



(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 I852936 B

(45)公告日：中華民國 113 (2024) 年 08 月 21 日

(21)申請案號：108122976

(22)申請日：中華民國 108 (2019) 年 06 月 28 日

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

C12Q1/70 (2006.01)

G16B30/10 (2019.01)

(30)優先權：2018/06/29 美國

62/691,890

(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 (CN)；林 偉棋 LAM, WAI KEI (HK)；張海強 ZHANG, HAIQIANG (CN)

(74)代理人：陳長文

(56)參考文獻：

CN 105874082A

期刊 Tennakoon, C., Sung, W.K. _Fast, sensitive and accurate detection of virus integrations_ 18 (Suppl 3), 71 BMC Bioinformatics 2017/03/14 pages 101-111.

審查人員：張維纓

申請專利範圍項數：19 項 圖式數：13 共 103 頁

(54)名稱

核酸重組及整合分析

(57)摘要

本文提供鑑別嵌合核酸片段，例如生物體-病原體嵌合核酸片段及染色體重組嵌合核酸片段之方法及系統。本文亦提供涉及測定生物樣本中之病原體整合型態或染色體重組及至少部分地基於生物樣本中之病原體整合型態或染色體重組來決定病理學分類的方法及系統。在本發明之某些態樣中，分析來自生物樣本之無細胞核酸分子。

Provided herein are methods and systems for identifying chimeric nucleic acid fragments, e.g., organism-pathogen chimeric nucleic acid fragments and chromosomal rearrangement chimeric nucleic acid fragments. Also provided herein are methods and systems relating to determining a pathogen integration profile or a chromosomal rearrangement in a biological sample and determining a classification of pathology based at least in part on a pathogen integration profile or a chromosomal rearrangement in a biological sample. In certain aspects of the present disclosure, cell-free nucleic acid molecules from a biological sample are analyzed.

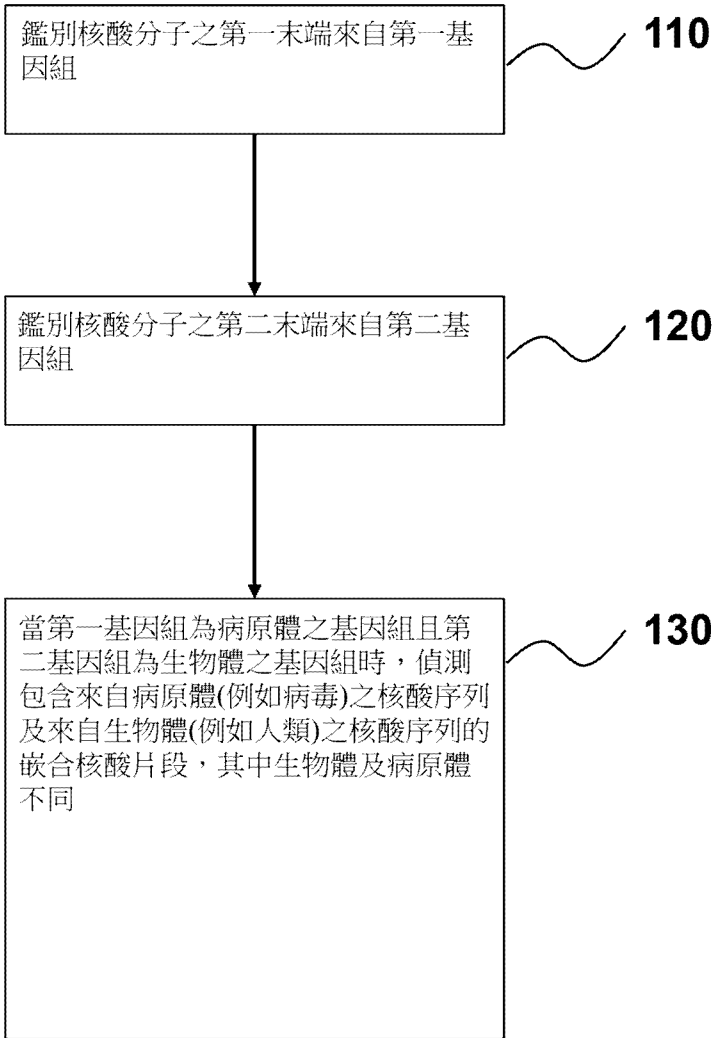
指定代表圖：

符號簡單說明：

110:步驟

120:步驟

130:步驟



【圖1】



公告本

I852936

【發明摘要】

【中文發明名稱】

核酸重組及整合分析

【英文發明名稱】

NUCLEIC ACID REARRANGEMENT AND INTEGRATION
ANALYSIS

【中文】

本文提供鑑別嵌合核酸片段，例如生物體-病原體嵌合核酸片段及染色體重組嵌合核酸片段之方法及系統。本文亦提供涉及測定生物樣本中之病原體整合型態或染色體重組及至少部分地基於生物樣本中之病原體整合型態或染色體重組來決定病理學分類的方法及系統。在本發明之某些態樣中，分析來自生物樣本之無細胞核酸分子。

【英文】

Provided herein are methods and systems for identifying chimeric nucleic acid fragments, e.g., organism-pathogen chimeric nucleic acid fragments and chromosomal rearrangement chimeric nucleic acid fragments. Also provided herein are methods and systems relating to determining a pathogen integration profile or a chromosomal rearrangement in a biological sample and determining a classification of pathology based at least in part on a pathogen integration profile or a chromosomal rearrangement in a biological sample. In certain aspects of the present disclosure, cell-free nucleic acid molecules from a biological sample are analyzed.

【指定代表圖】

圖1

【代表圖之符號簡單說明】

110	步驟
120	步驟
130	步驟

【發明說明書】

【中文發明名稱】

核酸重組及整合分析

【英文發明名稱】

NUCLEIC ACID REARRANGEMENT AND INTEGRATION
ANALYSIS

【技術領域】

【先前技術】

【0001】多種疾病及病況可能與病原體(例如病毒)核酸染色體重組或整合至宿主生物體(例如人類)基因組中相關。舉例而言，染色體間及染色體內重組，諸如易位可能與癌症相關。此外，病毒可能與約20%人類癌症病例相關，且約70%子宮頸癌症及癌症前期子宮頸損傷可能起因於兩種人類乳頭狀瘤病毒(HPV)類型(16及18)感染。目前，對於大部分實體腫瘤，主要使用經由組織切片獲得之腫瘤組織樣本偵測染色體重組或病原體整合，其可能涉及大針、內窺鏡或開放手術-且可為侵襲性、有風險、成本高且疼痛的。需要使用例如來自生物樣本之無細胞核酸分子鑑別染色體重排及核酸整合事件之改良方法、系統及電腦可讀取媒體。

【發明內容】

【0002】在一個態樣中，本文描述一種分析生物體之生物樣本以決定病理學分類之方法，該方法包含：(a)分析複數個來自生物樣本之無細胞核酸分子以鑑別生物體-病原體嵌合核酸片段，其中分析複數個無細胞核酸分子中之每一者包含：鑑別各別無細胞核酸分子之第一末端來自第一基因組，鑑別各別無細胞核酸分子之第二末端來自第二基因組，且當第一

基因組為病原體之基因組且第二基因組為生物體之基因組時，鑑別生物體-病原體嵌合核酸片段，其中生物體及病原體不同；及(b)至少部分地基於生物體-病原體嵌合核酸片段來決定病理學分類。

【0003】 在一些情況下，鑑別第一末端來自第一基因組包含獲得各別無細胞核酸分子之序列讀段且將序列讀段之第一末端之至少一部分與病原體之參考基因組進行比對。在一些情況下，鑑別第二末端來自第二基因組包含獲得各別無細胞核酸分子之序列讀段且將序列讀段之第二末端之至少一部分與生物體之參考基因組進行比對。在一些情況下，分析複數個無細胞核酸分子中之每一者包含獲得各別無細胞核酸分子之序列讀段，且當序列讀段之第一末端之至少一部分與病原體之參考基因組比對且序列讀段之第二末端之至少一部分與生物體之參考基因組比對時，鑑別生物體-病原體嵌合核酸片段。在一些實施例中，該方法進一步包含藉由配對末端測序獲得複數個無細胞核酸分子之序列讀段，且其中配對末端測序產生複數個無細胞核酸分子中之每一者之成對序列讀段。在一些情況下，成對序列讀段包含各別無細胞核酸分子之第一末端之第一序列讀段及各別無細胞核酸分子之第二末端之第二序列讀段。在一些情況下，鑑別生物體-病原體嵌合核酸片段包含將第一序列讀段或其部分與病原體之參考基因組進行比對及將第二序列讀段或其部分與生物體之參考基因組進行比對。在一些情況下，鑑別生物體-病原體嵌合核酸片段包含將第一序列讀段之至少20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95或100個連續核苷酸與病原體之參考基因組進行比對且將第二序列讀段之至少20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95或100個連續核苷酸與生物體之參考基因組進行比對。在一些情況下，

鑑別生物體-病原體嵌合核酸片段包含將第一序列讀段之至少50個連續核苷酸與病原體之參考基因組進行比對且將第二序列讀段之至少50個連續核苷酸與生物體之參考基因組進行比對。在一些情況下，該方法進一步包含基於來自生物樣本之生物體-病原體嵌合核酸片段之量測定病原體整合指數；及至少部分地基於病原體整合指數來決定病理學分類。在一些情況下，測定病原體整合指數包含測定包含來自病原體之基因組之第一末端及來自病原體之基因組之第二末端的複數個無細胞核酸分子之量。在一些情況下，測定病原體整合指數包含將生物體-病原體嵌合核酸片段之量與包含來自病原體之基因組之第一末端及來自病原體之基因組之第二末端的複數個無細胞核酸分子之量進行比較。在一些情況下，比較包含測定生物體-病原體嵌合核酸片段之量與包含來自病原體之基因組之第一末端及來自病原體之基因組之第二末端的複數個無細胞核酸分子之量的比率。在一些情況下，鑑別生物體-病原體嵌合核酸片段進一步包含分析複數個來自生物樣本之無細胞核酸分子之擴增反應。在一些情況下，擴增反應包含與病原體之基因組中之第一靶序列互補的第一引子及與生物體之基因組中之第二靶序列互補的第二引子。在一些情況下，擴增反應包含聚合酶鏈反應(PCR)。在一些情況下，該方法進一步包含分析擴增反應產生之擴增子之序列。在一些情況下，病原體包含病毒。在一些情況下，病毒包含艾伯斯坦-巴爾病毒(Epstein-Barr Virus) DNA、人類乳頭狀瘤病毒DNA、B型肝炎病毒DNA、C型肝炎病毒核酸或其片段。在一些情況下，病毒為人類乳頭狀瘤病毒。在一些情況下，病理學分類包含癌症之存在。在一些情況下，癌症選自由以下組成之群：膀胱癌、骨癌、腦瘤、乳癌、子宮頸癌、結腸直腸癌、食道癌、胃腸癌、造血惡性病、頭頸部鱗狀細胞癌、白血

病、肝癌、肺癌、淋巴瘤、骨髓瘤、鼻癌、鼻咽癌、口腔癌、口咽癌、卵巢癌、前列腺癌、肉瘤、胃癌及甲狀腺癌。在一些情況下，病理學分類包含癌症類型。在一些情況下，癌症類型包含子宮頸癌或頭頸部鱗狀細胞癌。在一些情況下，複數個無細胞核酸分子包含去氧核糖核酸分子。在一些情況下，生物體為動物。在一些情況下，動物為哺乳動物。在一些情況下，哺乳動物為人類。在一些情況下，生物樣本為血漿、血清或尿液。在一些情況下，生物樣本為血漿。在一些情況下，分析複數個來自生物樣本之無細胞核酸分子藉由電腦系統進行。

【0004】 在一個態樣中，本文描述一種鑑別來自生物體之生物樣本之生物體-病原體嵌合無細胞核片段的方法，該方法包含：**(a)**測定無細胞核酸分子之第一末端來自第一基因組且無細胞核酸分子之第二末端來自第二基因組；及**(b)**當第一基因組為病原體之基因組且第二基因組為生物體之基因組時，鑑別生物體-病原體無細胞核酸片段，其中生物體及病原體不同。

【0005】 在一些情況下，測定第一末端來自第一基因組包含獲得無細胞核酸分子之序列讀段且將序列讀段之第一末端之至少一部分與病原體之參考基因組進行比對。在一些情況下，測定第二末端來自第二基因組包含獲得無細胞核酸分子之序列讀段且將序列讀段之第二末端之至少一部分與生物體之參考基因組進行比對。在一些情況下，該方法進一步包含獲得無細胞核酸分子之序列讀段，且當序列讀段之第一末端之至少一部分與病原體之參考基因組配對且序列讀段之第二末端之至少一部分與生物體之參考基因組比對時，鑑別生物體-病原體嵌合核酸片段。在一些情況下，該方法進一步包含藉由配對末端測序獲得無細胞核酸分子之序列讀段，且其

中配對末端測序產生無細胞核酸分子之成對序列讀段。在一些情況下，成對序列讀段包含無細胞核酸分子之第一末端之第一序列讀段及無細胞核酸分子之第二末端之第二序列讀段。在一些情況下，鑑別生物體-病原體嵌合核酸片段包含將第一序列讀段或其部分與病原體之參考基因組進行比對且將第二序列讀段或其部分與生物體之參考基因組進行比對。在一些情況下，鑑別生物體-病原體嵌合核酸片段包含將第一序列讀段之至少20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95或100個連續核苷酸與病原體之參考基因組進行比對且將第二序列讀段之至少20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95或100個連續核苷酸與生物體之參考基因組進行比對。在一些情況下，鑑別生物體-病原體嵌合核酸片段包含將第一序列讀段之至少50個連續核苷酸與病原體之參考基因組進行比對且將第二序列讀段之至少50個連續核苷酸與生物體之參考基因組進行比對。在一些情況下，鑑別生物體-病原體嵌合核酸片段進一步包含分析來自生物樣本之無細胞核酸分子之擴增反應。在一些情況下，擴增反應包含與病原體之基因組中之第一靶序列互補的第一引子及與生物體之基因組中之第二靶序列互補的第二引子。在一些情況下，擴增反應包含聚合酶鏈反應(PCR)。在一些情況下，該方法進一步包含分析擴增反應產生之擴增子之序列。在一些情況下，病原體包含病毒。在一些情況下，病毒包含艾伯斯坦-巴爾病毒DNA、人類乳頭狀瘤病毒DNA、B型肝炎病毒DNA、C型肝炎病毒核酸或其片段。在一些情況下，病毒為人類乳頭狀瘤病毒。在一些情況下，無細胞核酸分子為去氧核糖核酸。在一些情況下，生物體為動物。在一些情況下，動物為哺乳動物。在一些情況下，哺乳動物為人類。在一些情況下，生物樣本為血漿、

血清或尿液。在一些情況下，生物樣本為血漿。在一些情況下，測定無細胞核酸分子之第一末端來自第一基因組且無細胞核酸分子之第二末端來自第二基因組藉由電腦系統進行。

【0006】 在一個態樣中，本文描述一種分析來自生物體之生物樣本之無細胞核酸分子以測定病理學之類型的方法，該方法包含：(a)分析來自生物樣本之無細胞核酸分子以測定病原體整合型態，該病原體整合型態包含整合在生物體之基因組中之病原體之基因組中之斷點位置；及(b)基於病原體整合型態測定病理學之類型。

【0007】 在一些情況下，當無細胞核酸分子包含來自病原體之基因組序列及來自生物體之基因組序列時，藉由偵測來自生物樣本之無細胞核酸分子中之生物體-病原體嵌合核酸片段來測定病原體整合型態。在一些情況下，偵測包含鑑別無細胞核酸分子之第一末端來自病原體之基因組，且鑑別無細胞核酸分子之第二末端來自生物體之基因組。在一些情況下，偵測包含自來自生物樣本之無細胞核酸分子獲得序列讀段，且分析序列讀段以偵測生物體-病原體嵌合核酸片段。在一些情況下，藉由無細胞核酸分子之配對末端測序獲得序列讀段，且其中配對末端測序產生無細胞核酸分子中之每一者之成對序列讀段。在一些情況下，成對序列讀段包含各別無細胞核酸分子之第一末端之第一序列讀段及各別無細胞核酸分子之第二末端之第二序列讀段。在一些情況下，偵測生物體-病原體嵌合核酸片段包含將第一序列讀段之至少20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95或100個連續核苷酸與病原體之參考基因組進行比對且將第二序列讀段之至少20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95或100個連續核苷酸與生物體之參考

基因組進行比對。在一些情況下，測定病原體整合型態進一步包含：基於生物體-病原體嵌合核酸片段之量測定病原體整合指數。在一些情況下，偵測生物體-病原體嵌合核酸片段包含分析來自生物樣本之無細胞核酸分子之擴增反應。在一些情況下，擴增反應包含與病原體之基因組中之第一靶序列互補的第一引子及與生物體之基因組中之第二靶序列互補的第二引子。在一些情況下，擴增反應包含聚合酶鏈反應(PCR)。在一些情況下，該方法進一步包含分析擴增反應產生之擴增子之序列。在一些情況下，病原體包含病毒。在一些情況下，病毒包含艾伯斯坦-巴爾病毒DNA、人類乳頭狀瘤病毒DNA、B型肝炎病毒DNA、C型肝炎病毒核酸或其片段。在一些情況下，病毒為人類乳頭狀瘤病毒。在一些情況下，病理學分類包含癌症類型。在一些情況下，癌症類型包含子宮頸癌或頭頸部鱗狀細胞癌。在一些情況下，無細胞核酸分子包含去氧核糖核酸分子。在一些情況下，生物體為動物。在一些情況下，動物為哺乳動物。在一些情況下，哺乳動物為人類。在一些情況下，生物樣本為血漿、血清或尿液。在一些情況下，生物樣本為血漿。在一些情況下，分析無細胞核酸分子藉由電腦系統進行。

【0008】 在一個態樣中，本文描述一種分析來自生物體之生物樣本之無細胞核酸分子以測定病原體整合型態的方法，該方法包含分析來自生物樣本之無細胞核酸分子以測定病原體整合型態，該病原體整合型態包含整合在生物體之基因組中之病原體之基因組中之整合斷點位置。

【0009】 在一些情況下，當無細胞核酸分子包含來自病原體之基因組序列及來自生物體之基因組序列時，藉由偵測來自生物樣本之無細胞核酸分子中之生物體-病原體嵌合核酸片段來測定病原體整合型態。在一些

情況下，偵測包含鑑別無細胞核酸分子之第一末端來自病原體之基因組，且鑑別無細胞核酸分子之第二末端來自生物體之基因組。在一些情況下，偵測包含自來自生物樣本之無細胞核酸分子獲得序列讀段，且分析序列讀段以偵測生物體-病原體嵌合核酸片段。在一些情況下，藉由無細胞核酸分子之配對末端測序獲得序列讀段，且其中配對末端測序產生無細胞核酸分子中之每一者之成對序列讀段。在一些情況下，成對序列讀段包含各別無細胞核酸分子之第一末端之第一序列讀段及各別無細胞核酸分子之第二末端之第二序列讀段。在一些情況下，偵測生物體-病原體嵌合核酸片段包含將第一序列讀段之至少20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95或100個連續核苷酸與病原體之參考基因組進行比對且將第二序列讀段之至少20、25、30、35、40、45、50、55、60、65、70、75、80、85、90、95或100個連續核苷酸與生物體之參考基因組進行比對。在一些情況下，測定病原體整合型態進一步包含：基於生物體-病原體嵌合核酸片段之量測定病原體整合指數。在一些情況下，偵測生物體-病原體嵌合核酸片段包含分析來自生物樣本之無細胞核酸分子之擴增反應。在一些情況下，擴增反應包含與病原體之基因組中之第一靶序列互補的第一引子及與生物體之基因組中之第二靶序列互補的第二引子。在一些情況下，擴增反應包含聚合酶鏈反應(PCR)。在一些情況下，該方法進一步包含分析擴增反應產生之擴增子之序列。在一些情況下，病原體包含病毒。在一些情況下，病毒包含艾伯斯坦-巴爾病毒DNA、人類乳頭狀瘤病毒DNA、B型肝炎病毒DNA、C型肝炎病毒核酸或其片段。在一些情況下，病毒為人類乳頭狀瘤病毒。在一些情況下，無細胞核酸分子包含去氧核糖核酸分子。在一些情況下，生物體為動物。在一些情況下，

動物為哺乳動物。在一些情況下，哺乳動物為人類。在一些情況下，生物樣本為血漿、血清或尿液。在一些情況下，生物樣本為血漿。在一些情況下，分析無細胞核酸分子藉由電腦系統進行。在一些情況下，分析包含對無細胞核酸分子進行測序。在一些情況下，分析包含：鑑別來自包含相同潛在整合斷點之生物樣本之無細胞核酸分子之序列讀段；及基於序列讀段偵測整合斷點。在一些情況下，基於序列讀段偵測整合斷點包含：評定與側接潛在整合斷點之基因組區域比對之序列讀段中之每一者之序列長度變化；及基於評定，偵測生物體中之整合斷點。在一些情況下，分析包含：鑑別來自包含相同潛在整合斷點之生物樣本之無細胞核酸分子之成對序列讀段；及基於成對序列讀段偵測整合斷點。在一些情況下，基於成對序列讀段偵測整合斷點包含：測定成對序列讀段中之每一者之第一序列讀段及第二序列讀段之股取向；濾除包含與大部分成對序列讀段之第一序列讀段及第二序列讀段之股取向不一致的第一序列讀段及第二序列讀段之股取向的成對序列讀段；及在濾除之後，基於成對序列讀段偵測整合斷點。

【0010】 在一些情況下，基於成對序列讀段偵測整合斷點包含：評定與側接潛在整合斷點之基因組區域比對之成對序列讀段之序列讀段之序列長度變化；及基於評定，偵測整合斷點。在一些情況下，鑑別成對序列讀段包含：鑑別由來自包含與生物體之參考基因組比對之第一序列讀段及與病原體之參考基因組比對之第二序列讀段的生物樣本之無細胞核酸分子之配對末端測序產生的生物體-病原體嵌合成對序列讀段，由此鑑別A型生物體-病原體嵌合成對序列讀段；自A型生物體-病原體嵌合成對序列讀段分出一組A型生物體-病原體嵌合成對序列讀段，其包含在生物體之參考基因組中之預定距離內重疊或間隔開之第一序列讀段及在病原體之參考基因

組中之預定距離內重疊或間隔開之第二序列讀段，由此鑑別生物體及病原體之參考基因組中之生物體-病原體候選整合區域；及鑑別由來自包含與生物體-病原體候選整合區域比對之第一序列讀段及包含與生物體之參考基因組比對之第一序列及與病原體之參考基因組比對之第二序列的第二序列讀段之生物樣本之無細胞核酸分子之配對末端測序產生的生物體-病原體嵌合成對序列讀段，由此鑑別B型生物體-病原體嵌合成對序列讀段。在一些情況下，預定距離為至多10、至多50、至多75、至多100、至多120、至多150、至多175、至多200、至多225、至多250、至多275、至多300、至多325、至多350、至多375、至多400、至多425、至多450、至多475或至多500個鹼基。在一些情況下，在生物體-病原體候選整合區域中，存在至少3、4、5、6、7、8、9、10、11、12、15、20、50、100、200、1000、10,000、50,000、100,000或更多個A型生物體-病原體嵌合成對序列讀段。在一些情況下，偵測整合斷點包含：測定A型生物體-病原體嵌合成對序列讀段及B型生物體-病原體嵌合成對序列讀段中之每一者之第一序列讀段及第二序列讀段的股取向；自B型生物體-病原體嵌合成對序列讀段濾除具有與生物體-病原體候選整合區域內大部分A型生物體-病原體嵌合成對序列讀段之第一序列讀段及第二序列讀段之股取向不一致的第一序列讀段及第二序列讀段股取向之B型生物體-病原體嵌合成對序列讀段；及在濾除之後，基於B型生物體-病原體嵌合成對序列讀段偵測整合斷點。在一些情況下，偵測整合斷點包含：測定B型生物體-病原體嵌合成對序列讀段之多樣性評分；其中多樣性評分之計算法為 $\frac{\sigma_1 - \sigma_2}{\max(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1})}$ ，其中 σ_1 為與生物體之參考基因組比對之B型生物體-病原體嵌合成對序列讀段之第一序列長度之標準差，且其中 σ_2 為與病原體之參考基因組比對之B型生物體

-病原體嵌合成對序列讀段之第二序列長度之標準差；及若多樣性評分等於或高於預定截斷值，則基於B型生物體-病原體嵌合成對序列讀段偵測整合斷點。在一些情況下，預定截斷值為至少1、至少1.6、至少2.0、至少2.4、至少2.8、至少3.0、至少3.4、至少3.8、至少4.0、至少4.2、至少4.4、至少4.6、至少4.8、至少5.0、至少5.4、至少5.8、至少6.0、至少6.5、至少7.0、至少8、至少9、至少10、至少20、至少50或至少100。

【0011】 在一個態樣中，本文描述一種分析生物體之生物樣本以偵測染色體重組之方法，該方法包含：鑑別來自包含相同潛在染色體重組序列之生物樣本之無細胞核酸分子之序列讀段；評定與側接潛在染色體重組之基因組區域比對之序列讀段中之每一者之序列長度變化；及基於評定，偵測生物體中之染色體重組。

【0012】 在一些情況下，生物體為人類。在一些情況下，染色體重組包含染色體易位、染色體缺失、染色體反轉或染色體擴增。在一些情況下，鑑別序列讀段包含：鑑別由來自生物樣本且包含序列讀段(其包含潛在染色體重組)之無細胞核酸分子之配對末端測序產生之染色體嵌合成對序列讀段。在一些情況下，鑑別染色體嵌合成對序列讀段包含：鑑別由包含與生物體之參考基因組之第一基因組區域比對之第一序列讀段及與參考基因組之第二基因組區域比對之第二序列讀段的配對末端測序產生之染色體嵌合成對序列讀段，由此鑑別A型染色體嵌合成對序列讀段，其中參考基因組中第一基因組區域及第二基因組區域之相對位置，指示潛在染色體重組；自A型染色體嵌合成對序列讀段分出一組A型染色體嵌合成對序列讀段，其包含在第一基因組區域中之預定距離內重疊或間隔開之第一序列讀段且包含在第二基因組區域中之預定距離內重疊或間隔開之第二序列讀

段，由此鑑別第一基因組區域及第二基因組區域中之候選重組區域；及鑑別由包含與候選整合區域比對之第一序列讀段及包含與第一基因組區域比對之第一序列及與第二基因組區域比對之第二序列的第二序列讀段之配對末端測序產生的染色體嵌合成對序列讀段，由此鑑別B型染色體嵌合成對序列讀段。在一些情況下，評定變化包含：測定B型染色體嵌合成對序列讀段之多樣性評分，其中多樣性評分之計算法為 $\frac{\sigma_1 + \sigma_2}{\max(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1})}$ ，其中 σ_1 為與第一基因組區域比對之B型染色體嵌合成對序列讀段之第一序列長度之標準差，且其中 σ_2 為與第二基因組區域比對之B型染色體嵌合成對序列讀段之第一序列長度之標準差。在一些情況下，若多樣性評分等於或高於預定截斷值，則偵測染色體重組包含基於B型染色體嵌合成對序列讀段偵測染色體重組。在一些情況下，預定截斷值為至少1、至少1.6、至少2.0、至少2.4、至少2.8、至少3.0、至少3.4、至少3.8、至少4.0、至少4.2、至少4.4、至少4.6、至少4.8、至少5.0、至少5.4、至少5.8、至少6.0、至少6.5、至少7.0、至少8、至少9、至少10、至少20、至少50或至少100。在一些情況下，第一基因組區域與第二基因組區域之間的距離為至少140個鹼基、至少180個鹼基、至少250個鹼基、至少350個鹼基、至少450個鹼基、至少550個鹼基、至少750個鹼基、至少900個鹼基、至少1100個鹼基、至少1250個鹼基、至少1800個鹼基、至少2500個鹼基、至少3500個鹼基、至少5500個鹼基、至少7500個鹼基、至少9000個鹼基或至少 10^4 個鹼基。在一些情況下，參考基因組中之第一基因組區域及第二基因組區域之相對5'至3'關係與A型染色體嵌合成對序列讀段中之每一者之各別無細胞核酸分子中之第一序列讀段及第二序列讀段之相對5'至3'關係相反。

【0013】 在一個態樣中，本文描述一種分析生物體之生物樣本以偵

測染色體重組之方法，該方法包含：鑑別來自包含相同潛在染色體重組之生物樣本之無細胞核酸分子之成對序列讀段；測定成對序列讀段中之每一者之第一序列讀段及第二序列讀段之股取向；濾除包含與大部分成對序列讀段之第一序列讀段及第二序列讀段之股取向不一致的第一序列讀段及第二序列讀段之股取向的成對序列讀段；及在濾除之後，基於成對序列讀段偵測生物體中之染色體重組。

【0014】 在一些情況下，生物體為人類。在一些情況下，染色體重組包含染色體易位、染色體缺失、染色體反轉或染色體擴增。在一些情況下，鑑別成對序列讀段包含：鑑別由來自包含與生物體之參考基因組之第一基因組區域比對之第一序列讀段及與參考基因組之第二基因組區域比對之第二序列讀段的生物樣本之無細胞核酸分子之配對末端測序產生的染色體嵌合成對序列讀段，由此鑑別A型染色體嵌合成對序列讀段，其中參考基因組中之第一基因組區域及第二基因組區域之相對位置，指示潛在染色體重組；自A型染色體嵌合成對序列讀段分出一組A型染色體嵌合成對序列讀段，其包含在第一基因組區域中之預定距離內重疊或間隔開之第一序列讀段及在第二基因組區域中之預定距離內重疊或間隔開之第二序列讀段，由此鑑別第一基因組區域及第二基因組區域中之候選重組區域；及鑑別由來自包含與候選重組區域比對之第一序列讀段及包含與第一基因組區域比對之第一序列及與第二基因組區域比對之第二序列的第二序列讀段之生物樣本之無細胞核酸分子之配對末端測序產生的染色體嵌合成對序列讀段，由此鑑別B型染色體嵌合成對序列讀段。在一些情況下，濾除包含：濾除具有與候選重組區域內大部分A型染色體嵌合成對序列讀段之第一序列讀段及第二序列讀段之股取向不一致的第一序列讀段及第二序列讀段之

股取向的B型染色體嵌合成對序列讀段。在一些情況下，第一基因組區域與第二基因組區域之間的距離為至少140個鹼基、至少180個鹼基、至少250個鹼基、至少350個鹼基、至少450個鹼基、至少550個鹼基、至少750個鹼基、至少900個鹼基、至少1100個鹼基、至少1250個鹼基、至少1800個鹼基、至少2500個鹼基、至少3500個鹼基、至少5500個鹼基、至少7500個鹼基、至少9000個鹼基或至少 10^4 個鹼基。在一些情況下，參考基因組中之第一基因組區域及第二基因組區域之相對5'至3'關係與A型染色體嵌合成對序列讀段中之每一者之各別無細胞核酸分子中之第一序列讀段及第二序列讀段之相對5'至3'關係相反。

【0015】 在一個態樣中，本文描述一種包含至少部分地基於藉由本文所描述之任何方法測定之染色體重組來決定病理學分類的方法。

【0016】 在一個態樣中，本文描述一種包含一或多個處理器及非暫時性電腦可讀取媒體之電腦系統，該非暫時性電腦可讀取媒體包含可操作的指令，當由一或多個電腦處理器執行時，可以使得電腦系統執行本文所描述之任何方法。

【0017】 在一個態樣中，本文描述一種非暫時性電腦可讀取媒體，其包含可操作的指令，當由電腦系統之一或多個處理器執行時，可以使得電腦系統執行本文所描述之任何方法。

【圖式簡單說明】

【0018】 本文所描述之新穎型態將在隨附申請專利範圍中細緻闡述。將參考闡述說明性實施例之以下具體實施方式獲得本文所描述之型態及優點之較佳理解，其中利用本文所描述之原理及其附圖：

【0019】 圖1為用於測定包含來自病原體(例如病毒)之核酸序列及來

自宿主生物體(例如人類)之核酸序列之嵌合核酸片段的工作流之示意圖。

【0020】 圖2為說明藉由來自兩種類型之核酸片段A型及B型之配對末端測序產生之成對序列讀段的示意圖。

【0021】 圖3A為人類乳頭狀瘤病毒(HPV)病毒DNA整合分析之例示性工作流的示意圖。

【0022】 圖3B為用於推導人類-HPV候選整合區域之例示性工作流的示意圖。

【0023】 圖4展示表明基於例示性股型之過濾法之示意圖及圖表。

【0024】 圖5為表明具有較低(<4)及較高(≥ 4)多樣性評分之候選整合斷點之示意圖。

【0025】 圖6展示自來自患有子宮頸癌(Ca Cer)及HPV陽性頭頸部鱗狀細胞癌(HNSCC)之患者之樣本的人類-HPV嵌合片段推導之HPV16基因組中之整合斷點之分佈圖案。

【0026】 圖7展示整個自來自患有子宮頸癌(Ca子宮頸-)及HPV陽性HNSCC (HPV+veSCC)之患者之樣本之無細胞DNA片段測定之HPV基因組中之整合指數。

【0027】 圖8展示可經程式化或以其他方式經配置以實施本文所提供之方法的電腦控制系統。

【0028】 圖9展示如本文所揭示之方法及系統之圖式。

【0029】 圖10為說明用於比較例示性算法與模擬數據集上之VIFI之間的偵測比率之工作流的圖式。

【0030】 圖11為概述例示性算法與模擬數據集上之VIFI之間的偵測比率之比較的圖表。

【0031】圖12展示概述藉由例示性算法對來自患有子宮頸癌(CaCx)、頭頸部鱗狀細胞癌(HNSCC)之患者之血漿樣本進行的病毒DNA整合分析之表格。

【0032】圖13展示概述藉由例示性算法對來自患有頭頸部鱗狀細胞癌(HNSCC)之患者之腫瘤組織樣本進行的病毒DNA整合分析之表格。

【實施方式】

交叉引用

【0033】本申請案主張2018年6月29日申請之美國臨時申請案第62/691,890號之權益，該申請案以全文引用之方式併入本文中。

引用併入

【0034】本說明書中所提及之所有公開案、專利及專利申請案均以引用的方式併入本文中，其引用的程度如各單獨的公開案、專利或專利申請經特定及單獨地指示以引用的方式併入一般。

I. 概述

【0035】本文提供用於分析生物體之生物樣本以鑑別包含來自病原體之核酸序列及來自生物體之核酸序列(生物體-病原體嵌合核酸片段)之嵌合核酸片段的方法及系統。該等方法及系統可包含例如藉由電腦系統分析複數個核酸分子，例如來自例如血漿之生物樣本之無細胞核酸(例如DNA)分子以偵測包含來自病原體之核酸序列及來自生物體之核酸序列的嵌合核酸片段。分析複數個核酸分子，例如無細胞核酸(例如DNA)分子中之每一者可包含鑑別各別核酸分子，例如無細胞核酸分子之第一末端來自第一基因組且鑑別各別核酸分子，例如無細胞核酸分子之第二末端來自第二基因組。當第一基因組為病原體(例如病毒，例如人類乳頭狀瘤病毒

(HPV))之基因組且第二基因組為生物體(例如人類)之基因組時，可偵測包含來自病原體之核酸序列及來自生物體(例如宿主生物體)之核酸序列之嵌合核酸片段。該等方法及系統可進一步包含至少部分地基於包含來自病原體(例如病毒)之核酸序列及來自生物體(例如宿主生物體，例如人類)之核酸序列的嵌合核酸片段來決定病理學分類(例如癌症)。癌症分類可為癌症類型，例如子宮頸癌或頭頸部鱗狀細胞癌。

【0036】 本文亦提供例如藉由電腦系統分析來自例如血漿之生物樣本之核酸分子，例如無細胞核酸分子以測定病原體整合型態的方法及系統。病原體整合型態可包含整合在生物體(例如宿主生物體，例如人類)之基因組中之病原體(例如病毒)之基因組中之斷點位置。該等方法及系統可進一步包含基於病原體整合型態測定病理學之類型。病理學之類型可為例如癌症類型，例如子宮頸癌或頭頸部鱗狀細胞癌(HNSCC)，或與癌症風險提高相關之個體狀態，例如子宮頸原位贅瘤或子宮頸上皮內贅瘤。

【0037】 本文亦提供用於偵測染色體重組之方法及系統。在一些情況下，該等方法包含分析來自例如血漿之生物樣本之核酸分子，例如無細胞核酸分子以鑑別包含染色體重組之嵌合核酸片段。該等方法及系統可包含例如藉由電腦系統分析複數個核酸分子，例如來自例如血漿之生物樣本之無細胞核酸(例如DNA)分子以偵測包含染色體重組之嵌合核酸片段。分析複數個核酸分子，例如無細胞核酸(例如DNA)分子中之每一者可包含鑑別各別核酸分子，例如無細胞核酸分子之第一末端來自生物體之參考基因組之第一基因組區域且鑑別各別核酸分子，例如無細胞核酸分子之第二末端來自生物體之參考基因組之第二基因組區域。當生物體之參考基因組中之第一及第二基因組區域之相對位置與各別無細胞核酸分子中之第一末端

及第二末端之相對位置不一致時，可偵測包含染色體重組之嵌合核酸片段。在一些情況下，如本文所描述之方法及系統包含分析包含染色體重組之嵌合核酸片段以偵測染色體重組。在一些情況下，該等方法及系統包含至少部分地基於染色體重組來決定病理學分類。

【0038】 本文所提供之方法可包括提高自包含病原體，例如病毒序列之無細胞樣本採集核酸片段之可能性的步驟。舉例而言，本文所提供之方法可利用用於靶向採集之雜交探針，其涵蓋完整病毒(諸如人類乳頭狀瘤病毒，例如HPV16、HPV18、HPV33；艾伯斯坦-巴爾病毒(EBV)；或B型肝炎病毒(HBV))之病原體(例如病毒)基因組，或完整病原體(例如病毒)基因組的至少99%、至少95%、至少90%或至少85%。使用此類雜交探針來靶向採集可增濃病毒核酸且提高獲得含有宿主(例如人類)及病毒核酸兩者之嵌合核酸片段之機會以便於病毒整合分析。

【0039】 本文所提供之方法可提供增強的敏感性。舉例而言，本文所提供之方法及系統可缺少用於濾除基因組區域(例如人類基因組區域)，諸如重複區域，例如短連續重複序列、短散佈核元件(SINE/Alu)或長末端重複序列/內源性反轉錄病毒(LTR/ERV1)的步驟。消除此類過濾可提高本文所提供之方法之敏感性。

【0040】 本文所提供之方法可提供增強的特異性。舉例而言，跨越宿主(例如人類)及病原體(例如病毒基因組)之成對末端讀段可用於構築可含有病原體(例如病毒)整合斷點之局部參考序列。彼等部分映射成對末端讀段之局部再比對可用於測定局部參考序列中之斷點。具有可相容映射定向或股資訊之成對末端讀段可用於病毒整合之進一步下游分析。可移除與比對(由於例如不相容的映射定向或股資訊)不相容的讀段以降低起因於映

射誤差或其他誤差之偽影。此類分析亦可排除來源於核酸，例如DNA庫製備(參見例如圖4)期間不同分子之交叉接合的彼等嵌合片段。本文所提供之方法可參考所關注區域內大部分讀段之比對來測試生物體-病原體嵌合讀段(亦即，在病原體(例如病毒)核酸序列與宿主(例如人類)核酸之間具有斷點之序列讀段)之比對相容性。舉例而言，若嵌合讀段中之病原體(例如病毒)整合斷點與大部分讀段中存在之股資訊不相容，則可較不確信地稱作此類病原體(例如病毒)整合斷點。

【0041】 本文所提供之方法可利用圍繞在片段長度、片段末端位置及嵌合核酸片段中之斷點相對位置方面具有足夠多樣性之整合斷點分佈之嵌合片段(參見例如圖5)。在一些情況下，對於特定整合斷點，片段長度、片段末端位置及嵌合片段中之斷點相對位置多樣性愈大，整合置信度愈高。

【0042】 在一些情況下，因為病毒游離型核酸(例如DNA)可比宿主(例如人類)核酸(例如DNA)短，由於病毒核酸(例如DNA)整合至宿主(例如人類)基因組中產生之嵌合核酸片段可預期比來源於游離基因體之病毒核酸(例如DNA)長。因此，包括病原體(例如病毒)整合斷點之嵌合核酸片段之平均或中值長度可大於來自游離基因體之其他病原體(例如病毒)核酸(例如DNA)。

【0043】 此外，在一些情況下，本文所提供之方法不包括涉及用於構築可用病毒基因組之種系發生學特徵之集合的數學建模(例如隱式馬爾可夫模型(Hidden Markov Model))之訓練步驟。在一些情況下，本文所提供之方法不包括來自組織樣本之病原體(例如病毒)核酸之分析。本文所提供之方法可利用無細胞生物樣本，例如血漿。核酸，例如血漿中之DNA

可經天然片段化，且可根據血漿核酸來源(例如DNA，例如肝臟DNA、病毒DNA及腫瘤DNA)改變片段化模式。在一些情況下，本文所提供之方法不包括在短讀段測序之前剪切來自無細胞生物樣本之核酸。

II. 工作流

【0044】 本文所提供之系統及方法可用於分析包含來自生物體(例如宿主生物體，例如人類)之基因組之序列及來自病原體(例如病毒，例如HPV)之基因組之序列的無細胞核酸分子。在人類個體，例如具有病毒相關惡性病之患者之血漿中，一些無細胞核酸分子可含有病毒基因組之序列及人類基因組之序列。血漿中之此等無細胞核酸分子可稱為嵌合核酸片段。

【0045】 圖1展示用於測定包含來自病原體之序列及來自生物體(其來自生物體之生物樣本)之序列的嵌合核酸片段之例示性工作流之示意圖。如所描繪，用於測定嵌合核酸片段之本文所提供之方法可包含鑑別各別無細胞核酸分子之第一末端來自第一基因組(110)。該等方法可進一步包含鑑別各別無細胞核酸分子之第二末端來自第二基因組(120)。該等方法可進一步包含當第一基因組為病原體之基因組且第二基因組為生物體之基因組時，偵測嵌合核酸片段，其中生物體及病原體不同(130)。核酸分子可為無細胞核酸分子。核酸分子可來自生物樣本。核酸分子之第一末端可為核酸分子之5'或3'端，而核酸分子之第二末端可為核酸分子之3'或5'端。生物體可為病原體，例如人類之宿主生物體。病原體可為病毒，例如HPV、EBV及HBV。

【0046】 本文亦提供藉由分析來自生物樣本之核酸分子測定病原體整合型態，例如宿主生物體之基因組或病原體之基因組中之一或多個病原

體整合斷點之位置的方法及系統。如圖2中所描繪，在配對末端測序中，可偵測宿主-生物體-病原體(例如人類-病毒)嵌合片段且呈現兩種不同形式之成對序列讀段(獲自相同核酸分子，例如無細胞DNA分子之兩端的成對序列讀段)。配對末端測序資料中之單一核酸片段之讀段對之兩個讀段可稱作配對讀段。一種類型(A型)可指具有可映射至宿主生物體參考基因組(例如人類參考基因組)之第一末端處之一個序列讀段及可映射至病原體參考基因組(例如病毒參考基因組)之第二末端處之另一讀段且反之亦然之無細胞核酸分子(例如無細胞DNA)。

【0047】 在一些情況下，當在完整序列讀段內序列讀段與參考基因組，例如人類參考基因組之特定區域具有至少80%、至少85%、至少90%、至少95%、至少98%、至少99%或100%序列一致性或互補性時，序列讀段映射至，可映射至，或比對至參考基因組。在一些情況下，當在完整序列讀段內序列讀段與參考基因組，例如人類參考基因組之特定區域具有至少80%序列一致性或互補性時，序列讀段映射至，可映射至，或比對至參考基因組。在一些情況下，當序列讀段與參考基因組，例如人類參考基因組之特定區域以不超過20、15、10、9、8、7、6、5、4、3、2或1個錯配或零個錯配一致或互補時，序列讀段映射至，可映射至，或比對至參考基因組。視本文所提供之方法及系統之應用之目的及上下文而定，最大錯配數目或百分比或最小類似性數目或百分比可隨著選擇準則改變。在一些情況下，當序列讀段與參考基因組，例如人類參考基因組之特定區域以不超過2個錯配一致或互補時，序列讀段映射至，可映射至，或比對至參考基因組。其他片段類型(B型)可指具有可映射至宿主生物體(例如人類)基因組之第一末端處之一個序列讀段之一部分及可映射至病原體(例如病

毒)基因組之序列讀段之其他部分(「嵌合讀段」)且具有可映射至宿主生物體基因組或病原體基因組之第二末端處之另一序列讀段的無細胞核酸分子(例如無細胞DNA)。

如圖2中所描繪，B型宿主生物體-病原體(例如人類-病毒)嵌合片段成對序列讀段可具有可映射至宿主生物體基因組(例如人類基因組或病原體基因組(例如病毒基因組)的一個序列讀段及具有可映射至宿主生物體(例如人類)基因組之其一部分及可映射至病原體(例如病毒)基因組之其其他部分的另一序列讀段。在B型成對序列讀段中，精確整合斷點可直接自嵌合讀段鑑別。在一個實例中，對於A型片段之鑑別，可使用具有兩個之最大錯配數目之SOAP算法作為選擇準則，將序列讀段與參考基因組，例如人類參考基因組及病毒參考基因組(例如HPV參考基因組)進行比對。在一個實例中，對於B型片段之鑑別，可使用具有三個之最大錯配數目之Bowtie2及局部比對程序作為選擇準則，將序列讀段與參考基因組，例如人類參考基因組及病毒參考基因組(例如HPV參考基因組)進行比對。

【0048】圖3A展示使用來自血漿之無細胞DNA片段分析整合至宿主生物體(例如人類)基因組之病原體(例如HPV)病毒DNA之例示性工作流之示意圖(310)。可藉由無細胞DNA片段之配對末端測序獲得來自血漿樣本之無細胞DNA片段之序列讀段(320)。在獲得來自血漿樣本之無細胞DNA片段之兩端之成對序列讀段之後，可藉由將一個序列讀段與人類之參考基因組且將其配對讀段與HPV之參考基因組進行比對來鑑別A型片段(330)。可藉由分別以人類及HPV參考基因組中之相鄰座標分組嵌合片段序列讀段(340)來偵測人類及HPV參考基因組中之候選整合區域。B型嵌合片段讀段與涵蓋潛在斷點之一個序列讀段配對且可隨後搜尋完全映射至相

同候選區域之另一序列讀段(350)。可進一步藉由測定A型及B型片段序列讀段之股型且藉由測定各候選整合斷點之B型片段(360)之嵌合讀段之多樣性評分過濾候選整合斷點以便測定整合斷點之可能性。

A. 鑑別生物體-病原體候選整合區域

【0049】 本文所提供之方法及系統可用於測定病原體(例如病毒，例如HPV)之基因組及宿主生物體(例如人類)之基因組中之生物體-病原體候選整合區域。

【0050】 在一些實例中，可藉由搜尋具有來自映射至人類基因組之片段之第一末端之一個序列讀段及來自映射至病原體(例如病毒，例如HPV)基因組之片段之第二末端之另一序列讀段的成對序列讀段來鑑別A型生物體-病原體嵌合片段。A型生物體-病原體嵌合片段序列讀段可分別用於測定宿主生物體及病原體之參考基因組中之候選整合區域。在一些實例中，當嵌合片段讀段重疊或分別在宿主及病原體之參考基因組中彼此鄰接時，可將所有A型嵌合片段讀段分組在一起以鑑別宿主及病原體兩者之參考基因組中之候選整合區域。在一些實例中，具有相同起點及末尾外部座標之嵌合片段可移除，因為其可能疑似為PCR複製品。在一些實例中，在移除推定PCR複製品之後，所有剩餘的A型嵌合片段讀段可隨後一起混合以分別將宿主及病原體之參考基因組中重疊或彼此鄰接之核苷酸座標分組，從而鑑別宿主及病原體兩者之參考基因組中之候選整合區域。在如圖3A中所描繪之實例中，人類基因組上各候選整合區域將具有一或多個HPV基因組上之對應整合區域，且反之亦然。序列讀段之間的距離可為參考基因組上兩個相鄰讀段上之兩個最近核苷酸之間的距離。分別在宿主生物體及病原體參考基因組中之預定距離內相鄰「宿主生物體讀段」(與宿

主生物體之參考基因組比對之序列讀段)或其對應相鄰「病原體讀段」(與病原體之參考基因組比對之序列讀段)可隨後視為屬於一個候選整合區域。預定距離之截斷值可為300個鹼基，如圖3A中所展現。在一個實例中，在一個候選整合區域中，任何兩個最近宿主序列讀段之間或任何兩個最近病原體序列讀段之間的距離不超過300個鹼基。在一個實例中，與其最近讀段之距離相隔超過300個鹼基之宿主序列讀段不包括於相同候選整合區域中。在一個實例中，與其最近讀段之距離相隔超過300個鹼基之病原體序列讀段不包括於相同候選整合區域中。在一些實例中，當測定候選整合區域時，在將A型嵌合片段讀段分組之後，候選斷點側上最外A型片段讀段之邊界可進一步朝向候選斷點擴增，例如擴增100個鹼基、200個鹼基、300個鹼基、400個鹼基、500個鹼基或600個鹼基，以使得在一些情況下，宿主生物體參考基因組、病原體參考基因組或兩者上之候選整合區域可覆蓋候選整合斷點。如上文所描述之擴增可為任何適當數目的鹼基。圖3B說明藉由合併宿主生物體候選整合區域及病原體候選整合區域推導生物體-病原體候選整合區域(例如人類-HPV候選整合區域)之例示性程序。如圖中所示，可分別藉由根據其在人類及HPV之參考基因組中之座標將A型嵌合片段讀段分組來推導人類候選整合區域及HPV候選整合區域。在此實例中，在相同候選整合區域內，任何相鄰人類讀段或HPV讀段之間的預定距離之截斷值為300個鹼基。此外，對於人類及HPV候選整合區域兩者，在此實例中朝向候選整合斷點側使邊界擴增300個鹼基。如圖中所示，在此情況下，300個鹼基擴增使得人類及HPV候選整合區域覆蓋候選斷點。隨後，可藉由沿著候選斷點側邊界合併人類及HPV候選整合區域來推導人類-HPV候選整合區域。在一些實例中，生物體-病原體候選整

合區域內A型嵌合片段成對序列讀段之數目亦可用以評估及測定候選整合區域。

1. 比對至參考基因組之序列讀段

【0051】 可藉由任何適當的生物資訊算法、程序、工具套組或封裝進行來自個體之樣本之無細胞核酸分子之序列讀段比對。舉例而言，吾人可使用短寡核苷酸分析封裝(SOAP)作為應用如本文所提供之方法及系統之比對工具。可用於本文所提供之方法及系統中之短序列讀段分析工具之實例包括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、Genallice MAP、Geneious Assembler、GensearchNGS、GMAP及GSNAP、GNUMAP、HIVE-hexagon、Isaac、LAST、MAQ、mrFAST、mrsFAST、MOM、MOSAİK、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及ZOOM。

【0052】 序列讀段中之多個連續核苷酸(「序列伸長段」)可用於比對至參考基因組以關於比對進行判讀。舉例而言，比對可包含將序列讀段之至少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，例如病原體之參考基因組或宿主生物體之參考基因組進行比對。在一些情況下，如本文所提及之比對可包含將序列讀段之至多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，例如病原體之參考基因組或宿主生物體之參考基因組進行比對。在一些情況下，如本文所提及之比對包含將序列讀段之約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，例如病原體之參考基因組，或宿主生物體之參考基因組進行比對。

【0053】 在一些情況下，當序列伸長段之完整序列讀段與參考基因組(例如人類參考基因組)之特定區域具有至少80%、至少85%、至少90%、至少95%、至少98%、至少99%或100%序列一致性或互補性時，則進行比對判讀。在一些情況下，當序列伸長段之完整序列讀段與參考基因組(例如人類參考基因組)之特定區域具有至少80%序列一致性或互補性時，則進行比對判讀。在一些情況下，當序列伸長段與參考基因組(例如人類參考基因組)之特定區域一致或以不超過20、15、10、9、8、7、6、5、4、3、2或1個錯配或零個錯配互補時，則進行比對判讀。在一些情況

下，當序列伸長段與參考基因組(例如人類參考基因組)之特定區域一致或以不超過2個錯配互補時，則進行比對判讀。視本文所提供之方法及系統之應用之目的及上下文而定，最大錯配數目或百分比或最小類似數目或百分比可隨著選擇準則改變。

2. 參考基因組之類型

【0054】 可將使用本文所提供之方法及系統產生之序列讀段與一或多個參考基因組(諸如病原體之參考基因組及宿主生物體之參考基因組)進行比對。在一些情況下，病原體之參考基因組可為病毒基因組，例如HPV病毒基因組。參考基因組可為共同基因組或特定病毒菌株之參考基因組。在一些情況下，病原體基因組包含RNA，與病原體之參考基因組之比對可包含將序列讀段之序列資訊轉化成RNA序列或將RNA基因組序列轉化成其互補DNA序列。宿主生物體可為人類或非人類動物。人類參考基因組之實例包括GRCh37、GRCh37、NCBI Build 34、NCBI Build 35及NCBI Build 36.1。人類參考基因組可為個別人類之基因組。人類基因組可為共同序列。

3. 整合區域之截斷值

【0055】 在界定候選整合區域時，宿主生物體讀段與宿主生物體讀段之間或病原體讀段與病原體讀段之間的預定距離之截斷值實例包括約10、約15、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120、約125、約130、約135、約140、約145、約150、約155、約160、約165、約170、約175、約180、約185、約190、約195、約200、約205、約210、約215、約220、約225、約230、約235、約240、

約245、約250、約255、約260、約265、約270、約275、約280、約285、約290、約295、約300、約305、約310、約315、約320、約325、約330、約335、約340、約345、約350、約355、約360、約365、約370、約375、約380、約385、約390、約395、約400、約405、約410、約415、約420、約425、約430、約435、約440、約445、約450、約460、約470、約480、約490或約500個鹼基。在界定候選整合區域時，宿主生物體讀段與宿主生物體讀段之間或病原體讀段與病原體讀段之間的預定距離之截斷值可為至少10、至少15、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、至少180、至少185、至少190、至少195、至少200、至少205、至少210、至少215、至少220、至少225、至少230、至少235、至少240、至少245、至少250、至少255、至少260、至少265、至少270、至少275、至少280、至少285、至少290、至少295、至少300、至少305、至少310、至少315、至少320、至少325、至少330、至少335、至少340、至少345、至少350、至少355、至少360、至少365、至少370、至少375、至少380、至少385、至少390、至少395、至少400、至少405、至少410、至少415、至少420、至少425、至少430、至少435、至少440、至少445、至少450、至少460、至少470、至少480、至少490或至少500個鹼基。在一些情況下，在界定候選整合區域時，宿主生物體讀段與宿主生物體讀段之間或病原體讀段與病原體讀段之間的預定距離可為至多10、至多15、至

多20、至多25、至多30、至多35、至多40、至多45、至多50、至多55、至多60、至多65、至多70、至多75、至多80、至多85、至多90、至多95、至多100、至多105、至多110、至多120、至多125、至多130、至多135、至多140、至多145、至多150、至多155、至多160、至多165、至多170、至多175、至多180、至多185、至多190、至多195、至多200、至多205、至多210、至多215、至多220、至多225、至多230、至多235、至多240、至多245、至多250、至多255、至多260、至多265、至多270、至多275、至多280、至多285、至多290、至多295、至多300、至多305、至多310、至多315、至多320、至多325、至多330、至多335、至多340、至多345、至多350、至多355、至多360、至多365、至多370、至多375、至多380、至多385、至多390、至多395、至多400、至多405、至多410、至多415、至多420、至多425、至多430、至多435、至多440、至多445、至多450、至多460、至多470、至多480、至多490或至多500個鹼基。

【0056】 生物體-病原體候選整合區域內嵌合片段成對序列讀段之數目可用於評估及篩選候選整合區域。舉例而言，各生物體-病原體候選整合區域可具有至少兩種生物體-病原體(例如人類-HPV)嵌合片段成對序列讀段，其不為PCR複製品，例如具有相同起點及末尾外部座標之讀段。在一些情況下，各生物體-病原體候選整合區域可具有至少3、4、5、6、7、8、9、10、11、12、15、20、50、100、200、1000、10,000、50,000、100,000或更多個生物體-病原體(例如人類-HPV)嵌合片段成對序列讀段，其不為PCR複製品，例如具有相同起點及末尾外部座標之讀段。

B. 股取向及推導整合斷點

【0057】 生物體-病原體候選整合區域內嵌合片段成對序列讀段之數目可用於評估及篩選候選整合區域。舉例而言，各生物體-病原體候選整合區域可具有至少兩種生物體-病原體(例如人類-HPV)嵌合片段成對序列讀段，其不為PCR複製品，例如具有相同起點及末尾外部座標之讀段。在一些情況下，各生物體-病原體候選整合區域可具有至少3、4、5、6、7、8、9、10、11、12、15、20、50、100、200、1000、10,000、50,000、100,000或更多個生物體-病原體(例如人類-HPV)嵌合片段成對序列讀段，其不為PCR複製品，例如具有相同起點及末尾外部座標之讀段。

【0058】 藉由配對末端測序產生之序列讀段之股型可用於測定整合斷點之可能性。在鑑別潛在候選整合區域之後，生物體-病原體嵌合片段成對序列讀段中之宿主生物體序列讀段及病原體序列讀段之股取向可用於消除活體外(例如藉由活體外接合)產生之生物體-病原體嵌合片段。可基於宿主生物體及病原體序列讀段上游或下游之宿主生物體序列讀段及病原體序列讀段之股取向推導宿主生物體及病原體基因組中之整合區域。可濾出具有相比於用於推導候選整合區域之A型嵌合片段成對序列讀段不一致之定向的嵌合讀段之B型嵌合片段成對序列讀段。當將A型嵌合片段成對序列讀段分組用於測定候選整合區域時，可存在分組在一起之具有不同股取向(或股型模式)之A型成對序列讀段。在一些情況下，給定樣本中可存在多達四種可能的不同股取向：「 $+/+$ 」、「 $-/-$ 」、「 $+/-$ 」、「 $-/+$ 」。其中當序列讀段匹配參考基因組中之序列時，+表示正股，且當序列讀段與參考基因組中之序列互補時，-表示反股；及與宿主生物體基因組比對之序列讀段在與病原體基因組比對序列讀段之前，表現術語如「 $+/+$ 」、「 $-/-$ 」、「 $+/-$ 」及「 $-/+$ 」。在一些情況下，計數具有四種可能的不同股取向中之每一者

之A型嵌合片段成對序列讀段之數目。在一些情況下，具有群組中之大部分股取向之A型嵌合片段成對序列讀段用於推導候選整合區域。在一些情況下，具有大部分股取向之A型嵌合片段成對序列讀段之數目不超過群組中之全部A型嵌合片段成對序列讀段的50%。在一些情況下，具有大部分股取向之A型嵌合片段成對序列讀段之數目超過任何其他股取向。**圖4**展示無細胞DNA片段之成對序列讀段之股取向之例示性分析。在此實例中，分析人類-HPV嵌合成對序列讀段。如**圖4**示意圖中所示，當大部分A型片段成對序列讀段(3/4)具有人類序列讀段之+股及HPV序列讀段之-股時，剩餘的A型片段成對序列讀段(其具有人類序列讀段之+股及HPV序列讀段之+股)可測定為以「假」讀段對形式濾出。可因此基於其股取向篩選B型嵌合片段成對序列讀段。在此實例中，具有與大部分A型片段成對序列讀段之股型不一致的股取向之嵌合讀段之B型成對序列讀段可作為「假」讀段濾出。**圖4**中之表列出候選整合區域中之大部分A型人類-HPV片段成對序列讀段之股取向及相同候選整合區域中之人類-HPV嵌合讀段之對應預期股取向的四種不同情境。

C. 鑑別整合斷點

【0059】 如**圖2**中所例示之B型嵌合片段讀段對或B型嵌合片段讀段對中之嵌合讀段可藉由使用補充有局部比對功能之短序列讀段分析工具偵測。舉例而言，Bowtie2軟體可用於該分析，且局部比對功能可適用於將短核苷酸伸長段(「分段」)與宿主生物體參考基因組或病原體參考基因組進行比對。短核苷酸伸長段可含有至少4、至少5、至少6、至少7、至少8、至少9、至少10、至少12或至少15個鹼基。在一些情況下，短核苷酸伸長段可含有至多50、至多30、至多25、至多24、至多23、至多22、至

多21、至多20、至多19、至多18、至多17、至多16、至多14、至多13、至多11、至多9、至多8、至多7、至多6、至多5或至多4個鹼基。短核苷酸伸長段可含有約4、約5、約6、約7、約8、約9、約10、約11、約12、約15、約18、約20、約22、約24、約25、約30或約50個鹼基。在一些情況下，最小映射伸長段可為4 bp。在一些情況下，可分步進行比對。舉例而言，比對可以相對更長伸長長度，例如約30個鹼基或約22個鹼基起始以使讀段局部映射至人類或HPV之參考基因組，以便鑑別嵌合讀段，其中之每一者具有與人類參考基因組比對之一個部分及與HPV參考基因組比對之其他部分。在鑑別出嵌合讀段後，嵌合讀段之較短「部分」，例如可映射至人類基因組或HPV基因組且比嵌合讀段之另一部分短之嵌合讀段之部分可隨後使用相對較短伸長長度，例如約4個鹼基與人類或HPV之參考基因組進一步比對，以便獲得與參考基因組更精細的比對，由此推導候選整合斷點。短核苷酸伸長長度可視各種參數而定進行調節，該等參數包括(但不限於)病原體及宿主生物體之物種、如擴增、銜接子接合、合成測序、測序深度及序列讀段長度之測序分析之各種步驟之變量以及測定病原體整合斷點之預期敏感性及特異性。在一些情況下，可在一個步驟中使用適當的局部序列比對算法進行比對。舉例而言，可藉由使用基於史密斯-沃特曼(Smith-Waterman)之動態算法進行之其序列之部分與宿主生物體基因組或另一部分與病原體基因組之比對來偵測嵌合讀段。

【0060】 可在候選整合區域內鑑別人類-HPV嵌合讀段以測定潛在整合斷點。在此實例中，整合斷點可定義為HPV DNA針對人類DNA拼接之邊界。如上文所論述，用於構築候選區域之嵌合讀段及A型嵌合片段讀段對之股取向可一致。利用此資訊，可進一步藉由組合候選區域及嵌合讀

段之股資訊來推導潛在斷點之相對位置。圖4中之表進一步展示在何處且如何嵌合片段之定向與具有人類-HPV嵌合讀段之嵌合片段具有一致股取向之例示性情境。亦針對各例示性情形給出斷點對人類或HPV基因組序列之預期相對位置。

【0061】 當鑑別及分析B型嵌合讀段時，若潛在人類-HPV嵌合讀段(配對末端測序中之讀段)之較長部分可映射於候選區域之人類側上，則可在嵌合讀段之較短部分與候選區域之HPV側之間測試比對。類似地，若潛在人類-HPV嵌合讀段(配對末端測序中之讀段)之較長部分可映射於候選區域之HPV側上，則可在嵌合讀段之較短部分與候選區域之人類側之間測試比對。因為嵌合讀段之較短部分之最小映射長度可較短，短映射伸長段偶然可能可映射至人類或HPV基因組上之多個區域。因此，當短序列錨之映射位置在潛在斷點之較短預定距離內時，映射位置可視為推定擊點，且因此，對應讀段可視為涵蓋整合斷點之嵌合讀段。

D. 高度確信整合斷點之篩選

【0062】 對於多個嵌合讀段所覆蓋之斷點，多樣性測試可用於評估在如上文所提及之協定之後鑑別出之整合斷點且濾出潛在假陽性斷點。真斷點可被具有比對至宿主生物體或病原體參考基因組之不同序列長度之多個嵌合片段覆蓋，其因此具有更大多樣性。可藉由檢驗涵蓋整合斷點之嵌合讀段之多樣性，例如比對至宿主生物體或病原體參考基因組之嵌合讀段之序列長度之多樣性來評估潛在整合斷點。用於檢驗嵌合讀段之多樣性之多樣性評分可計算如下：

$$\frac{\sigma_1 + \sigma_2}{\max(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1})},$$

其中

$\sigma 1$ ：與宿主生物體參考基因組比對之嵌合讀段部分之長度之標準差；

$\sigma 2$ ：與病原體參考基因組比對之嵌合讀段部分之長度之標準差。

【0063】 整合斷點可隨後藉由多樣性評分分選。在一些情況下，多樣性評分愈高，所鑑別之整合斷點可為真之可能性愈高(參見例如圖5)。在一些實例中，截止多樣性評分可設定為4，以使得排除具有低於4之多樣性評分之整合斷點，而具有至少4之多樣性評分之剩餘整合斷點可視為來自實際HPV整合事件。在一些情況下，截止多樣性評分設定為約0.1、約0.2、約0.4、約0.6、約0.8、約1、約1.2、約1.4、約1.6、約1.8、約2.0、約2.2、約2.4、約2.5、約2.6、約2.8、約3.0、約3.1、約3.2、約3.3、約3.4、約3.5、約3.6、約3.7、約3.8、約3.9、約4.0、約4.1、約4.2、約4.3、約4.4、約4.5、約4.6、約4.7、約4.8、約4.9、約5.0、約5.2、約5.4、約5.6、約5.8、約6.0、約6.5、約7.0、約7.5、約8、約8.5、約9、約9.5、約10、約20、約30、約40、約50、約60、約80或約100。在一些情況下，截止多樣性評分可為至多0.1、至多0.2、至多0.4、至多0.6、至多0.8、至多1、至多1.2、至多1.4、至多1.6、至多1.8、至多2.0、至多2.2、至多2.4、至多2.5、至多2.6、至多2.8、至多3.0、至多3.1、至多3.2、至多3.3、至多3.4、至多3.5、至多3.6、至多3.7、至多3.8、至多3.9、至多4.0、至多4.1、至多4.2、至多4.3、至多4.4、至多4.5、至多4.6、至多4.7、至多4.8、至多4.9、至多5.0、至多5.2、至多5.4、至多5.6、至多5.8、至多6.0、至多6.5、至多7.0、至多7.5、至多8、至多8.5、至多9、至多9.5或至多10。在其他情況下，截止多樣性評分可為至少1、至少1.2、至少1.4、至少1.6、至少1.8、至少2.0、至少2.2、至少2.4、至少

2.5、至少2.6、至少2.8、至少3.0、至少3.1、至少3.2、至少3.3、至少3.4、至少3.5、至少3.6、至少3.7、至少3.8、至少3.9、至少4.0、至少4.1、至少4.2、至少4.3、至少4.4、至少4.5、至少4.6、至少4.7、至少4.8、至少4.9、至少5.0、至少5.2、至少5.4、至少5.6、至少5.8、至少6.0、至少6.5、至少7.0、至少7.5、至少8、至少8.5、至少9、至少9.5、至少10、至少20、至少30、至少40、至少50、至少60、至少80或至少100。

【0064】如本文所提及之各種截斷值中之一或多者可視多種參數而定進行調節。可調節此等變量以便獲得所需評定結果，例如測定病原體整合斷點之敏感性及特異性。可根據本文所提供之方法調節之各種截斷值包括用於搜尋及鑑別A型嵌合片段成對序列讀段之比對長度、界定整合區域時讀段之間的距離之截斷值、搜尋及鑑別B型嵌合片段序列讀段期間用於比對之短核苷酸分段之長度及截止多樣性評分。此等變量中之一或多者可視多種參數而定進行調節，該等參數包括樣本分析之情境，例如病原體及宿主生物體之物種、待分析癌症類型、宿主生物體之年齡、性別及健康狀況、如擴增、銜接子接合、合成測序、測序深度及序列讀段長度之測序分析之各種步驟之變量以及樣本分析之預期結果，例如測定病原體整合斷點之敏感性及特異性。在一些情況下，可藉由其他分析評估及/或確認可變調節。舉例而言，在患有某種類型之癌症，例如子宮頸癌之患者中，可藉由根據本文所提供之方法及系統分析血漿樣本中之無細胞DNA來測定HPV整合斷點。候選HPV整合斷點之不同列表可經由調節如本文所論述之變量中之一或多者推導出。候選HPV整合斷點之此等列表可隨後與藉由分析來自同一患者之腫瘤組織樣本測定之HPV整合斷點進行比較。在一些

情況下，全基因組測序或其他下一代測序技術可用於此類分析。在一些情況下，可經由比較使變量最佳化以便獲得關於個體中之病原體整合型態的準確資訊。

E. 病原體整合指數

【0065】 如本文所描述之病原體整合型態可包含病原體整合指數。本發明之一些態樣亦提供藉由分析來自生物樣本之核酸分子來測定病原體整合指數之方法及系統。整合指數可為生物樣本中之核酸分子總數目內生物體-病原體嵌合核酸分子之幾率比。在一些情況下，基於如上文所論述之多種嵌合片段成對序列讀段而測定整合指數。舉例而言，可在樣本分析，例如測序分析期間測定嵌合讀段對(例如A型片段讀段對、B型片段讀段對或兩者)之數目。在一些實例中，藉由生物體-病原體嵌合讀段對(例如人類-HPV嵌合讀段對)之數目除以病原體片段讀段對(例如HPV片段讀段對)之數目計算病原體整合指數。HPV片段讀段對可為成對末端成對序列讀段，其兩者映射至病原體(例如HPV)之基因組。

【0066】 在一些實例中，整合指數可包含位置資訊。舉例而言，整合指數可運算病原體之基因組、宿主生物體之基因組或兩者中之位置中之每一者。在一些情況下，對於病原體之基因組中之位置(例如核苷酸座標)，可測定在該位置處終止之生物體-病原體嵌合片段讀段對之數目，且亦可測定在該位置處終止之病原體片段讀段對之數目。可因此藉由在該位置處終止之生物體-病原體嵌合片段讀段對之數目除以在該位置處終止之病原體片段讀段對之數目來計算該位置之整合指數。其他計算方法亦可用於測定整合指數，只要其反映生物樣本中之生物體-病原體嵌合片段之豐度即可。

【0067】 可在基因組譜或基因座特定水準下分析整合指數。在基因座特定水準下，可相對於沿著尺寸至少1 bp之病毒基因組之任何位點計算指數。在一個實例中，整合指數可用於預測癌症產生之風險或可能性。舉例而言，在一些情況下，整合指數愈高，提供樣本之個體愈可能罹患癌症。在另一實例中，病原體整合指數沿著病毒基因組(分析之基因座特定水準)之分佈可在整個不同腫瘤類型中不同。舉例而言，HPV整合指數展示為患有子宮頸癌及HNSCC 之患者之間的差異(圖7)，因此告知腫瘤來源。整合指數之計算可基於例如A型病毒-人類嵌合片段與全部病毒(僅DNA)之比率。在計算整合指數時可或可不應用用於鑑別整合斷點之上述過濾步驟。舉例而言，一端與病毒基因組比對且另一端與人類基因組比對之任何成對末端讀段，例如A型嵌合片段可用於在無任何過濾之情況下計算整合指數。在另一實例中，A型及B型嵌合片段之總數目比例可用於計算整合指數。

【0068】 應理解，當上文所描述之詳細例示性工作流程集中於分析來自生物體-病原體嵌合無細胞核酸片段，尤其人類-HPV嵌合無細胞DNA片段之配對末端測序的序列讀段時，類似概念及方法可應用於使用其他測序技術及不一定僅獲得末端序列資訊之情境，例如當使用如離子激流測序及奈米孔測序之測序技術以便測定整合區域及/或斷點時；類似概念及方法亦可應用於病原體及宿主生物體及/或其他核酸分子(非細胞或細胞)之其他組合有意義之情境。

III. 病原體整合型態

【0069】 病原體整合型態可包含病原體基因組、宿主生物體基因組或兩者中一或多個病原體整合斷點之位置。病原體整合型態可包含病原體

基因組或宿主生物體基因組中一或多個整合斷點之位置。病原體整合型態可用於決定病理學分類，例如癌症或癌前病變之存在或不存在、癌症類型及癌症分期。

【0070】 病原體整合型態可包含一或多個「整合斷點」。整合斷點可為病原體核酸分子(例如病毒DNA)整合至經病原體感染之宿主細胞之基因組(例如人類基因組)中的位點。

【0071】 病原體核酸之整合會破壞宿主生物體基因組之完整性，且在一些情況下，導致宿主細胞中之嚴重下游後果。宿主生物體基因組中之整合斷點可為可轉譯成蛋白質之表現序列。病原體整合會破壞經感染細胞中之蛋白質生產。在一些其他情況下宿主生物體基因組中之整合斷點可在轉錄調節區，例如強化子或啟動子中。病原體核酸序列之整合會導致通常藉由受影響之轉錄調節區調節之基因轉錄失調。在一些情況下，病原體整合可導致轉錄/轉譯橫向、細胞功能變化或正常細胞惡性轉化成癌細胞。病原體核酸分子可為病原體基因組DNA分子。當病原體基因組包含RNA分子時，例如當病原體為RNA病毒，例如反轉錄病毒時，病原體核酸分子可為病原體基因組RNA分子之互補DNA。在經HPV感染之個體之血液中，可存在在具有HPV DNA序列及人類DNA序列之血漿中循環之人類-HPV嵌合DNA片段。在經EBV感染之個體之血液中，可存在具有EBV DNA序列及人類DNA序列之血漿中循環之人類-EBV嵌合DNA片段。

【0072】 病原體整合型態可指受關注之核酸分子中存在或不存在來自病原體基因組之核苷酸序列與來自宿主生物體基因組之核苷酸序列之組合、受關注之核酸分子中來自病原體基因組之核苷酸序列之尺寸或長度、受關注之核酸分子中來自病原體基因組之核苷酸序列之部分貢獻、受關注

之核酸分子中來自病原體基因組之核苷酸序列之序列模式、受關注之核酸分子中病原體-宿主生物體接合點之特性、整合至宿主生物體之基因組中之病原體之基因組中之斷點位置或病原體核酸分子插入於宿主生物體基因組中之宿主生物體之基因組之斷點位置。在一些情況下，病原體整合型態包含整合在宿主生物體之基因組中之病原體之基因組中之斷點位置。

【0073】 在一些實例中，本文所提供之方法係關於分析生物體之生物樣本以測定病原體整合型態。在一些情況下，該等方法可包含分析來自生物樣本之無細胞核酸分子以測定病原體整合型態。病原體整合型態可包含整合在宿主生物體之基因組中之病原體之基因組中之整合斷點位置、宿主生物體之基因組中之整合斷點位置或兩者。一般而言，當來自第一生物體之核酸序列整合至第二生物體之基因組中時，斷點可指來自第一生物體之核酸序列拼接至來自第二生物體之核酸序列的位置。在一些情況下，如本文所論述之斷點可為針對來自宿主生物體之基因組之核酸序列拼接來自病原體之基因組之核酸序列的位置。舉例而言，當病原體基因組整合至宿主基因組中時，病原體之基因組中之斷點可指病原體基因組DNA斷裂且連接宿主生物體基因組DNA之病原體基因組中之位置。在一些情況下，整合過程可涉及病原體基因組中之一個斷點。在其他情況下，整合過程可涉及病原體基因組中之多個斷點，例如病原體基因組DNA破碎成多個片段，例如2、3、4、5、6、7、8、9、10、15、20、100或更多個片段，其分別插入宿主生物體基因組中。在一些情況下，本文所提供之方法及系統實現鑑別來自生物樣本之核酸分子中之一或多個整合斷點。在一些情況下，本文所提供之方法及系統實現鑑別來自生物樣本之核酸分子中之2、3、4、5、6、7、8、9或甚至更多個整合斷點。

IV. 染色體重組

A. 偵測染色體重組

【0074】 在一些態樣中，本發明提供偵測染色體重組之方法。該等方法及系統可包含藉由電腦系統分析複數個核酸分子，例如來自例如血漿之生物樣本之無細胞核酸(例如DNA)分子以偵測包含染色體重組之嵌合核酸片段(染色體重組嵌合核酸片段)。分析複數個核酸分子，例如無細胞核酸(例如DNA)分子中之每一者可包含鑑別各別核酸分子，例如無細胞核酸分子之第一末端來自生物體之參考基因組之第一基因組區域且鑑別各別核酸分子，例如無細胞核酸分子之第二末端來自生物體之參考基因組之第二基因組區域。當生物體之參考基因組中之第一及第二基因組區域之相對位置與各別無細胞核酸分子中之第一末端及第二末端之相對位置不一致時，可偵測包含染色體重組之嵌合核酸片段。

【0075】 相對位置可指參考基因組或無細胞核酸片段中之相對距離或相對5'至3'(或上游-下游)關係。相對距離可指精確相對距離或相對距離範圍。舉例而言，若參考基因組中之第一及第二基因組區域之間的相對距離及無細胞核酸片段中之第一末端及第二末端之間的相對距離在一些情況下恰好相同或在一些其他情況下在相同範圍(例如100-180個鹼基或無細胞核酸片段之典型長度範圍)內，則相對位置可一致。若參考基因組中之第一及第二基因組區域之間的相對距離及無細胞核酸片段中之第一末端及第二末端之間的相對距離在一些情況下不恰好相同或在一些其他情況下不在相同範圍(例如100-180個鹼基或無細胞核酸片段之典型長度範圍)內，則相對位置可不一致。不一致性可視生物體之參考基因組中之第一及第二基因組區域之相對位置之間的不同類型的染色體重組與各別無細胞核酸分子

中之第一末端及第二末端之相對位置不一致而變化。染色體重組可為易位、擴增、缺失、反轉、染色體臂缺失及染色體臂獲得。

【0076】 易位可涉及分別在不同染色體或在相同染色體上兩個先前各別基因組區域之核酸序列之染色體間或染色體內重組。若無細胞核酸片段起因於包涵易位事件之基因組區域，則無細胞核酸片段之兩端可來自生物體中之兩個先前各別基因組區域。在一些情況下，若無細胞核酸片段之第一末端與生物體之參考基因組中之第一基因組區域比對且無細胞核酸片段之第二末端與生物體之參考基因組中之第二基因組區域比對，則參考基因組中之第一及第二基因組區域可具有與無細胞核酸片段中之第一末端及第二末端之間的距離不一致的距離。舉例而言，若當無細胞核酸片段之兩端橫跨兩個原先不同染色體之部分時，參考基因組中之第一及第二基因組區域在不同染色體上，則參考基因組中之第一及第二基因組區域之間的距離可為不可測定的。在一些情況下，若無細胞核酸片段跨越發生染色體之染色體內易位之位置，則參考基因組中第一末端與第二基因組區域之間的距離可比無細胞核酸片段中第一末端及第二末端之間的距離長。在一些情況下，將參考基因組中之第一及第二基因組區域之相對距離與任意截斷值進行比較以便測定無細胞核酸是否跨越發生易位之位置。舉例而言，截斷值可為比無細胞核酸片段之預期長度更長的距離。在一些情況下，截斷值可為比無細胞核酸片段之預期長度減第一及第二序列讀段之長度更長的距離。截斷值可為約100個鹼基、約200個鹼基、約300個鹼基、約400個鹼基、約500個鹼基、約600個鹼基、約800個鹼基、約1000個鹼基、約1200個鹼基、約1500個鹼基、約2000個鹼基、約3000個鹼基、約4000個鹼基、約5000個鹼基、約6000個鹼基、約8000個鹼基或約 10^4 個鹼基。截斷值可

為至少80個鹼基、至少140個鹼基、至少180個鹼基、至少250個鹼基、至少350個鹼基、至少450個鹼基、至少550個鹼基、至少750個鹼基、至少900個鹼基、至少1100個鹼基、至少1250個鹼基、至少1800個鹼基、至少2500個鹼基、至少3500個鹼基、至少5500個鹼基、至少7500個鹼基、至少9000個鹼基或至少 10^4 個鹼基。

【0077】 反轉可為染色體鏈段端至端倒轉之染色體重組。當單一染色體在自身內經受斷裂及重組時，可出現反轉。若無細胞核酸片段起因於包涵反轉事件之基因組區域，則無細胞核酸片段之兩端可來自原先具有相反5'至3'關係之兩個基因組區域。在此等情況下，若無細胞核酸片段之第一末端與生物體之參考基因組中之第一基因組區域比對且無細胞核酸片段之第二末端與生物體之參考基因組中之第二基因組區域比對，則參考基因組中之第一及第二基因組區域可具有相比於無細胞核酸片段之第一末端及第二末端反向的5'至3'關係。

【0078】 擴增可為局部複製染色體鏈段之染色體重組。由包涵擴增之基因組區域產生之無細胞核酸片段可具有包含相同複製序列之兩端。缺失可為染色體鏈段缺失之染色體重組。由包涵缺失之基因組區域產生之無細胞核酸片段可具有來自先前藉由缺失鏈段間隔開之兩個基因組區域的兩端。

【0079】 該等方法可進一步包含分析染色體重組嵌合核酸片段以偵測染色體重組。染色體重組之偵測可包括測定染色體重組之類型，例如易位、擴增、缺失、反轉、染色體臂缺失或染色體臂獲得。染色體重組之偵測亦可包括測定染色體重組斷點。

【0080】 在一些情況下，該等方法包含鑑別來自複數個由來自生物

樣本之無細胞核酸分子之配對末端測序產生之成對序列讀段的A型染色體嵌合成對序列讀段。A型染色體嵌合成對序列讀段可具有與生物體之參考基因組之第一基因組區域比對的第一序列讀段及與生物體之參考基因組之第二基因組區域比對的第二序列讀段；且對於A型染色體嵌合成對序列讀段中之每一者，生物體之參考基因組中之第一及第二基因組區域之相對位置可與各別無細胞核酸分子中之第一末端及第二末端之相對位置不一致。該等方法可進一步包含將在生物體之參考基因組中之預定距離內重疊或間隔開之A型染色體嵌合成對序列讀段之第一序列讀段分組，且將在生物體之參考基因組中之預定距離內重疊或間隔開之A型染色體嵌合成對序列讀段之第二序列讀段分組，由此測定分別包含生物體之參考基因組之第一及第二基因組區域中之候選區域的候選重組區域。可基於A型染色體嵌合成對序列讀段推導染色體重組類型(例如易位或反轉)。

【0081】該等方法可進一步包含鑑別及分析B型染色體嵌合成對序列讀段。B型染色體嵌合成對序列讀段可具有與候選重組區域比對之第一序列讀段及為嵌合讀段之第二序列讀段，且其中與第一基因組區域比對之嵌合讀段之一部分及與第二基因組區域比對之嵌合讀段之其餘部分。基於B型染色體嵌合成對序列讀段，可測定染色體重組。舉例而言，染色體重組斷點可藉由將嵌合讀段之序列與參考基因組進行局部比對來測定。至少部分地基於染色體重組斷點，可進一步測定染色體重組之類型。

【0082】在一些情況下，該方法及系統包含測定B型染色體嵌合成對序列讀段之嵌合讀段之多樣性，由此基於多樣性篩選染色體重組。篩選可包含測定B型染色體嵌合成對序列讀段之多樣性評分。多樣性評分可計算為：

$$\frac{\sigma_1 + \sigma_2}{\max(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1})},$$

其中σ1為與第一基因組區域比對之嵌合讀段之部分之長度之標準差，且其中σ2為與第二基因組區域比對之嵌合讀段之其餘部分之長度之標準差。

【0083】該等方法可包括將多樣性評分與截斷值進行比較。在一些情況下，若多樣性評分等於或高於截斷值，則可基於嵌合讀段測定染色體重組。在一些情況下，多樣性評分低於截斷值，可排除候選染色體重組區域作為假陽性擊點。多樣性評分之截斷值可為約0.1、約0.2、約0.4、約0.6、約0.8、約1、約1.2、約1.4、約1.6、約1.8、約2.0、約2.2、約2.4、約2.5、約2.6、約2.8、約3.0、約3.1、約3.2、約3.3、約3.4、約3.5、約3.6、約3.7、約3.8、約3.9、約4.0、約4.1、約4.2、約4.3、約4.4、約4.5、約4.6、約4.7、約4.8、約4.9、約5.0、約5.2、約5.4、約5.6、約5.8、約6.0、約6.5、約7.0、約7.5、約8、約8.5、約9、約9.5、約10、約20、約30、約40、約50、約60、約80或約100。在一些情況下，截止多樣性評分可為至多0.1、至多0.2、至多0.4、至多0.6、至多0.8、至多1、至多1.2、至多1.4、至多1.6、至多1.8、至多2.0、至多2.2、至多2.4、至多2.5、至多2.6、至多2.8、至多3.0、至多3.1、至多3.2、至多3.3、至多3.4、至多3.5、至多3.6、至多3.7、至多3.8、至多3.9、至多4.0、至多4.1、至多4.2、至多4.3、至多4.4、至多4.5、至多4.6、至多4.7、至多4.8、至多4.9、至多5.0、至多5.2、至多5.4、至多5.6、至多5.8、至多6.0、至多6.5、至多7.0、至多7.5、至多8、至多8.5、至多9、至多9.5或至多10。在其他情況下，截止多樣性評分可為至少1、至少1.2、至少1.4、至少1.6、至少1.8、至少2.0、至少2.2、至少2.4、至少2.5、至少

2.6、至少2.8、至少3.0、至少3.1、至少3.2、至少3.3、至少3.4、至少3.5、至少3.6、至少3.7、至少3.8、至少3.9、至少4.0、至少4.1、至少4.2、至少4.3、至少4.4、至少4.5、至少4.6、至少4.7、至少4.8、至少4.9、至少5.0、至少5.2、至少5.4、至少5.6、至少5.8、至少6.0、至少6.5、至少7.0、至少7.5、至少8、至少8.5、至少9、至少9.5、至少10、至少20、至少30、至少40、至少50、至少60、至少80或至少100。

【0084】該等方法可包括藉由與偵測病原體整合斷點之方法類似的股型濾除「假陽性」成對序列讀段。該等方法可包含濾除具有與自測定染色體重組排除之大部分股型模式不一致的股型模式之B型染色體嵌合成對序列讀段。

B. 鑑別癌症基因組中之基因融合物

【0085】在一些態樣中，如本文所提供之方法及系統可應用於偵測人類癌症基因組中之基因融合物。基因融合物可包括分別不同或相同染色體上兩個先前各別基因組區域之核酸序列之染色體間或染色體內重組。基因融合事件會導致與自病原體整合所得之混雜DNA分子共有類似性的混雜基因組片段(「融合基因」)。基因融合可自易位、間質缺失或染色體反轉出現。基因融合可在癌症產生中起作用。鑑別基因融合可為診斷工具且融合基因可為抗癌治療中之治療標靶。如本文所論述之方法及系統係關於偵測來自腫瘤組織基因組釋放之無細胞核酸之融合基因。本文所提供之方法可賦予此類偵測無痛及非侵襲性，其可輔助精確癌症治療及實時監測腫瘤DNA之治療反應。

C. 鑑別體細胞結構變化

【0086】在一些態樣中，本文所提供之方法可應用於偵測基因組中

之體細胞結構變化。結構變化可包括缺失、複製、複本數目變型、插入、反轉及易位，其可促進複雜病症，包括自閉症及癌症。體細胞結構變化會導致釋放嵌合DNA分子至血漿中且可潛在地藉由如本文所描述之整合偵測方法來偵測。本文所提供之方法可允許精確鑑別各種結構變化且可促進探索病原性機制及與基因組結構變化相關之疾病之治療方法。可非侵襲性地基於本文所提供之方法偵測體細胞結構變化。

V. 測序方法

【0087】 可進行如本文所描述之生物樣本之測序分析以便測定病原體整合型態。本文所提供之方法可包含測序核酸分子，例如來自生物樣本之無細胞核酸分子、細胞核酸分子或兩者。在一些情況下，本文所提供之方法包含分析測序結果，例如來自生物樣本之核酸分子之測序讀段。本文所提供之方法及系統可涉及或不涉及測序之活性步驟。方法及系統可包含或提供用於自測序儀接受及處理測序資料之手段。方法及系統亦可包含或提供用於為測序儀提供命令以調節測序過程之參數，例如基於測序結果分析之命令的手段。

【0088】 可商購的測序設備可用於本發明中所提供之方法，諸如伊路米那(Illumina)測序平台及454/Roche平台。可使用此項技術中已知之任何方法進行測序核酸。舉例而言，測序可包括下一代測序。在一些情況下，可使用鏈終止測序、雜交測序、伊路米那測序(例如使用可逆終止物)、離子激流半導體測序、質量分光光度法測序、大規模平行標誌測序(MPSS)、Maxam-Gilbert測序、奈米孔測序、波羅泥(polony)測序、焦磷酸測序、鳥槍測序、單分子即時(SMRT)測序、SOLiD測序(使用四種經螢光標記之二-鹼基探針之雜交)、通用測序或其任何組合來進行核酸測序。

【0089】 可用於如本文所提供之方法中之一個測序方法可涉及配對末端測序，例如使用具有其基因組分析器之伊路米那「成對末端模組」。使用此模組，在基因組分析器已完成第一次測序讀段之後，成對末端模組可引導初始模板之再合成及第二輪簇生成。藉由使用本文所提供之方法中之成對末端讀段，吾人可自核酸分子之兩端獲得序列資訊且將兩端映射至參考基因組，例如病原體之基因組或宿主生物體之基因組。在映射兩端之後，吾人可根據如本文所提供之方法之一些實施例測定病原體整合型態。

【0090】 在配對末端測序期間，來自核酸分子之第一末端之序列讀段可包含至少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。來自核酸分子之第一末端之序列讀段可包含至多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。來自核酸分子之第一末端之序列讀段可包含約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。來自核酸分子之第二末端之序列讀段可包含至少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115、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。來自核酸分子之第二末端之序列讀段可包含至多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。來自核酸分子之第二末端之序列讀段可包含約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，來自核酸分子之第一末端之序列讀段可包含至少75個連續核苷酸。在一些情況下，來自核酸分子之第二末端之序列讀段可包含至少75個連續核苷酸。來自核酸分子之第一末端及第二末端之序列讀段可具有相同長度或不同長度。來自複數個來自生物樣本之核酸分子之序列讀段可具有相同長度或不同長度。

【0091】 可在各種測序深度下進行本文所提供之方法中之測序。測序深度可指基因座被與基因座比對之序列讀段覆蓋的次數。基因座可小至一個核苷酸，或大至一個染色體臂，或大至整個基因組。本文所提供之方法中之測序深度可為1x、2x、5x、10x、20x、25x、30x、40x、50x、

75x或100x等，其中「x」之前的數字係指基因座被序列讀段覆蓋的次數。測序深度亦可應用於多個基因座或全基因組，在此情況下，x可指基因座或單倍體基因組或全基因組分別定序之平均次數。在一些情況下，在本文所描述之方法中進行超深度測序，其可指進行至少100x測序深度。

【0092】在測序過程(例如測序深度)期間讀段特定核苷酸在核酸內之次數或平均次數可比所測序之核酸之長度大多倍。在一些情況下，當測序深度足夠比核酸之長度大(例如至少5倍)時，測序可稱作『深度測序』。在一些實例中，測序深度可平均比所測序之核酸之長度大至少約5倍、大至少約10倍、大至少約20倍、大至少約30倍、大至少約40倍、大至少約50倍、大至少約60倍、大至少約70倍、大至少約80倍、大至少約90倍、大至少約100倍。在一些情況下，樣本可富集特定分析物(例如核酸片段或癌症特定核酸片段)。

【0093】本文所提供之方法中產生之序列讀段(或定序讀段)可指自核酸分子之任何部分或所有測序之核苷酸串。舉例而言，序列讀段可為與核酸片段互補之核苷酸短串(例如20-150)、與核酸片段之末端互補之核苷酸串或與生物樣本中存在之整個核酸片段互補之核苷酸串。可在多種方法中，例如使用測序技術獲得序列讀段。

VI. 其他分析

【0094】在一些情況下，本發明之系統及方法係關於使用測序技術，繼而序列讀段之生物信息學分析以分析生物樣本中之病原體整合型態，例如分析整合斷點。可收集來自生物樣本之無細胞核酸且經受測序，例如配對末端測序。可藉由本文所提供之方法及系統分析無細胞核酸之序列讀段以便測定病原體之基因組、宿主生物體之基因組或兩者中之整合斷

點位置。其他技術亦可用作整合型態分析之替代或額外途徑，包括(但不限於)擴增分析及微陣列分析。

【0095】 在一些實例中，本文所提供之方法亦可包含分析來自生物樣本之核酸分子之擴增反應。擴增反應可包含與病原體之基因組中之第一靶序列互補的第一引子及與生物體之基因組中之第二靶序列互補的第二引子。擴增反應可經設計以擴增病原體整合區域，其中來自病原體之核酸序列連接來自宿主生物體之核酸序列。具有癌症之某些分類之整合斷點特性之知識可幫助設計此類擴增反應。在一些情況下，擴增反應包含聚合酶鏈反應(PCR)。PCR反應可以至少一對引子起始，其中之每一者與分別側接來自病原體基因組或宿主生物體基因組之整合斷點的序列互補。替代地，PCR反應可以隨機結合至模板序列之隨機探針起始。在一些情況下，擴增反應可提供關於整合斷點之有成本效益且快速的結果。PCR反應可為定量的，其可提供關於不同類型的病原體整合事件之部分貢獻之定量洞察。在一些情況下，亦可進行擴增反應所產生之擴增子序列之分析以便確認整合斷點。替代地或另外，可進行擴增反應及擴增子之序列分析作為靶向靶標特異性富集之步驟，因此提供潛在更穩固的讀數。可用於本文所描述之系統及方法之PCR反應之不同變型包括多重PCR、不對稱PCR、基於克列諾之PCR、巢式PCR、定量PCR、即時PCR、熱起始PCR、觸地PCR、組裝PCR、COLD-PCR(在較低變性溫度下共擴增-PCR)、雙尾PCR、接合介導PCR及甲基化特異性PCR。儘管本文已提供實例及實施例，與例如數位PCR及隨機測序相關之額外技術及實施例可見於2010年10月28日申請之美國專利8,722,334、2007年7月23日申請之美國臨時申請案60/951,438及2013年7月8日申請之美國專利9,121,069，其中之每一者全部以引用之方

式併入本文中。

【0096】 在一些其他情況下諸如微陣列分析之任何其他序列分析可用於獲得核酸分子，例如來自生物樣本之無細胞核酸之序列資訊。舉例而言，微陣列可包含靶向宿主基因組序列及病原體基因組序列兩者之探針以便偵測如本文所描述之嵌合片段。在其他實例中，微陣列可經設計以偵測施加至微陣列之核酸分子中之病原體整合斷點之位置。

【0097】 在一些情況下，如本文所描述之樣本分析可藉由利用不同原理及/或途徑之分析補充及/或確認。舉例而言，血漿樣本中之無細胞DNA之分析可產生展示與癌症之特定類型相關之較佳病原體整合斷點的一系列結果。在一些情況下，由無細胞DNA樣本產生之此等之敏感性及特異性可藉由諸如來自活檢體腫瘤樣本之特定類型之轉化癌細胞之測序，例如完整基因組測序評定。替代地，亦可應用相對更簡單且有成本效益的途徑，例如定性或定量PCR反應可藉由使用在鑑別候選斷點期間獲得之序列資訊經設計以評定斷點。

【0098】 在一些情況下，本文所提供之系統及方法係關於用於決定病理學分類之第一分析及第二分析。第一分析可包含測序分析，例如來自生物樣本之核酸分子之測序，繼而如本文所描述之序列讀段之生物資訊分析。第二分析可包含擴增分析、微陣列分析或其他類型之分析。在一些實例中，第二分析包含測序分析，而第一分析包含擴增分析、微陣列分析或其他類型之分析。在一些情況下，第一分析，例如測序分析用於測定病原體整合型態，例如病原體之基因組或宿主生物體之基因組中之一或多個病原體整合斷點之位置。來自第一分析之病原體整合型態可提供第二分析之引導，例如一或多個病原體整合斷點之位置可提供設計第二分析之PCR探

針或微陣列探針之引導。在一些實例中，人類個體，例如疑似經病原體，例如HPV感染之患者可藉由用於測定病原體整合型態之第一分析測試。病原體整合型態可用於決定病理學分類。至少部分地基於第一分析，患者可診斷患有疾病，例如癌症，例如子宮頸癌。患者可因此經受第二分析，其可用於例如在向患者投與療法之後或在對個體進行手術之後繼續監測來自患者之生物樣本中之病原體整合型態。病原體整合型態之持續監測可用於監測患者已經診斷患有之疾病，例如用於監測疾病之進展或消退。在一些實例中，第一分析可用作一或多種疾病/病況之篩選分析，而第二分析可用作一或多種疾病/病況之確認分析。

VII. 病理學分類

【0099】 整合在生物體之基因組中之病原體之基因組中之整合斷點之分析可用於決定病理學分類。舉例而言，本文所提供之方法可包含病毒DNA整合型態之分析，例如在人類宿主生物體基因組中整合之HPV病毒基因組中之斷點之分析，且可進一步包含基於病毒DNA整合型態測定癌症類型。病理學可一般涉及偏離健康情況之人類病況或疾病。與病原體整合相關且適用於本文所提供之方法及系統之病理學之非限制性實例可包括成人T細胞白血病、熱帶痙攣性截癱、子宮頸癌、頭頸癌、肛殖癌、卡堡氏肉瘤、伯基特氏淋巴瘤、肝細胞癌、梅克爾細胞癌、AIDS、腦癌、骨癌、間皮瘤、前列腺癌及B細胞淋巴瘤。

【0100】 癌症或腫瘤可指組織之異常質量，其中質量生長超越且與正常組織之生長不協調。癌症或腫瘤可取決於以下特徵而定義為「良性」或「惡性」：細胞分化程度，包含形態及功能；生長速率；局部侵襲；及轉移。「良性」腫瘤可為分化良好，生長典型地比惡性腫瘤慢，且保持侷

限於原發部位。另外，良性腫瘤沒有浸潤、侵襲或轉移至遠端之能力。

「惡性」腫瘤可以為分化不良(退行發育)，典型地快速生長，伴隨進行性浸潤、侵襲及破壞周圍組織。此外，惡性腫瘤具有轉移至遠端之能力。

【0101】 在一些態樣中，本文所提供之方法可包含染色體重組之分析，例如複本數變化、缺失、複製、反轉及易位，且可進一步包含測定與染色體重組相關之病理學。可採用本文所提供方法及系統之與染色體重組相關之病理學可包括幾乎所有類型之癌症及各種類型的遺傳疾病，例如沃夫-赫希霍恩(Wolf-Hirschhorn)症候群、雅各布森(Jacobsen)症候群、恰克-馬利-杜斯氏症(Charcot-Marie-Tooth disease)類型1A、安格爾曼氏症候群(Angelman syndrome)、特納症候群(Turner Syndrome)、22q11.2缺失症候群、三倍X症候群、威廉姆斯症候群(Williams Syndrome)、貓叫症候群(Cri du Chat Syndrome)、貓眼症候群(Cat Eye Syndrome)、1q21.1複製症候群、15q13.3微缺失症候群、16p11.2缺失症候群、17q23.1q23.2微缺失症候群、1q複製症候群、1q21.1微缺失症候群、22q11.2缺失症候群、22q11.2複製症候群、2q23.1微缺失症候群、2q37缺失症候群、47XXX症候群、47,XYY症候群、49,XXXXX症候群、二倍體-三倍體鑲嵌、末梢染色體18q缺失症候群、伊曼紐爾症候群(Emanuel syndrome)、克里斯塔(Kleefstra Syndrome)、庫冷德弗里斯(Koolen de Vries Syndrome)、鑲嵌型單染色體18、納布盧斯(Nablus)面具臉症候群、帕里斯特-基利安鑲嵌型症候群(Pallister-Killian Mosaic Syndrome)、史密斯-馬吉利氏症候群(Smith-Magenis Syndrome)、四體性(Tetrasomy) 9p症候群、四體性X症候群、三倍體性症候群(Triploidy Syndrome)、三染色體13症候群、三染色體17鑲嵌及三染色體2鑲嵌。

VIII. 癌症分類

【0102】 病毒DNA整合斷點可用作根據本文所提及之方法分類癌症之基礎。在一些情況下，其他病理學分類，例如癌症亦可利用本發明中所提供之方法。

【0103】 癌症分類可包含癌症之存在或不存在。本文所提供之方法可用於測定任何類型的癌症之分類，諸如(但不限於)膀胱癌、骨癌、腦瘤、乳癌、子宮頸癌、結腸直腸癌、食道癌、胃腸癌、造血惡性病、頭頸部鱗狀細胞癌、白血病、肝癌、肺癌、淋巴瘤、骨髓瘤、鼻癌、鼻咽癌、口腔癌、口咽癌、卵巢癌、前列腺癌、肉瘤、胃癌或甲狀腺癌。

【0104】 癌症分類可包含癌症程度或癌症分期。癌症程度或癌症分期(例如I期、II期、III期或IV期)可指腫瘤之尺寸、癌轉移之存在或不存在、身體之全部腫瘤負荷、及/或癌症嚴重程度之其他量度(例如癌症復發)。癌症程度可為數字或其他標誌，諸如符號、字母及顏色。程度可為零。癌症程度亦可包括與突變或突變數量有關的惡變前或癌變前病況(狀態)。

【0105】 癌症分類亦可用於各種方法中。在一些情況下，癌症分類包含癌症進展或癌症消退。在一些情況下，癌症分類包含在具有或不具有癌症治療之情況下癌症進展或消退之預言或預測。舉例而言，如本文所描述之方法及系統可用於判定癌症是否存在於先前尚未知曉患有癌症之某人中。評定可調查已經診斷患有癌症之某人以隨時間推移監測癌症之進展，研究療法有效性或確定預後。在一些情況下，預後可用患者死於癌症之機率或特定期限或時間之後癌症進展之機率或癌症轉移之機率表示。

【0106】 在一些情況下，癌症分類包含癌症類型，例如子宮頸癌

(CC)或頭頸部鱗狀細胞癌(HNSCC)。圖6展示基於來自分別患有子宮頸癌及HPV陽性頭頸部鱗狀細胞癌之患者之血漿樣本之無細胞DNA之分析，HPV16基因組中之病毒整合斷點之分佈。如圖中所示，在子宮頸癌與HPV-陽性頭頸部鱗狀細胞癌(區域A)之間可存在共有的較佳整合斷點。另一方面，亦可存在某一類型之癌症之較佳整合斷點特性，例如區域B展示HNSCC較佳整合斷點。

【0107】 某些癌症分類，例如癌症類型之病毒整合斷點特性可使用本文所提供之方法鑑別；在一些情況下，為了測定癌症分類，可應用類似方法以測定病毒整合斷點，其中分析來自生物樣本之核酸分子之序列讀段。

IX. 病原體整合：病毒整合

【0108】 本文所提供之方法及系統可用於偵測宿主生物體基因組中可能存在之不同類型的病毒整合，諸如必選整合。在一些病毒，諸如反轉錄病毒屬、假病毒科、轉座病毒科、一些肌尾噬菌體科及長尾病毒科之病毒複製期間，必選整合至宿主生物體基因組中可為必然事件。病毒DNA整合會導致病毒基因組永久性插入至宿主生物體染色體DNA中，該宿主生物體染色體DNA在反轉錄病毒之情況下稱為原病毒或在原核病毒之情況下稱為原噬菌體。本文所提供之方法及系統可用於偵測偶然整合。在一些病毒之病毒複製期間，偶然整合可能不為必需，但可賦予宿主/病毒偶合一些優點。本文所提供之方法及系統可用於偵測內源性病毒元件(EVE)。一些內源性病毒可經由罕見且有時意外的過程整合及「僵化」至宿主生物體基因組中，包括內源性反轉錄病毒及罕見的RNA病毒整合。EVE之一個實例包括整合於生殖系中之內源性反轉錄病毒。內源性逆病毒

之整合基因組可保持潛伏且與宿主生物體基因組一起被動複製且傳遞至細胞後代。然而，宿主生物體之環境條件變化可使病毒再活化，導致病毒轉錄及產生新感染性病毒(噬菌體感染)。

【0109】 本文所提供之方法及系統可用於分析DNA病毒及RNA病毒。

【0110】 本文所提供之方法及系統可用於分析DNA病毒整合。DNA病毒之基因組可為宿主生物體基因組整合之潛在受質，而不需要先前處理。在細胞感染後，DNA病毒之基因組可易位至細胞核，其中其可保持為游離基因體以確保病毒續存或整合至宿主細胞之基因組中。

【0111】 本文所提供之方法及系統可用於研究疱疹病毒整合。疱疹病毒可為DNA包膜病毒，且可基於其序列種系發生分類於三個家族中： α 、 β 及 γ 疱疹病毒。其可含有在病毒進入及環化後細胞核中遞送之線形雙股DNA。疱疹病毒可保持游離，例如呈染色體外環形DNA形式。一些疱疹病毒可將其基因組整合至宿主染色體中。疱疹病毒之實例包括艾伯斯坦-巴爾病毒(EBV)及人類疱疹病毒-6 (HHV-6)。

【0112】 本文所提供之方法及系統可用於分析人類乳突狀瘤病毒(HPV)整合。HPV DNA可整合至宿主細胞之基因組中。DNA破壞及可誘導DNA破壞(例如雙股斷裂(DSB))之藥劑可在HPV整合中起作用。

【0113】 本文所提供之方法及系統可用於分析B型肝炎病毒(HBV)之整合。HBV可導致肝細胞癌發展或在肝細胞癌發展中起作用。在急性感染期間，HBV可將其基因組整合至宿主染色體中且與反轉錄病毒整合存在若干類似性。

【0114】 本文所提供之方法及系統可用於分析2型腺相關病毒(AAV-

2)之整合。腺相關病毒AAV可為分類於細小病毒科家族中之普遍病毒。AAV之複製可嚴格地藉由輔助病毒，諸如腺病毒、人類乳突狀瘤病毒(HPV)或單純疱疹病毒(HSV)之相同經感染細胞之存在改良。在輔助病毒不存在下，AAV可以位點特異性方式整合其基因組。

【0115】 本文所提供之方法及系統可用於分析RNA病毒整合。反轉錄病毒之病毒RNA基因組可反轉錄至線形雙股DNA分子(病毒DNA中間物)中，其可因此為後續病毒基因組整合至宿主基因組中之受質。對於反轉錄病毒，整合可為噬菌體感染之必選步驟。其他RNA病毒之基因組亦可在宿主基因組中鑑別出。在此等情況下，整合可附帶出現，諸如對於淋巴球性脈絡叢腦膜炎病毒(LCMV)而言。本文所提供之方法及系統亦可幫助闡明病毒生命週期中RNA病毒整合之作用以及宿主生物體中之病理學發展。

X. 病毒整合及癌症

【0116】 在一些實例中，本文所提供之方法及系統可用於藉由測定病毒整合來測定癌症分類。

【0117】 在一些情況下，本文所提供之方法及系統可用於分析腫瘤病毒(或致癌病毒)，其可直接導致實驗動物或人類中之癌症。存在若干腫瘤病毒家族，包括B型肝炎病毒(肝癌)、乳突狀瘤病毒(子宮頸癌及其他肛殖癌)、艾伯斯坦-巴爾病毒(伯基特氏淋巴瘤及鼻咽癌)、卡波西氏肉瘤相關疱疹病毒(卡波西氏肉瘤)、多瘤病毒(梅克爾細胞癌)及人類T細胞嗜淋巴細胞病毒(成人T細胞白血病)。六個非限制性家族中之五個具有DNA基因組且稱為DNA腫瘤病毒，且另一家族為反轉錄病毒，其在病毒粒子中具有RNA基因組，但可經由經感染細胞中之DNA原病毒合成來複製。另

外，由於免疫缺陷，HIV可間接造成AIDS患者中之發生之癌症，且C型肝炎病毒(RNA病毒)可為由慢性組織破壞產生之肝癌之間接病因。

【0118】 本文所提供之方法及系統可用於分析HBV整合以便測定HBV相關病況/疾病，如肝細胞癌(HCC)。HBV可感染若干物種，包括鴨、土撥鼠、松鼠及人類之肝細胞。B型肝炎病毒感染會導致急性肝損傷。B型肝炎病毒感染可常見於亞洲及非洲部分地區，其中其可與每年高達一百萬肝癌病例相關(約全世界癌症發病率之10%)。B型肝炎病毒之細胞變形可藉由影響驅動異常細胞增殖及存活之多種細胞基因之表現的病毒基因(稱作X基因)介導。另外，藉由B型肝炎病毒誘發之癌症發展可由慢性組織破壞所導致的肝細胞連續增殖驅動。本文所提供之方法及系統可用於診斷、監測或預言與B型肝炎病毒之任何基因型或血清型，諸如基因型F、E、A、B、C及D及其各種次型相關之病理學。

【0119】 HBV可與HCC相關。HBV為具有反轉錄病毒特徵之部分雙股DNA嗜肝病毒。HBV基因組整合至細胞基因組中可存在於超過85%-90% HBV相關HCC中。然而，HBV之整合形式亦可存在於具有慢性HBV感染之患者之非腫瘤組織中。HBV基因組整合至肝細胞中可在持續性HBV感染期間出現且在HCC發展之前。HBV整合可導致提若干細胞癌症相關基因，諸如TERT、混合譜系白血病4 (MLL4)及CCNE1 (編碼週期素E1)之高的表現。HBV整合亦可與HCC提前發作及不佳結果有關。

【0120】 本文所提供之方法及系統可用於測定乳突狀瘤病毒整合型態以評定乳突狀瘤病毒相關病況/疾病，如子宮頸癌。乳突狀瘤病毒可為較小DNA病毒(約8 kb之基因組)，其可在人類及多種其他動物物種中誘發良性及惡性腫瘤兩者。已鑑別出約60種不同類型的人類乳突狀瘤病毒，其

可感染若干組織之上皮細胞。此等病毒中之一些可僅導致良性腫瘤(諸如疣)，而其他可為惡性癌瘤，尤其子宮頸癌、頭頸部鱗狀細胞癌(HNSCC)及其他肛殖癌之病原體。本文所提供之方法及系統可用於使不同類型的病理學分化，例如與HPV相關之癌症。本文所提供之方法及系統亦可用以分析任何類型的乳突狀瘤病毒，包括HPV類型6、11、16、18、31、33、35、39、45、51、52、56、58、59、68、73及82。

【0121】 HPV之子集可感染肛殖區域，且在此子集內，個別類型可歸類為高風險或低風險。高風險HPV可導致癌性損傷，而低風險HPV不會。高風險及低風險HPV之間的差異可為高風險HPV展示更大的整合至宿主基因組中之傾向，由此造成高級損傷及癌症，而低風險類型可優先維持染色體外環形游離基因體。HPV可導致幾乎所有子宮頸癌病例，其可為全世界女性中第二常見之癌症及全世界女性中第四常見之癌症死亡病因。HPV16、HPV18、HPV31及HPV33可佔所有子宮頸癌病例之90%。在此等高風險HPV中，HPV類型16為最普遍類型且本身可佔超過所有子宮頸癌病例之50%。高風險HPV感染亦可能與若干其他肛殖癌及口咽癌有關。舉例而言，認為造成超過90%肛門癌、70%陰道癌及外陰癌、60%陰莖癌及63%口咽癌。95% 具有子宮頸癌前病變之患者中可攜帶HPV，其中僅小部分可最終進展至侵襲性癌瘤。三個癌前階段(子宮頸上皮內贅瘤(CIN) 1、CIN2及CIN3)可在產生侵襲性癌瘤之前。CIN1損傷可自發消退，其中僅幾個損傷發展為CIN2/CIN3且最終發展為侵襲性癌瘤。經HPV感染之女性中之子宮頸癌進展可與整合狀態緊密相關，且發現整合於子宮頸癌症中之HPV之頻率可始終較高。舉例而言，100% HPV18、80% HPV16及81% HPV31陽性癌症可展示病毒整合。本文所提供之方法及系統可適用

於藉由測定個體中之HPV整合型態來預言及監測癌症，例如子宮頸癌。

【0122】 本文所提供之方法及系統可用於測定疱疹病毒整合型態以評定疱疹病毒相關病況/疾病，如鼻咽癌。疱疹病毒可在具有100至200 kb基因組之大部分複雜動物病毒中。若干疱疹病毒可誘發動物物種，包括青蛙、雞及猴中之腫瘤。另外，疱疹病毒家族之兩個成員卡波西氏肉瘤相關疱疹病毒及艾伯斯坦-巴爾病毒可與人類癌症相關。卡波西氏肉瘤相關疱疹病毒可在發展卡堡氏肉瘤中起關鍵作用，且艾伯斯坦-巴爾病毒可涉及若干人類惡性病，包括一些非洲地區的伯基特氏淋巴瘤、AIDS患者及其他免疫抑制個體中之B細胞淋巴瘤及中國的鼻咽癌。本文所提供之方法及系統可用於分析不同疱疹病毒屬之整合，如傳喉炎病毒(Iltovirus)、長鼻病毒屬(Proboscivirus)、細胞巨大病毒屬(Cytomegalovirus)、馬立克病毒屬(Mardivirus)、細長病毒屬(Rhadinovirus)、瑪卡病毒屬(Macavirus)、玫瑰疹病毒屬(Roseolovirus)、單純疱疹病毒屬(Simplexvirus)、壞血病病毒屬(Scutavirus)、水痘病毒屬(Varicellovirus)、馬疱疹病毒屬(Percavirus)、淋巴濾泡病毒屬(Lymphocryptovirus)及鼠巨細胞病毒屬(Muromegalovirus)；人類疱疹病毒，如單純疱疹病毒-1 (HSV-1)、單純疱疹病毒-2 (HSV-2)、水痘帶狀疱疹病毒(VZV)、艾伯斯坦-巴爾病毒(EBV)、淋巴濾泡病毒(lymphocryptovirus)、細胞巨大病毒(CMV)、玫瑰疹病毒、疱疹嗜淋巴細胞病毒及卡波西氏肉瘤相關疱疹病毒(KSHV)；及動物源性疱疹病毒，如獼猴疱疹病毒-1及鼠科動物疱疹病毒68 (MHV-68)。

【0123】 EBV為可主要與伯基特氏淋巴瘤、鼻咽癌及若干淋巴增生病症相關之雙股DNA疱疹病毒。伯基特氏淋巴瘤可以三個主要臨床變型出現-地方性、偶發性及免疫缺陷相關變型。可在96%涉及頷部之地方性

變型伯基特氏淋巴瘤病例中偵測到EBV，其為某些中非地區兒童之最常見惡性病。相比之下，EBV可能很少與伯基特氏淋巴瘤之偶發性變型有關，且較不常涉及頤部。EBV相關伯基特氏淋巴瘤可常見於不具有有效T細胞功能之個體，諸如AIDS患者或移植受體中。

【0124】 EBV可持續呈具有多個環形DNA複本之游離型狀態，但EBV整合至宿主染色體之脆性位點中可能與病毒基因組之部分缺失相關且可產生在宿主細胞基因組中增強的染色質不穩定性區域。此基因組不穩定性可能誘發宿主基因缺失，諸如BACH2，其可為推定腫瘤抑制基因，且此可促進淋巴瘤生成。

【0125】 除了伯基特氏淋巴瘤之外，EBV亦可能與鼻咽癌(NPC)有關。WHO分類為III型之NPC之未分化形式可展示與全世界EBV，例如尤其中國及東南亞地區一致相關。

【0126】 本文所提供之方法及系統可用於測定梅克爾細胞多瘤病毒(MCV)整合型態以評定MCV相關病況/疾病，如梅克爾細胞癌(MCC)。MCV可為雙股DNA多瘤病毒，其可能與MCC相關。此等腫瘤可在純系模式下腫瘤細胞基因組內呈現整合形式之MCV DNA。MCV感染及整合可促進腫瘤細胞之純系擴增。MCV基因組可編碼靶向若干腫瘤抑制蛋白，諸如pRB及p53之腫瘤(T)抗原蛋白複合物之多個剪接變異體。此等剪接變異體中之一者(較大腫瘤抗原)可在MCV陽性MCC腫瘤細胞中突變，且此選擇性突變可能影響細胞DNA損傷反應以阻止整合病毒複製之自身活化，破壞宿主基因組完整性且抑制細胞增殖，MCV可常常且選擇性地與MCV結合。

【0127】 RNA病毒之一個家族之成員(反轉錄病毒)可在多種動物物

種，包括人類中導致癌症。舉例而言，如人類T細胞白血病毒之反轉錄病毒可導致成人T細胞白血病，其可常見於日本、加勒比及非洲部分地區。不同反轉錄病毒在其致癌可能性方面可實質上不同。本文所提供之方法及系統可用於分化病理學，例如與反轉錄病毒相關之癌症。本文所提供之方法及系統可用於分析任何類型的反轉錄病毒，包括外源性反轉錄病毒，如 α 反轉錄病毒屬、 β 反轉錄病毒屬、 γ 反轉錄病毒屬、 δ 反轉錄病毒屬、 ϵ 反轉錄病毒屬、慢反轉錄病毒屬(Lentiretrovirus)及泡沫病毒屬及各種內源性反轉錄病毒屬之成員。

XI. 生物樣本

【0128】 用於如本文所提供之方法中之生物樣本可包括源自活個體或死個體之任何組織或物質。生物樣本可為無細胞樣本。生物樣本可包含核酸(例如DNA或RNA)或其片段。樣本中之核酸可為無細胞核酸。樣本可為液體樣本或固體樣本(例如細胞或組織樣本)。生物樣本可為體液，諸如血液、血漿、血清、尿液、陰道液、來自水囊腫(例如睪丸)之液體、陰道沖洗液、胸膜液、腹水、腦脊髓液、唾液、汗液、淚液、痰、支氣管肺泡灌洗液、乳頭溢液、來自身體不同部分(例如甲狀腺、乳房)之抽吸液等。亦可使用糞便樣本。在各種實施例中，無細胞DNA已富集之生物樣本(例如經由離心方案獲得之血漿樣本)中之大部分DNA可為無細胞的(例如大於50%、60%、70%、80%、90%、95%或99%之DNA可為無細胞的)。可處理生物樣本以便以物理方式使組織或細胞結構破裂(例如離心及/或細胞分解)，由此將胞內組分釋放至可另外含有酶、緩衝液、鹽、清潔劑及其用於製備樣本以供分析之類似者的溶液中。

【0129】 本文所提供之方法及系統可用於分析生物樣本中之核酸分

子。核酸分子可為細胞核酸分子、無細胞核酸分子或兩者。由如本文所提供之方法使用之無細胞核酸可為超出生物樣本中之細胞範圍外的核酸分子。無細胞核酸分子可存在於各種體液中，例如血液、唾液、精液及尿液。無細胞DNA分子可由於可起因於健康狀況及/或疾病之各種組織之細胞死亡(例如病毒感染及腫瘤生長)產生。無細胞核酸分子可包含由於病原體整合事件產生之序列。

【0130】 用於如本文所提供之方法中之無細胞核酸分子(例如無細胞DNA)可存在於血漿、尿液、唾液或血清中。無細胞DNA可天然地以短片段形式存在。不含細胞之DNA片段化可指藉以使高分子量DNA (諸如細胞核中之DNA)裂解、斷裂或消化成短片段(此時產生或釋放不含細胞之DNA分子)的過程。本文所提供之方法及系統可用於在一些情況下分析細胞核酸分子，例如在患者具有白血病、淋巴瘤或骨髓瘤時分析來自腫瘤組織之細胞DNA或來自白血球之細胞DNA。根據本發明之一些實例，可對獲自腫瘤組織之樣本進行檢定及分析。

XII. 個體

【0131】 本文所提供之方法及系統可用於分析來自個體(例如生物體，例如宿主生物體)之樣本。該個體可為任何人類患者，諸如癌症患者、有患癌風險之患者或具有家族或個人癌症病史之患者。在一些情況下，個體處於癌症治療之特定階段。在一些情況下，個體可患有癌症或可被懷疑患有癌症。在一些情況下，個體是否患有癌症係未知的。

【0132】 個體可患有任何類型之癌症或腫瘤。在一個實例中，個體可能患有鼻咽癌或鼻腔癌。在另一個實例中，個體可能患有口咽癌或口腔癌。癌症之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、血癌、骨癌、腦腫瘤、乳癌、支氣管癌、心血管系統癌、子宮頸癌、結腸癌、結腸直腸癌、消化系統癌、內分泌系統癌、子宮內膜癌、食道癌、眼癌、膽囊癌、胃腸腫瘤、肝細胞癌、腎癌、造血惡性腫瘤、喉癌、白血病、肝癌、肺癌、淋巴瘤、黑素瘤、間皮瘤、肌肉系統癌、骨髓發育不良症候群(MDS)、骨髓瘤、鼻腔癌、鼻咽癌、神經系統癌、淋巴系統癌、口腔癌、口咽癌、骨肉瘤、卵巢癌、胰臟癌、陰莖癌、垂體瘤、前列腺癌、直腸癌、腎盂癌、生殖系統癌、呼吸系統癌、肉瘤、唾液腺癌、骨骼系統癌、皮膚癌、小腸癌、胃癌、睪丸癌、咽喉癌、胸腺癌、甲狀腺癌、腫瘤、泌尿系統癌、子宮癌、陰道癌或外陰癌。淋巴瘤可為任何類型之淋巴瘤，包括B細胞淋巴瘤(例如瀰漫性大B細胞淋巴瘤、濾泡性淋巴瘤、小淋巴球性淋巴瘤、套細胞淋巴瘤、邊緣區B細胞淋巴瘤、伯基特淋巴瘤、淋巴漿細胞性淋巴瘤、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-淋巴母細胞性淋巴瘤或外周T細胞淋巴瘤)。白血病可為任何類型之白血病，包括急性白血病或慢性白血病。白血病之類型包括急性骨髓性白血病、慢性骨髓性白血病、急性淋巴球性白血病、急性未分化白血病或慢性淋巴球性白血病。在一些情況下，癌症患者未患特定類型之癌症。舉例而言，在一些情況下，患者可能患有不是乳癌的癌症。

【0133】 癌症之實例包括導致實體腫瘤之癌症以及不導致實體腫瘤之癌症。此外，本文所提及之任何癌症可為原發癌(例如以其首先開始生長之身體部位命名的癌症)或繼發癌或轉移癌(例如起源於身體之另一部位的癌症)。

【0134】 藉由本文所描述之任一方法診斷之個體可為任何年齡且可

為成人、嬰兒或兒童。在一些情況下，個體為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，或在其範圍內(例如在2至20歲之間、在20至40歲之間或在40至90歲之間)。可能受益之特定類別的患者為40歲以上之患者。可能受益之另一特定類別的患者可為兒科患者。此外，藉由本文所描述之任一方法或組合物診斷之個體可為男性或女性。

【0135】 在一些實施例中，本發明之方法可偵測個體之腫瘤或癌症，其中該腫瘤或癌症具有疾病之地理模式。在一個實例中，個體可能患有在華南(例如香港特區)盛行之EBV相關癌症(例如鼻咽癌)。在另一實例中，個體可能患有在美國及西歐盛行之HPV相關癌症(例如口咽癌)。在另一實例中，個體可能患有在日本南部、加勒比、中非、南美洲部分地區及美國東南部一些移民團體中盛行之人類T淋巴病毒-1 (HTLV-1)相關癌症(例如成人T細胞白血病/淋巴瘤)。

【0136】 亦可對非人類個體，諸如實驗室或農場動物，或源自本文所揭示之生物體之細胞樣本執行本文所揭示之方法中之任一種。非人類個體之非限制性實例包括犬、山羊、天竺鼠、倉鼠、小鼠、豬、非人類靈長類(例如大猩猩、猿、猩猩、狐猴或狒狒)、大鼠、綿羊、牛或斑馬魚。

XIII. 電腦系統

【0137】 本文所揭示之方法中之任一者可藉由一或多個電腦系統執行及/或控制。在一些實例中，本文所揭示之方法之任何步驟可全部地、單獨地或連續地由一或多個電腦系統執行及/或控制。本文所提及之電腦系統中之任一種可利用任何適合數目之子系統。在一些實施例中，電腦系統包括單個電腦設備，其中子系統可為電腦設備之組件。在其他實施例中，電腦系統可包括具有內部組件之多個電腦設備，其各自為子系統。電腦系統可包括桌上型及膝上型電腦、平板電腦、移動電話及其他移動裝置。

【0138】 子系統可經由系統匯流排互連。額外之子系統包括印表機、鍵盤、儲存裝置及耦接至顯示配接器之監視器。周圍裝置及耦接至I/O控制器之輸入/輸出(I/O)裝置可藉由此項技術中已知之任何數目之連接(諸如，輸入/輸出(I/O)埠(例如USB、FireWire[®]))連接至電腦系統。舉例而言，I/O埠或外部界面(例如乙太網路(Ethernet)、Wi-Fi等)可用於連接電腦系統與諸如網際網路之廣域網路、鼠標輸入器件或掃描儀。經由系統匯流排之互連允許中央處理器與各子系統通信並控制來自系統記憶體或儲存裝置(例如固定磁碟，諸如硬碟機，或光碟)之複數個指令的執行，以及資訊在子系統之間的交換。系統記憶體及/或儲存裝置可體現電腦可讀取媒體。另一子系統為資料收集裝置，諸如攝影機、麥克風、加速計及其類似者。本文所提及之任何資料可自一個組件向另一個組件輸出且可向使用者輸出。

【0139】 電腦系統可包括例如藉由外部界面或藉由內部界面連接在一起的複數個相同組件或子系統。在一些實施例中，電腦系統、子系統或設備可經網路通信。在此類情況下，可將一台電腦視為用戶端且另一台電

腦視為伺服器，其中每一者可為同一電腦系統之一部分。用戶端及伺服器各自可包括多個系統、子系統或組件。

【0140】 本發明提供經程式化以實施本發明之方法的電腦控制系統。**圖8**展示經程式化或以其他方式經配置以偵測嵌合核酸片段或以測定核酸分子中之病原體整合型態或以測定如本文所描述之病理學分類的電腦系統1101。電腦系統1101可實施及/或調節本文所提供之方法之各種態樣，諸如控制來自生物樣本之核酸分子之測序，執行如本文所描述之測序資料之生物資訊分析的各種步驟，整合資料收集、分析及結果報告，及資料管理。電腦系統1101可為使用者之電子裝置或相對於電子裝置遠距離定位之電腦系統。電子裝置可為行動電子裝置。

【0141】 電腦系統1101包括中央處理單元(CPU，本文中亦為「處理器」及「電腦處理器」) 1105，其可為單核或多核處理器，或用於並行處理之複數個處理器。電腦系統1101亦包括記憶體或記憶體位置1110 (例如，隨機存取記憶體、唯讀記憶體、快閃記憶體)電子儲存單元1115 (例如，硬碟)、用於與一或多個其他系統連通之通信接口1120 (例如，網路適配器)，及周邊裝置1125，諸如緩存、其他記憶體、資料儲存器及/或電子顯示適配器。記憶體1110、儲存單元1115、接口1120及周邊裝置1125經由通信匯流排(實線)，諸如母板與CPU 1105通信。儲存單元1115可為用於儲存資料之資料儲存單元(或資料儲存庫)。電腦系統1101可藉助於通信接口1120可操作地耦接至電腦網路(「網路」) 1130。網路1130可為網際網路，亦即網際網路及/或商際網路，或與網際網路通信之企業內部網路及/或商際網路。在一些情況下，該網路1130係電信及/或數據網路。網路1130可包括一或多個電腦伺服器，其可實現分佈式計算，諸如雲端計

算。網路1130在一些情況下藉助於電腦系統1101可實施點對點網路，其可實現耦接至電腦系統1101之裝置表現為用戶端或伺服器。

【0142】 CPU 1105可執行可體現於程式或軟體中之一連串機器可讀取指令。指令可儲存於記憶體位置，諸如記憶體1110中。指令可針對CPU 1105，其可隨後使CPU 1105程式化或以其他方式組態以實施本發明之方法。藉由CPU 1105執行之操作的實例可包括擷取、解碼、執行及寫回。

【0143】 CPU 1105可為諸如積體電路之電路之部分。系統1101之一或多個其他組件可包括於該電路中。在一些情況下，電路為特殊應用積體電路(ASIC)。

【0144】 儲存單元1115可儲存文件，諸如驅動程式、資料庫及保存的程式。儲存單元1115可儲存使用者資料，例如使用者偏好及使用者程式。電腦系統1101在一些情況下可包括在電腦系統1101外部，諸如位於經由企業內部網路或網際網路與電腦系統1101通信之遠端伺服器上之一或多個額外資料儲存單元。

【0145】 電腦系統1101可經由網路1130與一或多個遠端電腦系統通信。舉例而言，電腦系統1101可與使用者之遠端電腦系統(例如，安裝有接收或顯示自電腦系統1101發送之樣本分析之結果的應用程式的智慧型手機)通信。遠端電腦系統之實例包括個人電腦(例如便攜式PC)、平板(slate/tablet) PC (例如Apple® iPad、Samsung® Galaxy Tab)、電話、智慧型電話(例如Apple® iPhone、Android功能裝置、Blackberry®)或個人數位助理。使用者可經由網路1130獲取電腦系統1101。

【0146】 如本文所描述之方法可藉助於儲存於電腦系統1101之電子

儲存器位置上，諸如記憶體1110或電子儲存單元1115上之機器(例如，電腦處理器)可執行碼來實施。機器可執行或機器可讀取代碼可以軟體形式來提供。在使用期間，程式碼可藉由處理器1105執行。在一些情況下，程式碼可自儲存單元1115檢索且儲存於記憶體1110上以備由處理器1105獲取。在一些情形下，可排除電子儲存單元1115，且將機器可執行指令儲存於記憶體1110上。

【0147】 程式碼可進行預編譯且經組態用於與具有經調適以執行該程式碼之處理器之機器一起使用，或可在運行時間期間編譯。該程式碼可以可經選擇以使得程式碼能夠以預編譯或剛編譯方式來執行之程式設計語言形式來供應。

【0148】 本文所提供之系統及方法之態樣(諸如電腦系統1101)可體現於程式設計中。技術之各種態樣可被認為係通常呈機器(或處理器)可執行碼及/或承載於或體現在一種類型之機器可讀取媒體中的相關資料形式的「產物」或「製品」。機器可執行碼可儲存於電子儲存單元(諸如，記憶體(例如唯讀記憶體、隨機存取存儲器、快閃記憶體)或硬碟)上。「儲存」型媒體可包括電腦之有形記憶體、處理器或其類似物或其相關模組諸如各種半導體記憶體、磁帶機、磁碟機及其類似物中之任一者或全部，其可提供用於軟體程式設計之任何時間之非瞬時性儲存。軟體之全部或部分可時常經由網際網路或各種其他電信網路來通信。該等通信例如可致能軟體自一個電腦或處理器負載至另一者中，例如自管理伺服器或主電腦負載至應用程式伺服器之電腦平台中。因此，可攜帶軟體元件之另一類型之媒體包括諸如跨越本地器件之間之物理界面、經由有線及光學陸上通訊網路及越過各種空中鏈路來使用之光波、電波及電磁波。攜載此類波之物理元件，

諸如有線或無線鏈路、光學鏈路或類似物亦可視為攜帶軟體之媒體。如本文所使用，除非限於非暫時性有形「儲存」媒體，否則諸如電腦或機器「可讀取媒體」之術語係指參與向處理器提供指令以用於執行之任何媒體。

【0149】 因此，諸如電腦可執行代碼之機器可讀取媒體可採用多種形式，包括(但不限於)有形儲存媒體、載波媒體或物理傳輸媒體。非揮發性儲存媒體包括例如光碟或磁碟，諸如任何電腦或類似物中之存儲裝置中之任一者，諸如可用於實施附圖中所展示之資料庫等。揮發性儲存媒體包括動態記憶體，諸如此類電腦平台之主記憶體。有形傳輸媒體包括同軸電纜；銅線及光纖，包括包含電腦系統內之匯流排之電線。載波傳輸媒體可採用以下形式：電信號或電磁信號，或聲波或光波，諸如在射頻(RF)及紅外線(IR)資料通信期間所產生之聲波或光波。因此，電腦可讀取媒體之常見形式包括例如：軟碟、可撓性磁碟、硬碟、磁帶、任何其他磁性媒體、CD-ROM、DVD或DVD-ROM、任何其他光學媒體、打孔卡紙帶、任何其他具有洞圖案之實體儲存媒體、RAM、ROM、PROM及EPROM、FLASH-EPROM、任何其他記憶體晶片或盒、傳輸資料或指令之載波、傳輸此類載波之電纜或鏈路或電腦可自其中讀取程式碼及/或資料之任何其他媒體。此等形式之電腦可讀取媒體中之多者可涉及將一或多個序列之一或多個指令攜載至處理器以用於執行。

【0150】 電腦系統1101可包括或與包含使用者界面(UI) 1140之電子顯示器1135通信，以便提供例如樣本分析結果，諸如(但不限於)病原體整合型態之圖解顯示、病原體整合斷點之基因組位置、病理學分類(例如疾病或癌症之類型及癌症程度)及基於病理學分類之預防性步驟之治療暗示

或建議。UI之實例包括(但不限於)圖形使用者界面(GUI)及基於網站之使用者界面。

【0151】 本發明之方法及系統可藉助於一或多個演算法來實施。算法在由中央處理單元1105執行時可藉助於軟體來實施。演算法可例如控制來自樣本之核酸分子之測序、導入測序資料之收集、分析測序資料或基於測序資料之分析決定病理學分類。

【0152】 在一些情況下，如圖9中所示，樣本902可獲自個體901，諸如人類個體。樣本902可經受如本文所描述之一或多種方法，諸如進行分析。在一些情況下，分析可包含雜交、擴增、測序、標記、表觀遺傳修飾鹼基或其任何組合。可將方法產生的一或多個結果輸入至處理器904中。可將諸如樣本鑑別、個體鑑別、樣本類型、參考物或其他資訊之一或多個輸入參數輸入至處理器904中。來自分析之一或多個量度可輸入至處理器904中，使得該處理器可產生結果，諸如病理學分類(例如診斷)或治療建議。處理器可將結果、輸入參數、量度、參考或其任何組合發送至顯示器905，諸如視覺顯示器或圖形使用者界面。處理器904可(i)將結果、輸入參數、量度或其任何組合發送至服務器907中，(ii)自服務器907接收結果、輸入參數、量度或其任何組合，(iii)或其組合。

【0153】 本發明之態樣可以控制邏輯之形式使用硬體(例如，特殊應用積體電路或場可程式化閘陣列)及/或使用電腦軟體，利用大體可程式化處理器以模組化或整合方式實施。如本文所使用，處理器包括位於同一積體晶片上之單核處理器、多核處理器，或位於單一電路板上或網路化之多個處理單元。基於本發明及本文所提供之教示內容，一般熟習此項技術者應知道及瞭解使用硬體及硬體與軟體之組合來實施本文所描述之實施例的

其他方式及/或方法。

【0154】 本申請中所描述之軟體組件或功能中之任一種可以待使用例如習知或面向對象之技術使用任何合適的電腦語言(諸如(例如) Java、C、C++、C#、Objective-C、Swift)或腳本語言(諸如Perl或Python)由處理器執行之軟體代碼形式來實施。軟體程式碼可儲存為電腦可讀取媒體上之一系列指令或命令以用於儲存及/或傳輸。適合的非暫時性電腦可讀取媒體可包括隨機存取記憶體(RAM)、唯讀記憶體(ROM)、磁性媒體(諸如硬碟機或軟碟機)或光學媒體，諸如光盤(CD)或DVD (數位化通用光碟)、快閃記憶體及其類似者。電腦可讀取媒體可為此類儲存或傳輸裝置之任何組合。

【0155】 此類程式亦可使用適用於經由符合多種協定之有線、光學及/或無線網路(包括網際網路)傳輸的載波信號來編碼及傳輸。因此，電腦可讀取媒體可使用以此類程式編碼的資料信號建立。以程式碼編碼之電腦可讀取媒體可與相容裝置一起封裝或與其他裝置分開提供(例如藉助於網際網路下載)。任何此類電腦可讀取媒體可駐存於單個電腦產品(例如硬碟機、CD或整個電腦系統)上或其內部，且可存在於系統或網路內之不同電腦產品上或其內部。電腦系統可包括用於向使用者提供本文所提及之任何結果的監測器、印表機或其他合適之顯示器。

【0156】 本文所描述之任何方法可完全或部分地使用電腦系統來執行，該電腦系統包括一或多個處理器，該等處理器可經組態以執行該等步驟。因此，實施例可係關於經組態以執行本文所描述之方法中之任一者之步驟的電腦系統，其使用不同組件執行各別步驟或各別步驟組。儘管本文中方法之步驟以經編號步驟呈現，但其可同時或以不同次序執行。另外，

此等步驟之部分可與其他方法之其他步驟之部分一起使用。此外，所有或部分步驟可為視情況選用的。另外，任何方法中之任何步驟可使用用於執行此等步驟的模組、單元、電路或其他途徑來執行。

XIV. 其他實施例

【0157】 本文所使用之部分標題僅出於組織目的且不應理解為限制所描述之主題。

【0158】 應理解，本文所描述之方法不限於本文所描述之特定方法、協定、主題及測序技術且因此可變化。亦應理解，本文所使用之術語僅出於描述特定實施例之目的，而並不意欲限制本文所描述之方法及組合物之範疇，該範疇將僅受隨附申請專利範圍限制。雖然本文已顯示及描述本發明之一些實施例，但對於熟習此項技術者應顯而易見的是，此類實施例僅藉助於實例提供。熟習此項技術者現將在不背離本發明之情況下想到許多變化、改變及取代。應理解，本文所描述之本發明之實施例之各種替代方案可在實踐本發明時使用。預期以下申請專利範圍界定本揭示之範疇，且因此涵蓋此等申請專利範圍及其等效者之範疇內的方法及結構。

【0159】 參考用於說明之實例應用來描述若干態樣。除非另外指明，否則任何實施例可與任何其他實施例組合。應理解，闡述許多具體詳情、關係及方法以提供對本文所描述之特徵的充分理解。然而，熟習此項技術者應容易認識到，可在沒有一或多個具體詳情之情況下或使用其他方法來實踐本文所描述之特徵。本文所描述之特徵不受所示行為或事件之順序限制，因為一些行為可以不同的順序發生及/或與其他行為或事件同時發生。此外，實施根據本文所描述之特徵的方法不需要所有所示動作或事件。

XV. 實例

【0160】 以下實例係以說明而非限制之方式提供。

【0161】 實例1. HPV DNA整合之分析

【0162】 此實例藉由分析來自血漿樣本之無細胞DNA之序列讀段來測試用於測定HPV整合斷點之例示性途徑。實例亦檢驗患有子宮頸癌及HPV相關頭頸部鱗狀細胞癌之患者之樣本中之不同HPV整合圖譜，以及具有EBV感染之患者之樣本中之EBV整合型態。

【0163】 收集且處理來自具有HPV感染之患者之血漿樣本。隨後收集血漿樣本中之無細胞DNA分子且經受配對末端測序。在配對末端測序資料中，選擇來自如圖2中所說明之A型及B型人類-HPV嵌合片段之序列讀段。

【0164】 為了鑑別無細胞DNA樣本中之人類-HPV嵌合DNA，使用SOAP比對工具將所獲得之序列讀段與參考人類基因組及參考HPV基因組進行比對。藉由搜尋具有映射至參考人類基因組之一個讀段及映射至參考HPV基因組之另一讀段的讀段對(獲自相同漿細胞-游離DNA分子之兩端之成對序列讀段)鑑別A型人類-HPV嵌合片段讀段對。移除具有相同起點及末尾外部座標之A型嵌合片段作為疑似PCR複製品。所有其餘的A型嵌合讀段隨後覆疊以將人類及HPV序列中重疊或彼此鄰接之核苷酸座標分組以鑑別候選整合區域。人類基因組上之各候選整合區域將具有HPV基因組上之一或多個對應候選整合區域，且反之亦然。在300個鹼基之距離內之相鄰人類讀段分組至一個候選整合區域中，且在300個鹼基之距離內之相鄰HPV讀段分組至一個候選整合區域中(例如如圖3A中所說明)。與其最近相鄰序列讀段距離超過300個鹼基之任何人類序列讀段不包括於相同候選整

合區域中，且與其最近相鄰序列讀段距離相隔超過300個鹼基之任何HPV人類序列讀段不包括於相同候選整合區域中。在此實例中，在各人類-HPV候選整合區域中，必須存在不為PCR複製品之至少兩個人類-HPV嵌合片段。

【0165】來自第一比對之未映射讀段隨後使用具有22個鹼基之默認最小映射長度之局部比對功能之Bowtie2將人類及HPV基因組重新比對。隨後在A型嵌合片段所表明之候選整合區域中搜尋其可能的映射位置。隨後選擇4個鹼基之最小映射長度以將部分映射讀段之未映射部分映射至具有4 bp之最小映射長度之對應整合候選區域。因此，在候選整合區域內鑑別B型片段讀段對中之人類-HPV嵌合讀段以測定潛在整合斷點。在候選整合區域內，藉由用於構築候選整合區域之A型嵌合片段讀段對之股資訊推斷人類-HPV嵌合讀段之預期股取向。僅在潛在斷點之100 bp距離內之短序列錨定物之映射位置視為整合斷點之推定擊點。此等讀段視為涵蓋整合斷點之嵌合讀段。

【0166】進一步排除配對讀段不與庫插入尺寸及映射定向相容之人類-HPV嵌合讀段。隨後，排除多樣性評分低於4之候選整合位點。其餘整合斷點視為來自實際HPV整合事件。表1概述在具有或不具有多樣性評分過濾之情況下用來自患有子宮頸癌及HNSCC之患者之樣本來分析HPV整合斷點。

【0167】用來自具有EBV感染之鼻咽癌(NPC)患者之血漿樣本進行EBV整合斷點之類似分析。在具有或不具有多樣性評分過濾之情況下EBV整合斷點之分析概述於表2中。

【0168】圖6展示基於來自分別患有子宮頸癌及HPV陽性頭頸部鱗

狀細胞癌之患者之血漿無細胞DNA之分析，HPV16基因組中之病毒整合斷點之分佈。如圖中所示，在子宮頸癌與HPV-陽性頭頸部鱗狀細胞癌(區域A)之間存在共有的較佳整合斷點。另一方面，亦存在某一類型之癌症之較佳整合斷點特性，例如區域B展示HNSCC較佳整合斷點。

表1. HPV整合分析之概述

類型	樣本	候選區域之全部編號	不具有多樣性評分過濾		多樣性評分>=4	
			嵌合讀段之編號	斷點之編號	嵌合讀段之編號	斷點之編號
子宮頸癌	3485	458	28317	1054	10723	6
	C-788	43	77	26	0	0
	C-801	2	0	0	0	0
	C-819	2	0	0	0	0
	C-822	5	0	0	0	0
	C-877	9	0	0	0	0
HNSCC	TBR 1019	8	0	0	0	0
	TBR1245	1	0	0	0	0
	TBR1988	3	0	0	0	0
	TBR1989	12414	234667	96039	1521	41
	TBR2175	22	9	3	0	0

表2. EBV整合分析之概述

樣本	候選區域之全部編號	不具有多樣性評分過濾		多樣性評分>=4	
		嵌合讀段之編號	斷點之編號	嵌合讀段之編號	斷點之編號
TBR1358	1	1	1	0	0
TBR1378	45	117	59	0	0
TBR1390	188	1106	410	0	0
TBR1344	3153	33405	10426	0	0
TBR1360	913	14293	3550	0	0
TBR1379	1334	22027	4932	0	0

【0169】 實例2. HPV DNA整合指數之分析

【0170】 此實例檢驗來自患有子宮頸癌及HNSCC之患者之血漿樣本之無細胞DNA中之HPV DNA整合指數。

【0171】 為了分析展示HPV整合至癌症基因組內人類染色體中之血漿中之無細胞DNA片段，用此等無細胞DNA片段進行配對末端測序。使用讀取來自無細胞DNA片段之各端之75個核苷酸的配對末端測序方案。

具有映射至HPV基因組之一個讀段序列及映射至人類基因組之另一讀段序列的所有75個核苷酸之無細胞DNA片段視為人類-HPV嵌合片段。具有映射至HPV基因組之配對末端測序中之兩種讀段之無細胞DNA片段視為血漿HPV讀段。分析來自11名患有子宮頸癌之患者之樣本。

表3. CC樣本中之無漿細胞人類-HPV片段DNA

樣本	HPV類型	人類-HPV嵌合片段之讀段編號	血漿HPV讀段之編號
3276	HPV 16	12	62
3485	HPV 16	180892	2316992
3499	HPV 16	4	72
3542	HPV 16	5	50
3581	HPV 16	12	90
C-788	HPV 16	108848	1659430
C-801	HPV 16	10531	153660
C-803	HPV 16	3271	53544
C-819	HPV16	1466	20806
C-822	HPV 16	2461	59532
C-877	HPV 16	37126	616528

【0172】 使用用於界定無漿細胞人類-HPV嵌合片段之同一協定，在來自一個患有子宮頸癌之患者之樣本(CaCx3485)及來自一個患有HPV陽性頭頸部鱗狀細胞癌(HPV+ve HNSCC)之患者之樣本(TBR1989)中分析無漿細胞人類-HPV嵌合片段之基因組位置。在圖7中，各點代表具有HPV整合之無漿細胞DNA片段之精確末端位置。整合指數係指具有HPV整合之無漿細胞DNA片段之末端位置相比於血漿HPV片段之幾率比(映射至HPV基因組之讀段1及2兩者)。整合指數愈高，彼等DNA片段之位置愈可能以HPV整合末端讀取。在患有子宮頸癌及HPV陽性HNSCC (HPV+ve HNSCC)之患者之間的整個HPV基因組中具有HPV整合之無漿細胞DNA讀段之較佳末端位置中觀測到差異。

【0173】 實例3.利用公開軟體ViFi之比較

【0174】 在此實例中，將根據本發明之HPV整合偵測中之例示性算

法之效能與 ViFi (Nguyen 等人, Nucleic Acids Res. 2018 Apr 20;46(7):3309-3325. doi: 10.1093/nar/gkyl80)進行比較，其為經由譜系學方法及基於參考物之讀段比對之組合偵測病毒整合位點之軟體。具有人工構築整合位點之模擬數據集以及來自26名子宮頸上皮內贅瘤(CIN)、26名子宮頸癌(CaCx)及7名頭頸部鱗狀細胞癌(HNSCC)患者之組織樣本及成對血漿樣本之生物學資料用於評估例示性算法及ViFi之偵測效能。藉由例示性算法鑑別出之整合位點進一步藉由PCR擴增及桑格測序(Sanger sequencing)檢驗。

【0175】 方法

【0176】 公開軟體：ViFi

【0177】 ViFi自github.com/namphuon/ViFi下載且在其默認設置及來自fastq型式之子宮頸癌及HNSCC患者之組織及血漿樣本之模擬資料及真實測序資料之默認參數下執行。

【0178】 模擬數據集

【0179】 為了評估例示性算法，經由計算機構築包括A型及B型片段兩者之32,500個嵌合片段以人工產生500個HPV整合位點。此等經由計算機構築之嵌合片段隨後與一個人類子宮頸癌血漿樣本之真實測序讀段混合。人類子宮頸癌血漿樣本之測序讀段不含有HPV整合位點。用ViFi (Nguyen等人, Nucleic Acids Res. 2018 Apr 20;46(7):3309-3325. doi: 10.1093/nar/gky 180)比較例示性算法之偵測效能。

【0180】 來自子宮頸癌、HNSCC患者及NPC患者之組織及血漿DNA之測序資料

【0181】 為了探究HPV相關腫瘤中之HPV16整合橫向，對26個來自

子宮頸癌之血漿樣本及7個組織樣本及來自HNSCC患者之成對血漿樣本進行測序。應用例示性算法及ViFi兩者以便偵測此等樣本中之HPV整合位點。亦對6個來自鼻咽癌(NPC)患者之血漿樣本進行測序以便偵測人類基因組中之艾伯斯坦-巴爾病毒(EBV)整合事件。

【0182】 結果

【0183】 模擬資料

【0184】 為了直接評估例示性算法之偵測效能且將其與ViFi進行比較，如對應嵌合片段所支持人工構築500個整合位點。例示性算法及ViFi兩者均在此模擬數據集上執行且比較偵測比率(圖10)。如圖11中所示，ViFi以僅62.8%之偵測比率偵測全部500位點中之314個模擬整合區域。然而，使用例示性算法，以96%之顯著較高偵測比率，精確鑑別480個模擬整合位點，不僅包括整合區域且亦包括精確斷點。此等結果展示，相比於ViFi，例示性算法在偵測模擬數據集中之整合位點方面具有更高準確度。

【0185】 自血漿DNA鑑別之斷點

【0186】 對26個CIN、26個子宮頸癌及7個HNSCC患者之血漿DNA測序資料執行例示性算法，其中6個子宮頸癌及5個HNSCC血漿樣本展示具有至少1個候選整合區域(圖12)。在用4之截止多樣性評分過濾之後，1個子宮頸癌(CaCx3485)及1個HNSCC (TBR1989)病例分別保持6個及41個藉由例示性算法鑑別之斷點(圖12)。在此等斷點中，選擇6個及9個斷點以分別用於子宮頸癌及HNSCC病例中之實驗驗證，其中，子宮頸癌病例中之1個斷點及HNSCC病例中之3個斷點藉由PCR分析驗證。同時，對相同血漿測序數據集執行ViFi以用例示性算法進行比較。在藉由HPV16開始之491個推定整合斷點之情況下，ViFi僅鑑別一個病例CaCx3485。然而，藉

由ViFi鑑別之斷點中無一者與藉由例示性算法鑑別之斷點中之任一者(多樣性評分大於4)重疊，在此情況下包括經驗證的斷點。

【0187】 自腫瘤組織DNA鑑別之斷點

【0188】 另外，對前述HNSCC患者之7個組織樣本進行測序。4之多樣性評分用作例示性算法中之截斷值。**圖13**展示4/5樣本含有在應用截止評分之後保持至少一個鑑別整合斷點之至少一個候選整合區域。此等斷點進一步以實驗方式藉由PCR及桑格測序驗證。

【0189】 總共存在56個自多樣性評分 ≥ 4 之4個HNSCC樣本鑑別之整合斷點(**圖13**)。在與緊密連接位點(1-5 bp)合併26個斷點之後，斷點之總數目降低至43。針對側接尺寸小於200 bp之擴增子中之整合位點的32個斷點設計PCR分析。23個斷點已成功地自對應樣本之組織DNA擴增，展示預期擴增子尺寸且其中，進一步藉由桑格測序驗證22個斷點。因此，基於吾等算法，多樣性評分 ≥ 4 之斷點總體驗證比率為69.0% (22/32)。

【0190】 若多樣性評分截斷值設定為9，則藉由經設計之PCR分析靶向24個斷點，其中22個已成功地藉由桑格測序擴增及驗證。因此，在截止多樣性評分自4升高至9之後，總體驗證比率顯著提高至91.7% (22/24)。

【0191】 總之，使用組織測序資料將例示性算法之效能與ViFi進行比較。藉由ViFi任何組織樣本中均未鑑別出斷點。

【0192】 雖然已在本文中顯示並描述本發明之較佳實施例，但熟習此項技術者應顯而易見，此等實施例僅以舉例方式提供。熟習此項技術者現將在不背離本發明之情況下想到許多變化、改變及取代。應理解，本文所描述之實施例之各種替代方案可在實踐本發明時使用。預期以下申請專

利範圍界定本揭示之範疇，且因此涵蓋此等申請專利範圍及其等效者之範疇內的方法及結構。

【符號說明】

【0193】

110	步驟
120	步驟
130	步驟
901	個體
902	樣本
904	處理器
905	顯示器
907	服務器
1101	電腦系統
1105	中央處理單元/CPU/處理器/電腦處理器
1110	記憶體
1115	儲存單元
1120	通信接口
1125	周邊裝置
1130	網路
1135	電子顯示器
1140	使用者界面

【發明申請專利範圍】

【請求項1】

一種分析來自人類個體之生物樣本之無細胞核酸分子以鑑別該個體之基因組中之整合斷點位置之方法，該人類個體疑似在未知的基因組位置包含病原體整合，該方法包括：

(a) 對來自該個體之生物樣本之無細胞核酸分子進行測序以產生配對末端測序讀段；

(b) 將該配對末端測序讀段與(i)參考人類基因組和(ii)病原體的參考基因組進行比對，其中該比對使用比對演算法；

(c) 對一個或多個嵌合讀段對 (chimeric read pairs) 進行鑑別，其中嵌合讀段對包含(i)與該病原體的參考基因組比對的第一序列讀段及(ii)與該參考人類基因組比對的第二序列讀段；及

(d) 將該一個或多個嵌合讀段對之第二序列讀段的500個核苷酸內之個體的基因組中之未知的基因組位置鑑別為整合斷點位置。

【請求項2】

如請求項1之方法，其中該比對包括將該第一序列讀段之至少20個連續核苷酸與該病原體之該參考基因組進行比對，且將該第二序列讀段之至少20個連續核苷酸與該參考人類基因組進行比對。

【請求項3】

如請求項1之方法，其中測序該無細胞核酸分子包括將來自該生物樣本之無細胞核酸分子，與該病原體之基因組中之第一靶序列互補的第一引子及與該參考人類基因組中之第二靶序列互補的第二引子進行擴增。

【請求項4】

如請求項1之方法，其中該病原體包含病毒。

【請求項5】

如請求項4之方法，其中該病毒包含艾伯斯坦-巴爾病毒DNA、人類乳頭狀瘤病毒DNA、B型肝炎病毒DNA、C型肝炎病毒核酸、或其片段。

【請求項6】

如請求項1之方法，其中該生物樣本為血漿、血清或尿液。

【請求項7】

如請求項1之方法，其中步驟(c)另包括：

鑑別來自包含相同潛在整合斷點之生物樣本之無細胞核酸分子之複數個不同嵌合成對序列讀段。

【請求項8】

如請求項7之方法，其中步驟(d)另包括：

測定該等複數個不同嵌合成對序列讀段中之每一者之第一序列讀段及第二序列讀段之股取向；

濾除包含與大部分嵌合讀段對之第一序列讀段及第二序列讀段之股取向不一致的第一序列讀段及第二序列讀段之股取向的嵌合讀段對；及

在濾除之後，基於剩餘的嵌合讀段對偵測該整合斷點。

【請求項9】

如請求項7之方法，其中步驟(d)另包括：

評定與側接該潛在整合斷點之基因組區域比對之成對序列讀段之序列讀段之序列長度變化。

【請求項10】

如請求項7之方法，其中鑑別該等複數個嵌合讀段對包括：

鑑別由來自包含與該參考人類基因組比對之第一序列讀段及與該病原體之參考基因組比對之第二序列讀段的生物樣本之無細胞核酸分子之配對末端測序產生的人類-病原體嵌合成對序列讀段，由此鑑別A型人類-病原體嵌合成對序列讀段；

自該A型人類-病原體嵌合成對序列讀段分出一組A型人類-病原體嵌合成對序列讀段，其包含在該參考人類基因組中之預定距離內重疊或間隔開之第一序列讀段及在該病原體之參考基因組中之預定距離內重疊或間隔開之第二序列讀段，由此鑑別在參考基因組中之人類-病原體候選整合區域；及

鑑別由來自包含與該人類-病原體候選整合區域比對之第一序列讀段及包含與該參考人類基因組比對之第一序列及與該病原體之參考基因組比對之第二序列的第二序列讀段之生物樣本之無細胞核酸分子之配對末端測序產生的人類-病原體嵌合成對序列讀段，由此鑑別B型人類-病原體嵌合成對序列讀段。

【請求項11】

如請求項10之方法，其中該預定距離為至多300個鹼基。

【請求項12】

如請求項10之方法，其中鑑別該整合斷點，包括：

測定該等A型人類-病原體嵌合成對序列讀段及該等B型人類-病原體嵌合成對序列讀段中之每一者之該第一序列讀段及該第二序列讀段之股取向；

自該等B型人類-病原體嵌合成對序列讀段濾除具有與該人類-病原體候選整合區域內大部分A型人類-病原體嵌合成對序列讀段之第一序列讀段

及第二序列讀段之股取向不一致的第一序列讀段及第二序列讀段之股取向之B型人類-病原體嵌合成對序列讀段；及

在濾除之後，基於該等B型人類-病原體嵌合成對序列讀段，偵測該整合斷點。

【請求項13】

如請求項10之方法，其另包括：

測定該等B型人類-病原體嵌合成對序列讀段之多樣性評分；其中該多樣性評分之計算法為

$$\frac{\sigma_1 + \sigma_2}{\max(\frac{\sigma_1}{\sigma_2}, \frac{\sigma_2}{\sigma_1})},$$

其中 σ_1 為與該參考人類基因組比對之B型人類-病原體嵌合成對序列讀段之第一序列長度之標準差，且其中 σ_2 為與該參考人類基因組比對之B型人類-病原體嵌合成對序列讀段之第二序列長度之標準差；及

若該多樣性評分等於或高於預定截斷值，則基於該等B型人類-病原體嵌合成對序列讀段，偵測該整合斷點。

【請求項14】

如請求項13之方法，其中該預定截斷值為至少4.0。

【請求項15】

一種包含一或多個處理器及非暫時性電腦可讀取媒體之電腦系統，該非暫時性電腦可讀取媒體包含可操作的指令，當由一或多個電腦處理器執行時，可以使得該電腦系統執行如請求項1之方法。

【請求項16】

一種非暫時性電腦可讀取媒體，其包含可操作的指令，當由電腦系統之一或多個處理器執行時，可以使得該電腦系統執行如請求項1之方

法。

【請求項17】

如請求項1之方法，其另包括至少部分基於該整合斷點位置，測定在個體中的病理學分類。

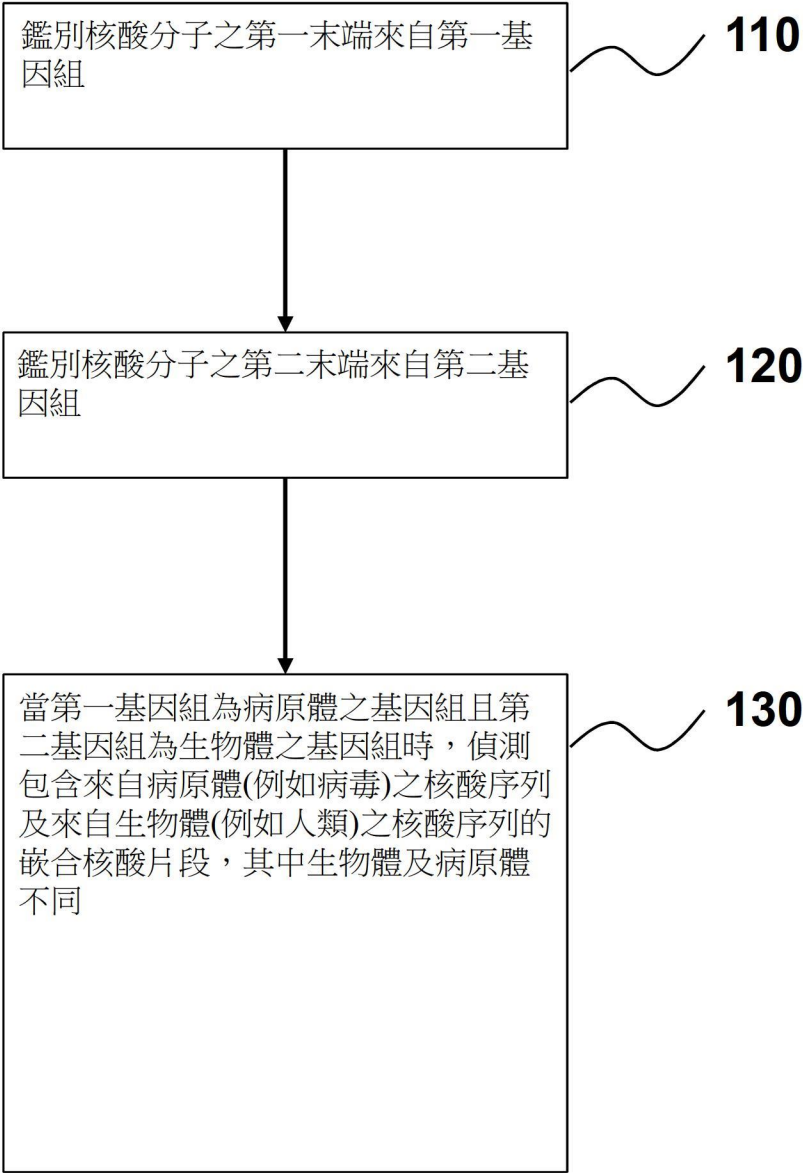
【請求項18】

如請求項17之方法，其中該病理學分類包含癌症之類型。

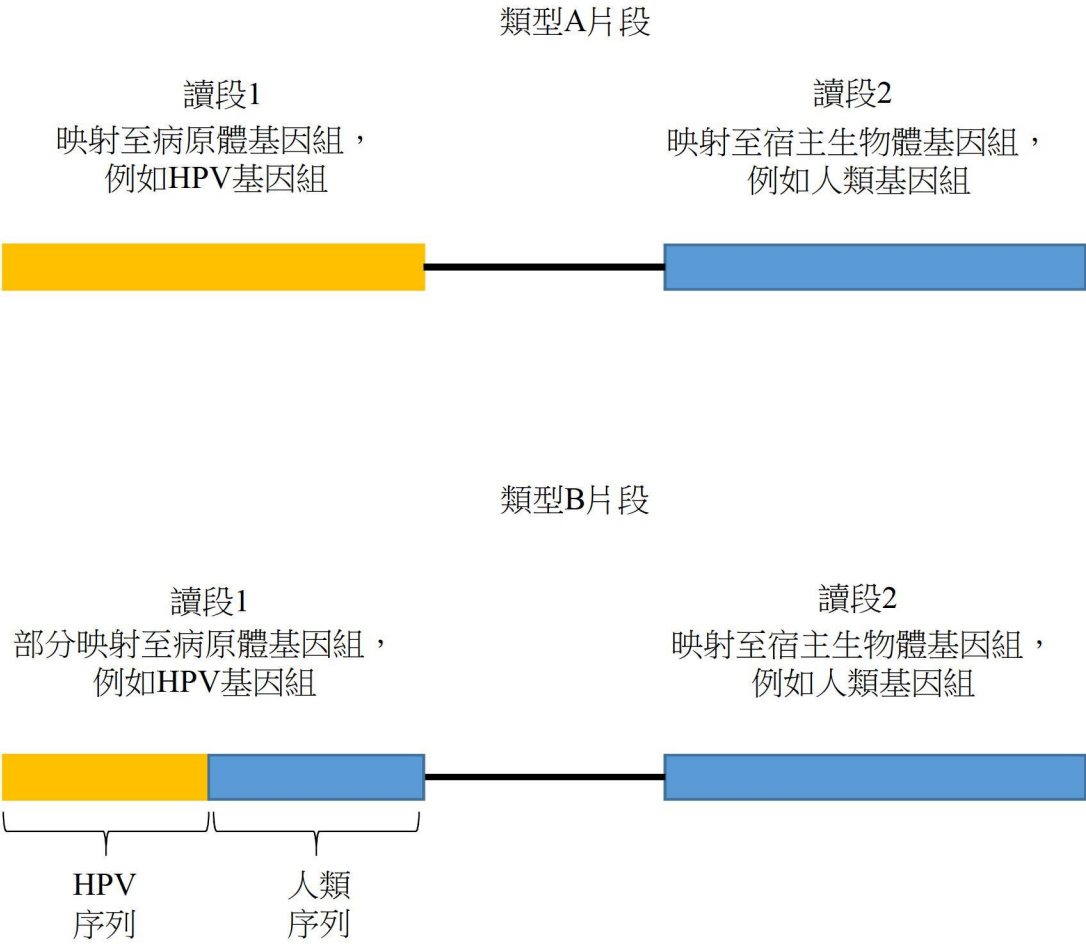
【請求項19】

如請求項18之方法，其中該癌症之類型包含子宮頸癌或頭頸部鱗狀細胞癌。

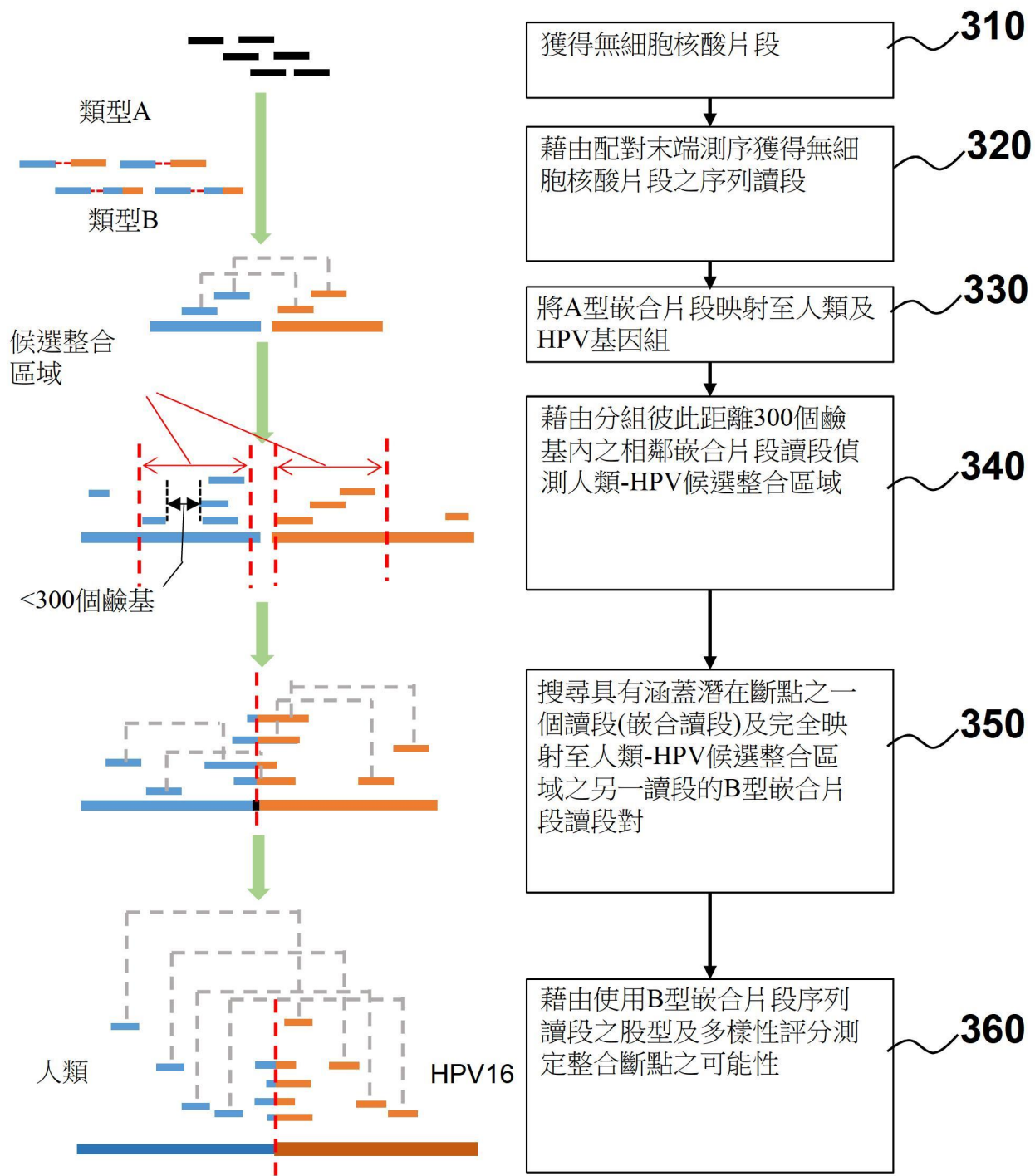
【發明圖式】



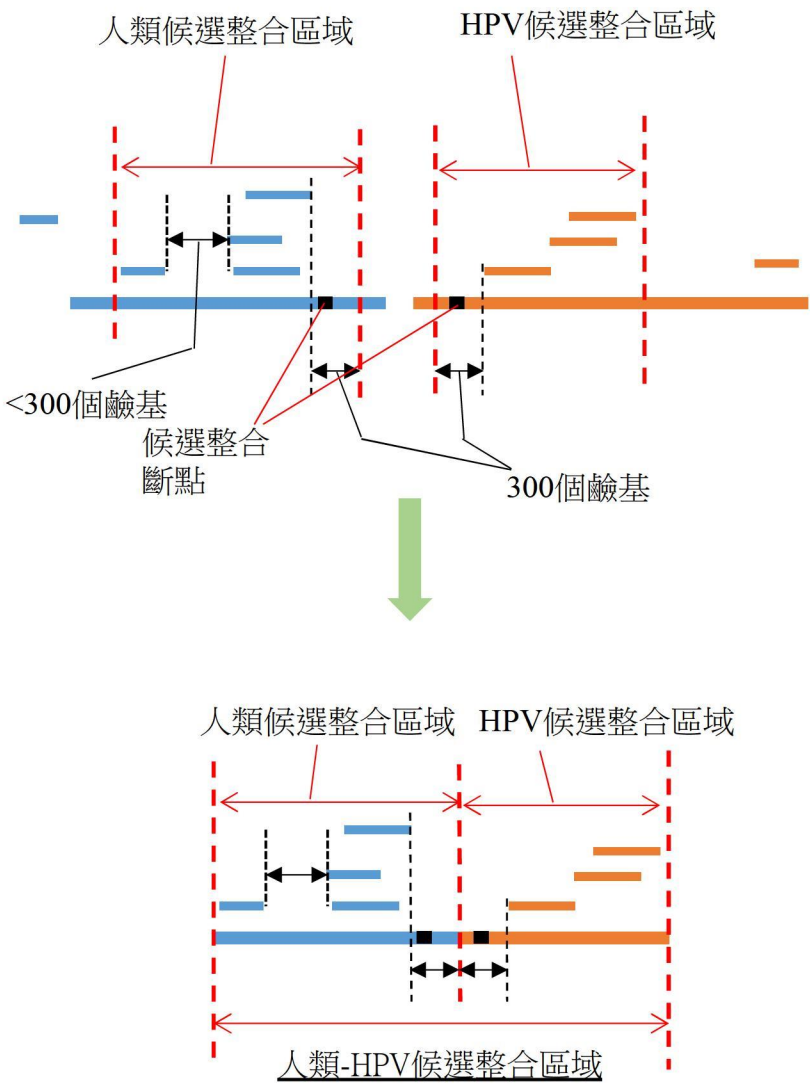
【圖1】



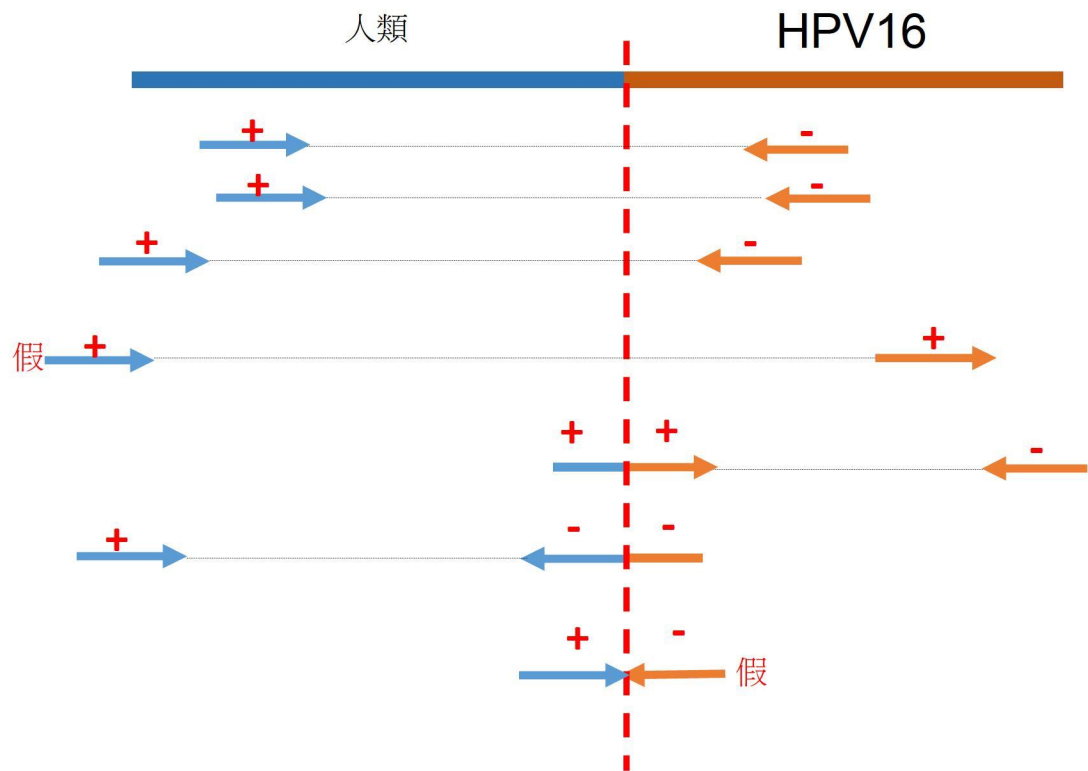
【圖2】



【圖3A】

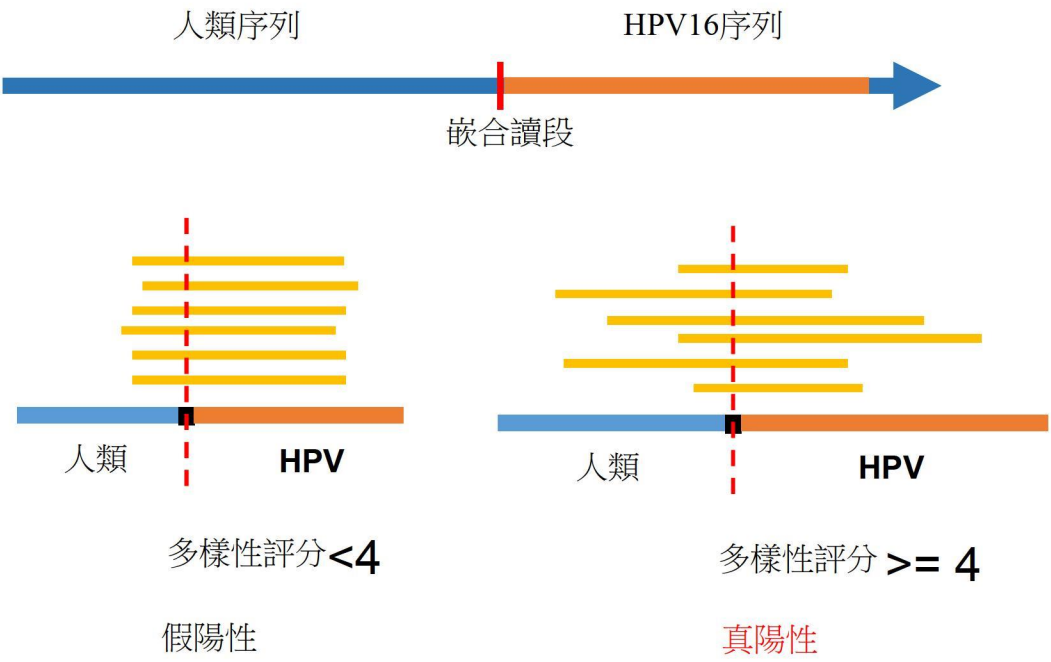


【圖3B】

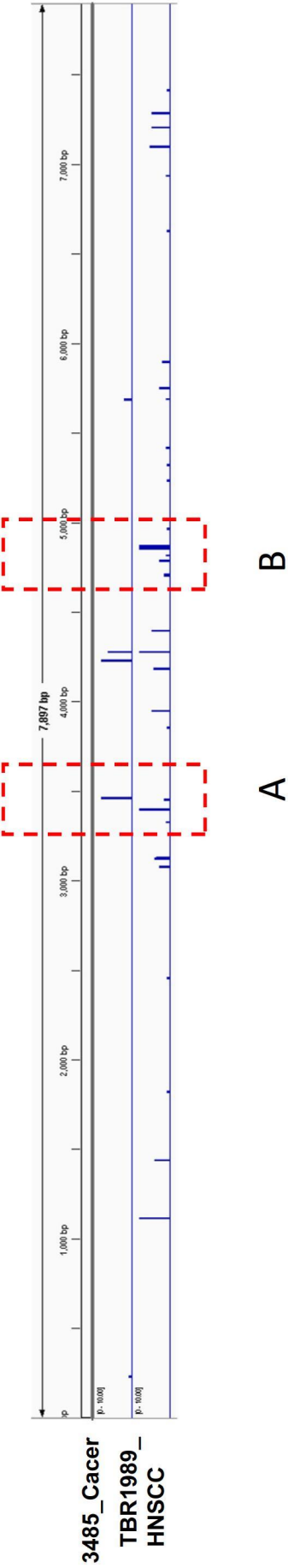


A型人類-HPV片段讀段對		人類-HPV嵌合 讀段之預期股	人類序列之斷 點之相對位置	HPV序列之斷 點之相對位置
映射至人類基 因組之讀段股	映射至HPV基 因組之讀段股			
+	+	+- 或 -+	下游	下游
+	-	++ 或 --	下游	上游
-	+	-- 或 ++	上游	下游
-	-	-+ 或 +-	上游	上游

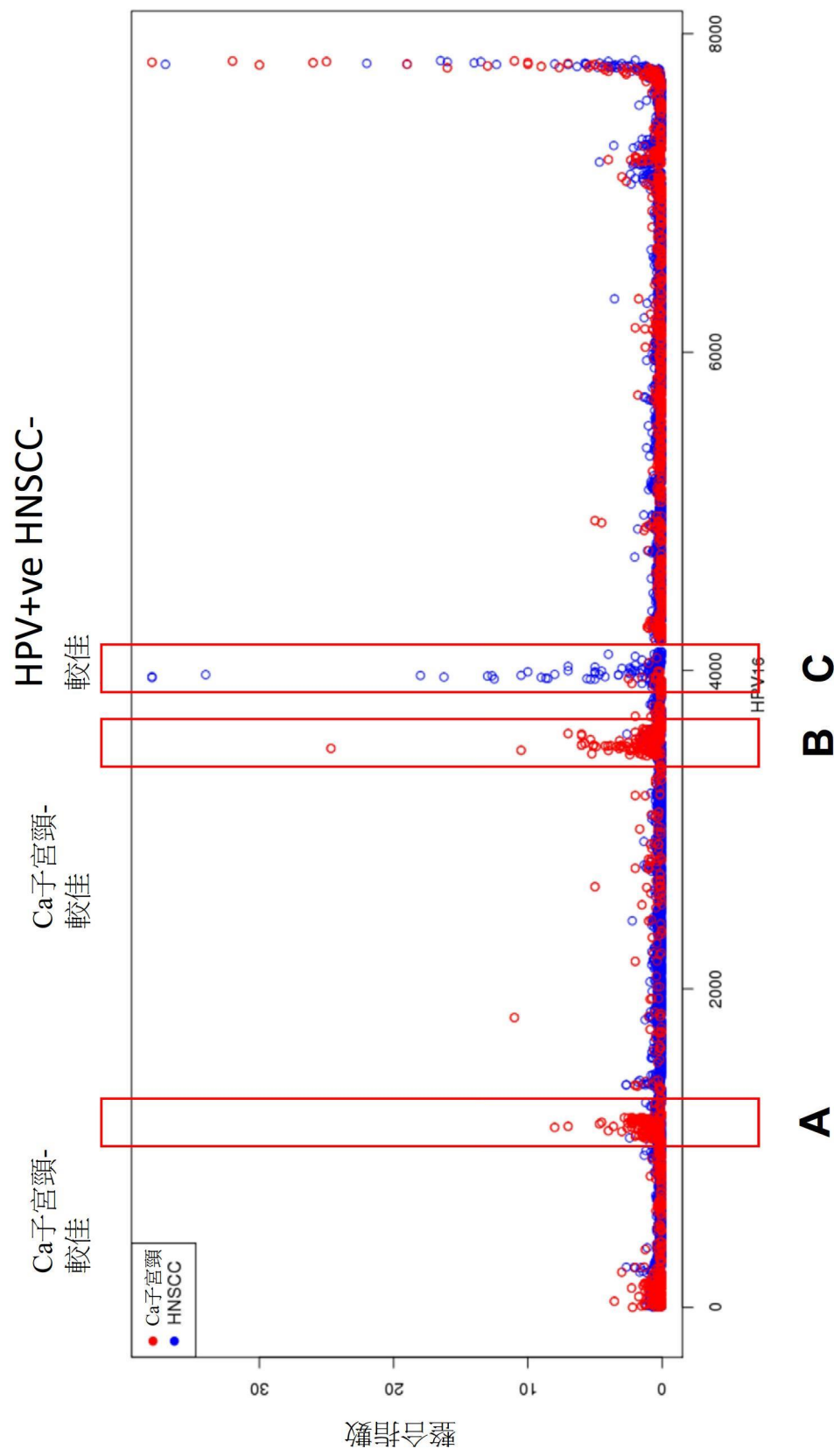
【圖4】



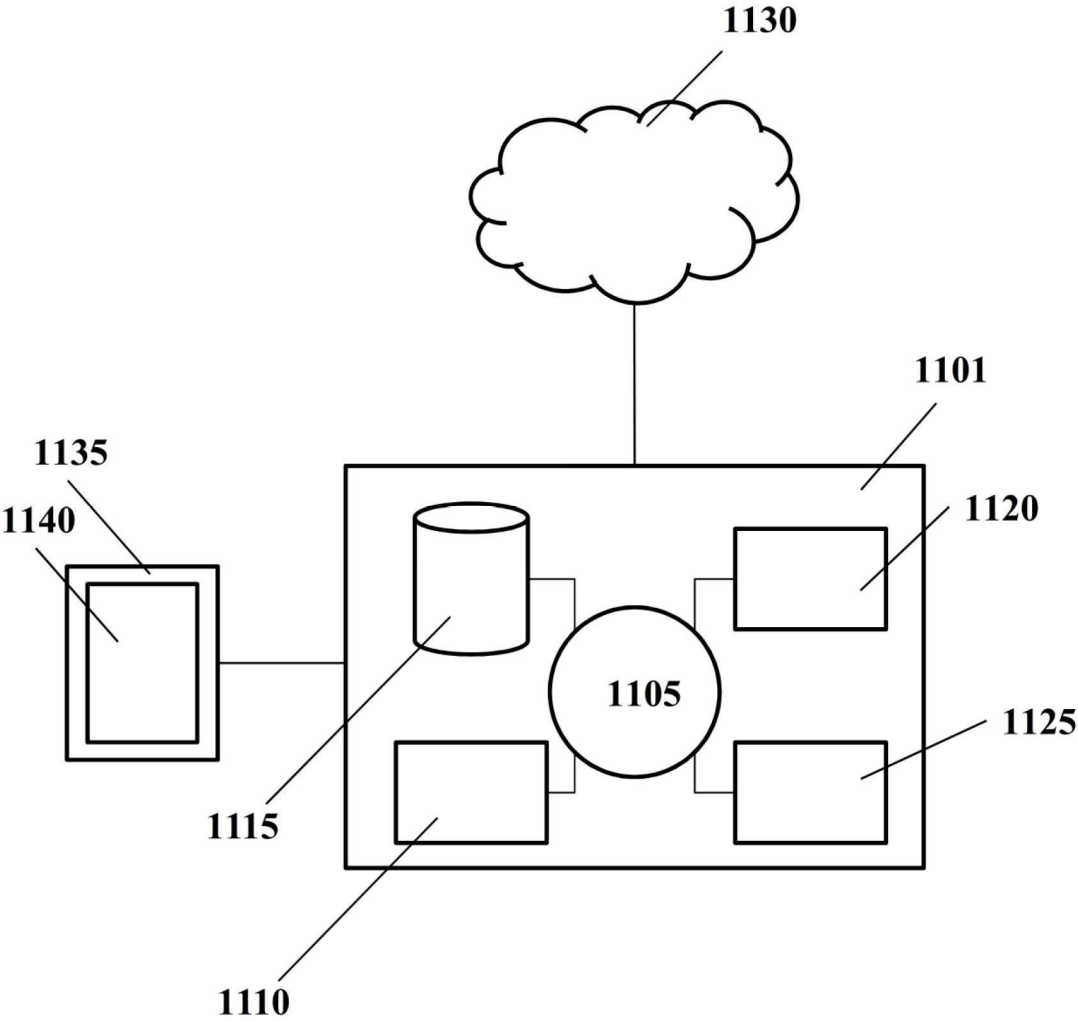
【圖5】



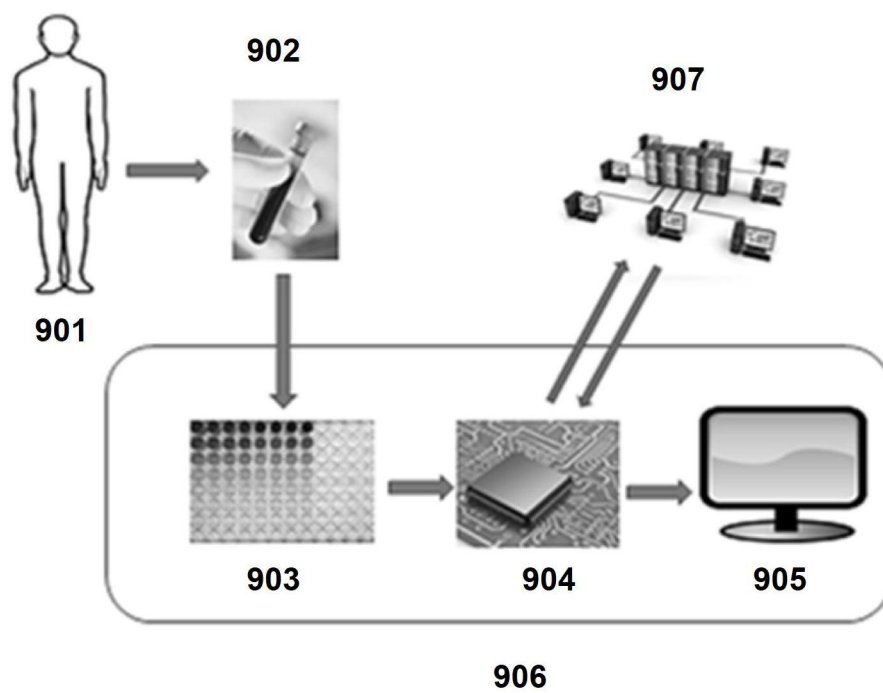
【圖6】



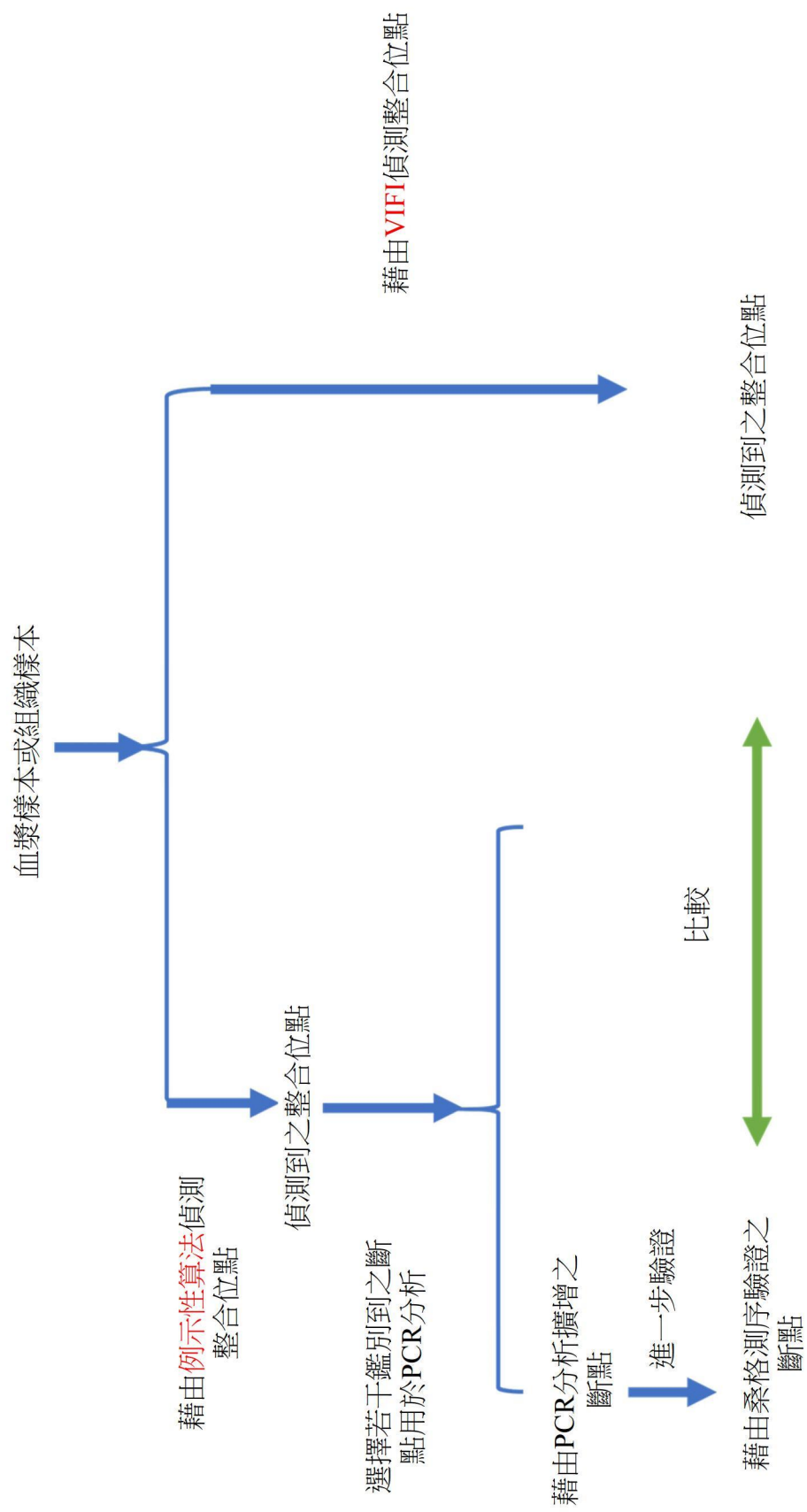
【圖7】



【圖8】

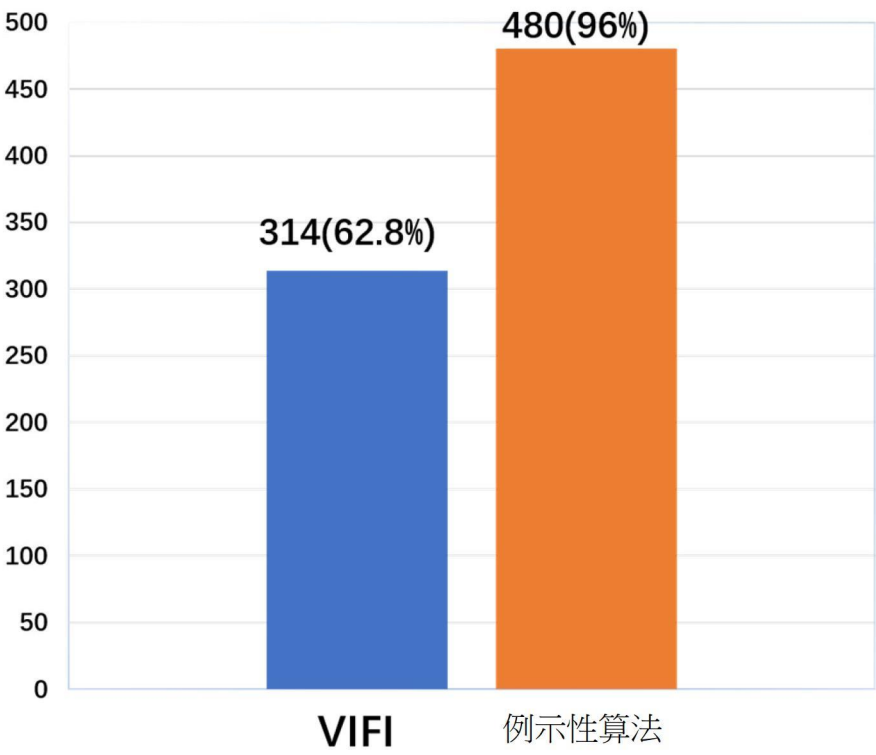


【圖9】



【圖10】

偵測到之模擬整合位點之數目



【圖11】

樣本類型	樣本ID	候選區域 之全部編號	不具有多樣性評分過濾		多樣性評分≥4		經設計之 PCR分析 之編號	用於驗證之 目標斷點之 編號	經驗證之斷 點之編號
			嵌合讀段之 編號	斷點之編號	嵌合讀段 之編號	斷點之編號			
CaCx	3485	458	28317	1054	10723	6	8	6	1
	C-788	43	77	26	0	0	-	-	-
	C-801	2	0	0	0	0	-	-	-
	C-819	2	0	0	0	0	-	-	-
	C-822	5	0	0	0	0	-	-	-
	C-877	9	0	0	0	0	-	-	-
HNSCC	TBR_1019	8	0	0	0	0	-	-	-
	TBR1245	1	0	0	0	0	-	-	-
	TBR1988	3	0	0	0	0	-	-	-
	TBR1989	12414	234667	96039	1521	41	10	9	3
	TBR2175	22	9	3	0	0	-	-	-

【圖12】

樣本類型	疾病分期	樣本ID	候選區域之編號	不具有多樣性評分過濾		多樣性評分≥4		在將斷點與靠近的基因組座標合併之後	經設計之PCR分析之編號	用於驗證之目標斷點之編號	經驗證之斷點之編號
				嵌合讀段之編號	斷點之編號	嵌合讀段之編號	斷點之編號				
HNSCC	II	TBR2175	1544	44970	1428	3399	31	23	17	14	13
	I	TBR1245	485	7614	447	1483	22	17	12	7	7
	II	TBR1019	889	27380	1344	34	2	2	2	1	1
	II	TBR1989	1339	12960	1115	15	1	1	1	1	1
	I	TBR2002	48	10	2	0	0	-	-	-	-
	III	TBR1067	0	0	0	0	0	-	-	-	-
	I	TBR1988	0	0	0	0	0	-	-	-	-

【圖13】