具有可檢測的血漿EBV DNA的受試者未來發展NPC的風險，以及使用分類模型的測試結果。

【0180】如實施例4所述，使用支持向量機(SVM)算法用訓練數據集構建分類器，該訓練數據集包括18名未患有NPC的受試者和8名NPC患

者。該測試數據集由5名NPC患者、5名未患有NPC的受試者和8個從4名受試者中採集的樣本組成，所述受試者在樣本採集時在內窺鏡檢查和核磁共振檢查中不具有可檢測的NPC，但隨後被診斷為NPC(標記為前NPC)，如實施例4所述。

【0181】 SVM分析方法描述如下：

給定一個包含n個樣本的訓練數據集：

$$(M_1, Y_1), \dots, (M_n, Y_n)$$

其中， Y_i 表示樣本 i 的NPC狀態。對於來自NPC患者的樣本， Y_i 為1；對於不具有NPC的受試者的樣本， Y_i 為-1； M_i 是包含樣本 i 的病毒變體模式的 p 維向量。例如， M_i 可以是一系列變體位點，例如與NPC相關的29個變體位點。或者， M_i 可以是相對存在於已知具有NPC的受試者中的參考EBV變體的一系列基於塊的變體相似性得分(例如，500 bp的非重疊窗口)。

【0182】 通過尋找滿足以下條件的一組係數(帶 p 維向量的 W)，識別出一個“超平面”，該超平面在訓練數據集中將非NPC組和NPC組盡可能準確地分開：

標準1：

$$W \cdot M_i - b \geq 1 \quad (\text{對於NPC 組中的任何受試者})$$

及

標準2

$$W \cdot M_i - b \leq -1 \quad (\text{對於非NPC 組中的任何受試者})$$

其中， W 是確定超平面的係數的 p 維向量； M 是具有 p 個變體(或基於塊的相似性得分)和 n 個樣本的矩陣($p \times n$ 維)； b 是截距。

【0183】這兩個標準(即標準1和2)也可以寫成：

$$Y_i(W * M_i - b) \geq 1 \text{ (標準3)}$$

其中 Y_i 是-1(非NPC)或1(NPC)。

【0184】標準1和標準2之間的邊距(D)為： $\frac{2}{\|W\|}$ ，

其中， $\|W\|$ 利用點到平面方程的距離計算。

【0185】根據標準3通過最小化 $\|W\|$ 而最大化D。

【0186】基於該原理，確定了分類器的參數(W和b)。然後使用訓練參數(W和b)計算每個測試樣本的NPC風險評分。

【0187】圖10A示出了使用經訓練的分類器而計算出的NPC風險評分，所述分類器基於使用基於塊的變體分析對所有EBV變體的分析。對於該分析，EBV基因組被分成344個500bp的片段，用於計算實施例4中所述的箱得分。箱分數被認為是機器學習的一個特徵。NPC樣本的NPC風險評分顯著高於從非NPC受試者收集的樣本(平均NPC風險評分：0.15 vs 0.53，p值 < 0.01，Student t檢驗)的NPC風險評分。同樣，與未患有NPC的樣本的NPC風險評分(平均風險評分：0.58 vs 0.15，p值 < 0.01，Student t檢驗)相比，從前NPC受試者收集的樣本的NPC風險評分明顯更高。使用0.32的截止值，可以以100%的敏感性和100%的特異性將來自NPC患者和前NPC受試者的樣本與未患有NPC的樣本區分開。

【0188】圖10B顯示了使用經訓練的分類器而計算出的NPC風險評分，所述分類器基於對Hui等人在(Hui等人，*Int J Cancer* 2019, doi.org/10.1002/ijc.32049)研究中報告的29個EBV變體的分析。NPC樣本的NPC風險評分顯著高於從非NPC受試者收集的樣本的NPC風險評分(平均NPC風險評分：0.89 vs 0.18，p值 < 0.01，Student t檢驗)。同樣，與

未患有NPC的樣品的NPC風險評分(平均風險評分：0.57 vs 0.18， p 值 = 0.02，Student t 檢驗)相比，從前NPC受試者收集的樣本的NPC風險評分明顯更高。使用0.6的截止值，可以以74%的敏感性和100%的特異性將來自NPC患者和前NPC受試者的樣本與未患有NPC的樣品區分開。

【0189】圖10C顯示使用經訓練的分類器而計算出的NPC風險評分，該分類器基於使用基於塊的變體分析對所有EBV變體(但不包括Hui等人先前報告(Hui等人 *Int J Cancer* 2019. doi: 10.1002/ijc.32049)的與NPC相關的29個變體)的分析。NPC樣本的NPC風險評分顯著高於從非NPC受試者收集的樣本的NPC風險評分(平均NPC風險評分：0.58 vs 0.15， p 值 < 0.01，Student t 檢驗)。同樣，與未患有NPC的樣品的NPC風險評分(平均風險評分：0.53 vs 0.15， p 值 < 0.01，Student t 檢驗)相比，從前NPC受試者收集的樣本的NPC風險評分明顯更高。使用0.31的截止值，可以以100%的敏感性和100%的特異性將來自NPC患者的樣本和後來發展為NPC的患者的樣本與未患有NPC的樣本區分開。這些結果表明，從分析中排除先前報告的29種EBV變體不會對該分析的準確性產生不利影響。

實施例6. 通過亞硫酸氫鹽測序法分析血漿EBV DNA的甲基化狀態

【0190】本實施例說明了使用亞硫酸氫鹽測序法區分NPC患者和非NPC受試者，但該非NPC受試者根據血漿EBV DNA的甲基化狀態具有可檢測的血漿EBV DNA。

【0191】採用亞硫酸氫鹽測序法測定NPC患者和未患有NPC的受試者血漿中EBV DNA的甲基化水平。亞硫酸氫鹽轉化能將未甲基化的胞嘧啶轉化為尿嘧啶。甲基化的胞嘧啶不能被亞硫酸氫鹽改變，可以作為胞嘧啶保留。在測序過程中，尿嘧啶可以被確定為胸腺嘧啶。測序後，可以通

過檢查胞嘧啶是否已改變為胸腺嘧啶來確定胞嘧啶在任何CpG二核苷酸環境中的甲基化狀態。

【0192】 檢測了10名NPC患者和40名未患有癌症但在血漿中具有可檢測的EBV DNA的受試者(非NPC受試者)的血漿EBV DNA的甲基化水平。對於40名非NPC受試者，4周後從他們每個受試者中採集另一份血液樣本。其中20例血漿EBV DNA變為陰性，將它們標記為具有一過性陽性的血漿EBV DNA。其中20例血漿EBV DNA仍呈陽性，將它們標記為具有持續性陽性的血漿EBV DNA。

【0193】 如圖11所示，與具有一過性陽性的血漿EBV DNA的非癌症受試者($P < 0.01$ ，Student t檢驗)和具有持續性陽性的血漿EBV DNA的非癌症受試者($P < 0.01$ ，Student t檢驗)相比，NPC患者的EBV DNA甲基化水平明顯更高。這些結果表明，對血漿EBV DNA的甲基化的分析有助於將NPC患者和未患有NPC但具有可檢測的血漿EBV DNA的受試者區分開。

實施例7.用甲基化敏感的限制性酶分析血漿EBV DNA的甲基化狀態

【0194】 該實施例描述了一個計算機模擬實驗，該實驗證明了使用甲基化敏感的限制性酶來分析血漿EBV DNA用於區分NPC患者和未患有NPC但具有可檢測的血漿EBV DNA的受試者。

【0195】 用一名來自非NPC的受試者和一名NPC患者的樣本對血漿DNA進行亞硫酸氫鹽測序。分別獲得了兩個受試者的血漿DNA的347516個和6271012個EBV DNA片段。他們血漿EBV DNA的甲基化水平分別為48.9%和86.3%。據測定，大約一半的血漿EBV DNA分子含有至少一個CCGG基序。

【0196】爲了模擬對血漿EBV DNA的限制性酶消化，根據從亞硫酸氫鹽測序結果推斷出的“CCGG”序列背景下的甲基化狀態，對血漿EBV DNA分子進行計算機內消化。由此獲得了使用和不使用甲基化敏感的限制性酶HpaII進行計算機內消化的血漿EBV DNA的模擬大小分布，如圖14所示。不進行酶消化，非NPC受試者的血漿EBV DNA的大小分布在NPC受試者的左側，說明非NPC受試者的大小分布較短。在酶消化的大小分布圖中也觀察到片段大小的這種分布，因爲與不進行酶消化的非NPC受試者相比，在非NPC受試者中，低於50bp的短DNA的豐度顯著增加。對於NPC患者，對於使用和不使用酶消化的樣品，< 50bp的DNA分子的比例分別爲5.87%和0.84%。但是對於非NPC受試者，對於使用和不使用酶消化的樣品，< 50bp的DNA分子比例分別爲22.24%和4.99%。對於NPC患者和非NPC受試者而言，酶消化後< 50bp的DNA的比例分別增長了17.2%和5.0%。圖15示出了NPC患者和非NPC受試者在進行和不進行甲基化敏感的限制性酶消化下，血漿EBV DNA的累積大小分布。利用相對於大小的累積頻率曲線可以更容易地瞭解酶消化程度的差異。進行酶消化和不進行酶消化兩條曲線之間的距離反映了消化的程度。距離越大，酶對血漿EBV DNA的消化程度越大，因此表明血漿EBV DNA中的甲基化水平越低。如圖所示，與NPC患者相比，非NPC受試者的距離更大。NPC患者與非NPC受試者不進行酶消化和進行酶消化的曲線之間的最大距離分別爲8.1和18.3；NPC患者和非NPC受試者的兩條曲線之間的面積分別爲2395和942.9。

實施例8.無細胞EBV DNA分子的SNV譜分析

【0197】 在包含63名NPC和88名非NPC受試者的血漿DNA測序數據的訓練數據集中分析兩組患者的EBV SNV譜的差異。在整個EBV基因組中識別出SNV。由這些SNV位點的基因型模式得出了NPC風險評分，然後在31個NPC和40個非NPC樣本的測試集中對這些NPC風險評分進行分析。在該實施例中，從訓練集(圖16D)中識別出了整個EBV基因組的661個顯著的SNV。在測試集中，NPC血漿樣本顯示有較高的NPC風險分數；可能存在NPC相關的EBV SNV譜。在非NPC樣本中，NPC的風險評分範圍很廣。非NPC受試者可以有多種EBV SNV譜。

【0198】 材料和方法。

【0199】 研究對象和設計。

【0200】 該研究涉及先前在Lam等人*Proc Natl Acad Sci U S A.* 2018;115:E5115-E5124中報告的NPC和非NPC血漿樣本(作為訓練集)以及來自NPC和非NPC受試者的最新測序的血漿DNA樣本(作為測試集)的測序數據集子集的分析。

【0201】 訓練數據集包括來自在Lam等人*Proc Natl Acad Sci U S A.* 2018;115:E5115-E5124所述的先前的前瞻性NPC篩查研究中的篩查測定的NPC患者和非NPC受試者的血漿樣本。這些非NPC受試者通過基於實時PCR檢測法具有可檢測水平的血漿EBV DNA。該數據集還包括來自獨立隊列的有症狀的NPC患者的樣本。研究了來自所有樣本的EBV分離株的EBV基因型信息用於建立NPC風險評分預測的訓練模型。在該研究中，對另外31名有症狀的NPC患者和40名非NPC患者的血漿樣本進行靶向捕獲測序，以作為測試集。這31名有症狀的NPC患者是從香港威爾斯親王醫院臨床腫瘤科招募的。非NPC受試者也來自先前提到的NPC篩查隊列(包括

20000多名受試者)，並從中隨機選擇。分析了來自這些NPC和非NPC樣本的EBV基因型的變體，並基於訓練模型得出其NPC風險評分。訓練集和測試集中的所有NPC樣本和非NPC樣本均沒有重疊。

【0202】 靶向捕獲測序。

【0203】 通過捕獲探針系統(myBaits 定制捕獲面板， Arbor Biosciences)富集血漿DNA文庫中的EBV DNA分子，對血漿樣本進行靶向捕獲測序。EBV捕獲探針設計成用於覆蓋整個病毒基因組。還包括靶向3,000個人類單核苷酸多態性(SNP)位點的探針用於參考。每次捕獲反應中使用含有EBV探針與常染色體DNA探針摩爾比為100:1的探針混合物。來自10個血漿樣本的DNA文庫在一次捕獲反應中被覆用，並使用來自每個樣本的相等數量的DNA文庫。表4A和4B列出了所有病例的測序統計數據，包括先前報告的用作當前訓練集的那些病例。

表4A.訓練集中所有NPC和非NPC病例的測序統計數據

訓練集					
樣本	組別 **	原始片段數	映射片段數 (mapped fragments)	映射率 (mapping rate) (%)	PCR重複率 (duplication rate) (%)
GG017	0	32715321	30223262	92.4	43.1
HL059	0	144554902	126762070	87.7	68.4
DN045	0	78914933	68428310	86.7	66.9
BP015	0	94168529	86145241	91.5	51.4
AB126	0	56541949	54346856	96.1	24
AC166	0	64450578	60439270	93.8	17.4
AD092	0	71510547	69046150	96.5	16.1
AE058	0	79728136	76825948	96.4	21.3
AQ104	0	96938063	84743586	87.4	16.4
BX011	0	72498952	70129591	96.7	14.9
CA062	0	72180027	69744659	96.6	15.3

訓練集					
樣本	組別 **	原始片段數	映射片段數 (mapped fragments)	映射率 (mapping rate) (%)	PCR重複率 (duplication rate) (%)
CH131	0	71459860	68990753	96.5	22.2
DC078	0	76239599	73238855	96.1	28.2
DF038	0	100612788	97254251	96.7	26.1
AG067	0	94932887	85387366	89.9	77.4
AR027	0	61611288	59001573	95.8	15.1
BL058	0	69559074	66513711	95.6	14.4
AF118	0	64803996	61659065	95.2	14.4
AF121	0	47656000	45104454	94.7	16
AO097	0	64803246	62335332	96.2	14
GV094	0	55594689	53398818	96	13.2
AL092	0	88202778	84617253	95.9	20.7
AM164	0	92235133	88753051	96.2	21.5
EI030	0	67332747	64898723	96.4	13.7
ER057	0	75611966	72851241	96.3	15.6
FF077	0	88728791	84934257	95.7	18.3
FF094	0	67950009	65456835	96.3	16.5
AO100	0	74073437	71534001	96.6	14.4
HE119	0	75939094	70594529	93	46.3
GC110	0	109911126	101627813	92.5	30
GT107	0	73134341	66124665	90.4	36.9
GZ039	0	58128740	54517308	93.8	26.1
AE151	0	118973652	109516490	92	21
AH116	0	97765995	88477724	90.5	28
AM095	0	87643692	80164284	91.5	19.6
BP065	0	84740540	80067572	94.5	37.4
EN086	0	32884093	31068440	94.5	38.3
GC038	0	52719658	49985247	94.8	38.1
AC106	0	46473277	43990963	94.7	82.5
AP080	0	38659615	36293332	93.9	60
GT123	0	90634113	82011875	90.5	65.1
AE011	0	64587311	59269827	91.8	49.2

訓練集					
樣本	組別 **	原始片段數	映射片段數 (mapped fragments)	映射率 (mapping rate) (%)	PCR重複率 (duplication rate) (%)
BV159	0	108366362	97270043	89.8	73.8
CZ031	0	104890395	93619970	89.3	73.4
AL071	0	35231149	32775649	93	74.6
AL122	0	132811199	123757690	93.2	76.6
AS079	0	33454154	31094045	93	74.3
AX070	0	82769034	77118993	93.2	75.8
DC125	0	82353895	76845022	93.3	64.2
DO041	0	98527392	91944421	93.3	63
DN037	0	73898976	66401716	89.8	69.3
DN131	0	85896965	77109501	89.8	68.8
DS050	0	97058938	87190650	89.8	68
DZ071	0	130632583	117555933	90	67.8
EH050	0	144211569	131747254	91.4	67.5
DZ026	0	63577798	60575778	95.3	24.9
HM142	0	74460599	71830670	96.5	28.9
HN068	0	58569268	56499964	96.5	27.6
HR120	0	78697168	75901684	96.5	28.7
CD005	0	67185044	64398576	95.8	18.9
DC146	0	67286289	64869690	96.4	20.4
DD090	0	72863832	69973561	96	18.9
DE103	0	74532024	71748839	96.3	20.1
DF112	0	80285807	77313233	96.3	16.6
DH045	0	73283371	70644621	96.4	21
DK016	0	98640353	95198449	96.5	22.8
DK057	0	65024042	62488386	96.1	19.8
DL055	0	64127942	61316770	95.6	18.9
CE144	0	55972062	53546313	95.7	15.4
CP042	0	67609649	64706108	95.7	15.2
CZ046	0	55236628	52985764	95.9	13.5
AP047	0	73544542	70437730	95.8	19.9
AS108	0	74546824	71474684	95.9	22.1

訓練集					
樣本	組別 **	原始片段數	映射片段數 (mapped fragments)	映射率 (mapping rate) (%)	PCR重複率 (duplication rate) (%)
BF137	0	87739825	83608642	95.3	19.2
AG020	0	67573799	63087296	93.4	17.6
AE055	0	62308055	59551554	95.6	11.4
AE105	0	59317164	56861140	95.9	10.2
AE107	0	69376388	66837992	96.3	13.3
AB004	0	69373853	66823399	96.3	12.4
AC153	0	83546018	80433313	96.3	13.4
AE026	0	80236204	77227885	96.2	13.8
AF091	0	79865448	76665569	96	12.4
HF020	0	73890276	69898875	94.6	11.9
BO049	0	54341974	49518640	91.1	12.2
CV094	0	69353920	62090890	89.5	11.9
DM146	0	86198122	83306628	96.7	13.7
DN054	0	57906125	55516552	95.9	21.6
DN092	0	65436665	62867803	96.1	16.7
AC173	1	77221448	69636427	90.2	53.5
AO050	1	94201867	84771216	90	51.9
AQ014	1	64826863	58371226	90	47.2
AZ118	1	75307129	67827313	90.1	47.7
AC088	1	76597786	55250665	72.1	47.2
AL038	1	76499430	55322894	72.3	45.7
AM086	1	84280496	61284379	72.7	43.4
AT038	1	64157394	46063166	71.8	45.8
BK041	1	61505610	44247376	71.9	44.8
CF028	1	97748094	88104244	90.1	59.1
CH047	1	123975141	112556783	90.8	56.6
CL037	1	106862473	96469537	90.3	60.7
CP006	1	61469649	54366171	88.4	59.4
CD007	1	103710165	93643893	90.3	61.9
DF120	1	96451355	89089726	92.4	51.6
DH101	1	73023724	67311149	92.2	60.3

訓練集					
樣本	組別 **	原始片段數	映射片段數 (mapped fragments)	映射率 (mapping rate) (%)	PCR重複率 (duplication rate) (%)
EG016	1	83087673	77307393	93	24.2
EN070	1	35732253	32582501	91.2	52.5
EV013	1	70202729	64881793	92.4	35.8
FD089	1	106149891	88230410	83.1	51.9
FG092	1	58840935	54320095	92.3	36.8
FM073	1	65062459	60232085	92.6	39.3
FZ037	1	46211337	42733248	92.5	37.6
GC137	1	73772882	68339539	92.6	62.9
GS059	1	103768139	95756898	92.3	64.4
GX170	1	112376826	104300963	92.8	60.7
HD083	1	80146546	74256782	92.7	59.8
HM169	1	69203940	64144652	92.7	59.7
AG006	1	73346449	68476847	93.4	22.9
FD163	1	62554476	58856976	94.1	27.7
CX027	1	88012245	80202542	91.1	67.7
CV009	1	60922871	56232165	92.3	45.6
TBR1433	2	77708246	70039392	90.1	30.2
TBR1470	2	73941394	67495510	91.3	21.6
TBR1572	2	71106989	64814893	91.2	23.6
TBR1605	2	115061297	94605333	82.2	47.8
TBR1606	2	60654197	55309308	91.2	32
TBR1607	2	75439582	69608132	92.3	28.1
TBR1650	2	83518964	76881089	92	21.8
TBR1665	2	73581524	68005926	92.4	26.7
TBR1685	2	64858923	59295059	91.4	28.4
TBR1794	2	77616481	72400504	93.3	31.9
TBR1795	2	84087680	78757703	93.7	25.2
TBR1821	2	89364373	83561953	93.5	25.2
TBR1822	2	74207438	69089332	93.1	32.3
TBR1841	2	76709226	71246483	92.9	27.6
TBR1857	2	93499651	85084161	91	29.1

訓練集					
樣本	組別 **	原始片段數	映射片段數 (mapped fragments)	映射率 (mapping rate) (%)	PCR重複率 (duplication rate) (%)
TBR1911	2	102778437	93039420	90.5	28.3
TBR1937	2	108092562	98448107	91.1	31.5
TBR1950	2	100931791	92237772	91.4	31.7
TBR1961	2	120837880	110269912	91.2	23.3
TBR2032	2	74713097	70057803	93.8	27.1
TBR2044	2	74572414	69808426	93.6	21.7
TBR2059	2	68180154	63969165	93.8	22.8
TBR2066	2	71590556	67039888	93.6	24.7
TBR2129	2	67520639	63360453	93.8	22.9
TBR1344	2	89830107	79295024	88.3	35.2
TBR1358	2	37407353	35051007	93.7	41.9
TBR1360	2	73282234	61715512	84.2	49.8
TBR1378	2	54841088	50538475	92.2	34.5
TBR1379	2	61335101	51046779	83.2	48.6
TBR1390	2	50153930	44313840	88.4	45
TBR1557	2	35803478	32801152	91.6	43.1

**：組別0 = 非NPC受試者，組別 1= NPC受試者(篩查隊列)，組別2 = NPC(外部隊列)。

表4B. 測試集中所有NPC與非NPC病例的測序統計數據

測試集						
樣本	組別 ##	原始片段數	映射片 段數	映射率 (%)	PCR 重 複率 (%)	NPC 風險 評分
AB069	0	62333414	56996119	91.4375	67.0529	0.25
AG102	0	50527076	47272142	93.558	79.7162	1.00
BF034	0	30900262	29069989	94.0768	79.9262	0.06
BH035	0	27968166	25683364	91.8307	78.2321	1.00
BM060	0	44571256	41656811	93.4612	82.7252	1.00
BN052	0	32654549	30177844	92.4154	77.7825	0.00
BO115	0	20605498	18891596	91.6823	76.3716	0.00

測試集						
樣本	組別 ##	原始片段 數	映射片 段數	映射率 (%)	PCR 重 複率 (%)	NPC 風險 評分
BR067	0	35222869	31942475	90.6867	10.9972	1.00
BS030	0	29488585	26961246	91.4294	66.5338	0.99
CB025	0	35335207	32498897	91.9731	81.8117	1.00
CI095	0	44920271	41857137	93.181	64.8167	0.00
CO003	0	22618823	20545705	90.8345	66.4679	1.00
DK129	0	26650610	24552495	92.1273	66.7223	1.00
DM162	0	46869923	42223785	90.0872	65.1806	0.99
DO001	0	35030693	32412652	92.5264	64.0082	1.00
DR058	0	33151251	30641021	92.4279	77.5861	0.41
DX145	0	30538948	28353858	92.8449	64.0698	0.00
DZ091	0	48775427	45509608	93.3044	79.647	0.00
EB064	0	15486333	14294637	92.3049	77.2137	0.52
EC056	0	44264275	41421171	93.577	64.8678	0.28
EI052	0	30414618	28373013	93.2874	79.4382	0.98
ER022	0	29318005	25814308	88.0493	64.2827	0.00
ET022	0	28303377	26549950	93.8049	79.5254	0.97
EZ015	0	34114519	31826767	93.2939	79.4083	0.65
FF159	0	27631827	25177560	91.118	66.2635	0.00
FH039	0	25047700	23182787	92.5546	73.199	1.00
FV078	0	59919758	55955981	93.3849	82.1063	1.00
GC157	0	22988959	21147818	91.9912	72.2857	0.00
GG040	0	58823944	53857823	91.5577	10.9781	0.14
GK072	0	28087271	26012505	92.6131	72.1235	0.99
GV071	0	30298816	27995522	92.3981	81.7554	1.00
GX058	0	52901878	47527912	89.8416	72.5617	0.00
GZ082	0	33025312	30743443	93.0905	76.508	0.00
HB042	0	39832106	37486823	94.1121	79.7558	0.59
HC056	0	27801939	25722722	92.5213	77.5543	0.80
HE176	0	26672711	24740453	92.7557	65.5094	0.00
HE181	0	20151536	18596587	92.2837	77.1676	0.00
HF010	0	36767150	34443572	93.6803	83.3378	0.99

測試集						
樣本	組別 ##	原始片段 數	映射片 段數	映射率 (%)	PCR 重 複率 (%)	NPC 風險 評分
HK068	0	24744347	22950199	92.7493	66.3875	0.02
HN102	0	18847144	17418641	92.4206	66.0707	0.00
p003704	1	24089077	22256290	92.3916	75.6729	1.00
p100405	1	27917819	25958361	92.9813	76.6278	1.00
p100742	1	33868828	31121633	91.8887	77.043	1.00
p101161	1	22077183	20555644	93.1081	76.2116	1.00
TBR2003	1	89502393	78014093	87.1643	67.8335	1.00
TBR2197	1	49274726	46072820	93.5019	79.8709	1.00
TBR2230	1	19463878	17991477	92.4352	77.7681	1.00
TBR2239	1	40477218	37931905	93.7117	79.5694	1.00
TBR2269	1	36732370	33345425	90.7794	10.8014	0.85
TBR2329	1	102625376	87445869	85.2088	79.1855	0.99
TBR2343	1	47646593	41027985	86.109	80.656	1.00
TBR2330	1	36942083	33822640	91.5559	11.0708	0.00
TBR2385	1	42000104	39181234	93.2884	81.8537	1.00
TBR2406	1	66799222	60524426	90.6065	83.3811	0.00
TBR2430	1	19062836	17515880	91.885	77.2878	1.00
TBR2466	1	39167493	35820959	91.4558	66.6063	1.00
TBR2553	1	20976134	19085605	90.9872	78.5291	1.00
TBR2605	1	28691106	26101695	90.9749	65.7645	1.00
TBR2615	1	33489016	29864524	89.1771	68.4423	1.00
TBR2641	1	113077610	94235991	83.3374	54.0705	0.98
TBR2647	1	52926587	46699098	88.2337	68.1284	1.00
TBR2655	1	44805097	41374955	92.3443	65.3989	1.00
TBR2669	1	43399057	39819658	91.7524	65.4329	1.00
TBR2682	1	35617499	32625124	91.5986	77.4284	1.00
TBR2699	1	78986032	67322508	85.2334	80.332	1.00
TBR2709	1	60912602	54630334	89.6864	78.8851	0.97
TBR2847	1	19610868	17657654	90.0401	52.1991	1.00
TBR2849	1	15220276	14043817	92.2704	51.0899	1.00
TBR2868	1	21065832	18609241	88.3385	53.7439	1.00

測試集						
樣本	組別 ##	原始片段 數	映射片 段數	映射率 (%)	PCR 重 複率 (%)	NPC 風險 評分
TBR2892	1	17905000	16600383	92.7137	51.5529	1.00
TBR2906	1	29385280	26298916	89.4969	53.0486	1.00

##：組別0 = 非NPC受試者，組別1 = NPC受試者

【0204】 EBV變體調用。

【0205】 使用在Li H等人，*Bioinformatics*. 2010;26:589-95描述的BWA比對器將測序讀數與人類(hg19)和EBV參考基因組(AJ507799.2型)比對，通過引用將Li H等人，*Bioinformatics*. 2010;26:589-95全文併入本文。當在EBV基因組位點檢測到與參考病毒基因組不同的替代等位基因時，用Samtools鑒定了EBV單核苷酸變體(SNV)，如Li H等人(*Bioinformatics*. 2009;25:2078-9所述，其通過引用將全文併入本文中。篩查出檢測到1個以上等位基因型的SNV位點(次要等位基因頻率截止值設定為5%)，用於隨後的NPC風險評分分析。

【0206】 NPC風險評分。

【0207】 在本實施例中，NPC風險評分是整個病毒基因組中一組固定的SNV位點的EBV基因型的加權總和(作為二元邏輯回歸模型中的解釋性變量)。首先通過分析訓練集中NPC樣本和非NPC樣本的EBV SNV譜的差異識別一組與NPC相關的SNV。利用Fisher精確檢驗分析整個EBV基因組中每個變體與NPC病例的關聯性。然後獲得一組固定的顯著的SNV，假髮現率(FDR)控制在5%。

【0208】 測試樣本的NPC風險評分可以通過其在訓練集確定的該特定組的顯著的SNV位點上的EBV基因型來確定。如前所述，由於血漿中EBV DNA分子濃度較低，通過測序的EBV DNA讀數可能無法完全覆蓋

整個EBV基因組。因此，將該評分制定成由血漿EBV DNA讀數(例如，具有可用的基因型信息)覆蓋的那些SNV位點上的基因型模式決定(圖16A、16B和16C)。為了得到NPC風險評分，首先識別顯著的SNV位點的子集，這些位點被檢測樣本中的血漿EBV DNA讀數所覆蓋。然後，在顯著的SNV位點的子集內確定每個位點上的基因型的權重(效應大小)。這是通過分析訓練數據集中NPC和非NPC樣本的每個位點的基因型模式來完成的(圖16B)。在此基礎上，建立邏輯回歸模型來報告NPC上每個SNV位點的風險基因型的效應大小。邏輯模型如下：

$$P = \frac{1}{1 + e^{-(\beta_0 + \sum_{k=1}^n \beta_k X_k)}}$$

也可作：

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) = \beta_0 + \sum_{k=1}^n \beta_k X_k$$

其中n是顯著的SNV位點的數目；並且 β_0 和 β_k 是可以通過最大似然估計確定的係數；P是EBV陽性患者患NPC的概率；該變量 X_k 表示基因組位置k處的SNV位點。如果樣本中存在與EBV參考基因組相同的變體，則將 X_k 編碼為-1。如果樣本中存在替代變體，則將 X_k 編碼為1。如果樣本中沒有分析的變體位點，則將 X_k 編碼為0。在python中使用邏輯回歸函數 (penalty = 'l2', C = 1, solver = 'saga', max_iter = 5000，且random_state = 0)來估計係數 β_0 和 β_k 。這是通過分析訓練數據集中NPC和非NPC樣本中每個位點的基因型模式來完成的。在python中輸入矩陣(c+d)×n，其中c是NPC樣本數，d是訓練集中非NPC樣本數，n是基因型變體數。每一行代表一個樣本(對於未患有NPC的患者為0；對於患有NPC患者為1)，每一列代表一個變體。然後可以推導出係數(β_0 和 β_k)。然後根據其自身在SNV位點的基因型，通過相應的係數 β_0 和 β_k 加權，從訓練模型中推導出測試

樣本的NPC風險評分。(圖16C)。

【0209】 結果

【0210】 建立NPC風險評分訓練模型。

【0211】 如上所述，先前報導的NPC和非NPC樣本的血漿EBV DNA測序數據被用於NPC風險評分訓練模型的開發。採用靶向捕獲測序法對血漿樣本中的EBV DNA進行了富集。本文研究了來自NPC和非NPC樣本的EBV分離株的病毒SNV譜。從該數據集中，選擇那些通過測序的EBV DNA讀數覆蓋EBV基因組至少30%的NPC和非NPC病例。之所以選擇這個截止值是因為訓練數據集中95%以上的NPC樣本的病毒基因組覆蓋率大於該截止值(表4A和4B)。表5詳細列出了這些選擇的NPC和非NPC受試者的人口統計數據，包括年齡和性別，以及NPC患者的癌症分期信息(AJCC第8版)。所選NPC和非NPC樣本的測序統計數據見(表4A和4B)。

表5. 訓練集中所有NPC和非NPC病例的受試者的特點

	NPC 患者	非NPC 受試者
數量	63	88
性別		
M	56	88
F	7	0
年齡中位數，年 (IQR)	53 (47.5 – 57.5)	54 (48 – 59)
腫瘤分期		
I	17	NA (不適用)
II	11	NA
III	26	NA
IV	9	NA

【0212】 對這些63個NPC和88個非NPC樣本的EBV SNV譜進行了分

析。所有樣本在整個EBV基因組上的中位數測序深度為2x(四分位間距(IQR)，1.0x - 9.2x)。從NPC樣本中識別出的EBV SNV的平均數為800(IQR，662 - 958)，非NPC樣本中的SNV平均數為539(範圍363 - 656)。所有樣本共識別出5678個不同的SNV。這些SNV在整個EBV基因組中的分布如圖16D所示。

【0213】 還用Fisher精確檢驗研究了訓練集中每個病毒SNV與NPC樣本的相關性。通過將假髮現率(FDR)控制在0.05(調整後的p值)，共識別出661個與NPC相關的顯著的SNV。表6列出了這661個SNV的基因組位置。隨後，根據這661個SNV位點的基因型模式，得出了NPC和非NPC受試者的血漿樣本的測試集的NPC風險評分。

表6. 661個示例性SNV的EBV基因組位置(相對於AJ507799.2型)

EBV基因組位置
46, 156, 158, 206, 212, 246, 390, 409, 475, 505, 536, 570, 612, 628, 631, 866, 1067, 1072, 1074, 1133, 1137, 1176, 1194, 1195, 1322, 1349, 1373, 1384, 1391, 1534, 1875, 1992, 2709, 2772, 3223, 3379, 3820, 3941, 4863, 5398, 5745, 5802, 5849, 6066, 6108, 6209, 6287, 6379, 6483, 6555, 6583, 6865, 6883, 6885, 6910, 6943, 6998, 7000, 7015, 7047, 7133, 7188, 7208, 7212, 7232, 7246, 7261, 7296, 7326, 7356, 7385, 8233, 8344, 8455, 8567, 8872, 10623, 11323, 11694, 35308, 35492, 35526, 35550, 35583, 35615, 35637, 35678, 35856, 35869, 35974, 36067, 36166, 36577, 36667, 36694, 36768, 36798, 36847, 36948, 36950, 37051, 37053, 37284, 37465, 37624, 37641, 37671, 37682, 37701, 37739, 37834, 37954, 40549, 40555, 40835, 41153, 41402, 42209, 42321, 42422, 42712, 42948, 42992, 43088, 43235, 43280, 43312, 43396, 43419, 43611, 43806, 43819, 44122, 44530, 44650, 45100, 45616, 45691, 45694, 45823, 46105, 46133, 46610, 46895, 47904, 48633, 48730, 48997, 50133, 50754, 50764, 50881, 50946, 51080, 51151, 51152, 51227, 51269, 51379, 51435, 51514, 51517, 51588, 51847, 52549, 53683, 57411, 58192, 58207, 59205, 59334, 59390, 59435, 59489, 59588, 60005, 60239, 60453, 60887, 60893, 61256, 62141, 62456, 62499, 62509, 62741, 62819, 63302, 63911, 64131, 64171, 64216, 64234, 64882, 64921, 65465, 66364, 66434, 66718, 66749, 66961, 67054, 67621, 67721, 67745, 67867, 68260, 68303, 68304, 68509, 68885, 69483, 75030, 75287, 75326, 76761, 76917, 77195, 77815, 77816, 78662, 79264, 79318, 79649, 79739, 80313, 80349, 80609, 80626, 80635, 80840, 80919, 80978, 81110, 81212, 81682, 81722, 82332, 82369, 83062, 83639, 84127, 84257, 84345, 84390, 84413, 84524, 84739, 84766, 84799, 84883, 84887, 84917, 84970, 85076, 85125, 85128, 85224, 85227, 85228, 85801, 85840, 86113, 86779, 86794, 87397, 87556, 88012, 88121, 88223, 88303, 88464, 88500, 88552, 88597, 88636, 88837, 88900, 89630, 89819, 89850, 89920, 90477, 90553, 90585, 90641, 91005, 91011, 91046, 91179, 91429, 91430, 91437, 91765, 93097, 93367, 93468, 94793, 95291, 95379, 95458, 95509, 95631, 98147, 98243, 98261, 98376, 98489, 98841, 98984, 98985, 99057, 99069, 99329, 99350, 99355, 99736, 99760, 99805, 100552, 101509, 101691, 101920, 101986, 102922, 103333, 103824, 104286, 104432, 104549, 104554, 104672, 104804, 105670, 106006, 106374, 106468, 107457, 107592, 108012, 108332, 108351, 108355, 108419, 109234, 109507, 109576,

EBV基因組位置
109775, 109939, 110032, 110477, 110687, 110773, 110873, 110939, 111026, 111694, 112486, 112980, 113691, 113718, 114468, 114762, 114811, 115371, 115462, 115574, 115639, 115711, 115726, 116058, 116310, 116393, 116394, 116501, 116583, 116807, 117030, 117291, 117456, 117564, 117994, 118097, 118210, 118349, 118432, 118460, 118505, 118955, 119031, 119295, 119381, 119417, 119786, 119804, 120294, 120318, 120360, 120672, 120866, 121160, 121164, 121230, 121383, 121473, 121689, 121719, 121737, 121776, 121893, 122140, 122208, 122340, 122343, 122361, 122443, 122481, 122490, 122607, 122610, 122820, 123174, 123312, 124938, 125271, 126135, 126225, 126442, 126601, 126681, 127197, 127408, 127465, 127597, 127615, 127840, 127991, 128036, 128268, 129730, 129835, 129904, 130450, 130453, 130687, 132047, 132182, 132224, 133635, 133648, 133779, 133947, 134155, 134157, 134199, 134349, 134371, 134385, 134718, 134729, 134760, 134766, 134788, 134874, 135060, 135078, 135102, 135108, 135117, 135354, 135606, 135866, 135949, 136053, 136077, 136185, 136554, 136645, 136914, 136932, 136974, 137080, 137142, 137315, 137346, 137480, 138869, 139209, 139440, 139495, 139683, 139945, 140001, 140059, 140227, 140254, 140256, 140305, 140492, 140569, 140600, 140688, 140744, 143451, 144072, 144086, 144354, 144564, 144684, 145144, 145245, 145538, 145736, 145918, 146158, 146237, 146241, 146242, 146249, 146270, 146557, 146627, 146690, 146744, 146756, 146764, 146887, 147059, 147060, 147068, 147088, 147102, 147310, 147426, 147478, 147492, 147607, 147651, 147663, 147681, 147698, 147708, 147731, 147773, 147783, 147849, 147882, 147899, 148050, 148230, 148283, 148488, 148627, 148636, 148930, 148971, 149130, 149318, 149354, 149643, 149835, 149925, 150021, 150027, 150171, 150356, 150470, 150749, 150777, 151139, 151146, 151202, 151255, 151337, 151352, 151370, 151643, 151821, 151876, 151942, 152023, 152086, 152244, 152611, 152945, 152946, 153011, 154386, 154614, 154971, 155084, 155388, 155390, 155608, 155919, 155988, 156012, 156132, 156138, 156153, 156183, 156282, 156636, 156695, 156797, 156809, 156818, 157052, 157124, 157229, 157427, 157466, 157805, 157823, 158015, 158142, 158407, 158429, 158480, 158777, 159219, 160803, 160826, 160970, 161035, 162116, 162146, 162194, 162214, 162236, 162463, 162475, 162506, 162851, 163106, 163286, 163292, 163363, 163403, 163421, 163463, 163610, 163628, 163685, 163925, 163994, 164723, 165086, 165850, 167201, 168172, 168176, 168411, 168432, 168466, 168559, 168593, 168596, 168659, 169008, 169428

【0214】 NPC風險評分訓練模型的評價。

【0215】 採用留一法對訓練模型進行評估，以分析訓練集中樣本的NPC風險評分。在留一法中，建立訓練模型並得出NPC風險評分的原則與方法中描述的相同。訓練集中除一個樣本外，其餘樣本均用於建立訓練模型，漏掉的樣本可用於分析其NPC風險評分。在留一法分析中，NPC組的NPC風險評分中位數為0.99(IQR，0.98 - 1.0)，非NPC組的NPC風險評分中位數為0.01(IQR，0.00 - 0.89)(圖17A)。採用受試者操作特徵(ROC)曲線分析法，通過NPC風險評分評價NPC與非NPC樣本的差異。曲線下面積值為0.91(圖17B)。

【0216】 測試集中的NPC風險評分分析。

【0217】 對另外31名NPC患者和45名非NPC受試者的血漿樣本進行靶向捕獲測序。其中所有31個NPC和40個非NPC樣本具有至少30%或以上的通過測序的EBV DNA讀數覆蓋的EBV基因組。表7總結了這些NPC和非NPC受試者的臨床特徵。表4A和4B中也示出了樣本的該測試集的測序統計數據。

表7. 測試集中所有NPC和非NPC病例的受試者統計數據特點

	NPC 患者	非NPC 受試者
數量	31	40
性別		
M	26	40
F	5	0
年齡中位數，年 (IQR)	53 (47 – 61.5)	53 (50 – 57)
腫瘤分期		
I	6	NA (不適用)
II	2	NA
III	12	NA
IV	11	NA

【0218】 根據所建立的訓練模型，對31個NPC樣本和40個非NPC樣本的測試集的NPC風險評分進行了分析。樣本的NPC風險評分可以通過其在訓練集中識別的661個顯著的SNV位置上的變體模式來確定。由於可能存在EBV基因組的不完全覆蓋，只有通過測序的EBV DNA讀數覆蓋並具有相應的等位基因信息的SNV位點才能被納入NPC風險評分分析(圖16A、16B和16C)。

【0219】 NPC 組 的 中 位 數 NPC 風 險 評 分 為 0.999(IQR ， 0.996-

0.999)，非NPC組的中位數NPC風險評分爲0.557(IQR，0.000-0.996)(圖18A)。同樣，在這31個NPC樣本中也發現了高NPC風險評分。測試集中的NPC樣本可以與訓練集中的那些NPC樣本共享相似的EBV SNV譜。通過NPC風險評分體現的NPC與非NPC樣本的差別還可以通過ROC曲線分析進行評估。曲線下面積值爲0.83(圖18B)。

【0220】 測試集中高危變體位點的基因型模式的分析。

【0221】 在EBER(EBV編碼的小RNA)區域存在與NPC相關的高危EBV變體。在EBER區域，Hui等人報告了23個顯著的SNV。在31個NPC樣本和40個非NPC樣本的測試集中，採用了類似的NPC風險預測方法，但僅基於EBER區域中23個報告的SNV的基因型模式進行了分析。

【0222】 在測試集中，71個NPC樣本和非NPC樣本中有31個樣本(44%)具有覆蓋了所有23個SNV位點的EBV DNA讀數。如表8所示，對於這23個SNV位點中的每一個，只有一部分樣本具有覆蓋SNV位點的讀數的可用基因型信息(即樣本中並非所有23個SNV位點都覆蓋了血漿EBV DNA讀數)。NPC樣本中23個SNV位點的每個位點的高危基因型的百分數在86%至97%之間。非NPC樣本中高危基因型的百分數爲35%至52%。分析的NPC樣本和非NPC樣本的數量是指具有可用基因型信息的樣本(例如，具有覆蓋SNV位點的EBV DNA讀數)。測試集中只有一部分樣本(31個NPC樣本和40個非NPC樣本)的讀數覆蓋SNV位點和相應位點上的可用基因型信息。通過僅用ROC曲線分析來分析EBER區域中的23個SNV的基因型模式也可評估NPC與非NPC樣本的差別。曲線下的面積值爲0.72(圖19A和19B)。該數值低於分析整個EBV基因組的基因型模式所獲得的數值(0.83)。對整個EBV基因組的基因型模式進行的分析，可以比在固定的病

毒基因組區域獲得更好的NPC樣本和非NPC樣本的差別。

表8 在EBER基因上的23個SNV位點的測試集中NPC和非NPC病例的基因型模式

SNV位置	危險等位基因	分析的NPC樣本數	分析的非NPC樣本數	具有危險等位基因的NPC樣本數(百分比)	具有危險等位基因的非NPC樣本數(百分比)
5398	A	29	31	25 (86%)	12 (39%)
5849	T	28	27	24 (86%)	11 (41%)
6483	T	29	19	25 (86%)	9 (47%)
6583	G	29	16	25 (86%)	7 (44%)
6865	A	29	25	26 (90%)	9 (36%)
6883	G	29	25	27 (93%)	11 (44%)
6885	T	29	23	26 (90%)	10 (43%)
6910	A	29	23	26 (90%)	8 (35%)
6943	G	29	23	28 (97%)	11 (48%)
6998	G	30	26	29 (97%)	11 (42%)
7000	T	30	25	29 (97%)	10 (40%)
7011	G	30	26	29 (97%)	11 (42%)
7015	T	30	25	29 (97%)	11 (44%)
7047	C	30	29	29 (97%)	14 (48%)
7124	G	29	28	28 (97%)	11 (39%)
7133	C	29	28	28 (97%)	12 (43%)
7197	T	28	26	27 (96%)	10 (38%)
7205	A	28	26	27 (96%)	11 (42%)
7212	C	28	27	27 (96%)	11 (41%)
7232	A	29	28	25 (86%)	11 (39%)
7261	A	29	27	28 (97%)	14 (52%)
7296	T	28	26	27 (96%)	13 (50%)
7326	C	28	26	27 (96%)	12 (46%)

【0223】 類似地，BALF2(BamHI 左框-2)基因上的3個高危SNV也被報導(Xu等人，*Nat Genet.* 2019;51:1131–6)。在測試集中，71個樣本

中有55個樣本(78%)的EBV DNA讀數覆蓋了所有3個SNV。對於這3個SNV位點中的每一個，測試集中只有一部分樣本的讀數覆蓋了具有可用基因型信息的SNV位點(表9)。在3個SNV位點的每個位點的高危基因型在NPC樣本中的百分數在86%至93%之間。在非NPC樣本中高危基因型的百分數為47%至65%。有4個沒有EBV DNA讀數覆蓋BALF2基因上的3個報告的SNV中任何一個的病例(1個NPC樣本和3個非NPC樣本)，這些病例無法分析。對來自測試集中的其餘30個NPC樣本和37個非NPC樣本採用了類似的NPC風險預測方法，只分析了在BALF2區域報告的3個SNV的基因型模式。還通過ROC曲線分析評估NPC與非NPC的差別。曲線下面積值為0.77分(圖20A和20B)。該數值低於分析整個EBV基因組的基因型模式得出的數值(0.83)。對整個EBV基因組的基因型模式進行的分析，可以比在固定的病毒基因組區域獲得更好的NPC和非NPC樣本的差別。

表9.在BALF2基因上的3個SNV位點的測試集中NPC和非NPC病例的基因型模式

SNV位置	危險等位基因	分析的NPC樣本數	分析的非NPC樣本數	具有危險等位基因的NPC樣本數(百分比)	具有危險等位基因的非NPC樣本數(百分比)
162214	C	30	31	28 (93%)	20 (65%)
162475	C	30	32	27 (90%)	17 (53%)
163363	T	29	32	25 (86%)	15 (47%)

【0224】 本實施例中描述的NPC風險評分分析允許基於整個EBV基因組上的一組661個顯著的SNV內浮動數量的隨機選擇的SNV上的基因型模式預測NPC風險(表6)。用於NPC風險評分分析的浮動數目的SNV位點可通過SNV位點是否被測序的EBV DNA讀數覆蓋並具有相應的等位基因信息來確定。已經對一組661個顯著的SNV進行了下採樣，並使用與下採

樣的一組SNV中的SNV的浮動數目相同的方法，在測試集中分析了樣本NPC預測的性能。對於下採樣分析，一定數量(例如，23，25，100，200或500)的SNV從661個顯著SNV中隨機選擇。然後，對於測試樣本，識別被EBV DNA測序讀數覆蓋的一組下採樣的SNV內的SNV位點。然後，通過在覆蓋的、下採樣的SNV位點的訓練集中訓練NPC和非NPC樣本的基因型模式，獲得NPC風險評分訓練模型。通過訓練，確定了訓練模型中各位點基因型的權重。然後，通過在這些覆蓋的、下採樣的SNV位點上的其自身基因型模式應用於對相同的下採樣SNV位點加權的NPC風險評分訓練模型，得出測試樣本的NPC風險評分。表10總結了不同SNV位點數的NPC風險評分訓練模型的預測性能。對於給定數量的SNV位點，隨機選擇SNV進行10次下採樣，表10中曲線下面積值為10次隨機下採樣中的平均值。整個EBV基因組中的一組SNV被下採樣至23個，與EBER區域中報告的SNV的數目相同。通過ROC曲線分析法評估NPC樣本與非NPC樣本的差別。曲線下面積值為0.78。該數值高於對EBER區域的23個報告的SNV的基因型模式的分析所得的數值(0.72)。

表10.基於不同數目的SNV的NPC預測性能

下採樣SNV的數目	曲線下面積(AUC)值
23	0.78
25	0.78
100	0.77
200	0.83
500	0.79
661(所有 SNV)	0.83

【0225】 本研究報告了通過血漿DNA測序分析EBV基因型信息。通過配對末端測序，識別出了攜帶血漿EBV DNA的NPC受試者和非NPC受

試者的血漿EBV DNA分子的差別分子特徵，包括數量和大小。結合這種基於數量和大小的血漿EBV DNA的分析，可以使目前基於PCR的方案陽性預測值幾乎翻倍，這可以成為第二代基於測序的篩查測試的基礎。對來自NPC和非NPC受試者的血漿樣本進行測序，可以進一步獲得EBV基因型信息，並可以提高其潛在的臨床應用價值。

【0226】 NPC風險評分可用於由病毒全基因組標記物而不是單個基因標記物來確定。此處的風險評分基於整個EBV基因組中不同的SNV位點的變體模式得出。EBV基因型信息的血漿測序可能涉及對具有低濃度EBV DNA分子的血漿樣本進行測序，從而導致EBV基因組的不完全覆蓋。在某些情況下，信息性SNV位點可能不被任何EBV DNA讀數所覆蓋，並且在某些情況下，無法判斷個體是否攜帶高危的EBV菌株類型。研究結果支持這一點，即對於EBER基因上23個SNV位點中的每一個，在測試集中的71個分析樣本中，只有部分具有覆蓋這些位點的讀數。測試集中的NPC樣本顯示有較高的NPC風險評分，這可以表明與NPC相關的EBV SNV譜的存在。採用捕獲探針法富集血漿樣本中的EBV DNA分子。還可以使用一種擴增子測序方法富集EBV DNA片段，該片段可以靶向基因型信息的高危變體區域。

【0227】 本文分析了在最近報導的EBER基因和BALF2基因上的高危變體位點上NPC和非NPC樣本的基因型模式。高危基因型在NPC和非NPC樣本中的分布與兩項分析細胞樣品的研究結果一致，即正常對照組的NPC腫瘤組織及唾液樣品。由於包括本研究在內的三項研究均在中國南方同一地區或鄰近地區進行，因此正常對照受試者的EBV基因型的分布可能相似。這為通過血漿樣本測序進行EBV基因分型分析的可行性提供了依

據。

【0228】在篩查的背景下，從血漿樣本中分析EBV SNV可能具有臨床實用性。如前所述，大約5%的篩查人群可以在血漿中攜帶EBV DNA，但未患有NPC(假陽性組)。此數據顯示，這些非NPC受試者的NPC風險評分是可變的，可能涉及不同的EBV SNV譜。可能存在異質的個體群體，他們具有未來發展NPC的不同風險。其中的一些攜帶高危EBV株的人具有更高的患有NPC的風險。NPC風險評分可用於根據病毒全基因組SNV譜將非NPC受試者分為不同的風險組。在一個實例中，可有理由為那些具有高NPC風險評分的人提供更頻繁的篩查。

【0229】通過對NPC患者和非NPC受試者血漿樣本的測序分析來分析其EBV基因型信息。雖然先前的研究集中在人群水平上識別與NPC相關的高危變體，但這項研究為病毒基因型分析的臨床應用提供了新思路。這樣的分析可以在個體層面通過表徵他們所攜帶的EBV基因型來告知患癌風險。

【0230】雖然本文已經示出並描述了本公開內容的優選實施方案，但是對於本領域技術人員顯而易見的是，這些實施方案僅以示例的方式提供。本領域技術人員現將在不脫離本公開內容的情況下想到多種變化、改變和替換。應當理解，本文所述的本公開內容的實施方案的各種替代方案均可用於實施本公開內容。以下述申請專利範圍旨在限定本公開內容的範圍，從而覆蓋這些申請專利範圍及其等同項範圍內的方法和結構。

【符號說明】

【0231】

1101:計算機系統

1105:中央處理單元/ CPU/處理器/計算機處理器

1110:存儲器

1115:電子存儲單元

1120:接口

1125:外圍裝置

1130:網絡

1135:電子顯示器

1140:用戶界面(UI)

1201:受試者

1202:樣本

1204:處理器

1205:顯示器

1207:服務器

【發明申請專利範圍】

【請求項1】

一種篩查受試者之癌症之方法，該方法包括：

接收在第一時間點進行的第一次測定的數據，該第一次測定包括確定在該受試者的生物樣品中的EB病毒(EBV)無細胞核酸分子的數量，其中該數量指示該受試者發展癌症的風險；以及

接收來自該受試者之無細胞DNA(cfDNA)樣品之第二次測定的數據，其中該第二次測定提供以下的測量：(i) EBV cfDNA之甲基化；或(ii) EBV cfDNA與源自具有癌症之受試者之參考EBV基因組的序列相似度，其中該參考EBV基因組係EBV基因組AJ507799.2之變體；

藉此篩查該受試者之癌症。

【請求項2】

一種在受試者中預測癌症之方法，該方法包括：

接收來自該受試者的無細胞DNA (cfDNA)樣品的測定的數據，其中該測定提供以下的測量：(i) EBV cfDNA之甲基化；或(ii) EBV cfDNA與源自具有癌症之受試者之參考EBV基因組的序列相似度，其中該參考EBV基因組係EBV基因組AJ507799.2之變體；以及

基於來自該測定的數據生成指示該受試者發展癌症的風險的報告。

【請求項3】

如請求項1之方法，其中該第一次測定的結果不會導致對該受試者進行癌症的醫學治療。

【請求項4】

如請求項1之方法，其中該篩查另包含接收具有假陽性率低於1%的

臨床診斷檢查的結果。

【請求項5】

如請求項1之方法，其另包括：

- (i)從該受試者獲得該生物樣品；以及
- (ii)進行該第一次測定。

【請求項6】

如請求項1之方法，其中確定該EBV無細胞核酸分子的數量包括測量(a)在該生物樣品中的該EBV無細胞核酸分子的拷貝數或(b)在該生物樣品中的該EBV無細胞核酸分子的百分比。

【請求項7】

如請求項1之方法，其中(a)該生物樣品係第一生物樣品；(b)該EBV無細胞核酸分子的數量係高於閾值；及(c)該第一次測定另包括：

(iii)從該受試者獲得第二生物樣品，並測量在該第二生物樣品中的EBV無細胞核酸分子的第二數量。

【請求項8】

如請求項7之方法，其中：

(a) 該第二生物樣品係在該第一生物樣品之後4周獲得；及/或

(b) 該受試者不具有癌症，且該方法另包含對該生物樣品之第一次測定2個月至1年後收集的一或多個額外樣品重複該第一次測定及該第二次測定。

【請求項9】

如請求項2之方法，其中該癌症係鼻咽癌。

【請求項10】

如請求項1之方法，其中該癌症係鼻咽癌。

【請求項11】

如請求項1之方法，其另包括：使用應用於數據輸入的分類器，計算該受試者發展癌症的風險評分，該數據輸入包含該第二次測定的測量，其中該分類器被配置為將函數應用於該數據輸入，以生成包含風險評分的輸出，該風險評分評估該受試者發展癌症的風險。

【請求項12】

如請求項1之方法，其中該第二次測定包括測量該EBV cfDNA之甲基化。

【請求項13】

一種分析來自受試者的生物樣品的核酸分子之方法，其包括：

在計算機系統中獲得來自該受試者的生物樣品的無細胞DNA分子的序列讀數，其中該生物樣品包含來自該受試者且可能來自EB病毒(EBV)的無細胞DNA分子；

在計算機系統中，將該無細胞DNA分子的序列讀數與EBV參考基因組比對，其中該EBV參考基因組係AJ507799.2；以及

在計算機系統中，識別EBV無細胞DNA分子的變體模式，其中該變體模式表徵在該EBV參考基因組的多個變體位點的每個點處映射到該EBV參考基因組的序列讀數的核苷酸變體，其中該多個變體位點包括跨越該EBV參考基因組上的至少30個位點，並且其中該變體模式指示在該受試者中癌症的狀態或風險。

【請求項14】

一種包括機器可執行代碼的非暫時性計算機可讀介質，該機器可執

行代碼在由一個或多個計算機處理器執行時實現如請求項1至13中任一項之方法。

【請求項15】

一種計算機產品，包括存儲用於控制計算機系統執行請求項1至13中任一項之方法的操作的多個指令的非暫時性計算機可讀介質。

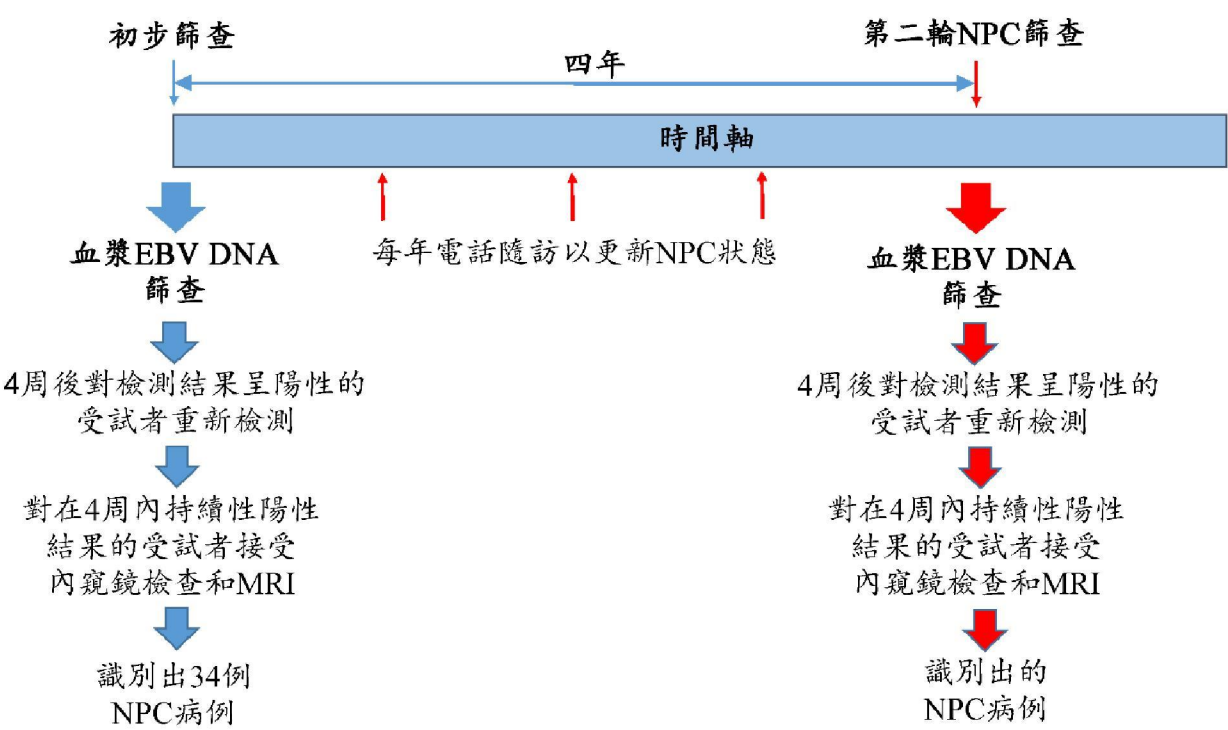
【請求項16】

一種系統，包括：

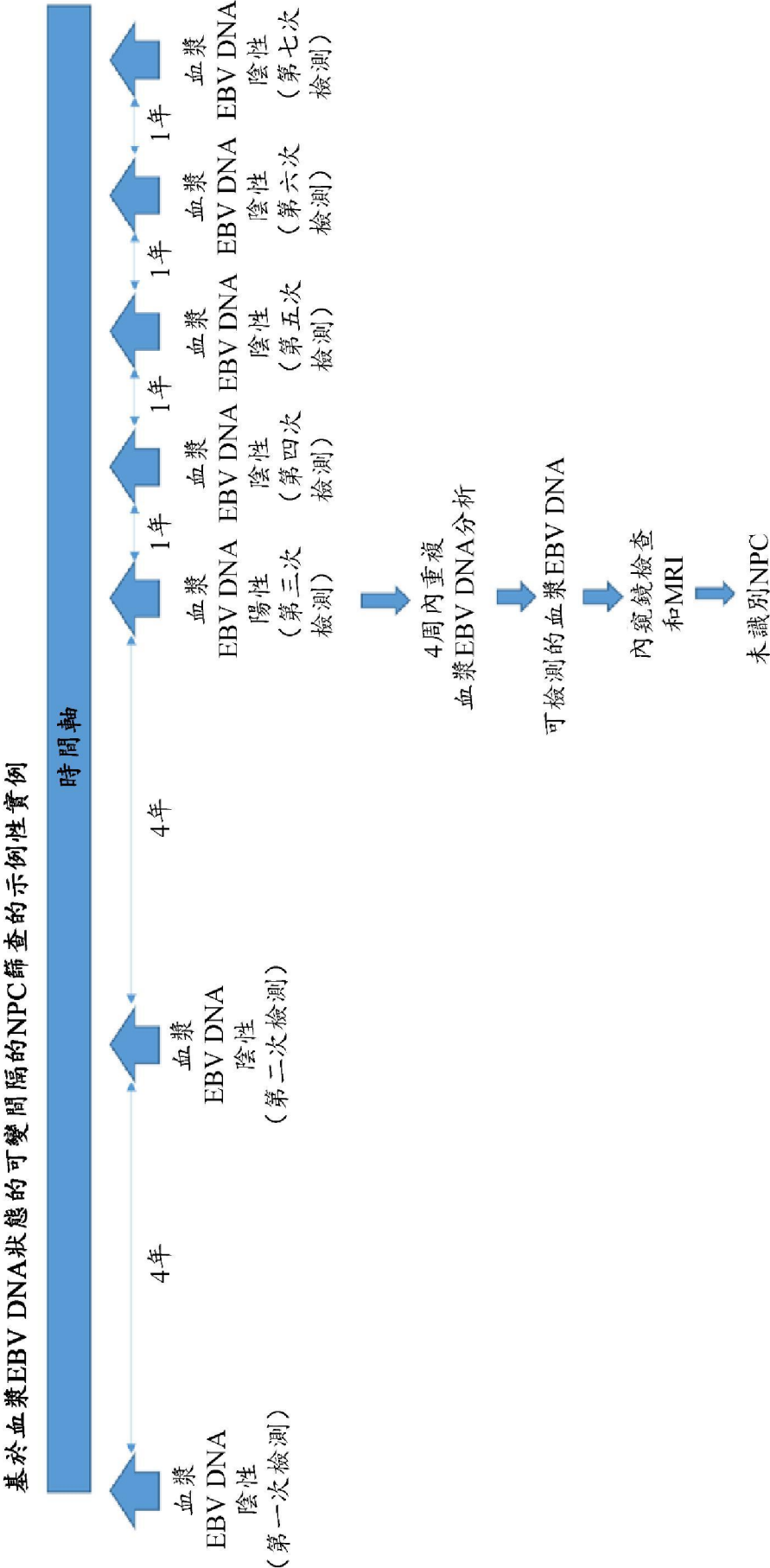
如請求項15之計算機產品；以及

一個或多個處理器，該處理器用於執行存儲在計算機可讀介質上的指令。

【發明圖式】

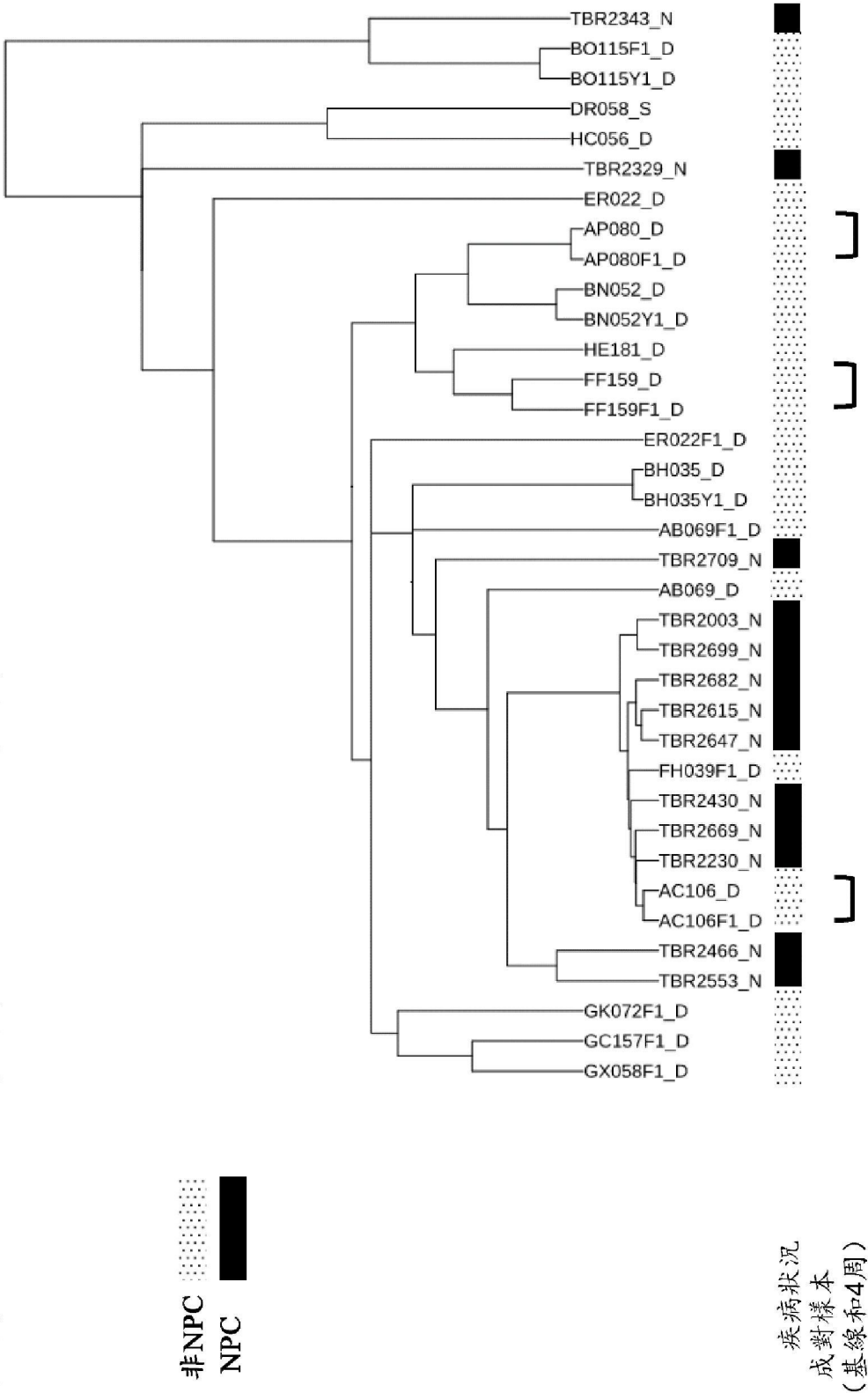


【圖 1】



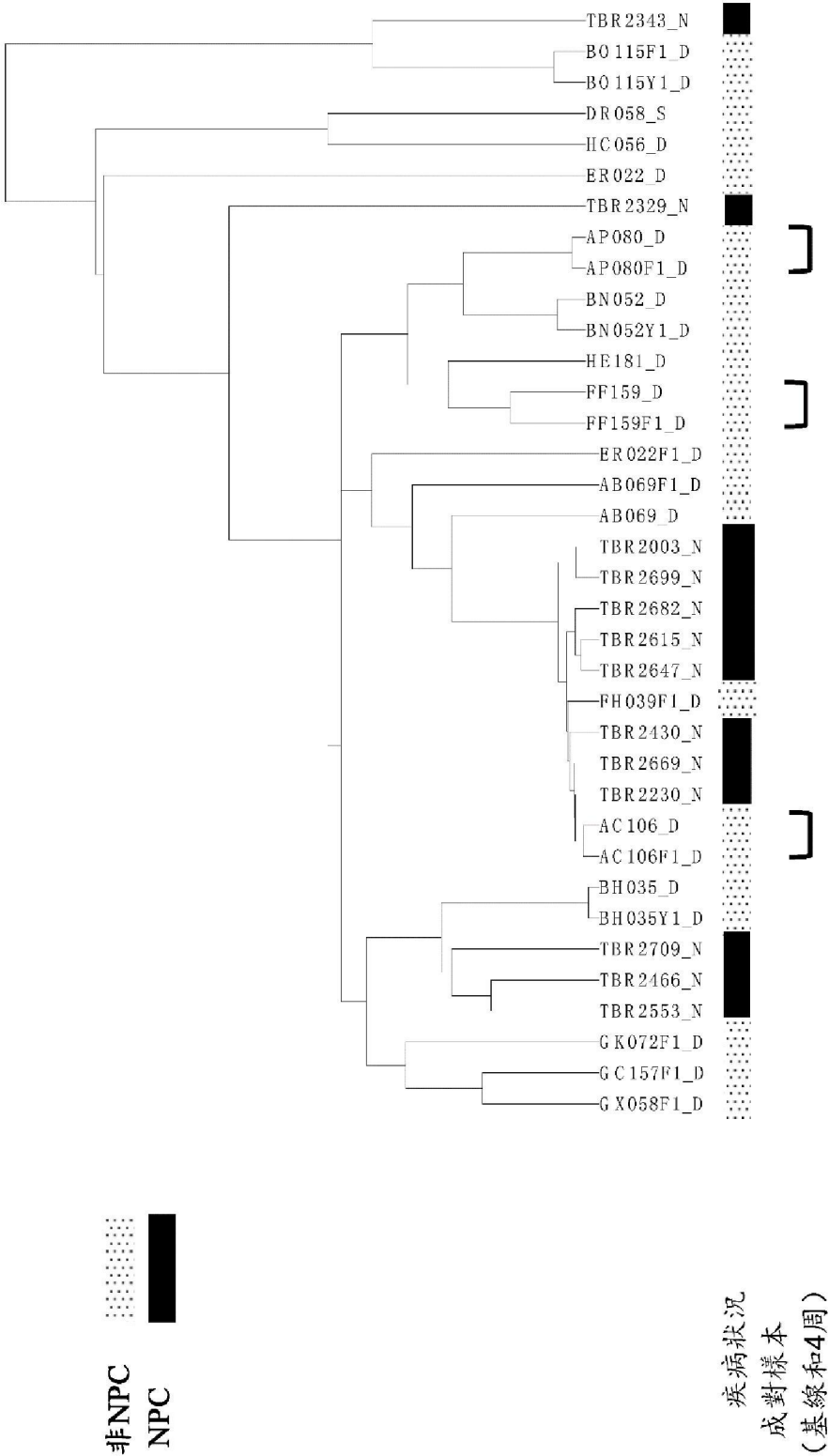
【圖 2】

測試集：13名NPC受試者+16名非NPC受試者



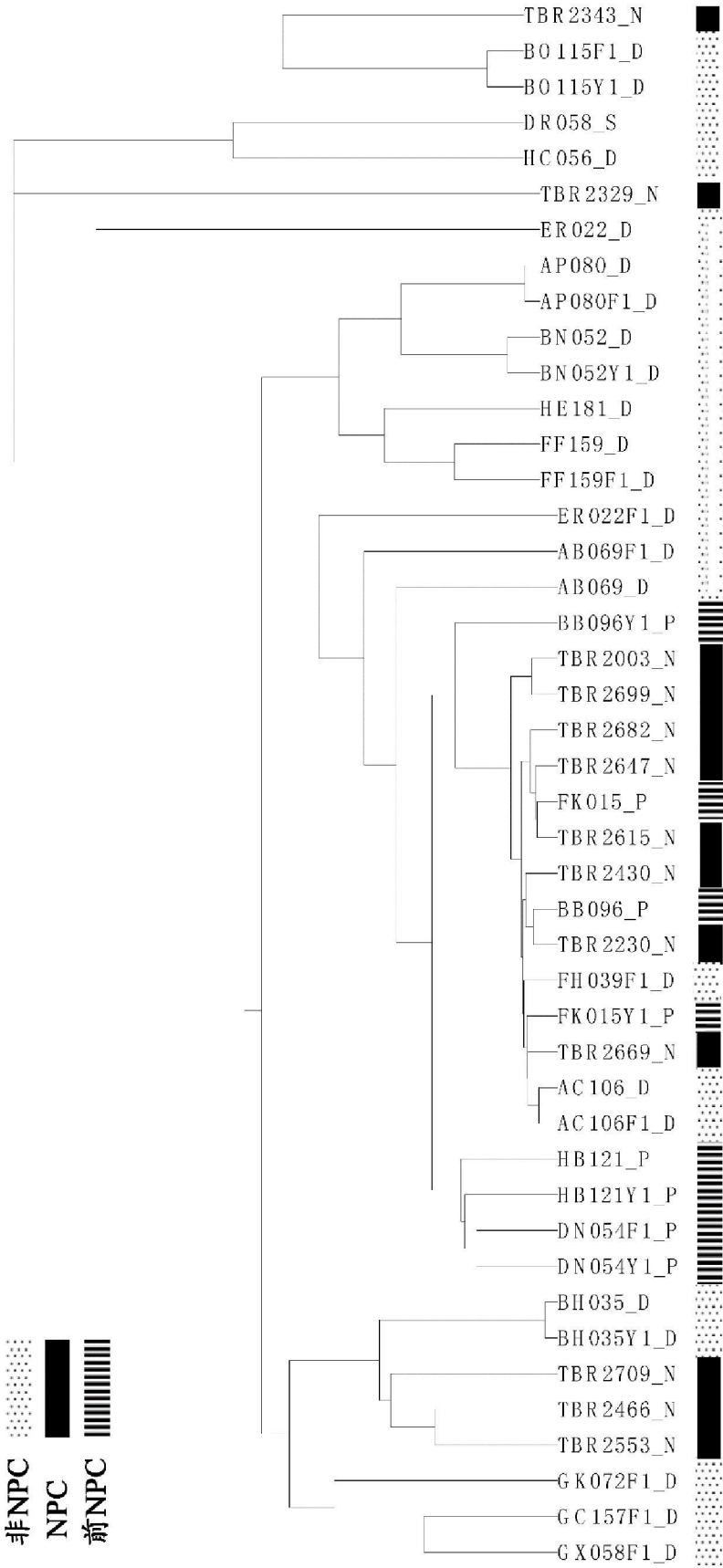
【圖3】

測試集：13名NPC受試者+16非NPC受試者

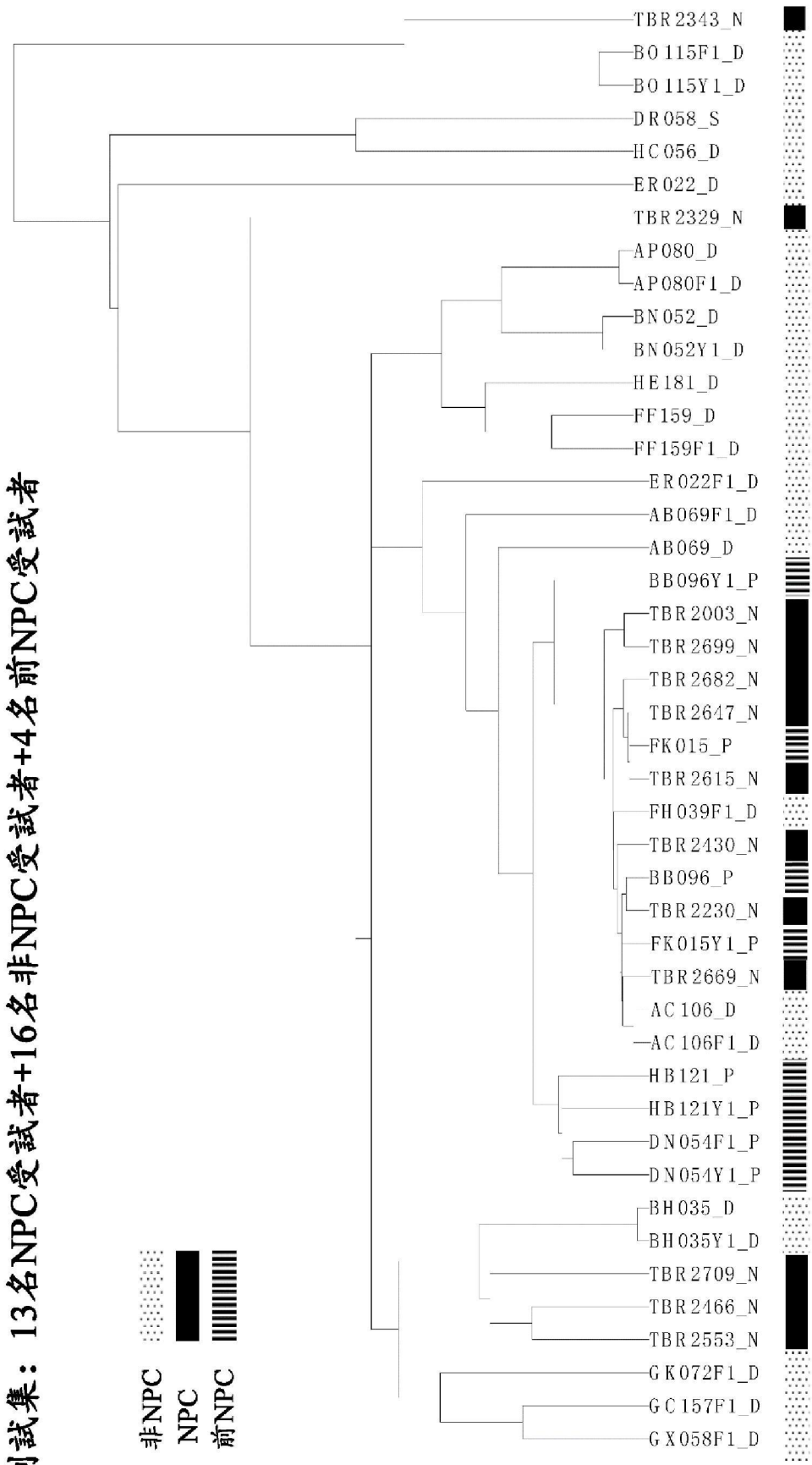


【圖 4】

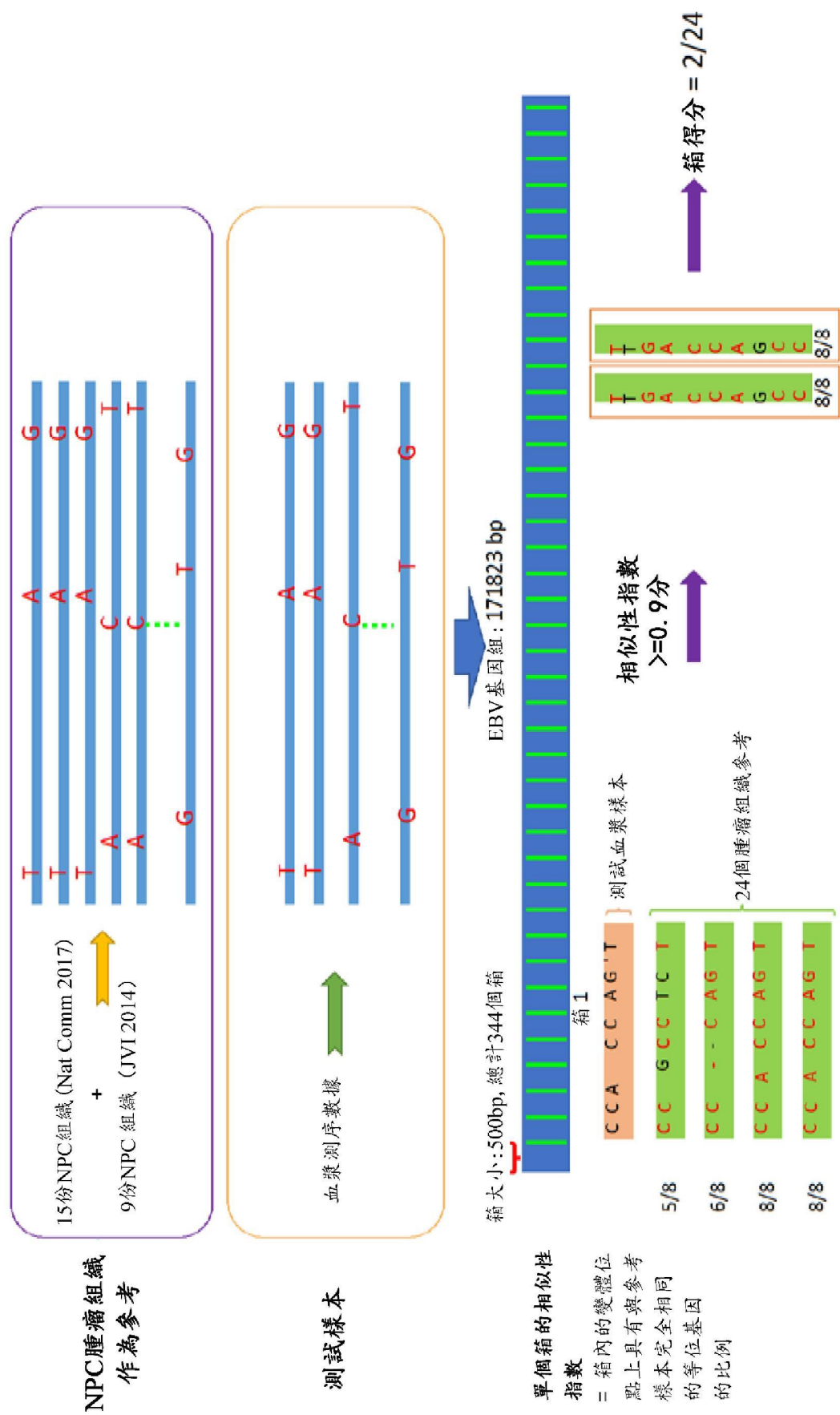
測試集：13名NPC受試者+16名非NPC受試者+4名前NPC受試者



【圖 5】



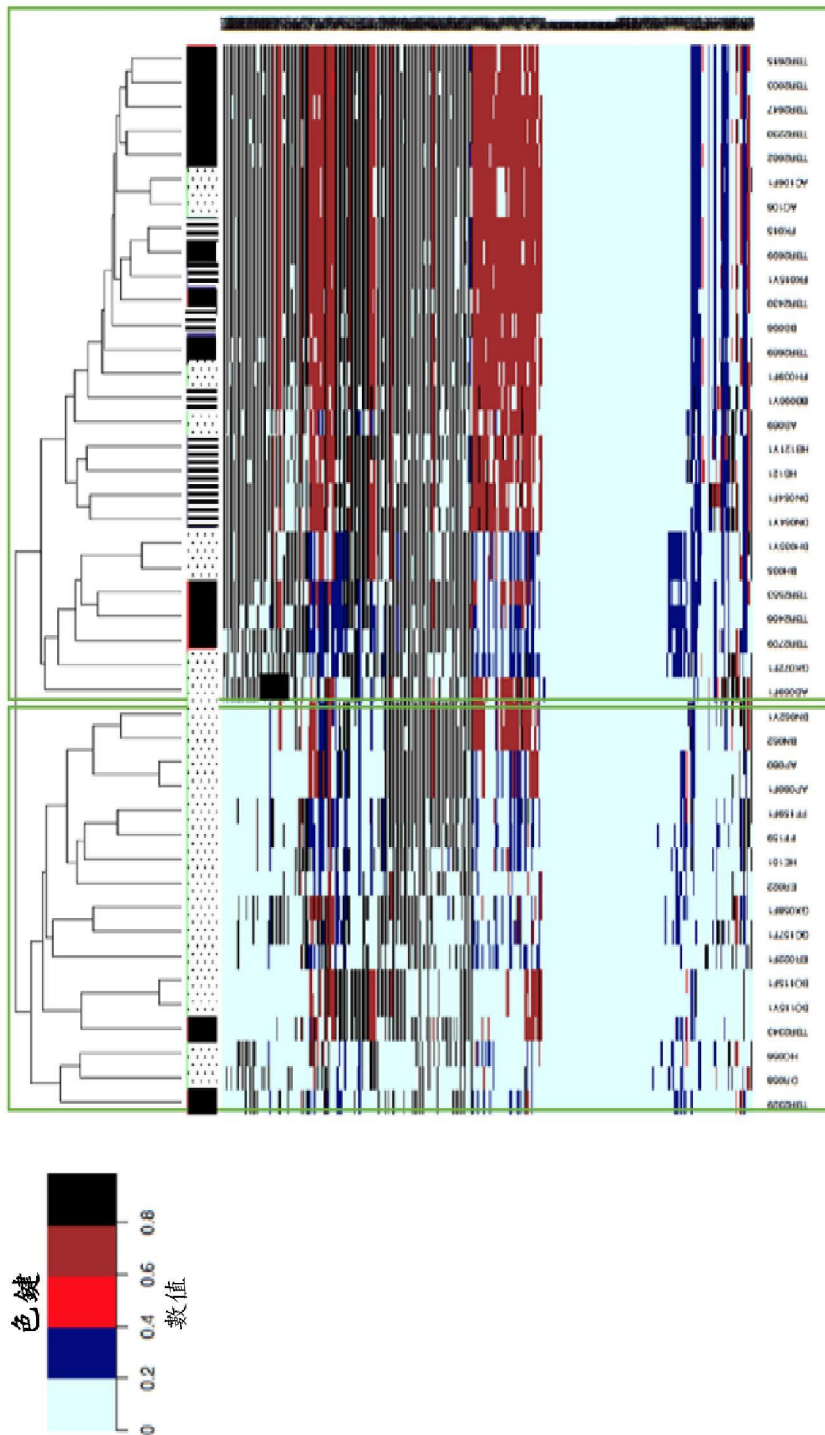
【圖6】



【圖 7】

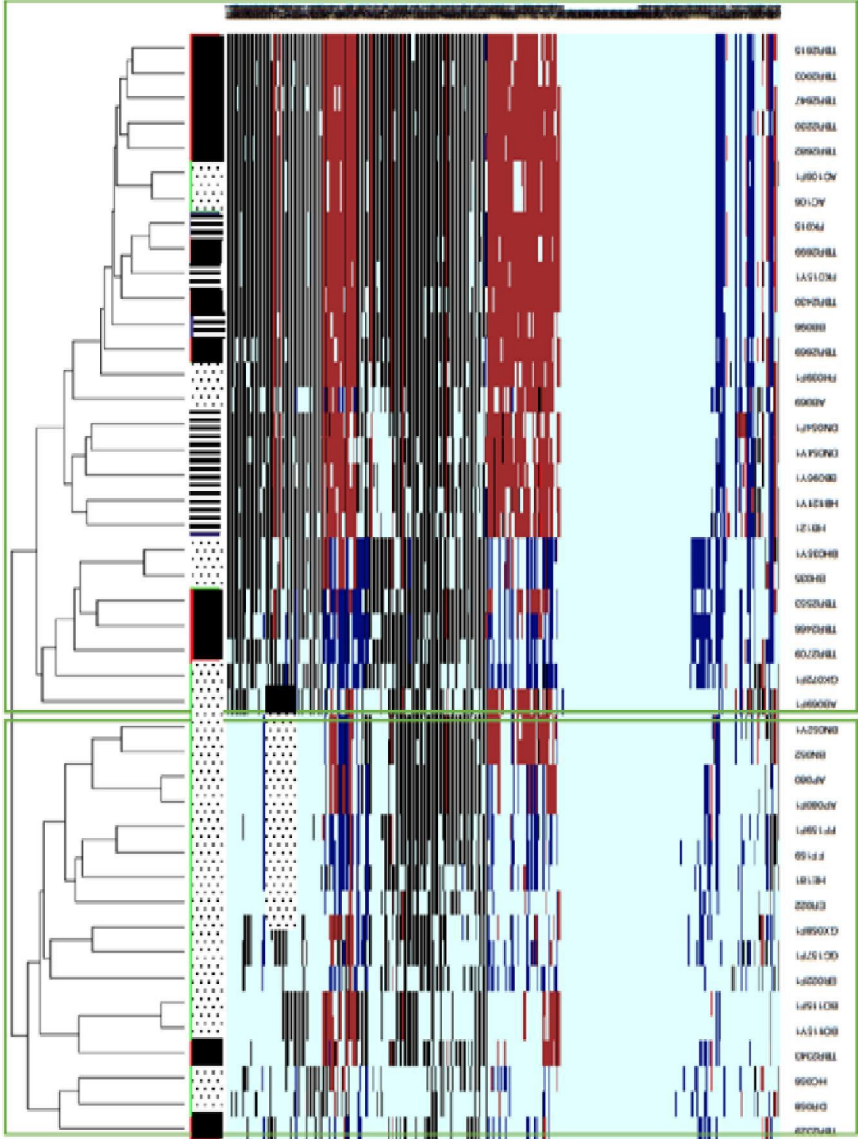
測試集：13名NPC受試者+16名非NPC受試者+4名前NPC受試者

非NPC
NPC
前NPC

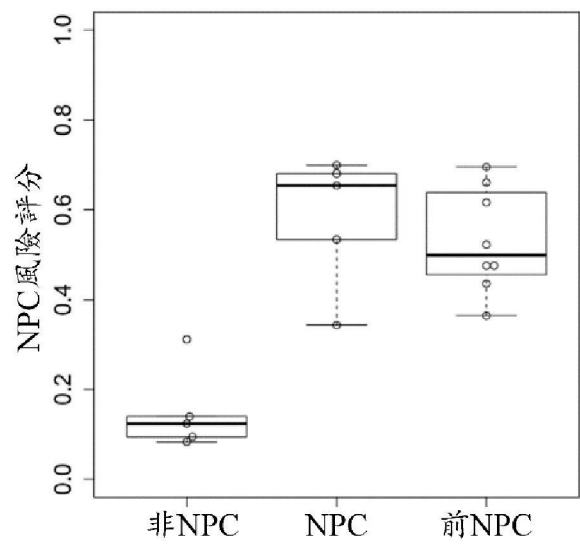


測試集：13名NPC受試者+16名非NPC受試者+4名前NPC受試者

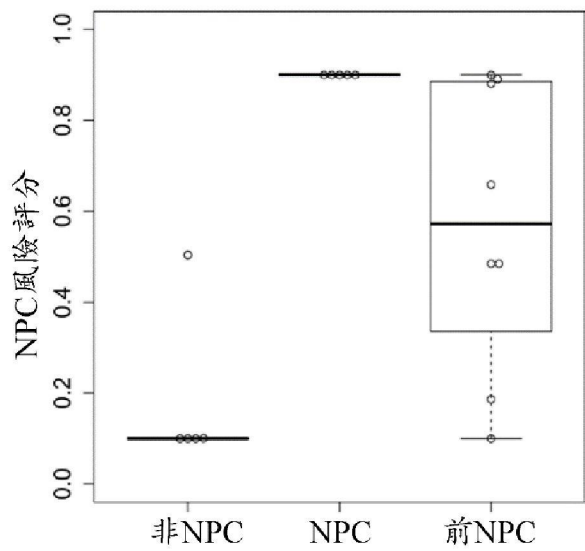
非NPC
NPC
前NPC



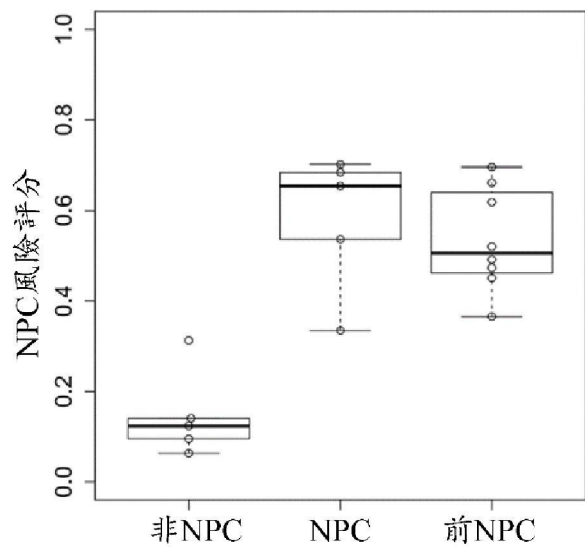
【圖9】



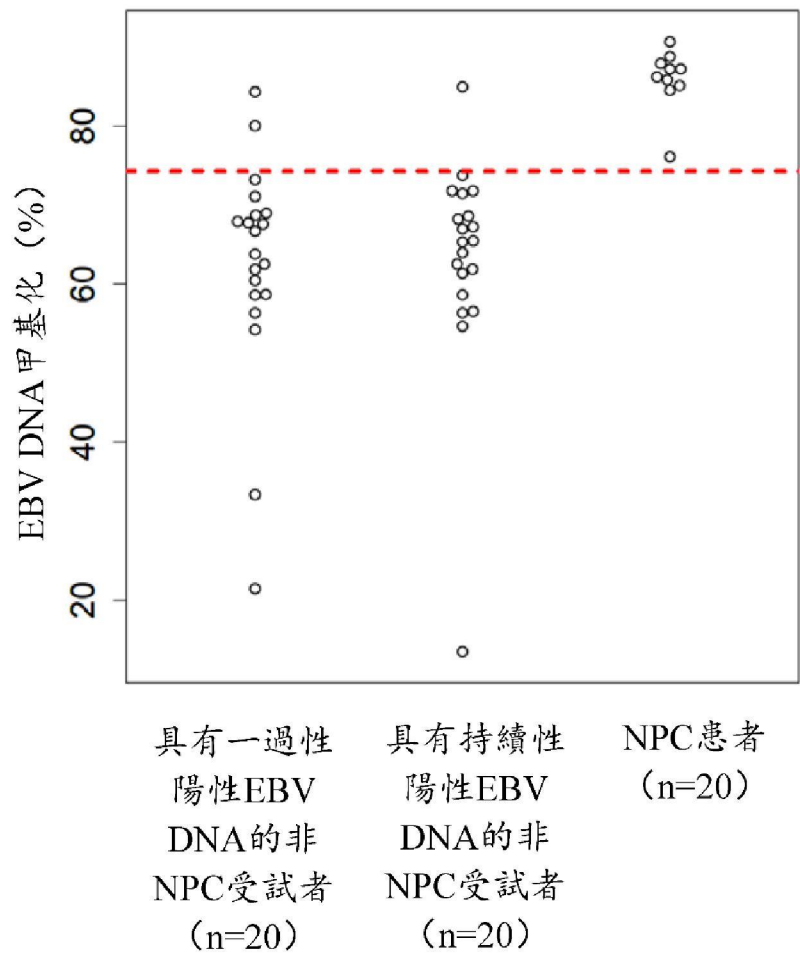
【圖 10A】



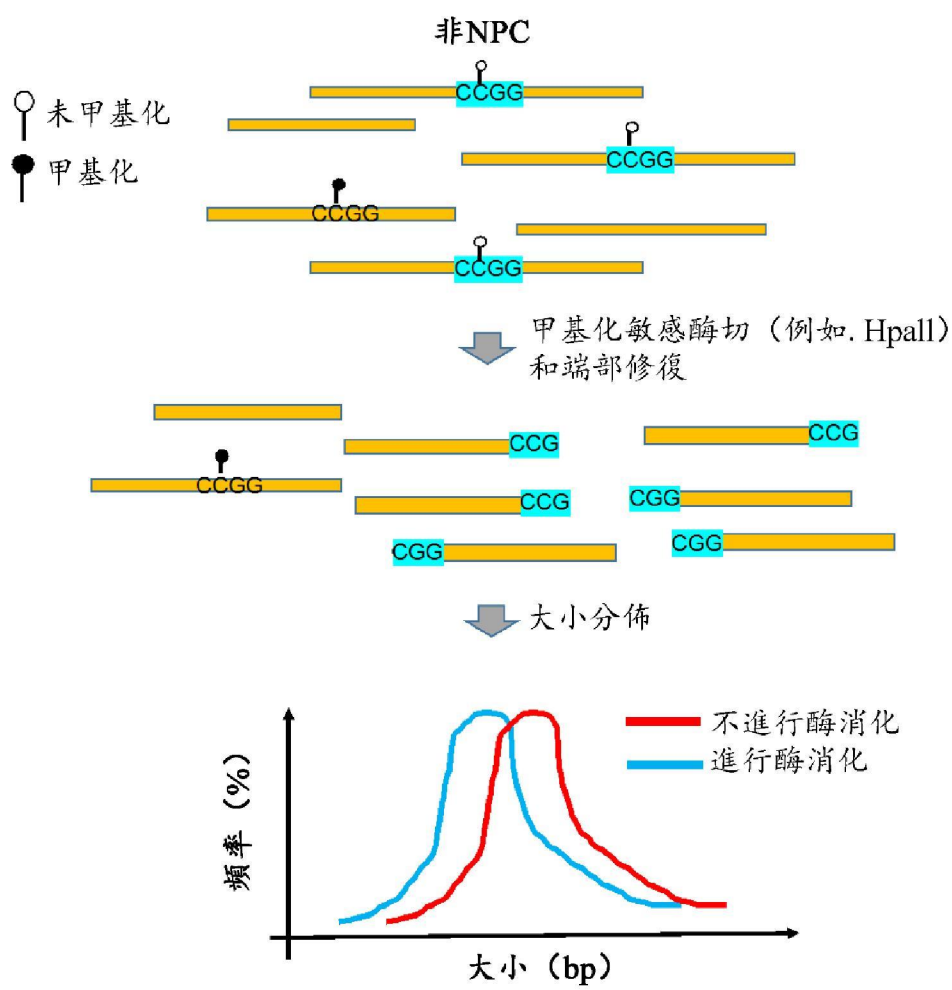
【圖 10B】



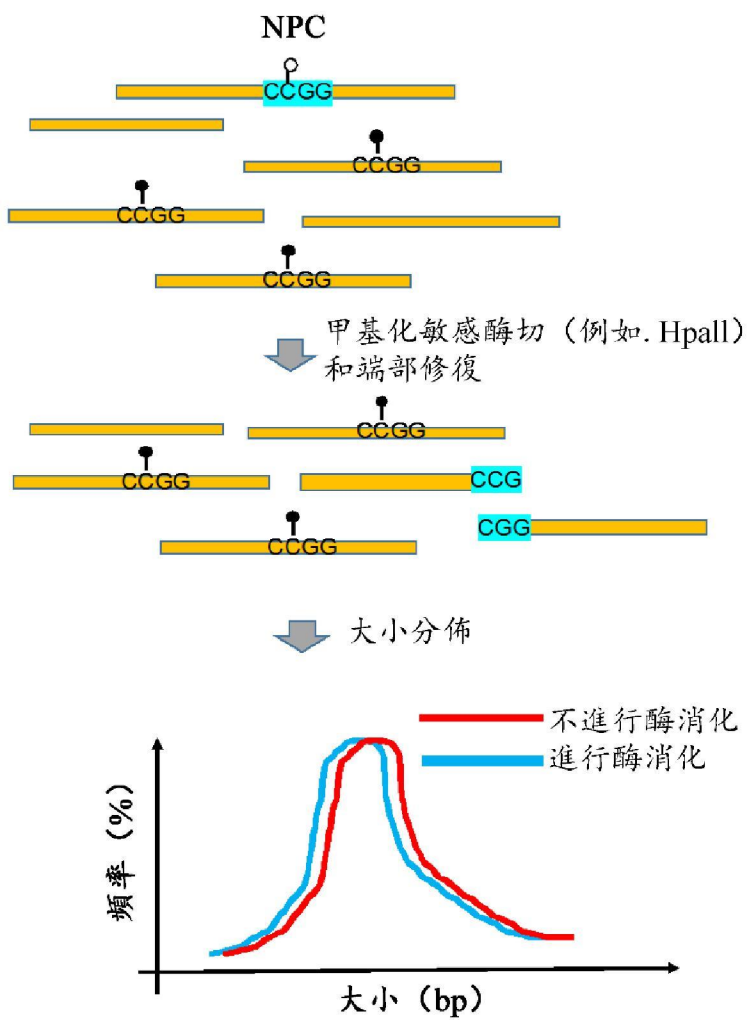
【圖 10C】



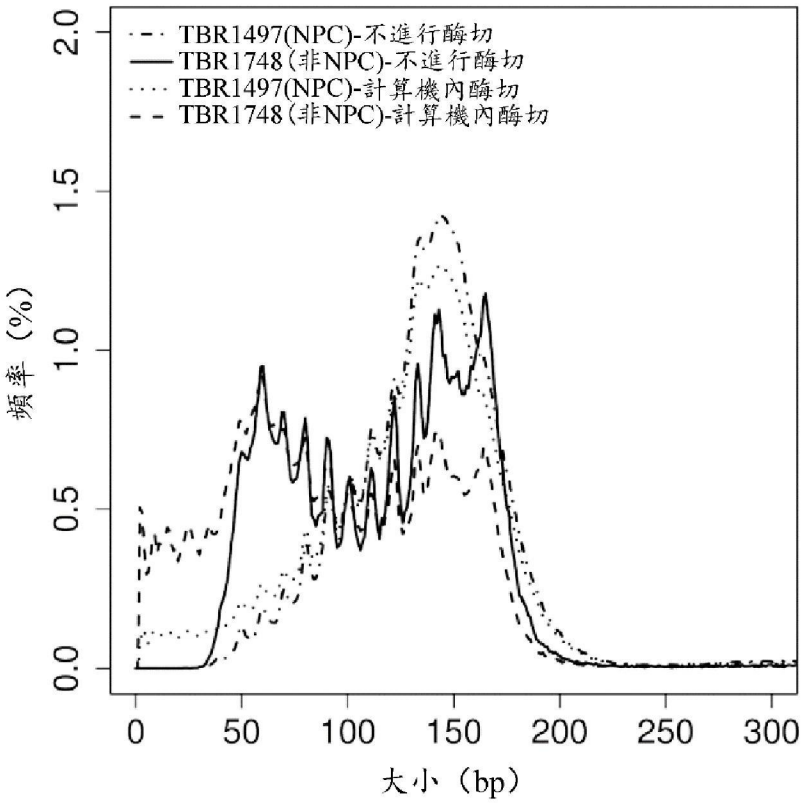
【圖 11】



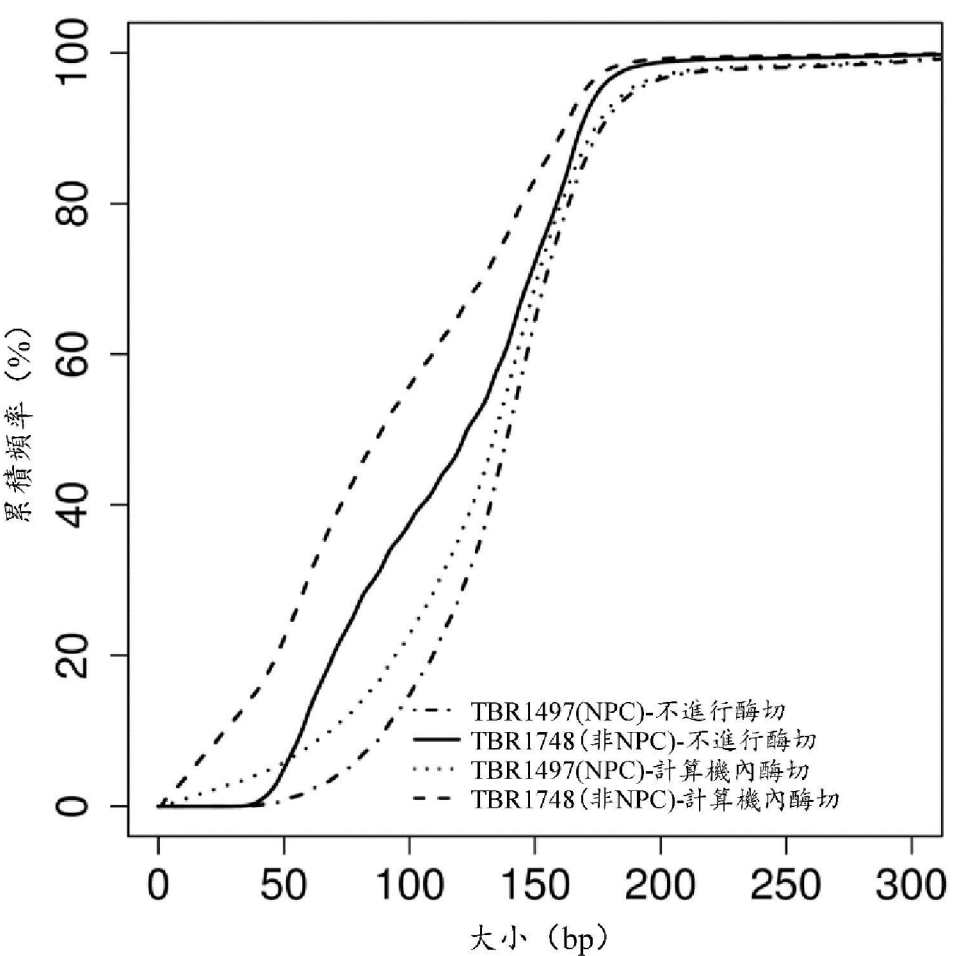
【圖 12】



【圖 13】



【圖 14】



【圖 15】

來自與EBV基因組比對的測試樣本
的血漿EBV DNA讀數

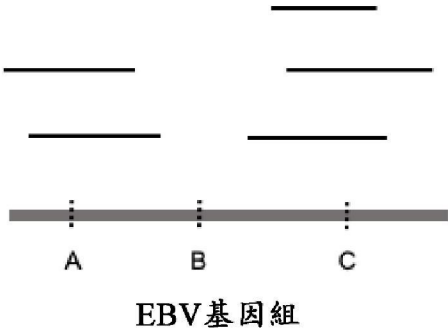


圖 16A

訓練模型

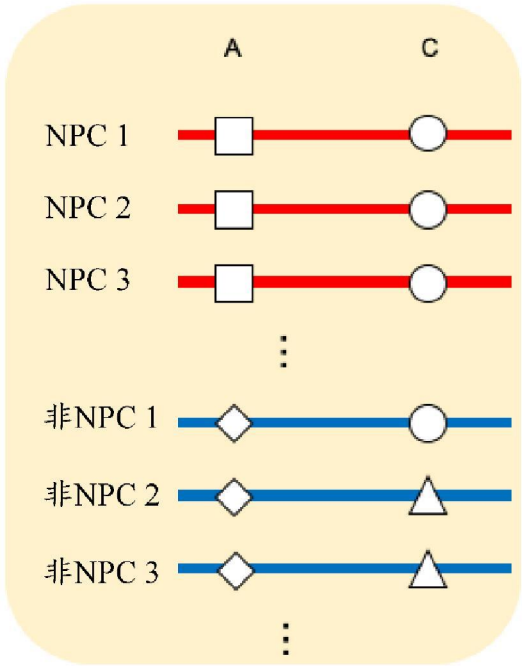
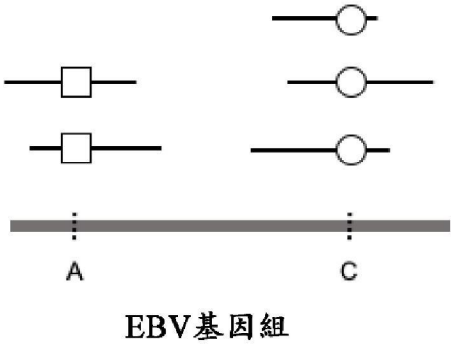
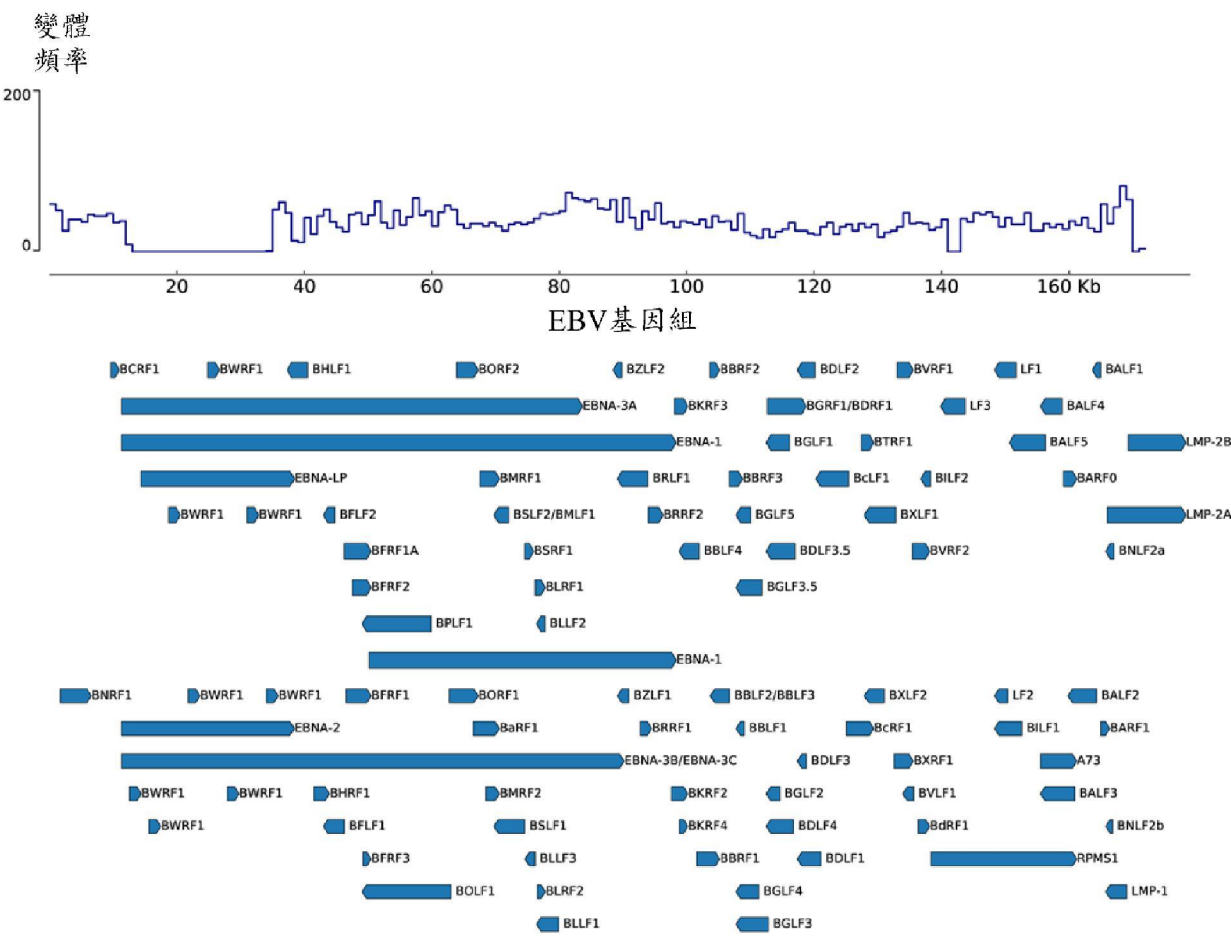


圖 16B

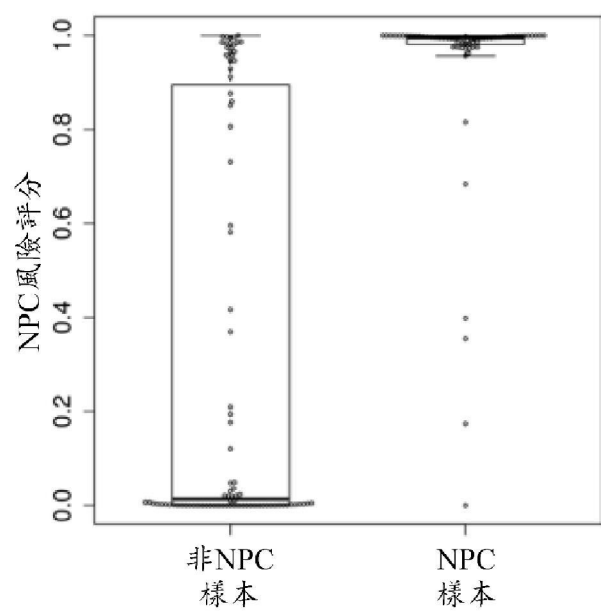


➡ NPC風險評分

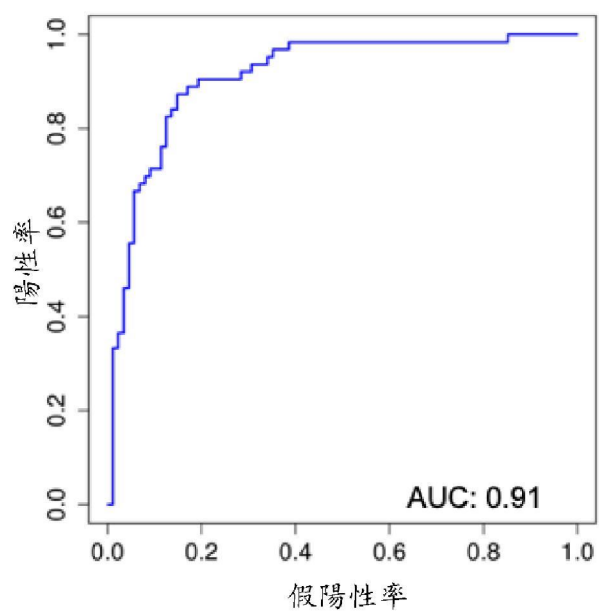
【圖 16C】



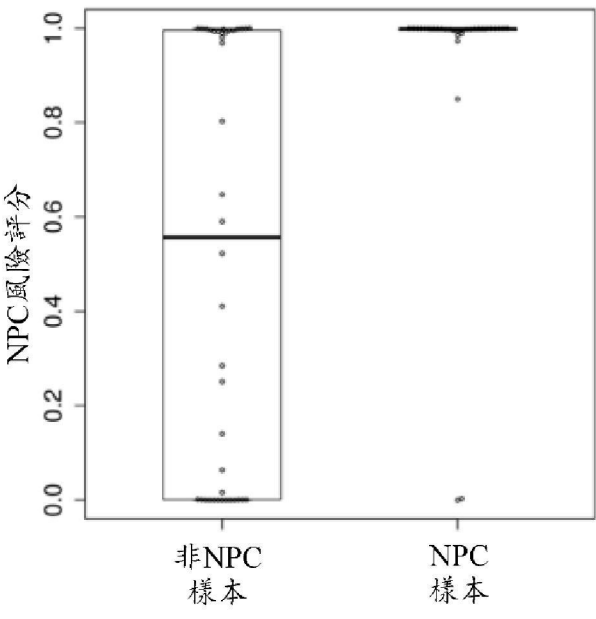
【圖 16D】



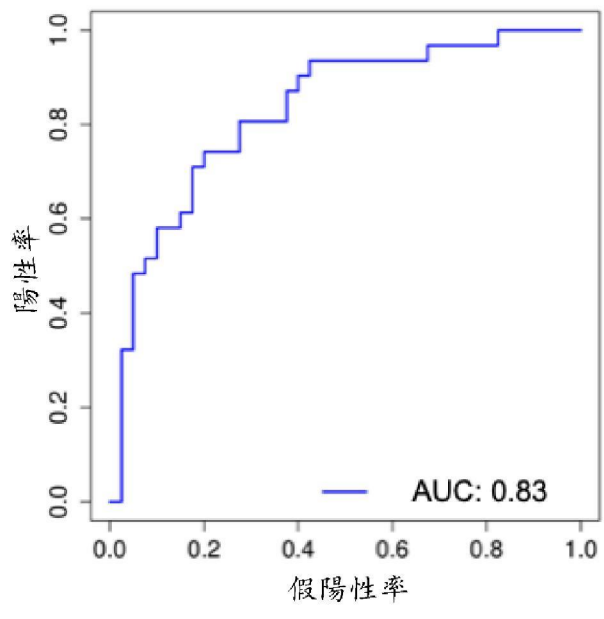
【圖 17A】



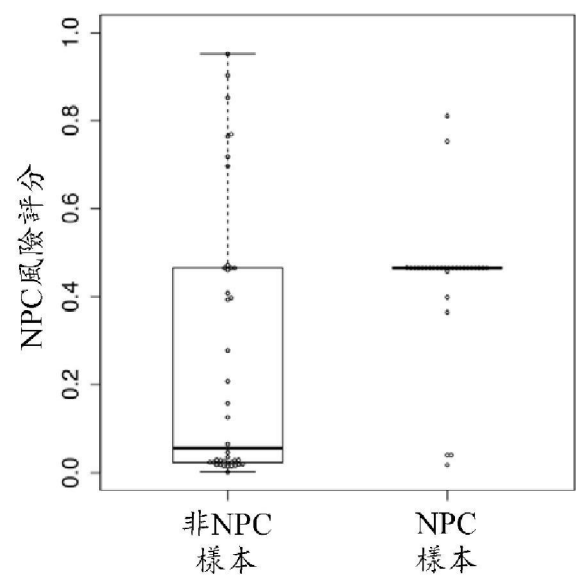
【圖 17B】



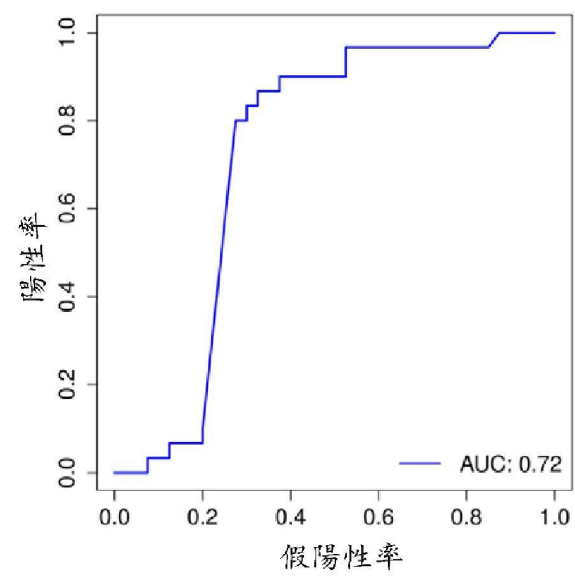
【圖 18A】



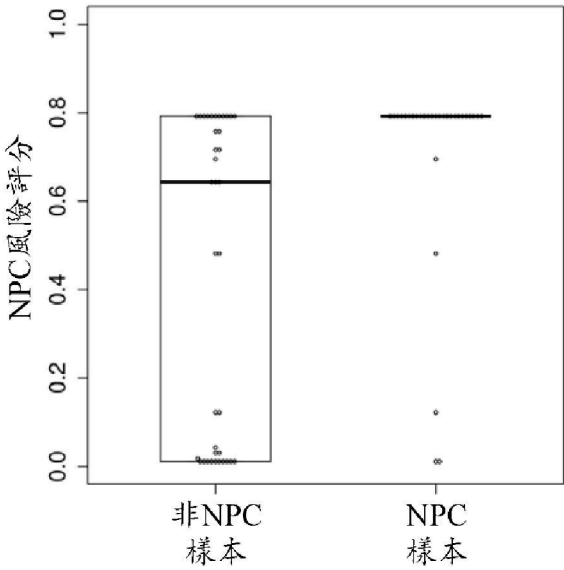
【圖 18B】



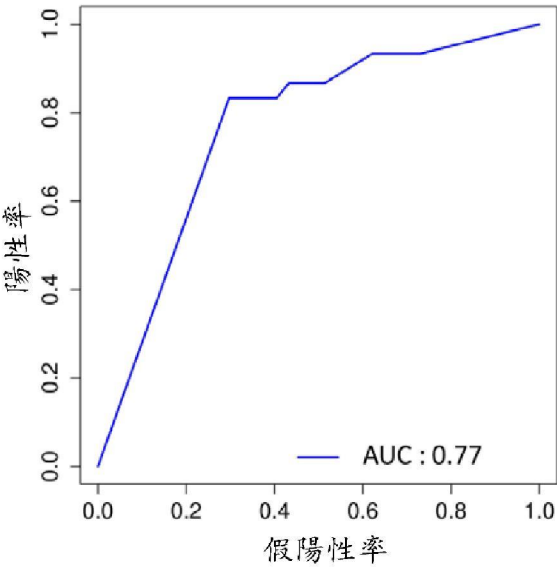
【圖 19A】



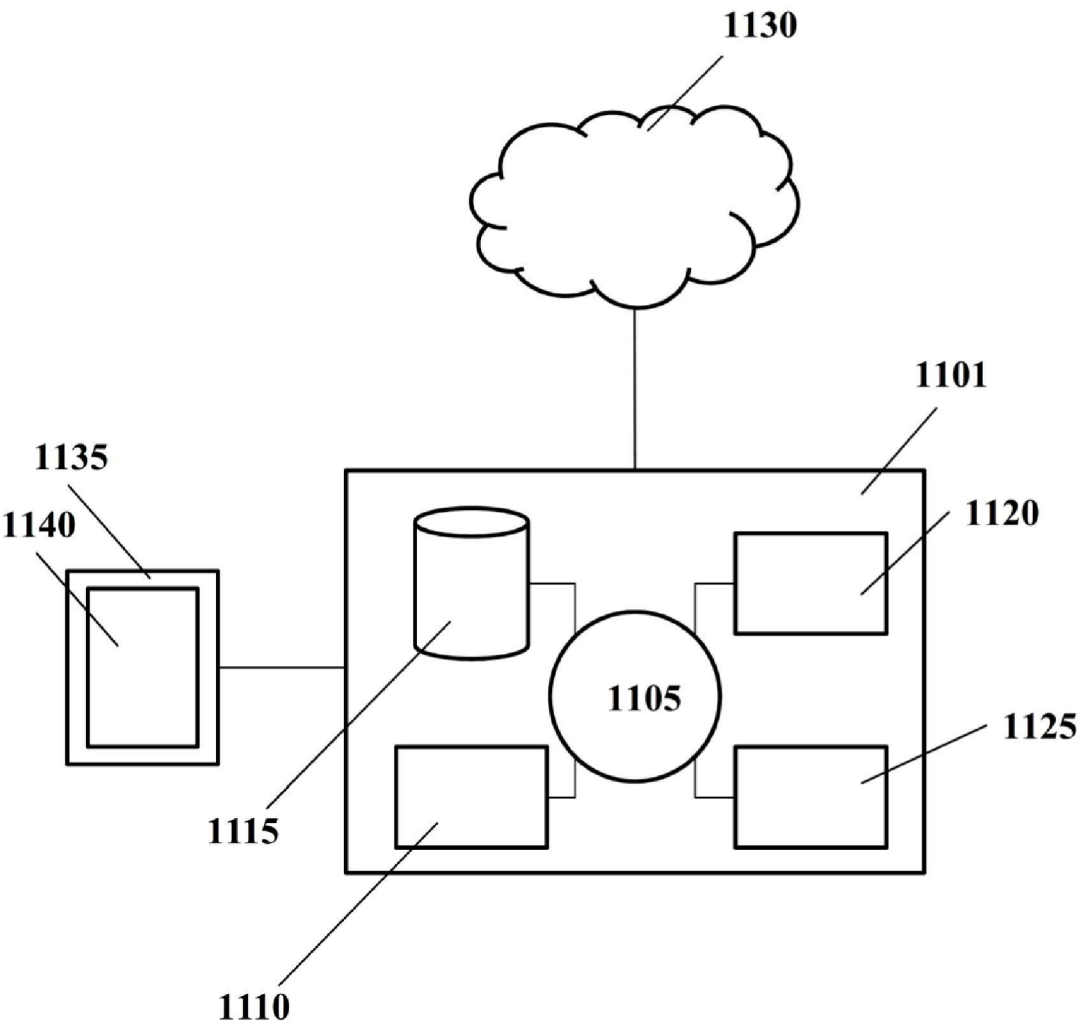
【圖 19B】



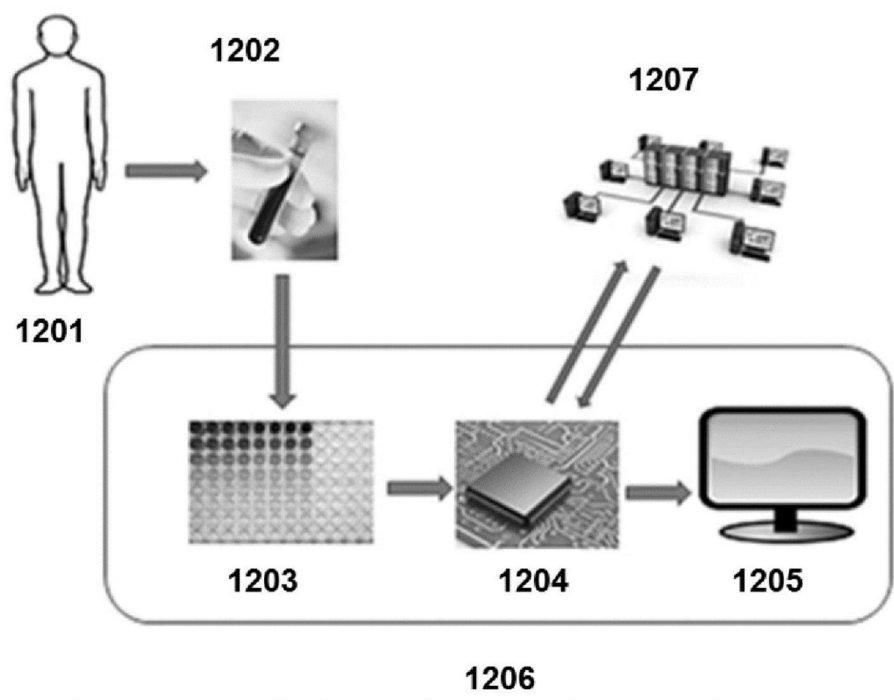
【圖 20A】



【圖 20B】



【圖 21】



【圖 22】