I837127

(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 I837127 B**

(45)公告日： 中華民國 113 (2024) 年 04 月 01 日

(21)申請案號：108115891 (22)申請日： 中華民國 108 (2019) 年 05 月 08 日

(51)Int. Cl.： ***C12N15/10 (2006.01)*** ***C12N15/66 (2006.01)***
C12N9/00 (2006.01) ***C40B20/04 (2006.01)***
C40B70/00 (2006.01)

(30)優先權：2018/05/08 美國 62/668,757
2018/05/16 美國 62/672,501
2018/06/19 美國 62/687,159

(71)申請人：大陸商深圳華大智造科技有限公司 (中國大陸) MGI TECH CO., LTD. (CN)
中國大陸
大陸商深圳華大生命科學研究院 (中國大陸) BGI SHENZHEN (CN)
中國大陸

(72)發明人：卓瑪奈克 雷都杰T DRMANAC, RADOJE T. (US)；彼得斯 布洛克A PETERS, BROCK A. (US)；王歐 WANG, OU (CN)

(74)代理人：李世章；彭國洋

(56)參考文獻：

US 2018/0044667A1 WO 2012/025250A1

期刊 Fan Zhang et al. Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. 35(9). Epub 2017 Jun 26. 852-857.

審查人員：張茜毓

申請專利範圍項數：42 項 圖式數：31 共 192 頁

(54)名稱

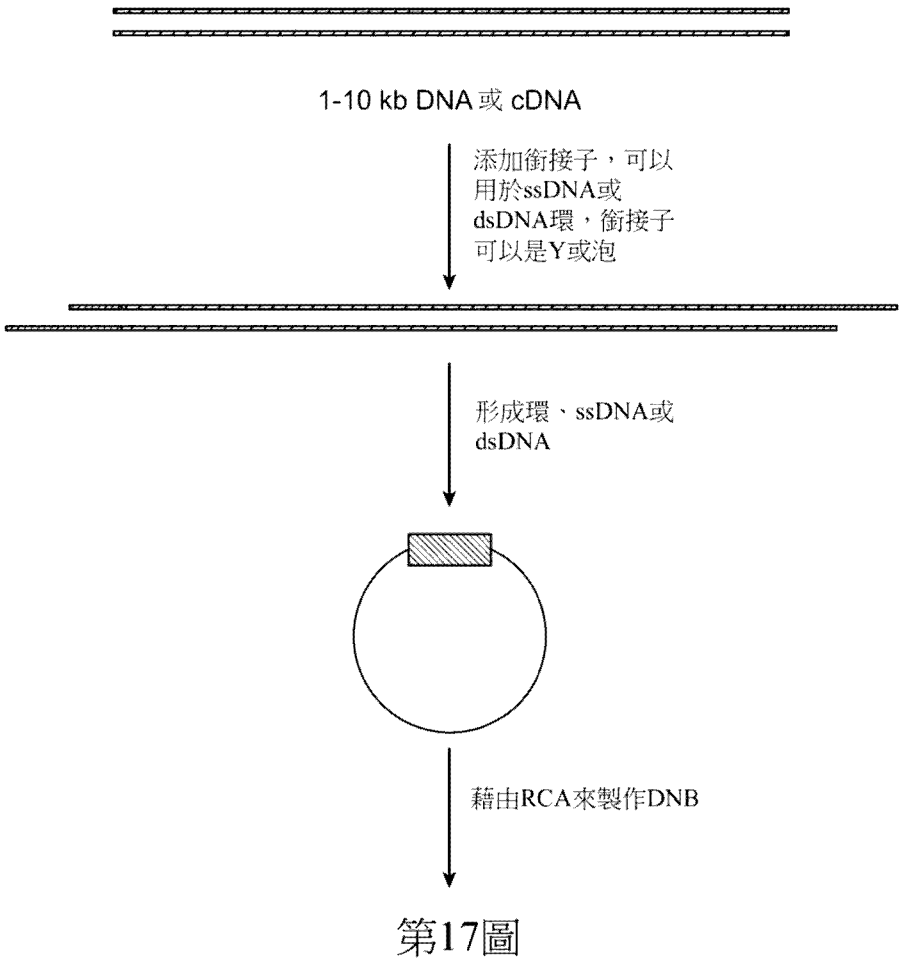
用於準確及具成本效益之定序、單倍體分類及組裝的基於單管珠之 DNA 共條碼編輯

(57)摘要

描述了用於製備核酸定序文庫之方法及組合物，包括(a)將插入序列轉座至靶核酸之第一片段中，其中插入序列包含雜交序列，並且其中轉座在第一片段中產生切口；(b)將以下各者組合在單一混合物中：(i)來自(a)之靶核酸之第一片段、(ii)夾板寡核苷酸，及(iii)珠子群體，其中每個珠子包含固定在其上之捕獲寡核苷酸，及(c)將各個珠子之捕獲寡核苷酸連接到各個第一片段之所插入雜交序列上。

Methods and compositions for preparing a nucleic acid sequencing library are described including (a) transposing an insertion sequence into first fragments of the target nucleic acid, wherein the insertion sequence comprises a hybridization sequence, and wherein the transposing produces nicks in the first fragments; (b) combining in a single mixture (i) the first fragments of the target nucleic acid from (a), (ii) a splint oligonucleotide, and (iii) a population of beads, wherein each bead comprises capture oligonucleotides immobilized thereon, and (c) ligating capture oligonucleotides of individual beads to inserted hybridization sequences of individual first fragments.

指定代表圖：


1-10 kb DNA 或 cDNA
添加銜接子，可以
用於ssDNA或
dsDNA環，銜接子
可以是Y或泡
形成環、ssDNA或
dsDNA
藉由RCA來製作DNB

第17圖

公告本

I837127

【發明摘要】

【中文發明名稱】用於準確及具成本效益之定序、單倍體分類及組裝的基於單管珠之DNA共條碼編輯

【英文發明名稱】SINGLE TUBE BEAD-BASED DNA CO-BARCODING FOR ACCURATE AND COST-EFFECTIVE SEQUENCING, HAPLOTYPING, AND ASSEMBLY

【中文】

描述了用於製備核酸定序文庫之方法及組合物，包括(a)將插入序列轉座至靶核酸之第一片段中，其中插入序列包含雜交序列，並且其中轉座在第一片段中產生切口；(b)將以下各者組合在單一混合物中：(i)來自(a)之靶核酸之第一片段、(ii)夾板寡核苷酸，及(iii)珠子群體，其中每個珠子包含固定在其上之捕獲寡核苷酸，及(c)將各個珠子之捕獲寡核苷酸連接到各個第一片段之所插入雜交序列上。

【英文】

Methods and compositions for preparing a nucleic acid sequencing library are described including (a)transposing an insertion sequence into first fragments of the target nucleic acid, wherein the insertion sequence comprises a hybridization sequence, and wherein the transposing produces nicks in the first fragments; (b) combining in a single mixture (i) the first fragments of the target nucleic acid from (a), (ii) a splint oligonucleotide, and (iii) a population of beads, wherein each bead comprises capture oligonucleotides immobilized thereon, and (c) ligating capture

oligonucleotides of individual beads to inserted hybridization sequences of individual first fragments.

【指定代表圖】第（17）圖。

【代表圖之符號簡單說明】

無

【特徵化學式】

無

【發明說明書】

【中文發明名稱】用於準確及具成本效益之定序、單倍體分類及組裝的基於單管珠之DNA共條碼編輯

【英文發明名稱】SINGLE TUBE BEAD-BASED DNA CO-BARCODING FOR ACCURATE AND COST-EFFECTIVE SEQUENCING, HAPLOTYPING, AND ASSEMBLY

【技術領域】

相關申請案之參考

【0001】 本申請案要求2018年5月8日提交之美國臨時專利申請案62/668,757；2018年5月16日提交之62/672,501；及2018年6月19日提交之62/687,159之優先權。上述先前申請案出於所有目的以全文引用之方式併入本文中。

【0002】 本發明係關於用於準確及具成本效益之定序、單倍體分類及組裝的基於單管珠之DNA共條碼編輯。

【先前技術】

【0003】 迄今為止，絕大多數單個全基因組序列缺乏關於在同源染色體上作為連續區塊傳遞之單鹼基至多鹼基變體之順序之資訊。最近已經開發了許多技術來實現此舉。大多數係基於共條碼編輯之過程(13)，亦即向單個長基因組DNA分子之亞片段添加相同之條碼。在定序後，條碼資訊可用於決定哪些讀段來自原始長DNA分

子。該過程首先由Drmanac(14)描述，並由Peters等人實施為384孔板測定(6)。然而，此等方法在技術上難以實現、昂貴、具有較低之資料品質、不提供獨特之共條碼編輯，或者所有四者之某種組合。在實踐中，大多數此等方法需要藉由標準方法產生單獨之全基因組序列以改良變體調用。此導致此等方法之有限使用，因為成本及易用性係WGS使用何種技術之主要因素。

【發明內容】

【0004】 在一個實施例中，一種製備定序文庫之方法，該文庫用於在不使用超微量分光光度計(nanodrops)之情況下對靶核酸定序，該方法包括：(a)將插入序列轉座至靶核酸之第一片段中，其中插入序列包含雜交序列，並且其中轉座在第一片段中產生切口；(b)將以下各者組合在單一混合物中：(i)來自(a)之靶核酸之第一片段、(ii)夾板寡核苷酸，及(iii)珠子群體，其中每個珠子包含固定在其上之捕獲寡核苷酸，該等捕獲寡核苷酸包含1)含有標籤(或條碼)之序列，其中每個含有標籤(或條碼)之序列包含條碼序列，其視情況為三部分條碼序列，其中固定在同一個別珠子上之寡核苷酸包含相同含有標籤(或條碼)之序列，並且大多數珠子具有不同之標籤(或條碼)序列，2)與夾板寡核苷酸之至少一部分互補之共同序列，其中該夾板寡核苷酸之第二部分與該雜交序列之至少一部分互補；3)視情況之第一PCR引子降溫貼合

(annealing)位點；及(c)將各個珠子之捕獲寡核苷酸連接到各個第一片段之所插入雜交序列上。

【0005】 在另一個實施例中，一種組合物包含(i)靶核酸之複數個第一片段，(ii)珠子之群體，該等珠子包含在其上雜交之捕獲寡核苷酸，其中每個珠子包含具有相同條碼之複數個寡核苷酸，並且該群體中之不同珠子包含不同條碼，其中該等第一片段包含轉座子整合之DNA，該等轉座子包含共同雜交序列；及(iii)與該雜交序列互補之夾板寡核苷酸。

【0006】 在又一個實施例中，一種組合物包含珠子群體，該等珠子包含附接在其上之複製捕獲寡核苷酸，其中每個珠子包含具有相同條碼之複數個捕獲寡核苷酸，並且該群體中之不同珠子包含不同條碼，其中複數個個別珠子包含與該珠子連接之插入轉座子之靶DNA；其中該連接係將各個珠子上之複數個捕獲寡核苷酸連接至插入轉座子之靶DNA中之複數個雜交序列；並且該組合物包含以下至少一者i)外切核酸酶；ii)DNA聚合酶；iii)尿嘧啶-DNA糖基化酶；iv)連接酶；iv)3'分支連接銜接子。

【圖式簡單說明】

附圖及表格

【0007】 第1(A)至1(D)圖。stLFR概述。第1(A)圖stLFR之第一步涉及將各自大約200-1000個鹼基對之雜交序列插入在長基因組DNA分子上。此係使用轉座子

實現的。然後將轉座子整合之DNA與珠子混合，每個珠子含有：約400,000個拷貝之銜接子序列，該序列含有珠子上所有銜接子共有之獨特條碼；共同之PCR引子位點；以及與整合之轉座子上之序列互補之共同捕獲序列。在將基因組DNA捕獲到珠子上後，將轉座子連接到條碼銜接子上。還有一些額外之文庫處理步驟，然後在BGISEQ-500或等效之定序儀上對共條碼編輯之亞片段進行定序。第1(B)圖藉由條碼來定位讀段資料導致讀段聚集在基因組之10至350 kb區域內。對於1 ng stLFR-1文庫，顯示了在Chr11之小區域中之總覆蓋率及4個條碼之條碼覆蓋率。大多數條碼僅與基因組中之一個讀段集群相關聯。第1(C)圖對於1 ng文庫stLFR-1及stLFR-2(橙色)及10 ng stLFR文庫stLFR-3及stLFR-4，將每個條碼之原始長DNA片段之數目作圖。來自1 ng stLFR文庫之超過80%之片段由單個獨特之條碼共同編碼。第1(D)圖針對1 ng stLFR-1文庫，將覆蓋每個原始長DNA片段之非重疊序列讀段及捕獲之亞片段(橙色)之分率作圖。亦參見第14圖，

【0008】 第2(A)至2(D)圖。SV偵測。先前報道之NA12878中之缺失也使用stLFR資料來發現。每個缺失之條碼共有之熱圖可以在第10圖中發現。第2(1)圖使用如前所述之Jaccard指數(12)，將在染色體8上具有約150 kb雜合缺失之區域之2kb窗口內條碼共有之熱圖作圖。高重疊區域以深紅色描繪。米色區域沒有重疊。箭頭

表明在染色體8上在空間上彼此遠離之區域之標記缺失位置之重疊如何增加。第2(B)圖共條碼編輯之讀段按照單倍型來分離，並按照y軸上之獨特條碼及x軸上之染色體8位置來作圖。雜合缺失存在於單個單倍型中。第2(C)圖還對於具有已知易位之患者細胞系(26)及第2(D)圖GM20759，在染色體2中具有已知顛換之細胞系(27)，將染色體5與12之間重疊條碼之熱圖作圖。

【0009】 第3(A)及3(B)圖：覆蓋分佈圖。將在BGISEQ500上定序之stLFR-2(A)及標準(B)文庫之覆蓋率作圖。兩個樣品之覆蓋率均降低取樣至30X。30X基因組之泊松分佈以藍色作圖。

【0010】 第4(A)及4(B)圖：文庫之間FP之重疊。(A)將來自每個stLFR文庫、BGISEQ-500標準文庫及藉由Illumina定序之無PCR文庫(從basespace下載之文庫「HiSeq2500-TruSeq_PCR-Free_DNA_2x251_NA12878」)之FP以維恩圖來作圖。在四個stLFR庫之間共有2,078個FP。(B)stLFR文庫FP與Chromium文庫FP之重疊表明，在兩種不同技術之間共有1,194個FP，此等兩種技術都使用從GM12878分離之DNA，而非NA12878之GIAB參考材料。884 FP係stLFR文庫獨有的。

【0011】 第5(A)至5(D)圖：stLFR-1變體指標。針對所有真陽性變體、假陽性變體及共有假陽性變體(綠色)，分析參考及變體等位基因之讀段深度及條碼深度。

將參考(A)及備選(B)等位基因之讀取深度以及參考(C)及備選(D)等位基因之條碼計數作圖。一般來說，共有之假陽性看起來更像真陽性，表明有一些過濾標準可以區分此等變體及非共有假陽性。

【0012】 第6(A)至6(D)圖：stLFR-3變體指標。針對所有真陽性變體、假陽性變體及共有假陽性變體(綠色)，分析參考及變體等位基因之讀段深度及條碼深度。將參考(A)及備選(B)等位基因之讀取深度以及參考(C)及備選(D)等位基因之條碼計數作圖。一般來說，共有之假陽性看起來更像真陽性，表明有一些過濾標準可以區分此等變體及非共有假陽性。

【0013】 第7(A)及7(B)圖：共有假陽性變體分佈。將分開2,078個共有FP變體之基因組距離在100 bp(深藍色)、1,000 bp(橙色)、10,000 bp、100,000 bp及100萬bp之連續區間內匯總。還將來自stLFR-1文庫之5組2,078個隨機選擇之變體作圖。對於每個樣品，將位置總數或變體總數作圖。只有發現2個或更多變體之區間或區間內之變體得以匯總。(A)在過濾之前，有219個共有FP似乎緊密聚集，可能係由於定位誤差引起的。其餘1,859個變體似乎共有與隨機變體集類似之分佈。(B)過濾後保留1,738個共有FP，但只有72個緊密聚集。

【0014】 第8(A)至8(T)圖：使用條碼共有熱圖之NA12878缺失偵測。在chr3:65189000-65213999處使用230 Gb (A)或100 Gb (B)，在

chr4:116167000-116176999處使用230 Gb (C)或100 Gb (D)，在chr4:187094000-187097999處使用230 Gb (E)或100 Gb (F)，在chr7:110182000-110187999處使用230 Gb (G)或100 Gb (H)，在chr16:62545000-62549999處在230 Gb (I)或100 Gb (J)，在chr1:189704509-189783359處使用230 Gb (K)或100 Gb (L)，在chr3:162512134-162569235處使用230 Gb (M)或100 Gb (N)，在chr5:104432113-104467893處使用230 Gb (O)或100 Gb (P)，在chr6:78967194-79001807處，以及在chr8:39232074-39309652處使用230 Gb (S)或100 Gb (T)之讀段資料，偵測stLFR-1文庫中之缺失。

【0015】 第9(A)至9(L)圖：用stLFR進行易位及倒位偵測。用stLFR分析患者細胞系及細胞系GM20759，該等細胞系分別在染色體5與12之間具有易位以及在染色體2上具有倒位。對於每個文庫，對總序列覆蓋率進行降低取樣以研究較低覆蓋率下之偵測能力。染色體12與5之間之易位在40 Gb(A)、20 Gb(B)、10 Gb(C)甚至5 Gb(D)之總序列覆蓋率下很容易偵測到。GM20759之倒位也很容易在46 Gb(E)、20 Gb(F)、10 Gb(G)及5 Gb(H)之總序列覆蓋率下偵測到。此外，我們研究了GM12878細胞系中之此等區域，此等區域已知不含有此

等SV中之任一者。染色體5與12之間之易位在具有230 Gb覆蓋率(I)之1 ng stLFR文庫或具有126 Gb覆蓋率(J)之10 ng文庫中均不明顯。在stLFR-1(K)或stLFR-4文庫(L)中也未發現顛換。

【0016】 第10(A)至10(C)圖：NA12878支架之比對點圖。將來自stLFR-1(A)及stLFR-4(B)文庫之SALSA支架相對於參考人類基因組之hg37來作圖。來自Dixon等人(29)之7.34億HiC讀段也用於產生支架，並且也相對於hg37來作圖(C)。在所有情況下，僅將覆蓋染色體之5%或5%以上之支架作圖。

【0017】 第11圖：LongHap定相。應用LongHap之定相演算法之完整描述可以在「方法及材料」部分中發現。

【0018】 第12圖：條碼序列組裝。需要使用三個連接來產生約36億個不同之條碼。條碼組裝之每個步驟之預期序列顯示為SEQ ID NO：1至13。

【0019】 第13圖：條碼序列組裝之流程圖。

【0020】 第14圖：條碼方案之示例性流程圖。

【0021】 第15圖：雜交步驟之草圖。

【0022】 第16圖：連接及降解步驟之草圖。顯示變性及C加尾之最終步驟係視情況的，本文不再進一步描述。

【0023】 第17圖：雙鏈DNB之產生及使用。

【0024】 第18圖：長分子之擴增。

【0025】 第19圖：隨機切口酶方法。

【0026】 第20圖：髮夾銜接子方法。

【0027】 第21(A)及21(B)圖：第21(A)圖係不同DNA受質上連接測定之示意圖。鈍端DNA供體係具有雙脫氧3'-末端(實心圓圈)之合成之部分dsDNA分子，以防止銜接子自身連接。銜接子之長臂係5'-磷酸化的。使用2或3個寡核苷酸(黑色、紅色及橙色線)組裝DNA受體以形成切口(不含磷酸鹽)、間隙(1或8 nt)或36 nt 5'-突出端。受質之所有鏈都係未磷酸化的，支架鏈係3'雙脫氧保護的。第21(B)圖顯示使用6%變性聚丙烯醯胺凝膠分析連接產物之尺寸偏移。將負的無連接酶對照(泳道1、3、4、6、7、9、10、12及13)以其相應實驗測試之1或0.5X體積加載。如果發生連接，則受質尺寸向上偏移22 nt。紅色箭頭對應於受質，藍色箭頭對應於銜接子連接之受質。M2 = Thermo Fisher之25 bp DNA梯(c)預期連接產物尺寸及使用ImageJ來估計連接效率之表。藉由將連接產物之強度除以連接及未連接產物之總強度來估計連接效率。

【0028】 第22(A)至22(D)圖：使用6%TBE聚丙烯醯胺凝膠對連接產物之尺寸偏移進行凝膠分析。紅色箭頭對應於受質，藍色箭頭對應於銜接子連接之受質：切口(A，左)，5'-突出端(A，右)，1nt間隙(B)，2nt間隙(C)及3nt間隙(D)。M2 = Thermo Fisher之25 bp DNA梯。比較了兩個銜接子(Ad1及Ad2)序列，並且還檢查了Ad2連接會合處之5'末端之不同鹼基(A或G)。**(e)使用ImageJ基於條帶強度計算之連接效率表。

【0029】 第23(A)及23(B)圖：第23(A)圖顯示了具有20 bp互補區之DNA/RNA雜交體上3'-分支連接之示意圖。我們測試了鈍端銜接子係否在5'-RNA突出端處連接到DNA之3'-末端，及/或在5'-DNA突出端處連接到RNA之3'-末端。(B)使用6%變性聚丙烯醯胺凝膠對連接產物之尺寸偏移進行凝膠分析。紅色箭頭對應於RNA受質(29 nt)，綠色箭頭對應於DNA受質(80 nt)。藍色箭頭對應於銜接子連接之RNA受質。如果發生連接，則受質尺寸將向上偏移20 nt。反應1及2係重複的。M2 = Thermo Fisher之25 bp DNA梯。

【0030】 第24(A)至24(C)圖：第24(A)圖係轉座子插入，然後3'-3'分支連接及使用Pr-A(藍色箭頭)及Pr-B(綠色箭頭)進行PCR擴增之示意圖。(B)用TnA及/或TnB插入轉座子及/或使用引子pr-A、pr-B或兩者，AdB之3'分支連接後之擴增產物。產物在6%聚丙烯醯胺凝膠上電泳。M1 = ThermoFisher MassRuler低範圍DNA梯。(C)在各種轉座子插入及3'分支連接條件後使用pr-A及pr-B之擴增信號圖。

【0031】 第25圖：中間長度標記。

【0032】 第26(A)及(B)圖：在藉由切口、間隙及突出端形成之非常規DNA末端處，藉由T4 DNA連接酶之3'分支連接。(A)不同DNA受體類型上之連接測定之示意圖。鈍端DNA供體係具有雙脫氧3'-末端(實心圓圈)之合成之部分dsDNA分子，以防止DNA供體自身連接。供體

之長臂係5'-磷酸化的。使用2或3個寡核苷酸組裝DNA受體以形成切口(不含磷酸鹽)、間隙(1或8 nt)或36-nt 3'-隱性末端。受質之所有鏈都係未磷酸化的，並且支架鏈係3'雙脫氧保護的。(B)使用6%變性聚丙烯醯胺凝膠分別分析受質1、2、3及4之連接產物之尺寸偏移。將負的無連接酶對照(泳道1、3、4、6、7、9、10、12及13)以相應實驗測試之1或0.5X體積加載。如果發生連接，則受質尺寸向上偏移22 nt。紅色箭頭對應於受質，紫色箭頭對應於供體連接之受質。使用Thermo Fisher之25-bp DNA梯。表S1中之供體及受質序列。表8顯示了每個實驗組中受質及連接產物之預期尺寸及近似之連接效率。使用ImageJ估計每個條帶之強度，並按照其預期尺寸來歸一化。藉由將連接產物之歸一化強度除以連接及未連接產物之歸一化總強度來估計連接效率。

【0033】 第27(A)至27(E)圖：使用6%TBE聚丙烯醯胺凝膠對連接產物之尺寸偏移進行凝膠分析。紅色箭頭對應於受質，紫色箭頭對應於供體連接之受質：受質5(切口)(A)，受質6(1-nt間隙)(B)，受質7(2-nt間隙)(C)，受質8(3-nt間隙)，(D)及受質9(3'-隱性末端)(E)。使用Thermo Fisher之25-bp DNA梯。檢查在連接會合處之5'末端具有不同鹼基(T、A或GA)之三個DNA供體。表9中顯示了使用ImageJ基於歸一化條帶強度計算之連接效率。

【0034】 第28(A)至28(D)圖。在DNA/RNA雜交體中RNA之3'末端處之3'分支連接。具有20-bp互補區之DNA/RNA雜交體上3'-分支連接之示意圖。我們測試了鈍端DNA供體係否會連接到DNA之3'-隱性末端及/或RNA之3'-隱性末端。DNA(ON-21)與RNA鏈(A)雜交，而DNA(ON-23)不能與RNA鏈(B)雜交。第28(C)及(D)圖顯示了使用6%變性聚丙烯醯胺凝膠對連接產物之尺寸偏移之凝膠分析。紅色箭頭對應於RNA受質(29 nt)，綠色箭頭對應於DNA受質(80 nt)。紫色箭頭對應於供體連接之RNA受質。如果發生連接，則受質尺寸將向上偏移20 nt。(c)泳道1及2，實驗重複；泳道7-10，無連接酶對照；與T4 DNA連接酶一起添加10%PEG。(d)泳道1，無連接酶對照；泳道2、3及8，T4 DNA連接酶與10%PEG；泳道4、5及9，T4 RNA連接酶1與20%DMSO；泳道6、7及10，T4 RNA連接酶2與20%DMSO。使用Thermo Fisher之25-bp DNA梯。可能對應於第23圖，但它在多個方面並不精確。

【0035】 第29(A)至29(C)圖顯示了三種轉座子標記方法，然後使用Pr-A(藍色箭頭)及Pr-B(綠色箭頭)進行PCR擴增之示意圖。雙轉座子方法(A)；一個Y轉座子標記與3'間隙填充(B)；一個轉座子方法，在3'-間隙處具有銜接子連接(C)。第29(D)圖係在各種標記及間隙連接條件後使用pr-A或pr-A與pr-B純化後之擴增信號圖。可能對應於第23圖，但它在多個方面並不精確。

【0036】 第30(A)至30(C)圖。Tn5-間隙連接(A)、兩個轉座子(B)及常規TA連接(C)之鹼基分佈偏差。僅提供連接兩端之前20個鹼基；腺嘌呤，藍色；胞嘧啶，橙色；鳥嘌呤，灰色；胸腺嘧啶，黃色；給出了五個獨立文庫之平均值及標準差。根本不存在。

【0037】 第31(A)及31(B)圖。具有不同添加條件之DNA 3'分支連接。(A)在滴定之ATP濃度下在5'-突出端DNA處連接。對0.01 mM(泳道4及5)及0.005 mM ATP(泳道6及7)進行重複。泳道9係無供體對照。(B)在具有或不具有SSB及連接酶的情況下，在切口、1-nt間隙、8-nt間隙、5'-突出端及鈍端處之DNA之3'分支連接。紅色箭頭對應於受質，紫色箭頭對應於供體連接之受質。根本不存在。

【0038】 表1：定相及變體調用統計。使用誘餌序列將讀段定位到Hg37，並使用GATK調用變體，除非另有說明，GATK具有針對所有文庫之預定設置。來自GIAB高置信變體調用VCF之SNP用作定相之輸入。

【0039】 表2：支架統計。

【0040】 表3：過濾減少假陽性調用。最終之FP調用係藉由從過濾之FP中減去1,666來計算的，除了STD文庫，根據定義，STD文庫不與stLFR文庫共有任何此等FP，因為它係用GIAB參考材料製作的。

【0041】 表4：LongHap SNP及插入缺失定相。

【0042】 表5：過濾標準。在材料及方法部分中解釋之各種過濾標準用於移除FP。

【0043】 表6：示例性序列。

【實施方式】

1. stLFR文庫過程

1.1 介紹

【0044】 在本文中，我們描述了單管長片段讀取(stLFR)技術(15)之實現，此係一種在單管中實現之用數百萬個條碼進行DNA共條碼編輯之有效方法。參見WO 2014/145820 A2(2014)，其出於所有目的以引用之方式併入本文中。此藉由使用微珠之表面作為隔室(例如，384孔板之孔)之替代物來實現。每個珠子攜帶獨特條碼序列之許多拷貝，該序列被轉移到每個長DNA分子之亞片段。然後在諸如BGISEQ-500或等同物之常見短讀段定序裝置上分析此等共條碼編輯之亞片段。在我們實施此方法時，我們使用基於連接之組合條碼生成策略，在三個連接步驟中建置超過18億個不同之條碼。對於單個樣品，我們使用約1000萬至5000萬個此等經條碼編輯之珠子在單管中捕獲約1000萬至1億個長DNA分子。很少有兩個珠子共有相同之條碼，因為我們從如此龐大之總條碼文庫中取樣了1000萬至5000萬個珠子。此外，在使用5000萬個珠子及1000萬個長基因組DNA片段之情況下，來自每個長DNA片段之絕大多數亞片段由獨特之條

碼共同編碼。此類似於長讀取單分子定序，並可能提供強大之資訊學方法以利於從頭組裝。重要地，stLFR易於實施並且可以用相對小的寡核苷酸投資來實施以產生經條碼編輯之珠子。此外，stLFR使用幾乎所有分子生物學實驗室中之標準設備，幾乎可以藉由任何定序策略進行分析。最後，stLFR取代了標準之NGS文庫製備方法，僅需要1 ng之DNA，並且不會顯著增加全基因組或整個外顯子組分析之成本，每個樣品之總成本低於30美元。

【0045】 如本文所用，「單管」係指分析大量單個DNA片段，而不需要在標記步驟期間將片段分離至單獨之管、容器、等分試樣、孔或液滴中。相反，微珠之表面用作隔室之替代物。

【0046】 stLFR之第一步係沿著基因組DNA片段插入雜交序列，較佳以規則之間隔插入。合適之間隔可隨應用及所需結果而變化，但通常在100-1500 bp之範圍內，通常為200-1000 bp。此經由藉由轉座來併入DNA序列而實現。在一個實施例中，轉座酶係Tn3、Tn5、Tn7或Mu。通常，使用Tn5轉座酶(參見Picelli等人，2014，其出於所有目的以引用方式併入本文)。轉座之DNA或插入序列包含用於雜交之單鏈區域(「雜交序列」)以及被酶識別並能夠實現轉座反應之雙鏈嵌合序列(第1A圖)。該轉座步驟在溶液中完成(而非使插入序列直接與珠子連接)。此使得能夠沿基因組DNA分子非常有效地併入雜交序列。如先前所觀察到的(10)，轉座酶具有在轉座事件

後保持與基因組DNA結合之性質，有效地使轉座子整合之長基因組DNA分子保持完整。

【0047】 在用例如Tn5處理DNA後，將DNA在雜交緩衝液中稀釋並與複製性地經條碼編輯之珠子組合。在一種方法(下面之實例)中，使用雜交緩衝液中之5000萬個約2.8 um複製性地經條碼編輯之珠子。每個珠子含有大約400,000個捕獲銜接子(也稱為捕獲寡核苷酸(capture oligo)或捕獲寡核苷酸(capture oligonucleotide))，每個銜接子都包含相同之條碼序列。捕獲銜接子之一部分含有尿嘧啶核苷酸，以便能夠在後續步驟中破壞未使用之銜接子。例如，捕獲銜接子可以係5-50%尿嘧啶，更常見係5-50%，更常見係5-20%。將混合物在優化之溫度及緩衝條件下溫育，在此期間，插入轉座子之DNA藉由雜交序列來捕獲到珠子中。

【0048】 有人提出，溶液中之基因組DNA會形成球，該等球具有伸出之兩個尾部(16)。此可以使得長DNA片段能夠在分子之一端附近得以捕獲，然後進行滾動運動，從而將基因組DNA分子圍繞珠子來捲繞。每個珠子表面上大約每7.8 nm有一個捕獲寡核苷酸。此使得能夠實現非常均勻且高速率之亞片段捕獲。100 kb之基因組片段將圍繞2.8 um珠子捲繞約3次。在我們的資料中，300 kb係捕獲之最長片段尺寸，表明可能需要更大之珠子來捕獲更長之DNA分子。

【0049】 在替代實施例中，諸如珠子尺寸、捕獲寡核苷酸間距或每種混合物之不同寡核苷酸之數量之參數可以變化。例如，所用珠子之直徑可以在1-20 um之範圍內，或者2-8 um、3-6 um或1-3 um。例如，珠子上之經條碼編輯之寡核苷酸之間距可以係至少1、至少2、至少3、至少4、至少5、至少6或至少7 nm。在實施例中，間距小於10 nm(例如，5-10 nm)、小於15 nm、小於20 nm、小於30 nm、小於40 nm或小於50 nm。在一些實施例中，每種混合物使用之不同條碼之數量可以係> 1M、> 10M、> 30M、> 100M、> 300M，或> 1B。如下所述，可以生產非常大量之條碼用於本發明，例如，使用本文所述之方法。在一些實施例中，每種混合物使用之不同條碼之數量可以係> 1M、> 10M、> 30M、> 100M、> 300M，或> 1B，並且它們從至少10倍更大之多樣性之池中取樣(例如，在珠子上> 10M、> 0.1B、0.3B、> 0.5B、> 1B、> 3B、> 10B不同之條碼。)

【0050】 經由藉由橋或夾板(術語可互換使用)寡核苷酸介導之將捕獲銜接子之3'末端連接到插入轉座子之雜交序列之5'末端來以規則的間隔轉移單個條碼序列，該寡核苷酸具有與捕獲銜接子互補之第一區域及與雜交序列互補之第二區域(第1A圖及第15圖)。收集珠子並將DNA/轉座酶複合物破壞，從而產生尺寸小於1 kb之亞片段。

【0051】 如果需要，可以在該步驟中實現樣品條碼編輯。使用在嵌合序列與雜交序列之間攜帶獨特條碼之轉座子。此等轉座子可以以96、384或1536板形式合成，每個孔含有攜帶相同條碼之轉座子之許多拷貝，並且每個條碼在各個孔之間係不同的。不同DNA樣品可以使用此等經條碼編輯之轉座子以96、384或1536板形式轉座子插入。用樣品條碼標記之樣品可以以任何方式複用。

【0052】 由於大量珠子及每個珠子具有高密度捕獲寡核苷酸，過量銜接子之量比產物量大四個數量級。此巨大之未使用之銜接子可以壓倒後續步驟。為了避免此情況，我們設計了具有藉由5'末端連接之捕獲寡核苷酸之珠子。此使得能夠開發出外切核酸酶策略，該外切核酸酶特異性降解過量之未使用之捕獲寡核苷酸。參見第14及16圖。尿嘧啶-DNA糖基化酶(Uracil-DNA Glycosylase; UDG)也可用於降解過量之銜接子。

【0053】 在一個態樣，該方法包括在單一混合物中組合(i)靶核酸之第一片段，及(ii)珠子群體，其中每個珠子包含固定在其上之寡核苷酸，該等寡核苷酸包含含有標籤之序列(或條碼銜接子)，其中每個含有標籤之序列包含標籤序列，其中固定在同一個別珠子上之寡核苷酸包含相同之含有標籤之序列，並且大多數珠子具有不同之標籤序列。在一些實施例中，DNA片段係至少2個、至少10個、至少30個，或至少100個拷貝之DNA或cDNA分子之多聯體。核酸單體之長度可以為0.5kb至10kb，或者>

1kb，或者長度> 10kb。在一些方法中，決定混合物中之序列為> 50%或> 70%> 90%、95%、> 99%、100%之DNA或cDNA分子之鹼基。

1.1.1 雙轉座子方法

【0054】 在stLFR之一種方法中，在初始插入步驟中使用兩種不同之轉座子，允許在核酸外切酶處理後進行PCR。然而，此方法導致每個長DNA分子之覆蓋率減少約50%，因為它需要將兩個不同之轉座子彼此相鄰地插入以產生合適之PCR產物。

1.1.2 使用3'分支連接之單轉座子方法

【0055】 為了獲得每個基因組DNA片段之最高覆蓋率，我們在初始插入步驟中使用單轉座子並藉由連接來添加額外之銜接子。此非常規連接，稱為3'分支連接，涉及將來自鈍端銜接子之5'磷酸共價連接到基因組DNA之凹入3'羥基上(第1A圖)。分枝連接描述於下文實例3中。亦參見美國專利公開案US2018/0044668及國際申請案WO 2016/037418，兩者均出於所有目的以引用方式併入。亦參見美國專利公開案2018/0044667，其出於所有目的以引用方式併入。使用該方法，理論上可以擴增及定序所捕獲之基因組分子之所有亞片段。

【0056】 此外，該連接步驟使得樣品條碼能夠與基因組序列相鄰放置以用於多重取樣。使用此等銜接子進行樣品條碼編碼之益處係條碼可以放置在基因組DNA附近，以

便相同之引子可用於定序條碼及基因組DNA，並且不需要額外之定序引子來讀取條碼。樣品條碼編輯允許在序列之前彙集來自多個樣品之製劑，並藉由條碼區分。3'分支連接銜接子可以96、384或1536板形式合成，每個孔含有攜帶相同條碼之銜接子之許多拷貝，並且每個條碼在各個孔之間係不同的。在珠子上捕獲後，此等銜接子可用於96、384或1536板形式之連接。

【0057】 在該連接步驟之後，進行PCR並且文庫準備好進入任何標準之下一代定序(next generation sequencing; NGS)工作流程。應當理解，可以使用與捕獲寡核苷酸或其互補序列上之位點雜交之第一引子(參見第1A圖)及與3'分支連接銜接子或其互補序列上之位點雜交之第二引子進行PCR(或其他擴增)。在BGISEQ-500之情況下，如前所述(17)將文庫進行環化。從單鏈環製備DNA奈米球並將其加載到圖案化之奈米陣列上(17)。然後將此等奈米陣列在BGISEQ-500上進行基於組合探針-錨定合成(combinatorial probe-anchor synthesis; cPAS)之定序(18-20)。定序後，提取條碼序列。藉由獨特條碼對讀段資料進行定位表明，具有相同條碼之大多數讀段聚集在基因組區域中，對應於文庫製備期間使用之DNA長度(第1B圖)。在實例1及2中描述了該方法之詳細描述以及製備珠子之方案。

【0058】 在一些實施例中，>50%、>70%、>80%，>90%或>95%之經條碼編輯之DNA片段用獨特之條碼

進行條碼化。在一些實施例中，將片段中>50%、>70%、>80%>90%之亞片段連接至條碼寡核苷酸。在一些實施例中，平均對長片段之>10%或>20%、>40%、>50%、>60%之亞片段進行定序。

1.2 stLFR讀段覆蓋率及變體調用

【0059】 為了證明stLFR定相及變體調用，我們使用來自NA12878之1 ng(stLFR-1及stLFR-2)及10 ng(stLFR-3及stLFR-4)DNA產生了四個文庫。珠子之數量係變化的，使用了1000萬(stLFR-3)、3000萬(stLFR-4)及5000萬(stLFR-1及stLFR-2)。最後，測試3'分支連接(stLFR-1、stLFR-2及stLFR-3)及雙轉座子(stLFR-4)方法。stLFR-1及stLFR-2分別定序至336 Gb及660 Gb之總鹼基覆蓋率之深度。我們還在降低取樣之覆蓋率下分析了此等方法。將stLFR-3及stLFR-4分別定序至117 Gb及126 Gb之更適度之水準。使用BWA-MEM將共條碼編輯之讀段定位至人參考基因組之構建37(21)。因為stLFR不需要任何預擴增步驟，所以跨基因組之讀段覆蓋分佈接近於泊松(第3圖)。非重複覆蓋率之範圍為34-58X，每條碼之長DNA分子數之範圍為1.2-6.8(表1及第1C圖)。如所預期的，由5000萬個珠子及1 ng基因組DNA製成之stLFR文庫具有超過80%之最高單一獨特條碼共條碼編碼率(第1C圖)。此等文庫還觀察到每個長DNA分子之最高平均非重

疊讀段覆蓋率為10.7-12.1%，並且每個長DNA分子之捕獲亞片段之最高平均非重疊鹼基覆蓋率為17.9-18.4%(第1d圖)。此覆蓋率比先前使用3 ng DNA及附著在珠子上之轉座子(12)所展示之覆蓋率高約10X。

【0060】 對於每個文庫，使用預定設置使用GATK(22)調用變體。藉由將SNP及插入缺失調用與瓶中之基因組(Genome in a Bottle; GIAB)(23)比較，可以決定假陽性(false positive; FP)及假陰性(FN)率(表1)。此外，對於由約1000倍更多之基因組DNA組成，並在BGISEQ-500(STD)上進行定序之標準非stLFR文庫，以及來自10X Genomics之Chromium文庫，我們使用GATK中相同之設置進行變體調用(11)。相對於Zhang等人之珠子單倍體分類文庫研究中報道之彼等(12)，我們還將精確度及靈敏度比率進行比較，該文獻出於所有目的以引用方式併入本文。我們的stLFR方法及Zhang等人描述之方法證明了比Chromium文庫更低之SNP及插入缺失FP率。stLFR之FP及FN率比STD文庫高2倍，並且取決於特定之stLFR文庫及過濾標準，FN率高於或低於Chromium文庫。與標準文庫相比，stLFR文庫中較高之FN率主要係由於較短之平均插入片段尺寸(約200 bp，而標準文庫中為300 bp)。亦即，對於SNP及插入缺失，stLFR之FN率比Zhang等人低得多，並且對於插入缺失，FN率比

Chromium文庫低得多(表1)。總體而言，與Zhang等人發表之結果或Chromium文庫相比，對於我們的stLFR文庫，變體調用之大多數指標更好，尤其當使用非優化之定位及變體調用過程時(表1，「無過濾器」)。

【0061】 使用GIAB資料測量FP率之一個潛在問題係我們無法使用GIAB參考材料(NIST RM 8398)，此歸因於分離之DNA之片段尺寸相當小。出於此原因，藉由使用能夠產生非常高分子量之DNA之基於透析之方法，我們使用GM12878細胞系及分離之DNA(參見方法)。然而，與GIAB參考材料相比，我們的GM12878細胞系分離株可能具有許多獨特之體細胞突變，因此導致在我們的stLFR文庫中，FP之數量膨脹。為了進一步對此檢查，我們比較了4個stLFR文庫及兩個非LFR文庫之間單核苷酸FP變體之重疊(第4a圖)。總體而言，六個文庫共有544個FP變體，並且2,078個FP為四個stLFR文庫所獨有。我們還將stLFR FP與Chromium文庫進行了比較，發現此等共有FP中有超過一半(1,194)也存在於Chromium文庫中(第4b圖)。對此等共有變體之讀段及條碼覆蓋率之檢查顯示它們更類似於TP變體(第5-6圖)。相比於2,078個隨機選擇之變體，我們還檢查了此等共有FP變體之跨基因組之分佈(第7a圖)。該分析顯示在集群中發現之219種變體，其中此等FP中之兩個或兩個以上彼此在100 bp內。然而，大多數(90%)變體具有看起來與隨機選擇之變體無法區分之分佈。此外，在

stLFR與Chromium文庫之間共有之彼等FP中，僅發現41個係聚集的(第7a圖)。最後，此等變體中之96個由GIAB調用，但是與stLFR文庫中調用之變體相比，具有不同接合性。

【0062】 如果我們接受此等共有之FP變體在很大程度上係真實的並且不存在於GIAB參考材料中之證據，則與表1中針對SNP偵測報道之FP率相比，stLFR之FP率可能低至少1,859個變體。此仍然比標準BGISEQ-500文庫多幾千個單核苷酸變體。為了進一步提高stLFR文庫中之FP率，我們測試了許多不同之過濾策略來消除誤差。最後，藉由應用基於參考及變體等位基因比率及條碼計數之一些過濾標準(參見實例)，我們能夠根據文庫及覆蓋量來移除3,647-13,840 FP變體。重要地，此係在stLFR文庫中僅將FN率提高0.10-0.29%之情況下實現的。在此過濾步驟之後，我們檢查了四個stLFR文庫之間之共有FP。過濾僅移除了340個共有FP變體，其中147個聚集在彼此100個鹼基對內，並且可能係不真實的(第7b圖)。此進一步表明此等共有FP中之大多數都係真實之變體。考慮到此等變體及過濾後FP變體數量之減少導致與用於SNP調用之過濾之STD文庫相比類似之FP率及高出2-3倍之FN率(表3)。此增加之FN率主要係由於stLFR文庫中具有短插入片段尺寸之配對之非獨特定位之增加。

1.3 stLFR定相效能

【0063】 為了評估變體定相效能，使用公開可用之軟體包HapCut2(24)對來自GIAB之高置信度變體進行定相。根據文庫類型及序列資料之量，超過99%之所有雜合SNP被置於具有0.6-15.1 Mb之N50之重疊群中(表1)。具有336 Gb總讀段覆蓋率(44X獨特基因組覆蓋率)之stLFR-1文庫實現了最高之定相效能，N50為15.1 Mb。N50長度似乎主要受長基因組片段之長度及覆蓋率之影響。此可以從stLFR-2之降低之N50中看出，因為用於該樣品之DNA比用於stLFR-1之材料稍微更老並且更碎片化(表1，平均片段長度為52.5 kb相比於62.2 kb)並且10 ng文庫(stLFR-3及4)之N50縮短10倍。與GIAB資料之比較表明，短及長切換誤差率較低，與先前之研究相當(11，12，25)。stLFR效能與Chromium文庫非常相似。因為Zhang等人珠子單倍體分類方法沒有可用之讀段資料，我們只能將我們的結果與他們為其資料編寫及優化之定相演算法之結果進行比較。此表明stLFR-1及stLFR-2文庫具有更長之N50、類似之短切換誤差率，但是更高之長切換誤差率。使用更多DNA之stLFR-3及stLFR-4具有與Zhang等人類似之N50。然而，由於DNA輸入及覆蓋率之差異，直接比較係困難的。

【0064】 應該注意地，該定相結果係使用未針對stLFR資料來編寫之程式實現的。為了看看此結果係否可以改良，我們開發了一個定相程式LongHap，並針對

stLFR資料進行了優化。使用GIAB變體，LongHap能夠將超過99%之SNP定相至具有18.1 Mb之N50之重疊群中(表1)。重要地，在減少短及長切換誤差之同時實現了此等增加之重疊群長度(表1)。LongHap也可以將插入缺失定相。使用GIAB SNP及插入缺失將LongHap應用於stLFR-1導致23.4 Mb N50，但也導致切換誤差率增加(表4)。

1.4 結構性變異偵測

【0065】 先前之研究表明，長片段資訊可以改良NA12878中之結構性變異(SV)及所描述大缺失(4-155 kb)之偵測(11，12)。為了證明stLFR偵測SV之能力，我們檢查了如前所述(12)關於stLFR-1及stLFR-4文庫之在此等區域中之條碼重疊資料。在每種情況下，即使在較低之覆蓋率下，也在stLFR-1資料中觀察到缺失(第2a圖及第8圖)。對包含染色體8中約150 kb缺失之共條碼編輯之序列讀段之更仔細檢查證明，該缺失係雜合的並且在單個單倍型中發現(第2b-c圖)。10 ng stLFR-4文庫也偵測到大部分缺失，但由於該文庫之每個片段之覆蓋率較低(因此條碼重疊較少)，因此難以識別三個最小之缺失。

【0066】 為了評估用於偵測其他類型SV之stLFR效能，我們從在染色體5與12之間具有已知易位之患者之細胞系(26)及GM20759中製備文庫，GM20759係在染色

體2上具有已知倒位之細胞系(27)。stLFR文庫能夠鑒定各細胞系中之倒位及易位(第2d-e圖)。對每個文庫之讀段量進行降低取樣顯示即使用少至5 Gb之讀段資料也偵測到易位之強信號(約1.7X總覆蓋率，第9a-h圖)。最後，對stLFR-1文庫中之兩種SV之檢查沒有產生明顯之模式(第9i-l圖)，表明偵測此等類型之SV之假陽性率較低。

1.5 使用stLFR來架設重疊群

【0067】 stLFR係一種強有力之方法，部分原因在於它使用非常大量(例如，約18億個)獨特之條碼並且能夠實現對每個單獨之長基因組DNA分子具有特異性之共條碼編輯。此類型之資料應該有利於從頭基因組裝配及改良之支架建立。為了證明stLFR如何用於改良基因組裝配，我們使用來自stLFR-1及stLFR-4文庫及SALSA(28)之讀段，以便架設NA12878(29)之單分子即時(SMRT)讀段組件，SALSA係針對染色質構象捕獲(Hi-C)資料來設計之程式。SALSA不係為stLFR資料設計的，因此必須將stLFR資料更改為類似於Hi-C之結構。此係藉由選擇共有相同條碼並且位於所捕獲之長DNA分子末端之讀段對來實現的。然後將此等標記為用於SALSA程式之讀段對。用stLFR資料代替Hi-C資料導致了優秀之支架建立。僅使用6千萬個stLFR讀段就能將1,411個重疊群連接成597個支架，N50為44.7 Mb。

此等支架覆蓋了2.84 Gb之基因組。與使用相同之重疊群及從人胚胎幹細胞產生之10倍(7.34億)Hi-C讀段對(30)在SALSA手稿中產生之指標相比，此等指標非常有利(表2)。stLFR支架之品質藉由將它們與人參考基因組之構建37比對並將其與程式dnadiff進行比較來進一步分析(31)。通常，stLFR支架與參考基因組密切一致，斷點、易位、重新定位及倒位之數量與用Hi-C讀段產生之支架之數量相似(表2)。比對點圖進一步證明了stLFR支架與參考基因組之間之高度連續性(第10圖)。

1.6 討論

【0068】 在本文中，我們描述了一種有效之全基因組定序文庫製備技術stLFR，它使得能夠在單管過程中用單一之獨特複製條碼對長基因組DNA分子之亞片段進行共條碼編輯。使用微珠作為小型化隔室允許每個樣品使用幾乎無限數量之複製條碼，成本可忽略不計。我們優化之基於雜交之將插入轉座子之DNA捕獲在珠子上，結合3'-分支連接及極端過量捕獲銜接子之核酸外切酶降解，成功地對長度長達300 kb之DNA分子中高達約20%之亞片段條碼化。重要地，此係在沒有初始長DNA片段之DNA擴增及隨之產生之表現偏差之情況下實現的。藉由此方式，stLFR解決了基於乳液之方法之成本及有限之共條碼編輯能力。

【0069】 使用stLFR之變體調用之品質非常高，並且藉由進一步優化，可能接近標準WGS方法之品質，但具有額外之益處，即共條碼編輯能夠實現進階資訊學應用。我們展示了在極低誤差率下高品質、接近完全地將基因組定相至長重疊群中，SV之偵測及重疊群之支架建立，以實現從頭組裝應用。所有此益處都係從不需要特殊設備，也不會顯著增加文庫製備成本之單一文庫來實現的。

【0070】 由於有效之條碼編輯，我們成功地使用了少至1 ng之人類DNA(600 X基因組覆蓋率)來製作stLFR文庫並獲得了高品質之WGS，其中大多數亞片段具有獨特之共條碼編輯。可以使用較少之DNA，但是stLFR在共條碼編碼期間不使用DNA擴增，因此不會從每個單獨之長DNA分子產生重疊之亞片段。因此，隨著DNA量降低，整體基因組覆蓋受到影響。此外，由於stLFR目前保留每個原始長DNA分子之10-20%，然後進行PCR擴增，因此產生了取樣問題。此導致相對高之讀取重複率並導致增加之定序成本，但是可以進行改良。一個明顯之解決方案係移除PCR步驟。此可以消除取樣，但也可以大大降低假陽性率及假陰性誤差率。此外，諸如優化轉座子之間之插入距離及增加配對末端200鹼基之定序讀段之長度之改良應該易於實現並且將增加覆蓋率及總體品質。對於一些應用，例如結構性變異偵測，可能需要使用較少之DNA及較少之覆蓋。正如我們在本文中所證明的，少至5 Gb之序列覆蓋率可以忠實地偵測染色體間及

染色體內易位，在此等情況下，重複率可以忽略不計。實際上，stLFR可以代表臨床環境中長配對文庫之簡單且具成本效益之替代品。

【0071】 此外，我們相信此類型之資料可以從單個stLFR文庫中實現完整之二倍體定相從頭組裝，而不需要長的物理讀取，例如由SMRT或奈米孔技術產生之彼等。轉座子插入之一個有趣特徵係它在相鄰之亞片段之間產生9鹼基序列重疊。通常，捕獲此等相鄰之亞片段並對其進行定序，使得讀段在長度上合成加倍(例如，對於200個鹼基讀段，兩個相鄰之捕獲之亞片段將產生具有9個鹼基重疊之兩個200鹼基讀段，或391個鹼基)。stLFR不需要特殊之設備，如基於液滴之微流體方法，每個樣品之成本很低。在本文中，我們展示了使用5000萬個珠子，但使用更多珠子係可能的。此將實現許多類型之具成本效益之分析，其中數億個條碼將係有用的。我們設想此類型之廉價大規模條碼編輯可藉由與單細胞技術或來自微生物樣品中16S RNA之深度群體定序組合，用於RNA分析，例如來自數千個細胞之全長mRNA定序。藉由測定轉座酶可達染色質(Assay for Transposase-Accessible Chromatin; ATAC-seq)(32)或甲基化研究之定相染色質定位也可以用stLFR進行。

1.7 靶核酸

【0072】 如本文所用，術語「靶核酸」(或多核苷酸)或「目標核酸」係指適於藉由本文所述方法來處理及定序之任何核酸(或多核苷酸)。核酸可以係單鏈或雙鏈的，並且可以包括DNA、RNA或其他已知之核酸。靶核酸可以係任何生物之核酸，包括但不限於病毒、細菌、酵母、植物、魚、爬行動物、兩棲動物、鳥類及哺乳動物(包括但不限於小鼠、大鼠、狗、貓、山羊、綿羊、牛、馬、豬、兔、猴子及其他非人類靈長類動物及人類)。靶核酸可以從個體或多個個體(即群體)獲得。獲得核酸之樣品可含有來自細胞或甚至生物之混合物之核酸，例如：包含人細胞及細菌細胞之人唾液樣品；小鼠異種移植物，該異種移植物包括小鼠細胞及來自移植之人腫瘤之細胞；等等。靶核酸可以係未擴增的，或者它們可以藉由此項技術中已知之任何合適之核酸擴增方法擴增。可以根據此項技術中已知之移除細胞及亞細胞污染物(脂質、蛋白質、碳水化合物，除了待定序之核酸之外之核酸等)之方法來純化靶核酸，或者它們可以係未純化的，即至少包括一些細胞及亞細胞污染物，包括但不限於被破壞以釋放其核酸用於處理及定序之完整細胞。可以使用此項技術中已知之方法從任何合適之樣品中獲得靶核酸。此等樣品包括但不限於：組織、分離之細胞或細胞培養物、體液(包括但不限於血液、尿液、血清、淋巴液、唾液、肛門及陰道分泌物、汗液及精液)；空氣、農業、水及土壤樣品等。靶核酸之非限制性實例包括「循環核酸」(circulating nucleic acid;

CNA)，該等核酸係在人血液或其他體液中循環之核酸，包括但不限於例如淋巴液、液體、腹水、乳汁、尿液、糞便及支氣管灌洗，並且可以區分為無細胞(cell-free;CF)或細胞相關核酸(在Pinzani等人，Methods 50：302-307,2010中綜述)。

【0073】 靶核酸可以係基因組DNA(例如，來自單個個體)、cDNA，及/或可以係複合核酸，包括來自多個個體或基因組之核酸。複合核酸之實例包括微生物組、在孕婦之血流中循環之胎兒細胞(參見，例如，Kavanagh等，J. Chromatol. B 878: 1905-1911, 2010)、來自癌症患者之血流之循環腫瘤細胞(CTC)(參見，例如，Allard等，Clin Cancer Res. 10: 6897-6904, 2004)。另一個實例係來自單個細胞或少量細胞之基因組DNA，例如來自活組織檢查(例如，從胚泡之滋養外胚層活檢之胎兒細胞；來自實體瘤之針抽吸物之癌細胞等)。另一個實例係在組織、血液或其他體液等中之病原體，例如細菌細胞、病毒或其他病原體。如本文所用，術語「複合核酸」係指大量不相同之核酸或多核苷酸。在某些實施例中，靶核酸係基因組DNA；外顯子組DNA(富含轉錄序列之全基因組DNA之一個子集，其中包含基因組中之外顯子組)；轉錄組(即在細胞或細胞群中產生之所有mRNA轉錄物之集合，或由此mRNA產生之cDNA)；甲基化組(即甲基化位點群及基因組中之甲基化模式)；外顯子組(即藉由外顯子捕獲或富集方法選擇之基因組之蛋白

質編碼區；微生物組；不同生物之基因組之混合物；生物之不同細胞類型之基因組之混合物；以及包含大量不同核酸分子之其他複合核酸混合物(實例包括但不限於微生物組、異種移植物、包含正常細胞及腫瘤細胞之實體腫瘤活組織檢查等)，包括上述類型之複合核酸之子集。在一個實施例中，此複合核酸具有包含至少一個千兆鹼基(Gb)之完整序列(二倍體人基因組包含約6 Gb之序列)。

【0074】 在一些情況下，靶核酸或第一片段係基因組片段。在一些實施例中，基因組片段長於10kb，例如10-100kb、10-500kb、20-300kb或長於100 kb。在單一混合物中使用之DNA(例如，人基因組DNA)之量可以係<10ng、<3ng、<1 ng、<0.3nm或<0.1ng之DNA。在一些情況下，靶核酸或第一片段之長度為5,000至100,000 KB

1.8 另外之方法

【0075】 儘管本文描述之工作實例使用聚合酶鏈反應，但是可以使用其他核酸擴增方法。熟習此項技術者能夠進行適合於合適之擴增技術之修飾。

【0076】 第17-25圖示出了另外之方法。第17圖顯示了雙鏈DNB之產生，雙鏈DNB可以係轉座子插入的並且被stLFR珠子捕獲。可以在同一DNA鏈上製備多達數千個拷貝(例如，10-10,000個拷貝，例如10-1000個拷貝或100-1000個拷貝)。此使得能夠藉由stLFR定序來高度覆蓋原始分子。第18圖說明當有限量之模板DNA可

用時，可在stLFR之前使用有限之預擴增步驟。第19圖描述了一種方法，其中使用低濃度之隨機切口酶、中等濃度之Klenow片段及高濃度之連接酶。珠子及DNA之濃度適合於stLFR。當產生切口並藉由Klenow來打開至間隙時，立即連接並將長片段鎖定在珠子上。允許產生切口並且打開更多間隙以使更多銜接子連接到間隙中。引子延伸產生約500個鹼基對片段。將第二個銜接子連接到鈍端，並對文庫進行定序。第20圖顯示髮夾銜接子在長DNA上之連接以及在環及Ph29或類似聚合酶中使用引子以在條碼編碼之前產生連鎖之dsDNA。除了改良每分子之讀段覆蓋率之外，該過程之有趣結果係，在聚合酶反應之0.5-3h結束時，每個多聯體之總「長度」(鹼基數)類似地與初始片段長度無關。此提供了使用經條碼編輯之珠子之選擇，該等珠子具有與多聯體之尺寸相對應之結合能力，從而防止每個珠子結合多個多聯體。此將減少每次反應所需之珠子數量，從而進一步降低成本。

【0077】 第25圖示出了中間長度標記之方法。在一種方法中，96個或更多個不同之經條碼編輯之轉座子藉由接頭部分(例如DNA，長惰性分子例如糊精或聚乙二醇(polyethylene glycol; PEG)，或長蛋白質例如角蛋白或膠原)以10個或更少之群組連接。雜交及連接可用於將轉座子附接至接頭DNA。其他方法可以藉由化學連接或藉由將抗生物素蛋白與此等分子連接並將生物素連接到轉座子上來附接。此實現了兩件事，它控制轉座子之間

插入之距離，並給出中間讀取接近資訊(10 kb或更小)。此對於重複序列之分析(串聯重複、三核苷酸定位等)係有用的。包含插入序列之DNA可以捕獲在珠子上，如本文及別處所述之其他stLFR方法。參見Joseph C. Mellor等人，「Phased NGS Library Generation via Tethered Synaptic Complexes,」 seqWell (2017)，可在全球資訊網(http://)上在seqwell.com/wp-content/uploads/2017/02/seqWell_LongBow_poster_ AGBT2017.pdf (2018年5月16日最後一次存取)處獲得。

1.9 第1部分之參考文獻

1. K. Zhang 等人, Long-range polony haplotyping of individual human chromosome molecules. *Nat Genet* 38, 382-387 (2006).

2. L. Ma 等人, Direct determination of molecular haplotypes by chromosome microdissection. *Nat Methods* 7, 299-301 (2010).

3. J. O. Kitzman 等人, Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol* 29, 59-63 (2011).

4. E. K. Suk等人, A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res* 21, 1672–1685 (2011).
5. H. C. Fan, J. Wang, A. Potanina, S. R. Quake, Whole-genome molecular haplotyping of single cells. *Nat Biotechnol* 29, 51-57 (2011).
6. B. A. Peters 等人, Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 487, 190-195 (2012).
7. J. Duitama等人, Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques. *Nucleic Acids Res* 40, 2041-2053 (2012).
8. S. Selvaraj, R. D. J, V. Bansal, B. Ren, Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol* 31, 1111-1118 (2013).
9. V. Kuleshov 等人, Whole-genome haplotyping using long reads and statistical methods. *Nat Biotechnol* 32, 261-266 (2014).
10. S. Amini 等人, Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and

combinatorial indexing. *Nat Genet* 46, 1343-1349 (2014).

11. G. X. Zheng等人, Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol*, (2016).

12. F. Zhang等人, Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat Biotechnol* 35, 852-857 (2017).

13. B. A. Peters, J. Liu, R. Drmanac, Co-barcoded sequence reads from long DNA fragments: a cost-effective solution for 「perfect genome」 sequencing. *Frontiers in genetics* 5, 466 (2014).

14. R. Drmanac. Nucleic Acid Analysis by Random Mixtures of Non-Overlapping Fragments. WO 2006/138284 A2 (2006).

15. R. Drmanac, Peters, B.A., Alexeev, A. Multiple tagging of long DNA fragments. WO 2014/145820 A2 (2014).

16. K. Jo, Y. L. Chen, J. J. de Pablo, D. C. Schwartz, Elongation and migration of single DNA molecules in microchannels using

oscillatory shear flows. *Lab Chip* 9, 2348-2355 (2009).

17. R. Drmanac 等人, Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327, 78-81 (2010).

18. T. Fehlmann等人, cPAS-based sequencing on the BGISEQ-500 to explore small non-coding RNAs. *Clin Epigenetics* 8, 123 (2016).

19. J. Huang等人, A reference human genome dataset of the BGISEQ-500 sequencer. *Gigascience* 6, 1-9 (2017).

20. S. S. T. Mak 等人, Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *Gigascience* 6, 1-13 (2017).

21. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-1760 (2009).

22. A. McKenna 等人, The Genome Analysis Toolkit: a MapReduce framework for analyzing

next-generation DNA sequencing data. *Genome Res* 20, 1297-1303 (2010).

23. J. M. Zook 等人, Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 32, 246-251 (2014).

24. P. Edge, V. Bafna, V. Bansal, HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* 27, 801-812 (2017).

25. Q. Mao等人, The whole genome sequences and experimentally phased haplotypes of over 100 personal genomes. *Gigascience* 5, 1-9 (2016).

26. Z. Dong 等人, Low-pass whole-genome sequencing in clinical cytogenetics: a validated approach. *Genet Med* 18, 940-948 (2016).

27. Z. Dong等人, Identification of balanced chromosomal rearrangements previously unknown among participants in the 1000 Genomes Project: implications for interpretation of structural variation in

genomes and the future of clinical cytogenetics. *Genet Med*, (2017).

28. J. Ghurye, M. Pop, S. Koren, D. Bickhart, C. S. Chin, Scaffolding of long read assemblies using long range contact information. *BMC Genomics* 18, 527 (2017).

29. M. Pendleton等人, Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* 12, 780-786 (2015).

30. J. R. Dixon等人, Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376-380 (2012).

31. A. M. Phillippy, M. C. Schatz, M. Pop, Genome assembly forensics: finding the elusive mis-assembly. *Genome biology* 9, R55 (2008).

32. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10, 1213-1218 (2013).

實例

2. 實例1：方法及材料

2.1. 高分子量DNA分離

【0078】 根據RecoverEase™ DNA分離套組(Agilent Technologies，La Jolla，CA)方案(1)之修改版本從細胞系中分離長基因組DNA。

【0079】 簡言之，將約1百萬個細胞沉澱並用500 ul裂解緩衝液裂解。在4℃溫育10分鐘後，將在4 mL消化緩衝液中之20 μL之RNase-IT核糖核酸酶混合物直接添加至裂解之細胞中，並在50℃加熱塊上溫育。5分鐘後，添加4.5 mL蛋白酶K溶液(約1.1 mg/mL蛋白酶K，0.56%SDS及0.89X TE)，將混合物在50℃下再溫育2小時。然後將基因組DNA轉移到具有1,000 kD截留分子量之透析管(Spectrum Laboratories，Inc.，Rancho Dominguez，CA)，並在室溫下在0.5X TE緩衝液中透析過夜。

2.2 經條碼編輯之珠子構建

【0080】 使用三組雙鏈經條碼編輯之DNA分子藉由基於分割及池連接之策略構建經條碼編輯之珠子。參見第12及13圖。將包含PCR引子降溫貼合位點之共同銜接子序列與具有5'雙生物素接頭之Dynabeads™ M-280鏈黴抗生物素蛋白(ThermoFisher，Waltham，MA)磁

珠連接。藉由Integrated DNA Technologies(Coralville，IA)構建了三組1,536個含有重疊序列區域之條碼寡核苷酸。在含有50 mM Tris-HCl(pH 7.5)、10 mM $MgCl_2$、1 mM ATP、2.5%PEG-8000、571單位T4連接酶、580 pmol條碼寡核苷酸及6500萬個M-280珠子之15 μL反應中在384孔板中進行連接。將連接反應在室溫下在旋轉器上溫育1小時。在連接之間藉由離心將珠子彙集到單個容器中，使用磁鐵收集到容器之側面，並用高鹽洗滌緩衝液(50 mM Tris-HCl(pH 7.5)、500 mM NaCl、0.1 mM EDTA，及0.05%吐溫20)洗滌一次並且用低鹽洗滌緩衝液(50 mM Tris-HCl(pH 7.5)、150 mM NaCl及0.05%吐溫20)洗滌兩次。將珠子重懸於1X連接緩衝液中並分佈在384孔板上，並重複連接步驟。

【0081】 本文提到之某些「條碼」係「三部分條碼」。三部分係指它們的結構及/或它們的合成。如第12圖所示，三部分條碼可以藉由較短(例如，4-20個核苷酸)序列之連續連接來合成。在一個實施例中，較短之條碼長度為10個鹼基。如圖所示，示例性結構包含CS1-BC1-CS2-BC2-CS3-BC3-CS4，其中CS係存在於所有捕獲銜接子上之恆定序列，並且BC序列係如本文所討論之不同之10鹼基條碼。可以使用部分雙鏈寡核苷酸來構建三部分條碼，該等寡核苷酸具有降溫貼合至作

為BC之互補序列(即BC')之較短寡核苷酸之結構CSa-BC-CSb，如圖中所示。

【0082】 在一個態樣，本發明提供了一種組合物，該組合物包含具有附著之複製條碼之捕獲寡核苷酸之珠子，其中該組合物包含超過30億個不同之條碼，並且其中條碼係具有結構5'-CS1-BC1-CS2-BC2-CS3 -BC3-CS4之三部分條碼。在一些實施例中，CS1及CS4比CS2及CS3長。在一些實施例中，CS2及CS3係4-20個鹼基，CS1及CS4係5或10到40個鹼基，例如20-30個，並且BC序列之長度係4-20個鹼基(例如，10個鹼基)。在一些實施例中，CS4與夾板寡核苷酸互補。在一些實施例中，組合物包含橋寡核苷酸。在一些實施例中，該組合物包含橋寡核苷酸，包含如上所述之三部分條碼之珠子，及包含具有與橋寡核苷酸互補之區域之雜交序列之基因組DNA。

2.3 使用兩個轉座子之stLFR

【0083】 在55℃下，在10 mM TAPS-NaOH(pH8.5)、5 mM $MgCl_2$及10%DMF之60 μL反應中，將2 pmol Tn5偶聯之轉座子插入到40 ng基因組DNA中，持續10分鐘。將1.5 μL插入轉座子之DNA轉移至248.5 μL由50 mM Tris-HCl(pH 7.5)、100 mM $MgCl_2$及0.05%TWEEN® 20組成之雜交緩衝液中。將1000萬至5000萬個經條碼編碼之珠子

重懸於同一雜交緩衝液中。將稀釋之DNA添加到條碼編輯之珠子中，並將混合物加熱至60℃持續10分鐘，偶爾進行輕微混合。將DNA-珠混合物轉移到實驗室烘箱中之管旋轉器中，並在45℃下溫育50分鐘。將500 uL含有50 mM Tris-HCl(pH 7.8)、10 mM DTT、1 mM ATP、2.5%PEG-8000及4,000單位T4連接酶之連接混合物直接添加到DNA-珠子混合物中。將連接反應物在室溫下在旋轉器上溫育1小時。添加110 μL之1%SDS，將混合物在室溫下溫育10分鐘以移除Tn5酶。藉由磁鐵將珠子收集到管之側面，並用低鹽洗滌緩衝液洗滌一次，用NEB2緩衝液(New England Biolabs，Ipswich，MA)洗滌一次。使用100 uL 1X NEB2緩衝液中之10單位之UDG(New England Biolabs，Ipswich，MA)、30單位之APE1(New England Biolabs，Ipswich，MA)及40單位之外切核酸酶1(New England Biolabs，Ipswich，MA)移除過量之條碼寡核苷酸。將該反應在37℃下溫育30分鐘。將珠子收集到管之側面並用低鹽洗滌緩衝液洗滌一次並用1X PCR緩衝液(1X PfuCx緩衝液(Agilent Technologies，La Jolla，CA)、5%DMSO、1 M甜菜鹼、6 mM $MgSO_4$及600 μM dNTP)洗滌一次。將含有1X PCR緩衝液，400pmol各引子及6μL PfuCx酶(Agilent Technologies，La Jolla，CA)之PCR混合物加熱至95℃3分鐘，然後冷卻至室溫。將該混合物用於重懸珠

子，將合併之混合物在72℃溫育10分鐘，然後進行12個循環之95℃10秒、58℃30秒及72℃2分鐘。

2.4 使用3'分支連接銜接子之stLFR

【0084】 該方法以相同之雜交插入條件開始，但僅使用一個轉座子而不係兩個轉座子。在捕獲及條碼連接步驟後，如上所述，將珠子收集到管之側面並用低鹽洗滌緩衝液洗滌。添加90單位外切核酸酶I(New England Biolabs，Ipswich，MA)及100單位外切核酸酶Ⅲ(New England Biolabs，Ipswich，MA)在100μL 1XTA緩衝液(Teknova，Hollister，CA)中之銜接子消化混合物。將珠子在37℃下溫育10分鐘。停止反應，藉由添加11 μL之1%SDS來移除Tn5酶。將珠子收集到管之側面並用低鹽洗滌緩衝液洗滌一次並用1X NEB2緩衝液(New England Biolabs，Ipswich，MA)洗滌一次。藉由添加100 uL 1X NEB2緩衝液(New England Biolabs，Ipswich，MA)中之10單位UDG(New England Biolabs，Ipswich，MA)及30單位APE1(New England Biolabs，Ipswich，MA)來移除過量之捕獲寡核苷酸，並且在37℃溫育30分鐘。將珠子收集到管之側面並用高鹽洗滌緩衝液洗滌一次並用低鹽洗滌緩衝液洗滌一次。在室溫下在旋轉器上，使用含有50 mM Tris-HCl(pH 7.8)、10 mM $MgCl_2$、0.5mM DTT、1 mM ATP及10%PEG-8000之100 uL

連接酶緩衝液中之4,000單位T4連接酶，將300 pmol第二銜接子與珠子結合之亞片段連接，持續2小時。將珠子收集到管之側面並在高鹽洗滌緩衝液中洗滌一次並在1X PCR緩衝液中洗滌一次。PCR混合物及條件與上述雙轉座子方法相同。

【0085】 示例性3'分支連接銜接子包含表6中所示之3'分支連接銜接子-F(/5Phos/CTGATGGCGCGAGGGAGGC)及3'分支連接銜接子-R(TCGCGCCATCA/3'dd/G)寡核苷酸。在該實例中，銜接子F序列包含PCR引子降溫貼合序列。視情況地，條碼(例如，樣品條碼)可以包括在5'磷酸與所示序列之間。在該實例中，銜接子R序列比引子降溫貼合序列短，使得它在PCR引子降溫貼合之條件下融化。

2.5 序列定位及變體調用

【0086】 使用條碼分割工具(可從GitHub https://github.com/stLFR/stLFR_read_demux獲得)，藉由相關條碼序列首先對原始讀段資料進行解複用。用BWA-MEM將分配條碼並且剪切之讀段定位到hs37d5參考基因組(2)。然後藉由SAMtools(3)按照染色體坐標對得到之BAM檔案進行分類，並用picard MarkDuplicate函數(http://broadinstitute.github.io/picard)標記重複物。使用GATK4.0.3.0(4)中之

HaplotypeCaller進行短變體(SNP及插入缺失)調用。然後，使用rtgtools vcfeval函數(6)，將從上述步驟生成之vcf檔案與瓶中基因組(GIAB)高置信度變體列表進行基準測試(ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/latest/GRCh37 /)(5)。基準測試後，使用GATK VariantRecalibrator分析stLFR文庫，GIAB真值集用於訓練高斯混合模型。然後使用GATK ApplyVQSR過濾VCF。在幾乎所有情況下，99.9份額應用於原始vcf，但100 Gb stLFR-1文庫及STD文庫除外，其中應用了100份額。然後，我們根據表5中列出之GQ得分、參考與替代深度之比率以及條碼支持來建立並應用了進一步硬過濾標準：

2.6 使用Hapcut2進行變體定相

【0087】 SNP使用其10X Genomics資料管道以Hapcut2(https://github.com/vibansal/HapCUT2)(7)來定相。BAM檔案首先被轉換為攜帶條碼資訊之格式，其格式與10X Genomics條碼編輯之BAM類似。具體地，向每行添加「BX」字段，反映該讀段之條碼資訊。GIAB變體或藉由GATK為每個文庫調用之變體用作定相之輸入，並且使用Hapcut2之calculate_haplotype_statistics.py工具總結定相結果並與GIAB定相之vcf檔案(5)進行比較。

2.7 LongHap

【0088】 種子延伸策略用於LongHap之定相過程。它最初從一對種子開始，該對種子由染色體中最上游之雜合變體組成。藉由連接其他下游候選變體來延長種子，直到不再向延伸種子添加變體(第11圖)。在該延伸過程中，不同基因座處之候選變體將不會被平等處理(即，跨越染色體，與下游變體相比，上游變體具有更高之優先級)。每兩個雜合基因座沿兩個不同之等位基因具有兩種可能之組合。以變體T_2/G_2及G_3/C_3為例(第11圖)，一種組合模式係T_2-G_3及G_2-C_3，而另一種模式係T_2-C_3及G_2-G_3。每個組合之得分藉由跨越兩個基因座之長DNA片段之數量來計算，該數量等於定位到此等兩個基因座之讀段之獨特條碼之數量。如第11圖所示，前一組合之最終得分為3，係後者之三倍。將變體T_2/G_2添加到延伸之種子中並重複該過程。值得注意地，如果任何條碼支持一個特定基因座上之兩個等位基因，則在計算連接得分時將忽略它。此有助於降低切換誤差率。當在連接下游候選變體之過程中發生衝突時，如第11圖中之變體A_4/C_4所示，將藉由比較連接之基因座編號進行簡單之決定以允許進一步延伸候選變體。在此情況下，左方情形中有兩個連接之基因座，而右方情形中只有一個。LongHap將選擇左組合模式作為最終定相結果。

2.8 SV偵測

【0089】 如前所述，藉由計算基因組區域之間之共有條碼來偵測結構變體(8)。首先移除重複讀段。沿著基因組使用滑動窗口(預定值為2 kb)來掃描所定位之共條碼讀段，每個窗口記錄在該2 kb窗口內發現了多少條碼，並且計算窗口對之間之共有條碼比率之Jaccard指數。藉由窗口對之間之Jaccard指數共有度量來識別結構變體事件。

【0090】 對於基因組中之每個窗口對(X，Y)，Jaccard指數計算如下：

$$X = (x_1, x_2, \ldots x_n);\ Y = (y_1, y_2, \ldots y_n)$$

$$\text{Jaccard\_指數}_{ij} = \begin{cases} \dfrac{x_i \cap y_j}{x_i \cup y_j} & (if\ x_i > 0\ or\ y_j > 0) \\ 0 & (if\ x_i = y_j = 0) \end{cases}$$

2.9 藉由SALSA之重疊群支架建立

【0091】 來自stLFR文庫之定序讀段用於架設含有18,903個重疊群之NA12878組裝體，NG50為26.83Mb(9)(使用支架程式SALSA從NCBI基因組網站下載之重疊群(10)。為了模擬適合於SALSA之HiC序列結構，從尺寸>= 5 kb之片段中選擇stLFR序列讀段。從長度>= 5 kb之每個片段中，選擇「第一」及「最後」讀段以形成讀段對。隨後，藉由以2 kb之間隔在此等片段上向內移動來選擇此等人工讀段對。然後將此等讀段對定位到NA12878重疊群上，並用SALSA進行支架建

立。然後使用MUMmer 4程式之nucmer及dnadiff將得到之支架與hg19參考基因組比對並進行比較(11)。

2.10 實例1之參考文獻

1. I. Agent Technologies, RecoverEase DNA Isolation Kit. Revision C.0, (2015).

2. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754-1760 (2009).

3. H. Li 等人, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079 (2009).

4. A. McKenna 等人, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297-1303 (2010).

5. J. M. Zook等人, Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 32, 246-251 (2014).

6. J. G. Cleary 等人, Comparing Variant Call Files for Performance Benchmarking of

Next-Generation Sequencing Variant Calling Pipelines. *bioRxiv*, (2015).

7. P. Edge, V. Bafna, V. Bansal, HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* 27, 801-812 (2017).

8. F. Zhang等人, Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat Biotechnol* 35, 852-857 (2017).

9. M. Pendleton等人, Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* 12, 780-786 (2015).

10. J. Ghurye, M. Pop, S. Koren, D. Bickhart, C. S. Chin, Scaffolding of long read assemblies using long range contact information. *BMC Genomics* 18, 527 (2017).

11. S. Kurtz等人, Versatile and open software for comparing large genomes. *Genome biology* 5, R12 (2004).

實例2：詳細方案

3.1 材料

【0092】 1 Kb Plus DNA梯（ThermoFisher, cat. no. 10787018）

100Kd MWCO Biotech CE透析管（Spectrum Labs, cat. no. 131486）

384-孔 Armadillo PCR 板（ThermoFisher, cat. no. AB2384）

Agencourt® AMPure XP珠子（Beckman Coulter, cat. no. A63882）

APE 1（10,000單位/mL）（New England Biolabs, cat. no. M0282L）

ATP（100 mM）（Teknova, cat. no. A1210）

條碼編輯之珠子構建寡核苷酸（IDT）（參見注釋）

甜菜鹼（5 M）（Sigma-Aldrich, cat. no. B0300-5VL）

BSA（20mg/mL）（New England Biolabs, cat. no. B9000S）

共同銜接子寡核苷酸（IDT）

DMF（約100%）（Sigma-Aldrich, cat. no. D4551-250ML）

DMSO（100%）（Sigma-Aldrich, cat. no. D9170-5VL）

dNTP（25 mM）（ThermoFisher, cat. no. R1121）

透析管（1,000 kD MWCO）（Spectrum Laboratories, Inc., cat. no. 131486）

DTT(Sigma-Aldrich, cat. no. 11583786001)

Dynabeads™ M-280 鏈黴抗生物素蛋白(ThermoFisher, cat. no. 60210)

EDTA(0.5 M, pH 8.0)(Sigma-Aldrich, cat. no. 03690-100ML)

外切核酸酶I(20,000單位/mL)(New England Biolabs, cat. no. M0293L)

外切核酸酶Ⅲ(100,000單位/mL)(New England Biolabs, cat. no. M0206L)

甲醯胺(100%, 250 mL)(Sigma-Aldrich, cat. no. 47671-250ML-F)

甘油(100%)(Sigma-Aldrich, cat. no. G5516-100ML)

KCl(Sigma-Aldrich, cat. no. P9333-1KG)

KH2PO4(Sigma-Aldrich, cat. no. 795488-1KG)

KOH(Sigma-Aldrich, cat. no. P5958-1KG)

MgCl2(1 M)(Sigma-Aldrich, cat. no. 63069-500ML)

MgSO4(1 M)(Sigma-Aldrich, cat. no. M3409-100ML)

MicroAmp透明黏合劑膜(ThermoFisher, cat. no. 4306311)

NaCl(5M)(ThermoFisher, cat. no. AM9760G)

Na2HPO4 (Sigma-Aldrich, cat. no. S7907-1KG)

NaOH (10M) (Sigma-Aldrich, cat. no. 72068-100ML)

NEB2緩衝液(10X) (New England Biolabs, cat. no. B7002S)

PEG-8000 (50%) (Rigaku, cat. no. 1008063)

Pfu Turbo Cx Hotstart DNA聚合酶(Agilent, cat. no. 600414)

蛋白酶K，重組，PCR等級溶液(14-22 mg/mL) (Roche, cat. no. 03115844001)

RiboRuler Low Range RNA梯 (Thermofisher, cat. no. SM1831)

RNase-IT核糖核酸酶混合物(Agilent, cat. no. 400720)

SDS (10%) (ThermoFisher, cat. no. 15553027)

蔗糖(Sigma-Aldrich, cat. no. S7903-1KG)

T4 DNA連接酶(2x106單位/mL) (New England Biolabs, cat. no. M0202M)

TA緩衝液(10X) (Teknova, cat. no. T0379)

TAPS-NaOH (1 M, pH 8.5) (Boston BioProducts, cat. no. BB-2375)

TBE (10X) (ThermoFisher, cat. no. 15581028)

TE緩衝液(10X)(Fisher Scientific, cat. no. BP24771)

Tn5酶

轉座子寡核苷酸(IDT)

Tris-HCl(1 M, pH 7.5)(ThermoFisher, cat. no. 15567027)

Tris-HCl(2 M, pH 7.8)(Amresco, cat. no. J837-500ML)

Triton™ X-100(10%)(Sigma-Aldrich, cat. no. 93443-100ML)

TWEEN® 20(10%)(Roche, cat. no. 11332465001)

UDG(5,000單位/mL)(New England Biolabs, cat. no. M0280L)

3.2 設備

【0093】 2.4 L高聚苯乙烯容器(Click Clack, cat. no. 659030)或等效物

DynaMag™-2磁鐵(ThermoFisher, cat. no. 12321D)

Easy 50 EasySep™磁鐵(Stem Cell Technologies, cat. no. 18002)或等效物

能夠容納管回轉器/旋轉器之實驗室烘箱

磁板攪拌器

中型磁力攪拌棒

標準實驗室渦旋器

Tetrad PCR熱循環儀(Bio-Rad, cat. no. PTC0240)或每孔能夠容納100 μL反應體積之等效物

管回轉器/旋轉器(Thermo Fisher, cat. no. 88881001)或等效物

【0094】 ***3.3 試劑設置***

【0095】 降溫貼合緩衝液(3X)

3 mL之1 M Tris-HCl, pH 7.5

6 mL之5 M NaCl

91 mL之無菌dH2O

在室溫下儲存1年。

【0096】 3.4 緩衝液D (10X)

224 mg之KOH

50 μL之0.5 M EDTA

2.45 mL之無菌dH2O

製作等分試樣並在-20℃下儲存1個月。

【0097】 3.5 偶聯緩衝液(1X)

5 mL之1X TE

5 mL之100%甘油

在-20℃下儲存1年。

【0098】 3.6 消化緩衝液(1X, pH 8.0)

1.75 g之Na2HPO4

0.2 g之KCl

0.2 g之KH2PO4

27.4 mL 5 M NaCl

20 mL之0.5 M EDTA (pH 8.0)

800 mL之無菌dH2O

用1 M NaOH將pH調節至8.0。

添加無菌dH2O至終體積為1升。

過濾消毒。

在室溫下儲存1年。

【0099】 3.7 3'分支連接緩衝液(3X)

6 mL之50% PEG-8000

0.75 mL 2 M Tris-HCl (pH 7.8)

0.3 mL 1 M MgCl2

0.3 mL 0.1 M ATP

15 μL 1 M DTT

75 μL 20 mg/mL BSA

2.560 mL之無菌dH2O

在-20℃下儲存1年。

【0100】 3.8 高鹽珠子結合緩衝液(1X)

5 mL之1 M Tris-HCl (pH7.5)

6 mL之5 M NaCl

20 μL之0.5 M EDTA

88.98 mL之無菌dH2O

在室溫下儲存1年。

【0101】 3.9 高鹽洗滌緩衝液(1X)

5 mL之1 M Tris-HCl, pH 7.5

10 mL之5 M NaCl

20 μL之0.5 M EDTA

0.5 mL之10% TWEEN® 20

84.48 mL之無菌dH2O

在室溫下儲存1年。

【0102】 3.10 雜交緩衝液(1X)

50 mL之1 M Tris-HCl, pH 7.5

100 mL之1 M MgCl2

5 mL之10% TWEEN® 20

845 mL之水

在室溫下儲存1年

【0103】 3.11 連接緩衝液(10X)

25 mL之50% PEG-8000

12.5 mL之2 M Tris-HCl (pH 7.8)

5 mL之100 mM ATP

5 mL之1 M MgCl2

2.5 mL之無菌dH2O

在-20℃下儲存1年。

【0104】 3.12 連接緩衝液，無MgCl2 (10X)

25 mL之50% PEG-8000

12.5 mL之2 M Tris-HCl (pH 7.8)

5 mL之100 mM ATP

5 mL之1 M DTT

2.5 mL之無菌dH2O

在-20℃下儲存1年。

【0105】 3.13 低鹽洗滌緩衝液(1X)

5 mL之1 M Tris-HCl, pH 7.5

3 mL之5 M NaCl

0.5 mL之10% TWEEN® 20

91.5 mL之無菌dH2O

在室溫下儲存1年。

【0106】 3.14 裂解緩衝液(1X, pH 8.3)

0.22 g之KCl

120 g蔗糖

13 mL之1 M Tris-HCl (pH 7.5)

2 mL之0.5 M EDTA (pH 8.0)

28 mL之5 M NaCl

10 mL之Triton® X-100

800 mL之無菌dH2O

將pH調節至8.3

添加無菌dH2O至終體積為1升。

過濾消毒

在4℃下儲存1年。

【0107】 3.15轉座酶緩衝液(5X)

0.5 mL之1 M TAPS-NaOH (pH 8.5)

0.25 mL之1 M MgCl2

5 mL之100% DMF

4.25 mL之無菌dH2O

在-20℃下儲存1年。

【0108】 3.16 PfuCx mix (2X)

2 mL之10X PfuCx緩衝液(與酶一起包括在內)

0.5 mL之100% DMSO

2 mL之5 M甜菜鹼

60 μL之1 M MgSO4

240 μL之25 mM dNTP

5.2 mL之無菌dH2O

【0109】 3.16 經條碼編輯之珠子構建寡核苷酸

【0110】 伴以標準脫鹽，所有條碼編輯之寡核苷酸以384孔形式以100 nmol規模合成，並藉由Integrated DNA Technologies(Coralville，IA)以1X TE(pH8.0)中之200 μM之濃度遞送。每個條碼集共有1,536種獨特之條碼寡核苷酸，並且有3個條碼集。此可以實現多達約36億個不同之條碼組合。對於一些應用來說，此可能係不必要的，並且藉由排序更少之寡核苷酸板可以實現更少之條碼組合。該特定設計確實需要使用來自每個集合之至少一種條碼寡核苷酸來產生適當之最終序列，然而，可以對條碼集之間之6個鹼基重疊序列進行輕微修改以移除整個條碼集。

3.2 程序

【0111】 從細胞中分離高分子量DNA

該方法基於RecoverEase™ DNA分離套組方案26，但是使用更大之體積進行，以降低所得溶液之黏度。

【0112】 1. 在15或50 mL錐形管(500xg，5分鐘)中沉澱至多1×107個分散之有核細胞。移除上清液。向細胞沉澱中添加500 μL裂解緩衝液，以中速短暫渦旋樣品3-5秒，將錐形管置於冰箱中約10分鐘，偶爾旋轉。

【0113】 2. 藉由組合250 μL 10%SDS、250 μL蛋白酶K及4 mL 1X TE製備蛋白酶K溶液。置於50℃加熱塊上並短暫加熱(約5分鐘)。

【0114】 3. 藉由將20 μL RNase-It核糖核酸酶混合物與4 mL消化緩衝液混合製備消化溶液。

【0115】 4. 將約4 mL製備之消化溶液添加到來自步驟1之裂解之細胞及緩衝液中並輕輕搖動錐形管。

【0116】 5. 5分鐘後將錐形管置於50℃之加熱塊中，將4.5 mL溫熱之蛋白酶K溶液添加到自由浮動之沉澱中。輕輕旋轉錐形管以混合。

【0117】 6. 重新蓋上試管，在50℃加熱塊中溫育2小時，每隔30分鐘輕輕旋轉試管。

【0118】 7. 切割約13 cm之透析管(其容量約為1 mL/cm)。在0.5X TE中平衡30分鐘。用透析夾密封一端。

【0119】 8. 將至少1 L之0.5X TE緩衝液倒入透析容器中。

【0120】 9. 小心地將黏性基因組DNA從錐形管倒入透析管之開口端。用透析夾密封透析管之開口端。將浮子附接到一個夾子。將帶有浮子之透析管放入透析容器中。

【0121】 10. 在室溫下透析基因組DNA 24至48小時，同時用磁力攪拌棒輕輕攪拌緩衝液。在透析期間更換TE緩衝液一次，以使回收之DNA之純度最大化。

【0122】 11. 完成透析後，從TE緩衝液中移除透析管，從透析管頂部移除浮子及夾子，輕輕倒入15 mL錐形管中。DNA可以立即使用而不會剪切。

3.3 條碼編輯之珠子

【0123】 使用分割及池策略以3組雙鏈條碼DNA分子來構建經條碼編輯之珠子。藉由連續連接來構建全長銜接子(第12圖及第13圖)。條碼寡核苷酸在384孔板中提供(參見試劑注釋)。共同之銜接子寡核苷酸以管形式提供。根據所使用之定序技術，可能需要改變共同銜接子寡核苷酸內之PCR引子序列。

【0124】 12. 在384孔PCR板中，將來自384孔板之每個孔之10 μL互補寡核苷酸與10 μL 3X降溫貼合緩衝液混合。將30 μL共同之銜接子寡核苷酸在8-孔PCR條管之一個孔中混合。

【0125】 13. 在70℃下溫育3分鐘，然後在PCR熱循環儀上以0.1℃/ s緩慢升溫至20℃。雜交之條碼寡核苷酸之終濃度為66 μM。

【0126】 14. 將含有5'雙生物素之4.725 mL(157.5 μmol)雜交珠子接頭與3.225 mL連接緩衝液(10X)、460.8 μL(921,600單位)T4 DNA連接酶及9.67 mL dH_2O混合，總體積為18.081 mL。

【0127】 15. 將11.2 μL連接混合物分配到四個新的384孔PCR板之每個孔中。然後從雜交之第一條碼板之每個孔向每個含有珠子接頭混合物之孔中添加8.8 μL(580 pmol)。用MicroAmp透明黏合劑膜密封、渦旋、離心，並在室溫下溫育1小時。

【0128】 16. 藉由將50 mL珠子轉移到空的50 mL離心管中，收集1000億(143 mL)M-280鏈黴抗生物素蛋白塗覆之磁珠。將帶有珠子之50 mL試管放入Easy 50 EasySep™磁鐵中5分鐘，將珠子收集到試管之側面。用移液管小心地移除上清液。將第二份50 mL珠子轉移到磁鐵上之管中。在磁鐵上靜置5分鐘，小心地移除上清液。將最終之43 mL珠子轉移到50 mL管中。在磁鐵上靜置5分鐘，小心地移除上清液。用低鹽洗滌緩衝液洗滌珠子兩次，然後在8 mL高鹽珠子結合緩衝液中重懸。

【0129】 17. 將高鹽珠子結合緩衝液中之5 μL珠子分配到含有連接產物之板之每個孔中。在分配期間偶爾渦旋珠子源管以保持珠子良好懸浮。

【0130】 18. 用MicroAmp透明黏合劑膜將板密封，渦旋，並置於管旋轉器上，在「振盪」模式下在室溫下溫育1小時。

【0131】 19. 將板以300×g離心5秒以使珠子解除密封，但不允許形成沉澱。解除密封，每孔添加2.8 μL 0.1%SDS。用MicroAmp透明黏合劑膜再次將板密封，短暫渦旋並在室溫下溫育10分鐘。

【0132】 20. 渦旋，然後以300×g將板離心5秒鐘以使珠子解除板密封。從每個板上解除密封，將板倒置到收集盤上。以500×g離心2分鐘。使用10 mL血清移液管，將珠子收集到一個新的50 mL管中。

【0133】 21. 將珠子收集在Easy 50 EasySep™磁鐵上之管子側面5分鐘。丟棄上清液。用10 mL高鹽洗滌緩衝液洗滌一次，然後用低鹽洗滌緩衝液洗滌兩次。將珠子重懸於8 mL 1X連接緩衝液中。

【0134】 22. 將5ul珠子分配到四個新的384孔PCR板之每個孔中。在分配期間偶爾渦旋珠子源管以保持珠子良好懸浮。

【0135】 23. 為了連接第二組條碼，製備含有3.225 mL連接緩衝液(10X)、460.8 μL(921,600單位)T4 DNA連接酶及6.33 mL dH2O之混合物至10.02 mL之總體積。將6.2 μL第二連接混合物分配到含有珠子之四個384孔PCR板之每個孔中。接下來，將雜交之第二條

碼板之每個孔中之8.8 μL(580pmol)添加至含有珠子及連接混合物之384孔PCR板之相應孔中。

【0136】 24. 重複步驟18-22。

【0137】 25. 為了連接第三組條碼，製備含有3.225 mL連接緩衝液(10X)、460.8 μL(921,600單位)T4 DNA連接酶及6.33 mL dH2O之連接混合物至10.02mL之總體積。將6.2 μL第三連接混合物分配到含有珠子之四個384孔PCR板之每個孔中。接下來，將雜交之第三條碼板之每個孔中之8.8 μL(580pmol)添加到含有珠子及連接混合物之384孔PCR板之相應孔中。

【0138】 26. 重複步驟18-22。珠子現在可以在4℃下儲存長達一年。在目前之形式中，珠子幾乎完全係雙鏈的，並且還沒有處於用於stLFR之正確形式下。

【0139】 27. 用血細胞計數器計數珠子並取出500萬個珠子用於QC步驟。將帶有珠子之試管放在DynaMag™-2磁鐵上5分鐘。丟棄上清液。添加5 μL 100%甲醯胺、4 μL dH2O及1 μL 10X上樣緩衝液。在PCR熱循環儀上將其在95℃溫育3分鐘。立即放在冰上2分鐘。將帶有珠子之試管放在DynaMag™-2磁鐵上5分鐘。收集上清液，加載到15%TBU凝膠上，並在200V下運行40分鐘以檢查寡核苷酸長度及量。或者，可以使用流式細胞儀藉由將螢光標記之寡核苷酸與珠子銜接子序列之3'末端雜交來檢查珠子。我們通常看到約25%之鏈黴抗生物素蛋白結合位點具有全長構建之銜接子序列。

3.20 用於stLFR之珠子製備

【0140】 為了製備用於stLFR之珠子，必須首先將它們變性為單鏈DNA，然後與橋接寡核苷酸重新雜交。

【0141】 28. 將來自前一部分之步驟26之5億個構建之經條碼編輯之珠子移取到標準之1.5 mL微量離心管中。

【0142】 29. 放置在DynaMag™-2磁鐵上2分鐘，將珠子收集到管子之側面。移除上清液。

【0143】 30. 添加1 mL緩衝液D之1X稀釋液。短暫渦旋並在室溫下溫育2分鐘。

【0144】 31. 放置在DynaMag™-2磁鐵上2分鐘，將珠子收集到管子之側面。移除上清液。

【0145】 32. 再重複步驟30及31。

【0146】 33. 在1X降溫貼合緩衝液中洗滌一次。放置在DynaMag™-2磁鐵上2分鐘，將珠子收集到管子之側面。移除上清液。

【0147】 34. 混合36 μL之100 μM橋接寡核苷酸、333.33 μL之降溫貼合緩衝液(3X)及630.67 μL之dH2O，最終體積為1 mL。將混合物添加到珠子中。短暫地渦旋。

【0148】 35. 在60℃溫育5分鐘，在室溫下溫育50分鐘。

【0149】 36. 放置在DynaMag™-2磁鐵上2分鐘，將珠子收集到管子之側面。移除上清液，重懸於500 μL低

鹽洗滌緩衝液中。此等珠子現在已準備好用於stLFR，可在4℃下儲存3個月。

3.21兩個轉座子stLFR方案

【0150】 該方案利用兩個轉座子沿著基因組DNA分子之長度產生雜交序列及PCR引子位點。此係最簡化及最快速之stLFR方法，但每個長DNA片段之覆蓋率可能比3'分支連接方案低50%。為了與BGISEQ-500以外之定序技術相容，可能需要改變在嵌合區域之後之一些轉座子序列。在排序此等寡核苷酸之前檢查所用之定序引子。有關所有寡核苷酸序列之資訊可在補充材料中獲得。

【0151】 37. 將藉由在8孔PCR條管之第一個孔中將10 μL轉座子1T(100 μM)、10 μL轉座子B(100 μM)、10 μL降溫貼合緩衝液(3X)組合所獲得之捕獲轉座子寡核苷酸與藉由在相同PCR條管之第二個孔中將10 μL轉座子1T(100 μM)、10 μL轉座子B(100 μM)、10 μL降溫貼合緩衝液(3X)組合所獲得之未捕獲之轉座子寡核苷酸雜交。

【0152】 38. 在70℃下溫育3分鐘，然後在PCR熱循環儀上以0.1℃/s緩慢升溫至20℃。將兩個轉座子組合到PCR條管之第三個孔中。

【0153】 39. 藉由將9.6 μL混合轉座子與23.53 μL Tn5(13.6 pmol/μL)及46.87 μL偶聯緩衝液(1X)組合，將Tn5酶偶聯至轉座子混合物。

【0154】 40.在30℃溫育1小時。立即使用或在-20℃下儲存長達1個月。為了獲得最佳效能及實驗之間之一致性，我們建議在儲存前製作等分試樣。

【0155】 41.藉由將12 μL轉座酶緩衝液(5X)、來自步驟40之0.5 μL偶聯轉座子及40 ng之DNA以60 μL之總體積組合在8孔條管之一個孔中，將轉座子整合到長基因組DNA中。注意：在此步驟中可以調整該DNA量及偶聯轉座子之量。由於批次之間可能存在差異，因此有必要滴定所用Tn5酶之量。此外，從較少之DNA開始係可能的，但是為了滴定之目的，使用40 ng係有用的，使得一些材料可以在瓊脂糖凝膠上運行以決定轉座子併入之效率(參見後面之步驟)。

【0156】 42.在55℃溫育10分鐘。

【0157】 43.將40 μL轉座子併入之材料轉移到新的8孔條管之一個孔中。添加4 μL 1%SDS，在室溫下溫育10分鐘。

【0158】 44.將來自步驟43之材料裝載在0.5X TBE 1%瓊脂糖凝膠上，並在150 V下運行40分鐘。轉座之DNA應在凝膠上運行200至1,500 bp。我們通常希望看到DNA塗片之最亮部分大約為600 bp，根據選擇之定序技術，此可能會有所不同。我們通常加載執行相同步驟但缺少轉座子、Tn5酶或基因組DNA之對照。如果轉座子整合產物之尺寸看起來正確，則進行到步驟45。如果沒

有，重複上述步驟，但調整偶聯產物之濃度，直到塗片達到所需之尺寸。

【0159】 45. 用248.5 μL之1x雜交緩衝液稀釋1.5 μL步驟42之剩餘產物。

【0160】 46. 將來自步驟36之50 μL珠子(5000萬)轉移到1.5 mL微量離心管中。放置在DynaMag™-2磁鐵上2分鐘，將珠子收集到管子之側面。移除上清液並重懸於250 μL雜交緩衝液(1X)中。

【0161】 47. 在60℃下分別加熱DNA及珠子30秒。

【0162】 48. 向250 μL珠子中添加250 μL稀釋之DNA，藉由用手指輕輕敲打管底，輕輕混勻，繼續在60℃溫育10分鐘。每隔幾分鐘用手指輕輕混合管子。

【0163】 49. 放置在管旋轉器上，在「振盪」模式下在45℃之烘箱中溫育50分鐘。

【0164】 50. 藉由組合100 μL連接緩衝液、無MgCl2(10X)、2 μL之T4DNA連接酶(2x106單位/mL)及398 μL之dH2O來製備連接混合物。從旋轉器中移除管並添加連接混合物，總體積為1 mL。

【0165】 51. 在室溫下以「振盪」模式在管旋轉器上溫育1小時。

【0166】 52. 向管中添加110 μL之1%SDS，在室溫下溫育10分鐘。

【0167】 53. 放置在DynaMag™-2磁鐵上2分鐘，將珠子收集到管子之側面。移除上清液，用500 μL低鹽洗

滌緩衝液洗滌一次，用500 μL NEB2緩衝液(1X)洗滌一次。

【0168】 54. 藉由組合10 μL NEB2緩衝液(10X)、2 μL UDG(5,000 U/mL)、3 μL APE1(10,000 U/mL)、2 μL外切核酸酶1(20,000單位/mL)及83 μL dH2O，製備捕獲寡核苷酸消化混合物。移除洗滌緩衝液並將消化混合物添加到珠子中。

【0169】 55. 輕輕渦旋以重懸珠子並在37℃下溫育30分鐘。

【0170】 56. 放置在DynaMag™-2磁鐵上2分鐘，將珠子收集到管子之側面。移除上清液，用500 μL低鹽洗滌緩衝液洗滌一次，用500 μL PfuCx緩衝液(1X)洗滌一次。

【0171】 57. 藉由添加150 μL PCR混合物(2X)、4 μL PCR引子1(100 μM)、4 μL PCR引子2(100 μM)、6 μL PfuCx酶及136 μL dH2O來製備PCR主混合物。將PCR主混合物在95℃下預熱3分鐘。放置在DynaMag™-2磁鐵上2分鐘，將珠子收集到管子之側面。移除洗滌緩衝液並將PCR主混合物添加到珠子中。

【0172】 58. 在以下條件下輕輕渦旋以重懸珠子並循環PCR反應：

步驟1	72 ℃	10分鐘
步驟2	95 ℃	10秒
步驟3	58 ℃	30秒

步驟4　　72 ℃　　2分鐘

步驟6　　重複步驟2-5 10-12次

【0173】 59.PCR應該產生約500 ng DNA，在0.5X TBE 1%瓊脂糖凝膠上在150 V下運行20 ng產物40分鐘。該材料應為塗片，峰值約為500 bp。

【0174】 60. 按照製造商之方案，用300 μL Agencourt XP珠子純化PCR產物。現在，此純化之產物已準備好進入定序過程。

【0175】 單轉座子3'分支連接stLFR方案

【0176】 該方案基於DNA間隙中之單轉座子插入及新型銜接子連接方法，並且可以實現每個片段更高之覆蓋率，此對於一些定序策略(例如從頭組裝)可能係重要的。由於額外之試劑，此策略稍微昂貴一些。它還需要2.5小時之更長時間。

【0177】 61. 將藉由在8孔PCR條管之第一個孔中將10 μL轉座子1T(100 μM)、10 μL轉座子B(100 μM)、10 μL降溫貼合緩衝液(3X)組合所獲得之捕獲轉座子寡核苷酸與藉由在相同PCR條管之第二個孔中將10 μL分支T(100 μM)、10 μL分支B(100 μM)、10 μL降溫貼合緩衝液(3X)組合所獲得之間隙連接銜接子雜交。

【0178】 62. 在70℃下溫育3分鐘，然後在PCR熱循環儀上以0.1℃/ s緩慢升溫至20℃。

【0179】 63. 藉由將步驟61中之9.6 μL雜交捕獲轉座子與23.53 μL Tn5(13.6 pmol/μL)及46.87 μL偶聯緩衝液(1X)組合，將Tn5酶偶聯至轉座子。

【0180】 64. 在30℃溫育1小時。立即使用或在-20℃下儲存長達1個月。

【0181】 65. 按照步驟41-51。

【0182】 66. 放置在DynaMag™-2磁鐵上2分鐘，將珠子收集到管子之側面。移除上清液，用500 μL低鹽洗滌緩衝液洗滌一次。

【0183】 67. 藉由組合10 μL TA緩衝液(10X)、4.5 μL外切核酸酶I(20,000 U/mL)、1 μL外切核酸酶Ⅲ(100,000 U/mL)及74.5 μL dH2O來製備銜接子寡核苷酸消化混合物。移除洗滌緩衝液並將消化混合物添加到珠子中。

【0184】 68. 輕輕渦旋以重新懸浮珠子並在37℃下以「振盪」模式在管旋轉器上溫育10分鐘。

【0185】 69. 添加11 μL 1%SDS，在室溫下溫育10分鐘。

【0186】 70. 放置在DynaMag™-2磁鐵上2分鐘，將珠子收集到管子之側面。移除上清液，用500 μL低鹽洗滌緩衝液洗滌一次，用500 μL NEB2緩衝液(1X)洗滌一次。

【0187】 71. 藉由組合10 μL NEB2緩衝液(10X)、2 μL UDG(5,000 U/mL)、3 μL APE1(10,000

U/mL)及85 μL dH2O來製備捕獲寡核苷酸消化混合物。移除洗滌緩衝液並將消化混合物添加到珠子中。

【0188】 72.輕輕渦旋以重懸珠子並在37℃下溫育30分鐘。

【0189】 73.放置在DynaMag™-2磁鐵上2分鐘，將珠子收集到管子之側面。移除上清液，用500 μL高鹽洗滌緩衝液洗滌一次，用500 μL低鹽洗滌緩衝液(1X)洗滌一次。

【0190】 74.藉由組合33.4 μL 3'分支連接緩衝液(3X)、18 μL在步驟61中製備之3'分支連接銜接子(16.7 μM)、2 μL T4 DNA連接酶(2x106單位/mL)及46.6 μL之dH2O來製備3'分支連接混合物。移除洗滌緩衝液並將連接混合物添加珠子中。

【0191】 75.輕輕渦旋以重新懸浮珠子並在「振盪」模式下在25℃下在管旋轉器上溫育2小時。

【0192】 76.放置在DynaMag™-2磁鐵上2分鐘，將珠子收集到管子之側面。移除上清液，用500 μL低鹽洗滌緩衝液洗滌一次，用500 μL PCR緩衝液(1X)洗滌一次。

【0193】 77.藉由添加150 μL 2X PCR緩衝液、4 μL PCR引子1(100 μM)、4 μL PCR引子2(100 μM)、6 μL PCR酶及136 μL dH2O來製備PCR主混合物。移除洗滌緩衝液並將PCR主混合物添加到珠子中。

【0194】 78. 在以下條件下輕輕渦旋以重懸珠子並循環PCR反應：

步驟1	95	C	3分鐘
步驟2	95	C	10秒
步驟3	58	C	30秒
步驟4	72	C	2分鐘
步驟5	重複步驟2-4 10-12次		

【0195】 79. 按照上面之步驟59-60。

3.4 分析stLFR資料

【0196】 此過程之起點係FASTQ檔案。此係用於大多數定序技術生成之讀段資料之標準格式。我們用來解卷積條碼資訊之軟體採用FASTQ檔案，並期望將條碼及共同銜接子序列之42個鹼基附加到第一讀段之末尾。它將條碼讀段資料與每個條碼位置之預期1536序列相匹配。stLFR使用之條碼策略可以糾正具有單個鹼基錯配之條碼。我們軟體之最終輸出係FASTQ檔案，條碼資訊附加到讀段ID之末尾，格式為#Barcode1ID_Barcode2ID_ Barcode3ID，其中BarcodeID係0-1536之間之數字。條碼ID為零意味著它與我們使用BWA-mem27進行定位、使用GATK28進行變體調用，及使用HapCUT229進行定相來建議之任何預期條碼序列都不匹配。我們還建議使用誘餌序列定位到Hg19。

3.5 實例2之參考文獻

1 Zhang, K. 等人 Long-range polony haplotyping of individual human chromosome molecules. Nat Genet 38, 382-387 (2006).

2 Ma, L. 等人 Direct determination of molecular haplotypes by chromosome microdissection. Nat Methods 7, 299-301 (2010).

3 Kitzman, J. O. 等人 Haplotype-resolved genome sequencing of a Gujarati Indian individual. Nat Biotechnol 29, 59-63 (2011).

4 Suk, E. K.等人 A comprehensively molecular haplotype-resolved genome of a European individual. Genome Res 21, 1672–1685 (2011).

5 Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. Nat Biotechnol 29, 51-57 (2011).

6 Peters, B. A.等人 Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. Nature 487, 190-195 (2012).

7 Duitama, J.等人 Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping

techniques. Nucleic Acids Res 40, 2041-2053 (2012).

8 Selvaraj, S., J, R. D., Bansal, V. & Ren, B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat Biotechnol 31, 1111-1118 (2013).

9 Kuleshov, V. 等人 Whole-genome haplotyping using long reads and statistical methods. Nat Biotechnol 32, 261-266 (2014).

10 Amini, S. 等人 Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. Nat Genet 46, 1343-1349 (2014).

11 Zheng, G. X.等人 Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. Nat Biotechnol (2016).

12 Zhang, F.等人 Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. Nat Biotechnol 35, 852-857 (2017).

13 Peters, B. A., Liu, J. & Drmanac, R. Co-barcoded sequence reads from long DNA

fragments: a cost-effective solution for 「perfect genome」 sequencing. Frontiers in genetics 5, 466 (2014).

14 Drmanac, R. Nucleic Acid Analysis by Random Mixtures of Non-Overlapping Fragments. WO 2006/138284 A2 (2006).

15 McElwain, M. A., Zhang, R. Y., Drmanac, R. & Peters, B. A. Long Fragment Read (LFR) Technology: Cost-Effective, High-Quality Genome-Wide Molecular Haplotyping. Methods Mol Biol 1551, 191-205 (2017).

16 Schaaf, C. P.等人 Truncating mutations of MAGEL2 cause Prader-Willi phenotypes and autism. Nat Genet 45, 1405-1408 (2013).

17 Peters, B. A.等人 Detection and phasing of single base de novo mutations in biopsies from human in vitro fertilized embryos by advanced whole-genome sequencing. Genome Res 25, 426-434 (2015).

18 Ciotlos, S.等人 Whole genome sequence analysis of BT-474 using complete Genomics' standard and long fragment read technologies. Gigascience 5, 8 (2016).

19 Hellner, K. 等人 Premalignant SOX2 overexpression in the fallopian tubes of ovarian cancer patients: Discovery and validation studies. EBioMedicine 10, 137-149 (2016).

20 Mao, Q.等人 The whole genome sequences and experimentally phased haplotypes of over 100 personal genomes. Gigascience 5, 1-9 (2016).

21 Gulbahce, N. 等人 Quantitative Whole Genome Sequencing of Circulating Tumor Cells Enables Personalized Combination Therapy of Metastatic Cancer. Cancer Res 77, 4530-4541 (2017).

22 Walker, R. F.等人 Clinical and genetic analysis of a rare syndrome associated with neoteny. Genetics In Medicine (2017).

23 Mao, Q. 等人 Advanced Whole-Genome Sequencing and Analysis of Fetal Genomes from Amniotic Fluid. Clinical chemistry (2018).

24 Drmanac, R., Peters, B.A., Alexeev, A. Multiple tagging of individual long DNA fragments. WO 2014/145820 A2 (2013).

25 Picelli, S.等人 Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res 24, 2033-2040 (2014).

26 Agent Technologies, I. RecoverEase DNA Isolation Kit. Revision C.0 (2015).

27 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760 (2009).

28 McKenna, A.等人 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297-1303 (2010).

29 Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. Genome Res 27, 801-812 (2017).

【0197】 來自最近之Chromium資料集之BAM檔案「NA12878_WGS_v2_phased_possorted_bam.bam」從10X Genomics網站下載並以與stLFR文庫相同之方式處理。對於過濾結果，我們使用來自同一Chromium文庫之VCF檔案

「NA12878_WGS_v2_phased_variants.vcf.gz」。此VCF包含藉由10X Genomics優化管道處理之資料。Chromium文庫之片段尺寸係從10X Genomics網站複製的。10 Genomics使用長度加權平均值來計算片段尺寸，此可能導致比平均片段尺寸更大之尺寸。[2]讀段資料不可用，此係Zhang等人w(12)報道的。[3]來自BGISEQ-500處理之標準文庫之資料。

【0198】 表6顯示了可用於本文所述之stLFR方法之示例性序列。

實例3：3'分支連接，一種將DNA連接到DNA或RNA之3'OH末端之新方法及其應用

4.1 介紹

【0199】 此實例通常描述3'分支連接。3'分支連接用於在本文所述之stLFR實施例中添加另外之銜接子(3'分支連接銜接子)。參見例如§1.1.2。

【0200】 連接酶將核酸中之斷裂接合，此對細胞生存力及活力至關重要。DNA連接酶催化DNA末端之間磷酸二酯鍵之形成，並在活體內DNA修復、重組及複製中起關鍵作用。RNA連接酶藉由磷酸二酯鍵連接5'-磷醯基(5'PO4)及3'-羥基(3'OH)RNA末端，並參與RNA修復、剪接及編輯。來自所有三個生物界(細菌、古細菌及真核生物)之連接酶可以在活體外用作用於諸如選殖、基

於連接酶之擴增或偵測、合成生物學等之應用之重要分子工具。

【0201】 活體外最廣泛使用之連接酶之一係噬菌體T4 DNA連接酶，該連接酶係單個55-kDA多肽並且需要ATP作為能量來源。T4 DNA連接酶通常連接雙鏈DNA之相鄰5'PO4及3'OH末端。除了密封切口或連接黏性末端外，T4 DNA連接酶還可以有效地催化鈍端連接，此在所有其他DNA連接酶中都沒有觀察到。先前報道了該連接酶之一些不尋常之催化性質，例如密封雙螺旋DNA中之單鏈間隙、密封雙鏈DNA(dsDNA)中無鹼基位點附近之切口、促進部分雙鏈DNA之分子內環形成，以及連接含有3'分支延伸之DNA鏈。(Nilsson及Magnusson, Nucleic Acids Res 10:1425–1437, 1982；Goffin等人, Nucleic Acids Res 15:8755–8771, 1987；Mendel-Hartvig等人, Nucleic Acids Res. 32:e2, 2004；Western and Rose, Nucleic Acids Res., 19:809–813, 1991)。研究人員還觀察到由T4連接酶介導之模板非依賴性連接，例如dsDNA中錯配之切口密封(Alexander, 2003, Nucleic Acids Res. 2003 Jun 15；31(12):3208-16)或甚至單鏈DNA(ssDNA)連接，儘管效率非常低(H. Kuhn, 2005, FEBS J. 2005 Dec；272(23):5991-6000)。此等結果表明，對於一些非常規之T4 DNA連接酶活性，在連接會合處或附近之完美

互補鹼基配對並不係至關重要的。T4 RNA連接酶1及2分別係T4噬菌體之基因63及24之產物。它們都需要在ATP水解成AMP及PPi的情況下相鄰之5'PO4及3'OH末端以成功連接。T4 RNA連接酶1之受質包括單鏈RNA及DNA，而T4 RNA連接酶2優先密封dsRNA上之切口而不係連接ssRNA之末端。

【0202】 在本文中，我們展示了由T4 DNA連接酶介導之非常規末端連接事件，我們將該事件稱為3'-分支連接(3'BL)。它可以在切口、單鏈間隙或5'-突出端區域處連接DNA或DNA/RNA片段以形成分支結構。該報告廣泛研究了各種連接輔助因子及激活劑，並優化了此新型連接之連接條件。使用我們的3'BL方案，不需要鹼基配對，即使對於1-nt間隙，連接也可以完成90%以上。其應用之一係在NGS文庫製備中將銜接子附接到DNA或RNA。先前被認為係不可連接之幾種基因組結構現在可以成為3'BL之受質，導致輸入DNA轉化為銜接子連接分子之高轉化率，同時避免嵌合體。我們證明3'BL可以與轉座子插入結合。我們提出之定向轉座子插入策略理論上可以產生100%可用於定序微小RNA應用之模板。我們的研究證明了此新技術在NGS文庫製備中之價值，以及推進許多其他分子應用之潛力，例如RNA之3'末端之放射性標記。

4.2 3'分支連接，一種連接DNA末端之新方法

【0203】 通常，DNA連接涉及連接黏性或鈍端片段之5'PO4及3'OH DNA末端。與鈍端連接相比，黏性末端連接通常更快並且更少依賴於酶濃度。兩種過程都可以藉由噬菌體T4 DNA連接酶催化，該連接酶使用ATP作為能量產生輔助因子並且需要Mg2+。還報道了T4 DNA連接酶藉由雜交將特異性或簡並單鏈寡核苷酸連接到部分單鏈受質上。在本文中，我們展示了前所未有之T4 DNA連接酶介導之連接，它不需要互補之鹼基配對，並且可以在切口、間隙或5'突出端處將鈍端之DNA供體連接到雙鏈DNA受體之3'OH末端，形成分支結構(第21a圖)。因此，我們使用術語3'-分支連接(3'BL)來描述此等連接。我們使用之合成供體DNA含有一個鈍端雙鏈末端及一個ssDNA末端。受體受質含有以下結構之一：去磷酸化之切口、1-或8-核苷酸(nt)間隙，或36 nt之5'-突出端。T4連接酶有助於將銜接子鏈之5'PO4連接到受質鏈之唯一可連接之3'OH上，以形成叉形連接產物。

【0204】 為了優化連接效率，我們廣泛測試了影響一般連接效率之許多因素，包括銜接子::DNA受質比例、T4連接酶量、最終ATP濃度、Mg^{2+}濃度、pH、溫育時間及不同添加劑之存在，諸如聚乙二醇-8000(PEG-8000)及單鏈結合蛋白(SSB)(補充第1及2圖)。我們發現添加PEG-8000至終濃度為10%可以將連接效率從小於10%顯著增加至大於90%(第21圖)。大範圍之ATP濃度(從1uM到1mM)及Mg^{2+}濃度(3mM到

10mM)也與3'分支連接一起起作用。3'BL所需之連接酶量與鈍端連接相當。在我們的優化條件下，我們使用10比100之銜接子::受質DNA莫耳比，並在pH 7.8下用1mM ATP、10mM $MgCl_2$及10%PEG-8000在37℃下進行反應1小時。將相同之銜接子連接到鈍端受質以及無連接酶反應用作對照。為了測定連接產物之產率，在變性聚丙烯醯胺凝膠(第21b圖)或TBE凝膠(第22a-d圖)上進行反應。藉由ImageJ，將產物與受質強度之比率用於定量連接效率(第21b-c圖)。5'-突出端連接(第1b圖中之泳道11)似乎完成超過90%，甚至高於鈍端連接對照(泳道14，76.9%)，表明DNA 5'-突出端之連接效率非常高。1或8個nt間隙受質(泳道5及8)顯示出約60%之良好連接效率。然而，切口連接(泳道2)效率最低，約為20%。但是如果我們將切口連接溫育12小時，則可以提高連接產率，表明切口連接反應之動力學較慢(第22圖)。

【0205】 我們還將研究擴展到不同之銜接子及受質序列(第22圖)。三個不同銜接子(Ad-T、Ad-A或Ad-GA)之5'PO_4末端在作為連接之共有CTGCTGA序列之前含有單個T或A或二核苷酸GA。藉由在連接會合處之T，將它們分別連接到受體模板之3'OH末端。總體而言，在除切口連接外之所有情況下，使用Ad-T及Ad-A比Ad-GA觀察到更高之連接效率(70-90%)(第22圖)，表明在連接會合處T4 DNA連接酶之一些核苷酸偏好。儘管有銜接子及受質序列，5'突出端或3'分支連接總係顯示出更好

之效率(60-90%)，而切口連接在溫育1小時時效率相當低。我們假設連接效率之此等差異係由於DNA彎曲，其中切口/間隙/突出端開始並且暴露3'OH基團用於連接。較長之ssDNA區域可能使3'末端在連接中更容易獲得並導致更高之連接效率。我們還測試了類似之末端連接事件係否可能發生為5'分支連接。相反，在間隙或3'突出端處沒有觀察到鈍端銜接子與5'PO_4末端之明顯連接，表明與3'末端相比，T4 DNA連接酶在供體5'末端處可能具有更嚴格之三級結構要求。

4.3. 3'分支連接，一種將DNA連接到RNA之新方法

【0206】 我們進一步研究了在形成一個DNA及一個RNA 5'-突出端之DNA/RNA雜交體(ON21/22)上之3'BL(第23a圖)。陰性連接對照包括DNA/RNA雜交體、ssDNA或ssRNA寡核苷酸，該等寡核苷酸係單獨的或與銜接子一起溫育(第23b圖中之泳道3、4及5)。有趣地，當DNA/RNA雜交體與銜接子一起溫育時，我們看到RNA寡核苷酸之尺寸從最初之29 nt改變到49 nt，效率> 90%，此表明T4 DNA連接酶可以有效地將銜接子連接到RNA上。然而，DNA受質保持不變(第23b圖中之泳道1及2)。此表明鈍端DNA銜接子在5'-DNA突出端處連接到RNA之3'-末端，而不係在5'-RNA突出端處連接到DNA之3'-末端。為了確認3'BL需要5'突出端結構，我們進行相同之連接反應，用另一種與ON22 RNA

不互補之長DNA模板(ON23)替換原始DNA寡核苷酸(ON21)。毫不奇怪，使用ON23 DNA模板未觀察到連接，表明3'BL僅在5'突出端發生。我們的發現表明T4 DNA連接酶具有某些受質偏好，可能係由蛋白質-受質結合親和力之差異引起的。

【0207】 先前之研究表明，T4 DNA連接酶及T4 RNA連接酶2，但不係T4 RNA連接酶1，可以將5'PO_4 DNA末端連接到RNA/DNA雙鏈雜交體上並列之3'OH DNA或RNA末端，但不能連接到RNA 3'OH(Bullard 2006，Biochem J 398：135-144)。我們使用T4 RNA連接酶1及2進行相同之連接測試(第24C圖)。似乎R4 RNA連接酶1及2可以將鈍端銜接子與RNA連接，但連接效率非常低(<10%)。

4.4 定向轉座子插入文庫構建

【0208】 由於3'-分支連接已被證明可用於將銜接子高效連接到幾個基因組結構，因此我們探索了其在NGS工作流程中之應用。基於轉座子之文庫構建方法具有時間效率，並且比傳統之NGS文庫製備消耗更少之輸入DNA。然而，使用商業基於轉座子之文庫製備系統，只有一半標記分子側接有兩個不同之銜接子序列，標記之DNA側接有自身互補區域，可能形成穩定之髮夾結構，可能影響定序品質(Gorbacheva, 2015, Biotechniques Apr; 58(4): 200–202)。此外，PCR介導之銜接子序列併入

不適用於全基因組亞硫酸氫鹽定序，也不適用於無PCR之NGS文庫構建。

【0209】 為了克服此等局限性，我們開發了一種新的基於轉座子之NGS文庫構建方案，該方案包含3'BL。Tn5及MuA轉座子均藉由「剪切及黏貼」機制起作用，其中轉座子銜接子序列與靶DNA之5'末端來末端連接，在基因組DNA之3'末端分別產生9 bp或5 bp之間隙(第24圖)。然後，使用3'BL將另一個銜接子序列添加到間隙處之基因組DNA之3'末端以完成定向銜接子連接。我們比較了3'BL方法與雙轉座子插入方法之效率，後者使用兩種不同之基於Tn5之銜接子TnA及TnB。將人基因組DNA與單獨之TnA轉座體複合物或與等莫耳量之TnA及TnB轉座體複合物一起溫育。TnA轉座體片段化之產物進一步用於具有鈍端銜接子AdB之3'BL，該AdB與TnB共有共同之銜接子序列。使用分別針對TnA及AdB/TnB銜接子設計之兩種引子Pr-A及Pr-B進行之PCR擴增顯示出相似之PCR產率(第4b圖，泳道9及10以及第4c圖)，表明此等兩種方法效率相同。當僅使用一種對TnA或AdB/TnB銜接子特異之引子時，未觀察到顯著之擴增(第4b圖及第4c圖)。正如預期，與具有單獨僅TnA或TnB轉座體複合物之轉座子插入反應相比，由於PCR抑制，3'-連接方法及雙轉座子插入方法顯示出顯著更高之PCR效率(第24b圖，泳道3及泳道8，及第24c圖)。

4.5 材料及方法

用於雙鏈DNA之3'-分支連接

【0210】用於3'BL之受質由在pH 8 Tris-EDTA(TE)緩衝液(Life Technologies)中與各4 pmol之一種或兩種另外之寡核苷酸混合之2 pmol ON1或ON9組成。受質1及5(切口):ON1/2/3及ON9/10/11;受質2及6(1 bp間隙):ON1/2/4及ON9/10/12;受質3(8 bp間隙):ON1/4/5;受質4及9(5'-突出端):ON1/2及ON9/10;受質7(2 bp間隙):ON9/10/13;受質8(3 bp間隙):ON9/10/14;鈍端對照:ON1/6(第1圖,補充表1)。使用3'BL緩衝液[0.05 mg/ml BSA (New England Biolabs)、50 mM Tris-Cl pH7.8 (Amresco)、10 mM MgCl2 (EMD Millipore)、0.5 mM DTT (VWR Scientific), 10% PEG-8000 (Sigma Aldrich)及1 mM ATP (Sigma Aldrich)]中之2400單位之T4連接酶(Enzymatics Inc),將模板連接至180 pmol之銜接子(Ad-C:ON7/8,Ad-T:ON15/16,Ad-A:ON17/8或Ad-GA:ON19/20)。藉由將ATP濃度從1uM改變為1mM、Mg^{2+}濃度、pH值、溫度從12至42℃及添加劑如PEG-8000從2.5%至10%及SSB從2.5至20ng/ul進行優化測試。在冰上製備連接混合物並在37℃下溫育1至12小時,然後在65℃下加熱滅活15分鐘。使用Axygen

珠子(Corning)純化樣品，並溶離到40 μL TE緩衝液中。所有連接反應均在6%TBE或變性聚丙烯醯胺凝膠(Life Technologies)上進行，並在Alpha Imager(Alpha Innotech)上顯現。輸入對照以用於連接之模板之相等或一半之量加載。藉由使用ImageJ軟體(NIH)將連接產物之強度除以連接及未連接產物之總強度來估計連接效率。

用於DNA/RNA雜交體之3'-分支連接

【0211】用於3'BL之受質由與2 pmol ON21或ON23 DNA寡核苷酸混合之10 pmol ON22 RNA寡核苷酸組成。對於T4 DNA連接酶介導之3'BL，如上所述將受質與3'BL緩衝液中之Ad-T(ON15/16)一起溫育，並在37℃溫育1小時。使用T4 RNA連接酶1或2之3'BL在其自身之1x RNA連接酶緩衝液(NEB)與20%DMSO中進行。在6%變性聚丙烯醯胺凝膠上測定所有連接產物。

定向轉座子插入文庫構建

【0212】本實驗中使用之轉座子寡核苷酸由Sangon Biotech合成。對於使用TnA及TnB之2個轉座子實驗，TnA、TnB及MErev寡核苷酸以1:1:2之比例降溫貼合。對於tn1之單轉座子實驗，tn1及MErev以1:1之比例降溫貼合。

【0213】 藉由混合15 pmol預降溫貼合之銜接子、7 ul Tn5轉座酶(Vazemy)及5.5 ul甘油進行轉座體組裝，以獲得20ul反應物，將該反應物在30℃溫育1小時。基因組DNA(Coriell 19240)之轉座子插入在含有100 ng gDNA、TAG緩衝液(Vazyme)及2 ul組裝之轉座體之20 ul反應中進行。將反應物在55℃溫育10分鐘，然後添加100μl PB緩衝液(Qiagen)以從標記之DNA中移除轉座體複合物並使用Agencourt AMPure XP珠子(Beckman Coulter)純化。在含有100 pmol銜接子、600 U之T4 DNA連接酶(Enzymatics Inc.)及3'BL緩衝液之反應中，將AdB(ONB1，ONB2)3'分支連接到經標記之DNA上，在25℃下溫育1小時。使用AMPure XP珠子來純化反應。標記及間隙連接DNA之PCR擴增在含有2 ul標記或間隙連接DNA、TAB緩衝液、1 μl TruePrep擴增酶(Vazyme)、200 mM dNTP(Enzymatics Inc.)及各自400 mM之引子Pr-A及Pr-B之50 ul反應中進行。標記之反應在72℃下進行3分鐘；98℃30秒；98℃10秒，58℃30秒，72℃2分鐘，8個循環；及72℃，延長10分鐘。使用相同之程式進行間隙連接反應，沒有在72℃下之最初3分鐘延伸。使用AMPure XP珠子以單步尺寸選擇或藉由雙重分級來純化PCR反應。使用Qubit High-Sensitivity DNA套組(Invitrogen)定量純化之產物。

實例4：3'分支連接：將非互補DNA連接到DNA或RNA中之凹入或內部3'OH末端之新方法

【0214】 核酸連接酶係在合成、修復及重組過程中修復DNA或RNA斷裂之關鍵酶。已經使用DNA/RNA連接酶之多種活性開發了各種分子工具。然而，還有待發現其他連接酶活性。在本文中，我們證明了T4 DNA連接酶在3'隱性末端、間隙或切口處連接5'磷酸化之鈍端雙鏈DNA至DNA斷裂以形成3'分支結構之非常規能力。因此，此不依賴鹼基配對之連接被稱為3'分支連接(3'BL)。在最佳連接條件之廣泛研究中，類似於鈍端連接，在連接緩衝液中存在10%PEG-8000顯著提高了連接效率。使用不同之合成DNA在會合位點處觀察到一些核苷酸偏好，此表明3'BL之連接偏差水準。此外，我們發現T4 DNA連接酶有效地將DNA連接到DNA/RNA雜交體中之RNA之3'末端，而RNA連接酶在該反應中效率較低。T4 DNA連接酶之此等新特性可用作許多重要應用中之廣泛分子技術。我們對用於下一代定序(next generation sequencing; NGS)文庫構建之新的定向標記方案進行了概念驗證研究，該方案消除了倒位銜接子並允許樣品條碼插入基因組DNA附近。單轉座子標記後3'BL理論上可以達到100%可用之模板，我們的經驗資料表明，與傳統之雙轉座子或Y轉座子標記相比，新方法產生更高之產量。我們進一步探索3'BL用於製備靶向

RNA NGS文庫之潛在用途，該用途具有減輕之基於結構之偏差及銜接子二聚體問題。

5.1 介紹

【0215】 連接酶修復核酸中之斷裂，此對細胞生存力及活力至關重要。DNA連接酶催化DNA末端之間磷酸二酯鍵之形成，並在活體內DNA修復、重組及複製中起關鍵作用1-3。RNA連接酶藉由磷酸二酯鍵連接5'-磷醯基(5'PO4)及3'-羥基(3'OH)RNA末端，並參與RNA修復、剪接及編輯4。來自所有三個生物界(細菌、古細菌及真核生物)之連接酶可以在活體外用作用於諸如選殖、基於連接酶之擴增或偵測及合成生物學之應用之重要分子工具5–7。

【0216】 活體外最廣泛使用之連接酶之一係噬菌體T4 DNA連接酶，它係一種需要ATP作為能量來源之單一55-kDA多肽8。T4 DNA連接酶通常連接雙鏈DNA之相鄰5'PO4及3'OH末端。除了密封切口及連接黏性末端外，T4 DNA連接酶還可以有效地催化鈍端連接，此在任何其他DNA連接酶中都沒有觀察到9,10。先前報道了該連接酶之一些不尋常之催化性質，例如密封雙螺旋DNA中之單鏈間隙、密封雙鏈DNA(dsDNA)中無鹼基位點附近之切口、促進部分雙鏈DNA之分子內環形成，以及連接含有3'分支延伸之DNA鏈11-13。研究人員還觀察到由T4連接酶介導之模板非依賴性連接，例如

dsDNA14中之錯配切口密封或甚至單鏈DNA(ssDNA)連接，儘管效率非常低15。此等結果表明，對於一些非常規之T4 DNA連接酶活性，在連接會合處或附近之完美互補鹼基配對並不係至關重要的。T4 RNA連接酶1及2分別係T4噬菌體之基因63及24之產物。兩者都需要相鄰之5'PO4及3'OH末端以成功連接，同時將ATP水解為AMP及PPi。T4 RNA連接酶1之受質包括單鏈RNA及DNA，而T4 RNA連接酶2優先密封dsRNA上之切口而不係連接ssRNA之末端16,17。

【0217】 在本文中，我們證明了由T4 DNA連接酶介導之非常規末端連接事件，我們稱之為3'-分支連接(3'BL)。該方法可以在切口、單鏈間隙或3'隱性末端處連接DNA或DNA/RNA片段以形成分支結構。該報告包括對各種連接輔助因子及激活劑以及此新型連接之連接條件之優化之廣泛研究。使用我們的3'BL方案，不需要鹼基配對，並且在大多數情況下連接可以達到70-90%完成，包括1-nt間隙。該方法之一個應用係在NGS文庫製備期間將銜接子連接到DNA或RNA。先前被認為係不可連接之幾種基因組結構現在可用作3'BL之受質，導致輸入DNA轉化為銜接子連接之分子之高轉化率，同時避免嵌合體。我們證明3'BL可與轉座子標記結合以增加文庫產量。我們提出之定向標記策略理論上將產生模板，該模板100%可用於定序。我們的研究證明了此新技術在NGS文庫製備中之價值以及推動許多其他分子應用之潛力。

5.2 結果：3'分支連接，一種連接DNA末端之新方法

【0218】 通常，DNA連接涉及連接黏性或鈍端片段之5'PO4及3'OH DNA末端。與鈍端連接相比，黏性末端連接通常更快並且更少依賴於酶濃度。兩種過程都可以藉由噬菌體T4 DNA連接酶催化，該連接酶使用ATP作為能量產生輔因子並且需要Mg2+ 8。還報道了T4 DNA連接酶藉由雜交將特異性或簡並單鏈寡核苷酸連接到部分單鏈受質18,19。在本文中，我們證明了非常規之T4 DNA連接酶介導之連接，該連接不需要互補之鹼基配對，並且可以將鈍端DNA供體連接到3'凹入鏈、間隙或切口處之雙鏈DNA受體之3'OH末端(第26a圖)。因此，我們使用術語3'-分支連接(3'BL)來描述此等連接。我們使用之合成供體DNA含有5'鈍端雙鏈末端及3'ssDNA末端。受體受質含有以下結構之一：去磷酸化之切口、1-或8-核苷酸(nt)間隙，或3'36-nt凹入端(補充表1)。T4連接酶有助於將供體鏈之5'PO_4連接到受體鏈之唯一可連接之3'OH，以形成分支形狀之連接產物。

【0219】 為了優化連接效率，我們廣泛測試了影響一般連接效率之許多因素，包括銜接子:DNA受質比例、T4連接酶數量、最終ATP濃度、Mg2+濃度、pH、溫育時間及不同之添加劑，諸如聚乙二醇-8000(PEG-8000)及單鏈結合(single-stranded binding; SSB)蛋白。添加PEG-8000至終濃度為10%，將連接效率從小於

10%顯著提高到大於80%(第26及27圖)。大範圍之ATP濃度(從1 μM到1 mM)及Mg2+濃度(3 mM到10 mM)與3'BL相容。3'BL所需之連接酶量與鈍端連接相當。在我們的優化條件下,我們使用30比100之供體:受質DNA莫耳比,並且我們使用1 mM ATP、10 mM MgC12及10%PEG-8000在pH 7.8下在37℃下進行反應1小時。將相同供體與鈍端受質連接及無連接酶反應分別用作陽性及陰性對照。

【0220】 連接供體(Ad-G)在一端係雙鏈的(5'磷酸化及3'雙脫氧保護)及在另一端係單鏈的(3'雙脫氧保護)(第26圖)。連接受質由具有不同頂部鏈之相同底部鏈(ON1)組成,以構成切口、間隙及突出端結構。為了量化連接產物產率,將反應產物在6%變性聚丙烯醯胺凝膠上分離(第26b圖)。使用ImageJ將連接效率計算為產物與受質強度之比率(第26b-c圖)。3'-隱性連接(第26b圖中之泳道11)顯示約90%完成,其甚至高於鈍端連接對照(泳道14,72.74%)並且表明具有3'-隱性DNA末端之非常高之連接效率。1-或8-nt間隙受質(泳道5及8)顯示出約45%之良好連接效率。切口連接(泳道2)效率最低,約為13%。然而,當切口連接反應溫育更長時間時,此連接產率得到改良,表明切口連接反應之動力學較慢。

【0221】 我們還將研究擴展到不同之銜接子及受質序列(第27圖)。三個不同銜接子之5'PO4末端(補充表1中之Ad-T、Ad-A或Ad-GA)在共有CTGCTGA序列之前

在連接會合處含有單個T或A或二核苷酸GA。藉由在連接會合處之T，將此等5'PO4末端分別連接到受體模板之3'OH末端。總體而言，在除了切口連接或3'BL之外的大多數情況下，使用Ad-GA觀察到高連接效率(70-90%)(第27f圖)，因此表明在連接會合處T4 DNA連接酶之一些核苷酸偏好。獨立於銜接子及受質序列，3'-隱性末端或間隙連接總係顯示出更好之效率(60-90%)，而切口連接在1小時溫育中效率相當低。我們假設連接效率之此等差異係由於DNA彎曲，其中切口/間隙/突出端開始並且暴露3'OH基團用於連接。較長之ssDNA區域可能使3'末端在連接中更容易獲得並導致更高之連接效率。我們還測試了類似之末端連接事件係否可能作為5'分支連接來發生。與3'BL相比，沒有觀察到鈍端銜接子在間隙或5'-隱性末端處與5'PO4末端之明顯連接。該結果表明，與3'末端相比，供體5'末端之T4 DNA連接酶之空間位阻更大。

5.3： 3'分支連接以將DNA連接到RNA

【0222】 我們進一步研究了在形成一個DNA及一個RNA 5'-突出端之DNA/RNA雜交體(表3中之ON-21/ON-23)上之3'BL(第28a圖)。DNA/DNA雜交體上之連接用作陽性對照，而陰性連接對照包括單獨或與銜接子溫育之DNA/RNA雜交體、ssDNA或ssRNA寡核苷酸(第28c及圖中之泳道3、4及5)。有趣地，當

DNA/RNA雜交體與鈍端dsDNA供體一起溫育時，我們觀察到連接後RNA寡核苷酸之尺寸從原來29 nt變化到49 nt。然而，DNA受質保持不變(第28c圖中之泳道1及2)。該結果表明鈍端dsDNA供體在3'-隱性DNA末端連接到RNA之3'-末端，但不在3'-隱性RNA末端連接到DNA之3'-末端。作為陽性對照，每側具有3'隱性末端之DNA/DNA雜交體顯示兩條鏈上之較大種類之條帶移位，效率接近100%。為了確認3'BL需要3'隱性結構，我們進行相同之連接反應，同時用另一個與ON-22 RNA不相關之長DNA模板(ON-23)替換原始DNA寡核苷酸(ON-21)(第28b圖)。不出所料，使用ON-23 DNA模板未觀察到連接(第28c圖中之泳道10-13)。我們的發現表明，T4 DNA連接酶可以促進DNA/RNA雜交體上之3'BL，並且該活性具有某些空間受質偏好，此等偏好可能受到T4 DNA連接酶-受質結合親和力差異之影響。

【0223】 先前之研究報道，當互補鏈係RNA但不係DNA時，為了密封DNA/RNA雜交中之切口，T4 DNA連接酶及T4 RNA連接酶2，而不係T4 RNA連接酶1，可以有效地將5'PO4 DNA末端連接到並列之3'OH DNA或RNA末端17。因此，我們使用T4 RNA連接酶1及2在20%DMSO(第28d圖)或10%PEG中進行相同之連接測試。在兩個測試中，T4 RNA連接酶1及T4 RNA連接酶2將鈍端銜接子輕微連接到DNA/RNA雜交體中RNA之3'末端。值得注意地，在僅含RNA之對照中，T4

RNA連接酶2可以將鈍端dsDNA銜接子連接到ssRNA。總之，T4 DNA連接酶，但不係T4 RNA連接酶，能夠藉由3'BL有效地將鈍端dsDNA連接到RNA之3'末端。

5.4 定向標記文庫構造

【0224】 因為3'BL可用於將銜接子高效連接到幾個基因組結構，我們探索了它在NGS工作流程中之應用。與常規NGS文庫製備相比，基於轉座子之文庫構建快速且消耗較少之輸入DNA。然而，使用商業基於轉座子之文庫製備系統，只有一半標記分子側接有兩個不同之銜接子序列(第29a圖)，標記之DNA側接有自身互補區域，可形成穩定之髮夾結構並影響定序品質20。此外，PCR介導之銜接子序列併入尚未適用於全基因組亞硫酸氫鹽定序或無PCR之NGS文庫構建。

【0225】 為了克服此等限制，我們藉由結合3'BL開發了基於轉座子之NGS文庫構建之新方案。Tn5及MuA轉座子都藉由「剪切及黏貼」機制起作用，其中轉座子銜接子序列與靶DNA之5'末端來末端連接，在基因組DNA之3'末端分別產生9-bp或5-bp間隙(第29a圖)。隨後，3'BL可用於在間隙處將另一銜接子序列添加到基因組DNA之3'末端以完成定向銜接子連接(第29c圖)。我們在該手稿中使用Tn5轉座子來比較單標記+ 3'BL方法(第29c圖)與使用兩種不同之基於Tn5之銜接子TnA及

TnB之雙標記方法(第29a圖),以及使用包含兩個不同銜接子序列之Y銜接子之另一個定向單標記策略(第29b圖)之效率。將人基因組DNA與等莫耳量之TnA及TnB轉座體複合物、與單獨TnA轉座體複合物,或與TnY(TnA/B)轉座體複合物一起溫育。

【0226】 僅TnA轉座體片段化之產物進一步用作具有鈍端銜接子AdB之3'BL之模板,其與TnB共有共同之銜接子序列。使用兩種引子Pr-A及Pr-B進行PCR擴增,該等引子分別設計用於識別TnA及AdB/TnB銜接子。定量資料表明,與TnA及TnB及TnY(TnA/B)相比,TnA及AdB具有最高效率(第29d圖)。當僅使用一種對TnA銜接子特異之引子時,未觀察到顯著之擴增(第29d圖)。如所預期的,與僅用單獨TnA或TnB轉座體複合物之標記反應相比,由於PCR抑制,TnA-3'BL方法、雙標記方法及TnY方法均顯示出顯著更高之PCR效率(第29d圖)。

【0227】 我們還使用BGISEQ-500對此等文庫進行定序,並比較轉座子干擾末端、3'BL末端及常規TA連接末端之間之鹼基位置偏差(第30圖)。顯然,3'BL末端之位置偏差小於Tn5末端之位置偏差(第30a-b圖),此係因為3'BL末端受轉座子中斷及3'BL兩者之影響而發生。因為只有3'BL末端之前6個核苷酸(位置1-6)顯示鹼基偏差,並且偏差與其雜交之Tn5末端相似但不完全相同(位置30-35,在9-nt突出端之後),我們得出結論,我們在

3'BL末端觀察到之位置偏差主要係由Tn5轉座子引起的。因此，3'BL引起最小偏差並且類似於常規TA連接(第30c圖)。

5.5 討論

T4 DNA連接酶之一個重要特性係其有效連接鈍端dsDNA21,22，此在其他DNA連接酶中未被觀察到。據報道，該連接酶也介導一些不尋常之催化事件，例如連接雙鏈DNA中之單鏈間隙或錯配鹼基11,12，從部分雙鏈DNA形成莖環分子13，或以不依賴於模板之方式低效連接ssDNA 20。

【0228】 在本文中，我們證明了T4 DNA連接酶催化鈍端dsDNA與具有切口之dsDNA之3'OH末端之連接，及具有間隙或5'突出端之部分單鏈雙螺旋DNA之連接。相反，沒有觀察到在5'凹入末端或間隙中5'PO4末端之連接，此表明在與dsDNA銜接子之5'PO4末端結合後，T4 DNA連接酶可以在DNA彎曲時接近凹入之3'末端。使用我們的3'BL方法，不需要鹼基配對，即使對於1-nt間隙，使用優化條件也可完成大於70%之完成。然而，觀察到將5'T、A或GA連接至3'T之不同連接效率(第2圖)，此表明在連接會合處有一些序列偏好。儘管識別出連接偏差23，但在NGS文庫製備期間，T4 DNA連接酶通常用於銜接子添加步驟。由於其能夠進行3'BL，T4連接酶可以將銜接子連接到先前被認為係不可連接之幾

種基因組結構，從而導致更高之模板使用率。3'BL也可以與轉座子標記結合。傳統之雙轉座子策略僅有50%之標記分子適合於隨後之擴增步驟。然而，當使用一個轉座子及隨後之3'BL進行DNA標記時，可以獲得在每個插入片段末端具有不同銜接子之分子之增加之產量(第4圖)。此外，標記之3'BL產物可以作為無PCR之WGS文庫直接加載到Illumina之流動池上，此係使用雙轉座子策略難以實現的。

【0229】 已經提出了使用由兩個不同銜接子序列組成之Y轉座子或用第二銜接子寡核苷酸替換來自單一轉座子之未連接之鏈，然後進行間隙填充及連接的其他定向轉座子方案24。然而，此等方法繼續保留倒位銜接子序列，並且不能像經標記之3'BL方案那樣插入與基因組DNA相鄰之樣品條碼。基於NGS資料，3'BL連接之基因組末端也顯示較少之位置具有位置鹼基組成偏差，並且第一個6-nt偏差係輕微的並且主要由轉座子中斷引起，表明3'BL具有最小之位置偏差。使用此新的文庫構建方法，Wang等人成功實現了WGS中高度準確及完整之變體調用，並且對於長片段讀取，實現近似完美地將變體定相到長重疊片段，其中N50尺寸高達23.4 Mb(BioRxiv，https://doi.org/10.1101/324392)。

【0230】 在該研究中，我們還使用嵌合DNA/RNA雙鏈體之模板研究3'BL，該雙鏈體形成5'DNA及5'RNA突出端(第3圖)。出乎意料地，鈍端dsDNA有效連接到

RNA之3'末端，但不係DNA，此表明T4連接酶具有三元複合物形成偏好。如果使用T4 RNA連接酶I或II連接末端，則連接效率大大降低。3'BL與T4 DNA連接酶之另一個初步但重要之應用係mRNA之富集或靶向RNA文庫之構建，尤其對於miRNA，此較小調節RNA之不受控制之表達導致許多疾病25,26。因此，我們的3'BL技術可以很容易地應用於使用miRNA偵測癌症及阿爾茨海默病。與靶向Poly(A)尾部或特定miRNA序列之DNA探針雜交可用於產生具有DNA 5'-突出端之DNA-RNA雜交體，然後藉由3'BL連接到具有樣品及/或UID條碼之銜接子序列。然後可以逆轉錄此等共同序列以產生靶RNA序列之cDNA。與目前之miRNA捕獲技術相比，T4 DNA連接酶介導之3'BL之使用可能為NGS RNA文庫構建提供若干優勢。首先，與DNA鏈之雜交將阻止RNA鏈形成二級結構，因此減輕了其他方案引入之偏差。其次，T4 DNA連接酶藉由3'BL來實現高效之銜接子添加，此避免了RNA連接酶可以促進之分子內RNA相互作用。第三，可以有效地消除銜接子二聚體，可能不需要進行不希望之凝膠純化。此新方法可以藉由簡單且可擴展之工作流程來產生經改良之無偏差之微小RNA表達譜，因此，大規模研究將變得更加實惠。

【0231】 該研究之發現增加了對T4 DNA連接酶活性之日益瞭解。我們設想3'分支連接成為分子生物學之一般

工具，將推動新的DNA工程方法之發展超越所描述之NGS應用。

5.6 材料及方法

用於雙鏈DNA之3'-分支連接

【0232】用於3'BL之受質由在pH 8 Tris-EDTA(TE)緩衝液(Life Technologies)中與各4 pmol之一種或兩種另外之寡核苷酸混合之2 pmol ON1或ON9組成，如下所述：受質1及5(切口)，ON-1/2/3及ON-9/10/11；受質2及6(1-nt間隙)，ON1/2/4及ON9/10/12；受質3(8-nt間隙)，ON1/4/5；受質4及9(5'突出端)，ON1/2及ON9/10；受質7(2-nt間隙)，ON9/10/13；受質8(3-nt間隙)，ON9/10/14；鈍端對照，ON1及ON6(第26圖，補充表1)。使用3'BL緩衝液[0.05 mg/ml BSA (New England Biolabs)、50 mM Tris-Cl pH 7.8 (Amresco)、10 mM MgCl2 (EMD Millipore)、0.5 mM DTT (VWR Scientific), 10% PEG-8000 (Sigma Aldrich)及1 mM ATP (Sigma Aldrich)]中之2,400單位之T4連接酶(Enzymatics Inc)，將模板連接至180 pmol之銜接子(Ad-G：ON7/8，Ad-T：ON15/16，Ad-A：ON17/8或Ad-GA：ON19/20)。優化測試藉由將ATP濃度從1 μM改變為1 mM，Mg2+濃度從3到10 mM，pH值從3到9，溫度從12到42℃，

以及將諸如PEG-8000之添加劑從2.5%調整到10%，及SSB從2.5到20 ng/μL來進行。在冰上製備連接混合物並在37℃下溫育1至12小時，然後在65℃下加熱滅活15分鐘。使用Axygen珠子(Corning)純化樣品，並溶離到40 μL TE緩衝液中。所有連接反應均在6%TBE或變性聚丙烯醯胺凝膠(Life Technologies)上進行，並在Alpha Imager(Alpha Innotech)上顯現。輸入對照以用於連接之模板之相等或一半之量加載。藉由使用ImageJ軟體(NIH)將連接產物之強度除以連接及未連接產物之總強度來估計連接效率。

用於DNA/RNA雜交體之3'-分支連接

【0233】 用於3'BL之受質由與2 pmol ON-21或ON-23 DNA寡核苷酸混合之10 pmol ON-21 RNA寡核苷酸組成。對於T4 DNA連接酶介導之3'BL，如上所述將受質與3'BL緩衝液中之Ad-T(ON15/16)一起溫育，並在37℃溫育1小時。使用T4 RNA連接酶1或2之3'BL在具有20%DMSO或25%PEG之1x RNA連接酶緩衝液(NEB)中進行。在6%變性聚丙烯醯胺凝膠上測定所有連接產物。

定向標記文庫構建

【0234】 本實驗中使用之轉座子寡核苷酸由Sangon Biotech合成。對於使用TnA/TnB之2個轉座子實驗，

將TnA(ON24)、TnB(ON25)及MErev(ON26)之寡核苷酸以1:1:2之比例降溫貼合。對於使用TnA之單轉座子實驗，ON24及ON26以1:1之比例降溫貼合。對於Y(TnA及TnB)轉座子實驗，ON24及ON27以1:1之比例降溫貼合。

【0235】 藉由混合100 pmol預降溫貼合之銜接子，7 μL之Tn5轉座酶及足夠之甘油進行轉座子組裝，以獲得總共20 μL之反應物，將該反應物在30℃下溫育1小時。基因組DNA(Coriell 12878)之標記在含有100 ng gDNA、TAG緩衝液(自製)及1 μL組裝之轉座子之20-μL反應中進行。將反應在55℃溫育10分鐘；然後添加40 μL之6 M鹽酸胍(Sigma)以從標記之DNA中移除轉座子複合物，並使用Agencourt AMPure XP珠子(Beckman Coulter)純化DNA。在含有100 pmol銜接子、600 U T4 DNA連接酶(Enzymatics Inc.)及3'BL緩衝液之反應中，在25℃下進行AdB(ON28及ON29)與標記DNA之間隙連接1小時。使用AMPure XP珠子純化反應。標記及間隙連接之DNA之PCR擴增在含有2 μL標記或間隙連接之DNA、TAB緩衝液、1 μL TruePrep擴增酶(Vazyme)、200 mM dNTP(Enzymatics Inc.)及各自400 mM之引子Pr-A及Pr-B之50 μl反應中進行。將標記之反應溫育如下：72℃3分鐘；98℃30秒；8個循環之98℃10秒、58℃30秒及72℃2分鐘；及72℃，延長10分鐘。使用相

同之程式進行間隙連接反應，沒有在72℃下之最初3分鐘延伸。使用AMPure XP珠子純化使用prA(ON30)或prA及prB(ON31)兩者之PCR反應。使用Qubit High-Sensitivity DNA套組(Invitrogen)定量純化之產物。

5.6 實例4之參考文獻

1. Lehnman, I. R. DNA ligase: structure, mechanism, and function. Science (80-.). 186, 790–797 (1974).

2. Tomkinson, A. E. & Mackey, Z. B. Structure and function of mammalian DNA ligases. Mutat. Res. Repair 407, 1–9 (1998).

3. Timson, D. J., Singleton, M. R. & Wigley, D. B. DNA ligases in the repair and replication of DNA. Mutat. Res. Repair 460, 301–318 (2000).

4. Ho, C. K., Wang, L. K., Lima, C. D. & Shuman, S. Structure and mechanism of RNA ligase. Structure 12, 327–339 (2004).

5. Tomkinson, A. E., Vijayakumar, S., Pascal, J. M. & Ellenberger, T. DNA ligases: structure, reaction mechanism, and function. Chem. Rev. 106, 687–699 (2006).

6. Pascal, J. M. DNA and RNA ligases: structural variations and shared mechanisms. Curr. Opin. Struct. Biol. 18, 96–105 (2008).

7. Shuman, S. DNA ligases: progress and prospects. J. Biol. Chem. 284, 17365–17369 (2009).

8. Dickson, K. S., Burns, C. M. & Richardson, J. P. Determination of the free-energy change for repair of a DNA phosphodiester bond. J. Biol. Chem. 275, 15828–15831 (2000).

9. Cai, L., Hu, C., Shen, S., Wang, W. & Huang, W. Characterization of bacteriophage T3 DNA ligase. J. Biochem. 135, 397–403 (2004).

10. Ampligase® Thermostable DNA Ligase. Available at: http://www.epibio.com/enzymes/ligases-kinases-phosphatases/dna-ligases/ampligase-thermostable-dna-ligase?details.

11. Nilsson, S. V & Magnusson, G. Sealing of gaps in duplex DNA by T4 DNA ligase. Nucleic Acids Res. 10, 1425–1437 (1982).

12. Goffin, C., Bailly, V. & Verly, W. G. Nicks 3′ or 5′ to AP sites or to mispaired bases, and one-nucleotide gaps can be sealed by T4

DNA ligase. Nucleic Acids Res. 15, 8755–8771 (1987).

13. Mendel-Hartvig, M., Kumar, A. & Landegren, U. Ligase-mediated construction of branched DNA strands: a novel DNA joining activity catalyzed by T4 DNA ligase. Nucleic Acids Res. 32, e2–e2 (2004).

14. Alexander, R. C., Johnson, A. K., Thorpe, J. A., Gevedon, T. & Testa, S. M. Canonical nucleosides can be utilized by T4 DNA ligase as universal template bases at ligation junctions. Nucleic Acids Res. 31, 3208–3216 (2003).

15. Kuhn, H. & Frank-Kamenetskii, M. D. Template-independent ligation of single-stranded DNA by T4 DNA ligase. FEBS J. 272, 5991–6000 (2005).

16. Ho, C. K. & Shuman, S. Bacteriophage T4 RNA ligase 2 (gp24.1) exemplifies a family of RNA ligases found in all phylogenetic domains. Proc. Natl. Acad. Sci. 99, 12709–12714 (2002).

17. Bullard, D. R. & Bowater, R. P. Direct comparison of nick-joining activity of the

nucleic acid ligases from bacteriophage T4. Biochem. J. 398, 135–144 (2006).

18. Broude, N. E., Sano, T., Smith, C. L. & Cantor, C. R. Enhanced DNA sequencing by hybridization. Proc. Natl. Acad. Sci. 91, 3072–3076 (1994).

19. Gunderson, K. L.等人 Mutation detection by ligation to complete n-mer DNA arrays. Genome Res. 8, 1142–1153 (1998).

20. Gorbacheva, T., Quispe-Tintaya, W., Popov, V. N., Vijg, J. & Maslov, A. Y. Improved transposon-based library preparation for the Ion Torrent platform. Biotechniques 58, 200 (2015).

21. Sgaramella, V. & Khorana, H. G. CXII. Total synthesis of the structural gene for an alanine transfer RNA from yeast. Enzymic joining of the chemically synthesized polydeoxynucleotides to form the DNA duplex representing nucleotide sequence 1 to 20. J. Mol. Biol. 72, 427–444 (1972).

22. SGARAMELLA, V. & EHRLICH, S. D. Use of the T4 Polynucleotide Ligase in The Joining of Flush-Ended DNA Segments Generated by

Restriction Endonucleases. FEBS J. 86, 531–537 (1978).

23. Seguin-Orlando, A.等人 Ligation bias in illumina next-generation DNA libraries: implications for sequencing ancient genomes. PLoS One 8, e78575 (2013).

24. Goryshin, I., Baas, B., Vaidyanathan, R. & Maffitt, M. Oligonucleotide replacement for di-tagged and directional libraries. (2016).

25. Bushati, N. & Cohen, S. M. microRNA functions. Annu. Rev. Cell Dev. Biol. 23, 175–205 (2007).

26. Mallory, A. C. & Vaucheret, H. Functions of microRNAs and related small RNAs in plants. Nat. Genet. 38, S31 (2006).

* * *

【0236】 雖然本發明參照特定態樣及實施例來揭示，但是顯而易知本發明之其他實施例及變化可由熟習此項技術者設計而不脫離本發明之真實精神及範圍。

【0237】 出於在美利堅合眾國之所有目的，本發明中引用之每個公開案及專利文獻以引用方式併入本文，如同每個此公開案或文獻被具體及單獨地指出以引用方式併入

本文。公開案及專利文獻之引用並不意味著任何此類文獻係相關之先前技術，也不構成對其內容或日期之承認。

表1A：定相及變體調用統計

		stLFR-1			stLFR-2			stLFR-3	stLFR-4
定序統計	定序之鹼基對(Gb)	336	230	100	660	200	100	117	126
	輸入基因組DNA(ng)	1	1	1	1	1	1	10	10
	平均基因組片段尺寸(kb)	66.2	66.3	66.4	52.5	52.7	52.6	30.2	46.8
	獨特基因組覆蓋	44X	38X	24X	58X	37X	23X	37X	34X
	重複率	59.4%	49.6%	29.4%	70.88%	41.05%	25.37%	5.4%	15.0%
	讀段長度	PE100	PE100	PE100	PE100	PE100	PE100	PE100	PE100
	獨特隔室	10,186,086	10,007,746	9,427,999	11,823,872	10,932,966	10,297,180	30,544,841	10,577,590
	每個隔室之平均片段	1.18	1.18	1.17	1.25	1.23	1.22	2.87	6.84
	每個片段之平均共條碼讀段	80.7	71.5	47.4	88.3	60.2	40.7	7.5	8.9
基礎濾器	SNP精度	0.997	0.997	0.995	0.997	0.997	0.995	0.997	0.993
	SNP靈敏度	0.996	0.995	0.988	0.997	0.994	0.986	0.996	0.991
	插入缺失精度	0.934	0.935	0.924	0.938	0.938	0.924	0.960	0.948
	插入缺失靈敏度	0.956	0.951	0.914	0.965	0.950	0.912	0.961	0.925
經過濾	SNP精度	0.999	0.998	0.997	0.999	0.998	0.996	0.999	0.997
	SNP靈敏度	0.995	0.994	0.985	0.995	0.993	0.985	0.995	0.989
	插入缺失精度	0.971	0.965	0.943	0.974	0.964	0.942	0.978	0.964
	插入缺失靈敏度	0.943	0.940	0.902	0.958	0.940	0.902	0.952	0.917
HapCut2	經定相之雜合SNP %	99.9%	99.9%	99.8%	99.9%	99.7%	99.7%	98.9%	98.7%
	單體群N50尺寸(Mb)	15.1	12.9	8.6	6.4	4.2	2.6	0.6	1.2
	短切換誤差率	0.00273	0.00272	0.00272	0.00261	0.00272	0.00271	0.00272	0.00571
	長切換誤差率	0.00571	0.00571	0.00570	0.00553	0.00570	0.00570	0.00574	0.00276
LongHap	經定相之雜合SNP %	0.999	0.9988	0.9966	0.9991	0.9984	0.9952	0.9895	0.9879
	單體群N50尺寸(Mb)	18.1	15.6	10.7	8	5.2	3.3	1.1	1.9
	短切換誤差率	0.0025748	0.0025949	0.0026139	0.0025228	0.0025307	0.0025773	0.0027524	0.0030534
	長切換誤差率	0.0017183	0.0017073	0.0017638	0.0017197	0.0017038	0.0017101	0.0019273	0.0020666

表 1 B

10X Genomics[1]	Illumina珠子單倍體分類[2]	BGISEQ500 STD[3]
128	99	132
1.25	3	1,000
85.7	-	N/A
33X	19X	43X
6.0%	21.0%	3.7%
PE150	PE76	PE100
1,538,345	147,456	N/A
8.32	~100	N/A
49.8	5	N/A
0.952	0.997	0.998
0.996	0.952	0.998
0.639	0.932	0.960
0.864	0.832	0.972
0.994	-	0.999
0.997	-	0.997
0.916	-	0.991
0.871	-	0.962
99.9%	98.0%	N/A
12.8	1.14	N/A
0.00273	0.0013	N/A
0.00572	0.000085	N/A
N/A	N/A	N/A
N/A	N/A	N/A
N/A	N/A	N/A
N/A	N/A	N/A

表2. 支架統計

	stLFR-1	stLFR-4	HiC[1]	HiC[2]
讀段對(M)	60	134	734	734
總支架長度(Gb)	2.84	2.72	2.92	2.92
支架N50(Mb)	44.7	42.8	68.3	60.02
比對鹼基%	98.61%	98.56%	98.22%	94.52%
支架計數	597	699	1,411	1,555
支架中之重疊群	1,411	1,586	3,096	18,903
斷點	31,386	30,501	35,132	33,079
重定位	296	327	430	136
易位	179	189	406	96
倒位	624	656	898	408

[1] 下載來自人胚胎幹細胞(hESC)之HiC讀段對(30)並用於使用SALSA(28)及與用於stLFR文庫之相同過程來架設SMRT讀段。

[2] 使用相同之HiC讀段對以使用SALSA來架設SMRT讀段，由Ghurye等人(28)報道之結果。

表 3

	stLFR-1			stLFR-2			stLFR-3	stLFR-4	BGISEQ-500 STD
定序之總鹼基(Gb)	336	230	100	660	200	100	117	126	132
FP調用	10,579	10,498	14,602	11,068	11,012	15,022	8,422	22,404	5,438
FN調用	13,023	15,106	40,088	11,218	18,511	46,182	14,205	27,792	7,816
過濾之FP調用	4,491	5,443	9,503	4,606	6,326	11,326	4,775	8,564	3,111
過濾之FN調用	16,988	19,014	49,330	15,302	22,152	49,443	17,436	34,482	8,984
FP調用中之改變	-6,088	-5,055	-5,099	-6,462	-4,686	-3,696	-3,647	-13,840	-2,327
FN調用中之改變	3,965	3,908	9,242	4,084	3,641	3,261	3,231	6,690	1,168
移除共用FP之最終FP調用	2,825	3,777	7,837	2,940	4,660	9,660	3,109	6,898	3,111

表 4

	stLFR-1			stLFR-2			stLFR-3	stLFR-4
定序之總鹼基(Gb)	336	230	100	660	200	100	117	126
定相之雜合SNP %	99.9%	99.9%	99.7%	99.9%	99.9%	99.6%	99.1%	99.0%
定相之雜合插入缺失%	96.8%	96.6%	94.9%	97.1%	96.2%	94.1%	93.9%	90.9%
重疊群N50尺寸(Mb)	23.4	19.7	13	10.5	7.3	4.1	1.2	2.1
短切換誤差率	0.00939	0.00938	0.00988	0.00943	0.00935	0.01002	0.01171	0.01212
長切換誤差率	0.00332	0.00337	0.00340	0.00313	0.00337	0.00321	0.00390	0.00426

表 5

	標準	stLFR-1 336 Gb	stLFR-1 230 Gb	stLFR-1 100 Gb	stLFR-2 660 Gb	stLFR-2 200 Gb	stLFR-2 100 Gb	stLFR-3	stLFR-4	STD
SNP	GQ	23	18	18	41	12	0	13	3	41
SNP	最小參考/備選	0.125	0.125	0.15	0.2	0.1	0.07	0.105	0.11	0.22
SNP	最大參考/備選	6.68	6.68	5	6.7	6.68	6.67	6.5	4.8	5.3
SNP	條碼	ref < 1	ref < 1	ref < 1	ref < 2	ref < 2	ref < 2	ref < 1	alt < 1	NA
插入缺失	GQ	70	60	45	80	65	40	60	50	95
插入缺失	最小參考/備選	0.3	0.27	0.2	0.27	0.28	0.2	0.3	0.22	0.4
插入缺失	最大參考/備選	3.2	3.5	5	3.2	4.2	5	3.5	5	3

表 6

珠子共同T	SEQ. ID NO:14	/52-Bio/AAAAAAAAAATGTGAGCCAAGGAGTTG
珠子共同B	SEQ. ID NO:15	CCAGAGCAACTCCTTGGCTCACA
橋	SEQ. ID NO:16	GCACUGACGACAUGAUCACCAAGGAUCGCCAUAGUCCAUGCUA
對於 BGISEQ-500		
轉座子1T	SEQ. ID NO:17	/5Phos/CGATCCTTGGTGATCATGTCGTCAGTGCTTGTCTTCCTAAGATGTGTATAAGAGACAG
轉座子2T	SEQ. ID NO:18	GCCTCCCTCGCGCCATCAGAGATGTGTATAAGAGACAG
轉座子B	SEQ. ID NO:19	/5Phos/CTGUCTCUTATACACAUCT
PCR1	SEQ. ID NO:20	TGTGAGCCAAGGAGTTG
PCR2	SEQ. ID NO:21	GCCTCCCTCGCGCCATCAG
定序引子		
BGI R1定序引子	SEQ. ID NO:22	GCCTCCCTCGCGCCATCAGAGATGTGTATAAGAGACAG
BGI stLFR條碼定序引子	SEQ. ID NO:23	CGAGAACGTCTTGTGAGCCAAGGAGTTGCTCTGG
BGI R2定序引子	SEQ. ID NO:24	CGTCAGTGCTTGTCTTCCTAAGATGTGTATAAGAGACAG
BGI MDA引子1	SEQ. ID NO:25	TGATCACCAAGGATCGCCATAGTCCATGCTA
BGI MDA引子2	SEQ. ID NO:26	CTGTCTCTTATACACATCTTAGGAAGACAAGCACTGACGA
對於 3' 分支連接		
3' 分支連接銜接子-F	SEQ. ID NO:27	/5Phos/CTGATGGCGCGAGGGAGGC
3' 分支連接銜接子-R	SEQ. ID NO:28	TCGCGCCATCA/3'dd/G
定序引子		
R1 定序引子間隙	SEQ. ID NO:29	CAACTCCTTGGCTCACACGGAGGGAGCGCGGTAGTC

表 7 ： 連 接 效 率

連接效率	受質				
	1	2	3	4	5
銜接子	切口	突出端	1 nt間隙	2 nt間隙	3 nt間隙
Ad-T	15.2%	79.5%	89.6%	88.9%	83.9%
Ad-A	12.0%	88.6%	77.5%	68.3%	83.9%
Ad-GA	7.7%	58.9%	80.5%	56.4%	59.2%

泳道#	受質類型	受質			連接產物			連接效率
		尺寸 (nt)	強度 (像素)	歸一化強度 (像素/nt)	尺寸 (nt)	強度 (像素)	歸一化強度 (像素/nt)	
2	切口	27	19044.75	705.36	49	5062.00	103.31	12.77%
5	1-nt間隙	27	13120.29	485.94	49	22807.69	465.46	48.92%
8	8-nt間隙	25	14042.49	561.70	47	19060.60	405.54	41.93%
11	3'-隱性末端	27	1376.23	50.97	49	17684.29	360.90	87.62%
14	鈍端對照	40	5311.44	132.79	62	21973.00	354.40	72.74%

表 8

受質類型	供體類型	受質			連接產物			連接效率
		尺寸(nt)	強度(像素)	歸一化強度 (像素/nt)	尺寸(nt)	強度(像素)	歸一化強度 (像素/nt)	
切口	Ad T	124	11801.78	95.18	156	1744.01	11.18	10.51%
	Ad A	124	13130.49	105.89	182	1091.70	6.00	5.36%
	Ad GA	124	12810.37	103.31	184	603.87	3.28	3.08%
1-nt間隙	Ad T	123	2561.08	20.82	155	23719.00	153.03	88.02%
	Ad A	123	2058.55	16.74	181	7034.96	38.87	69.90%
	Ad GA	123	1709.67	13.90	183	8340.98	45.58	76.63%
2-nt間隙	Ad T	122	1164.89	9.55	154	6909.36	44.87	82.45%
	Ad A	122	3882.74	31.83	180	7688.03	42.71	57.30%
	Ad GA	122	6573.57	53.88	182	8495.74	46.68	46.42%
3-nt間隙	Ad T	121	2344.08	19.37	153	11764.83	76.89	79.88%
	Ad A	121	1974.72	16.32	179	9738.98	54.41	76.93%
	Ad GA	121	8896.47	73.52	181	10145.81	56.05	43.26%
3'-隱性末端	Ad T	108	1934.79	17.91	140	8791.10	62.79	77.80%
	Ad A	108	1070.23	9.91	166	7834.38	47.20	82.65%
	Ad GA	108	5675.05	52.55	168	7206.26	42.89	44.94%

【符號說明】

【 0 2 3 8 】

無

【生物材料寄存】

【 0 2 3 9 】　　國內寄存資訊 (請依寄存機構、日期、號碼順序註記)

無

【０２４０】　國外寄存資訊(請依寄存國家、機構、日期、號碼順序註記)

無

【發明申請專利範圍】

【第1項】 一種製備定序文庫之方法，該文庫用於在不使用超微量分光光度計(nanodrops)之情況下對靶核酸定序，該方法包括以下步驟：

(a) 藉由一轉座酶將插入序列轉座至該靶核酸之第一片段中，其中該插入序列包含雜交序列，並且其中該轉座在該等第一片段中產生切口；

(b) 將以下各者組合在單一混合物中：(i)來自(a)之該靶核酸之該等第一片段、(ii)夾板寡核苷酸，及(iii)珠子群體，其中每個珠子包含固定在其上之捕獲寡核苷酸，該等捕獲寡核苷酸包含：

1) 含條碼序列，其中固定在同一個別珠子上之該等寡核苷酸包含相同的含條碼序列，並且大多數珠子具有不同的含條碼序列，

2) 與夾板寡核苷酸之至少一部分互補之共同序列，其中該夾板寡核苷酸之第二部分與該雜交序列之至少一部分互補；

3) 第一 PCR 引子降溫貼合位點(PCR primer annealing site)；

(c)將各個珠子之捕獲寡核苷酸連接到各個第一片段之所插入雜交序列；以及

(d) 將 3'-分支連接(3'BL)銜接子與步驟(c)的連

接產物結合，其中該 3'BL 銜接子包含第一寡核苷酸及第二寡核苷酸，經降溫貼合以形成一鈍端雙鏈末端(blunt duplex end)及一單鏈 DNA 末端，其中該第一寡核苷酸包含位在該鈍端雙鏈末端處之一 5'磷酸；以及

將該第一寡核苷酸在該等第一片段中之該等切口處連接到該等第一片段連接到該等第一片段，其中該連接為 3'分支連接將該第一寡核苷酸的該 5'磷酸與該等第一片段的該等切口處之該 3'羥基共價連接，且其中該第一寡核苷酸包含第二 PCR 引子降溫貼合位點。

【第2項】 如請求項 1 所述之方法，其中該第一 PCR 引子降溫貼合位點及該第二 PCR 引子降溫貼合位點具有不同之序列。

【第3項】 如請求項 1 所述之方法，其中該 3'分支連接銜接子寡核苷酸包含條碼序列，其係樣品條碼序列。

【第4項】 如請求項 1 所述之方法，其中在將該插入序列轉座至該等第一片段之步驟中，該轉座酶保持與該等第一片段結合。

【第5項】 如請求項 1 至 4 中任一項所述之方法，其包括以下步驟：從該等第一片段中移除該轉座酶，從而產生亞片段。

【第6項】 如請求項 5 所述之方法，其包括以下步驟：擴增該等亞片段以產生擴增子。

【第7項】 如請求項 6 所述之方法，其包括以下步驟：對該等擴增子進行定序以產生序列讀段，其中具有相同條碼序列之序列讀段來自相同第一片段。

【第8項】 如請求項 1 至 4 中任一項所述之方法，其中該靶核酸係基因組 DNA。

【第9項】 如請求項 8 所述之方法，其中該基因組 DNA 係人基因組 DNA。

【第10項】 如請求項 1 所述之方法，其中將兩個不同之轉座子插入該等第一片段中。

【第11項】 如請求項 8 所述之方法，其中該基因組 DNA 來自真核生物。

【第12項】 如請求項 8 所述之方法，其中該基因組 DNA 來自單個個體生物。

【第13項】 如請求項 1 至 4 中任一項所述之方法，其中該靶核酸來自原核生物或原核生物之異質群體。

【第14項】 如請求項 1 至 4 中任一項所述之方法，其中在步驟(a)之前不擴增該靶核酸。

【第15項】 如請求項 1 至 4 中任一項所述之方法，其中該等珠子包含該捕獲寡核苷酸之至少 100,000 個拷貝。

【第16項】 如請求項15所述之方法，其中在反應中使用超過10億個珠子。

【第17項】 如請求項6所述之方法，其中在擴增(i)之前，酶促移除至少一些該等捕獲寡核苷酸及/或酶促移除至少一些夾板寡核苷酸及/或酶促移除至少一些嵌合末端(mosaic end; ME)序列。

【第18項】 如請求項1至4中任一項所述之方法，其中在步驟(c)後，使用一或多種外切核酸酶來酶促移除至少一些捕獲寡核苷酸。

【第19項】 如請求項18所述之方法，其中該一或多種外切核酸酶包含外切核酸酶Ⅰ、Ⅲ或兩者。

【第20項】 如請求項1至4中任一項所述之方法，其中該等捕獲寡核苷酸及/或該等夾板寡核苷酸包含尿嘧啶。

【第21項】 如請求項20所述之方法，其中用尿嘧啶-DNA糖基化酶(Uracil-DNA Glycosylase; UDG)處理使含尿嘧啶之寡核苷酸降解。

【第22項】 如請求項1所述之方法，其中該等捕獲寡核苷酸在5'末端與珠子連接。

【第23項】 如請求項1至4中任一項所述之方法，其中>50%之DNA第一片段用獨特條碼進行條碼編碼。

【第24項】 如請求項1至4中任一項所述之方法，其中片段中之>50%之亞片段與條碼寡核苷酸連接。

【第25項】 如請求項1至4中任一項所述之方法，其中，平均而言，該等珠子上之經條碼編輯之捕獲寡核苷酸之間距係<50nm。

【第26項】 如請求項5所述之方法，其包括對至少一些該等亞片段定序，以產生序列讀段。

【第27項】 如請求項6所述之方法，其包括對至少一些該等擴增子定序以產生序列讀段。

【第28項】 如請求項26所述之方法，其中具有相同條碼序列之序列讀段來自相同第一片段。

【第29項】 如請求項28所述之方法，其中對該等亞片段之長片段之>10%之亞片段進行定序。

【第30項】 如請求項7或27所述之方法，包括另外步驟：

(e) 將大多數該等序列讀段分配給相應第一片段；及

(f) 組裝該等序列讀段以產生該靶標之組裝序列。

【第31項】 如請求項1至4中任一項所述之方法，其中步驟(a)中之大多數第一片段長於20 kb。

【第32項】 如請求項1至4中任一項所述之方法，其中步驟(a)中之大多數第一片段長於50 kb。

【第33項】 如請求項1至4中任一項所述之方法，其中步驟(a)中之大多數第一片段長於100 kb。

【第34項】 如請求項1至4中任一項所述之方法，其中大多數第一片段之長度在50千鹼基至200千鹼基之範圍內。

【第35項】 如請求項1至4中任一項所述之方法，其中該等複數個第一片段來自單個細胞。

【第36項】 如請求項1至4中任一項所述之方法，其中該等複數個長基因組DNA片段來自1至100個真核細胞。

【第37項】 如請求項1至4中任一項所述之方法，其中該等複數個長基因組DNA片段來自3至30個人細胞。

【第38項】 如請求項1至4中任一項所述之方法，其中單一容器或混合物含有5至100個基因組當量之人DNA。

【第39項】 如請求項1至4中任一項所述之方法，其中單一容器或混合物含有50至1000個基因組當量之人DNA。

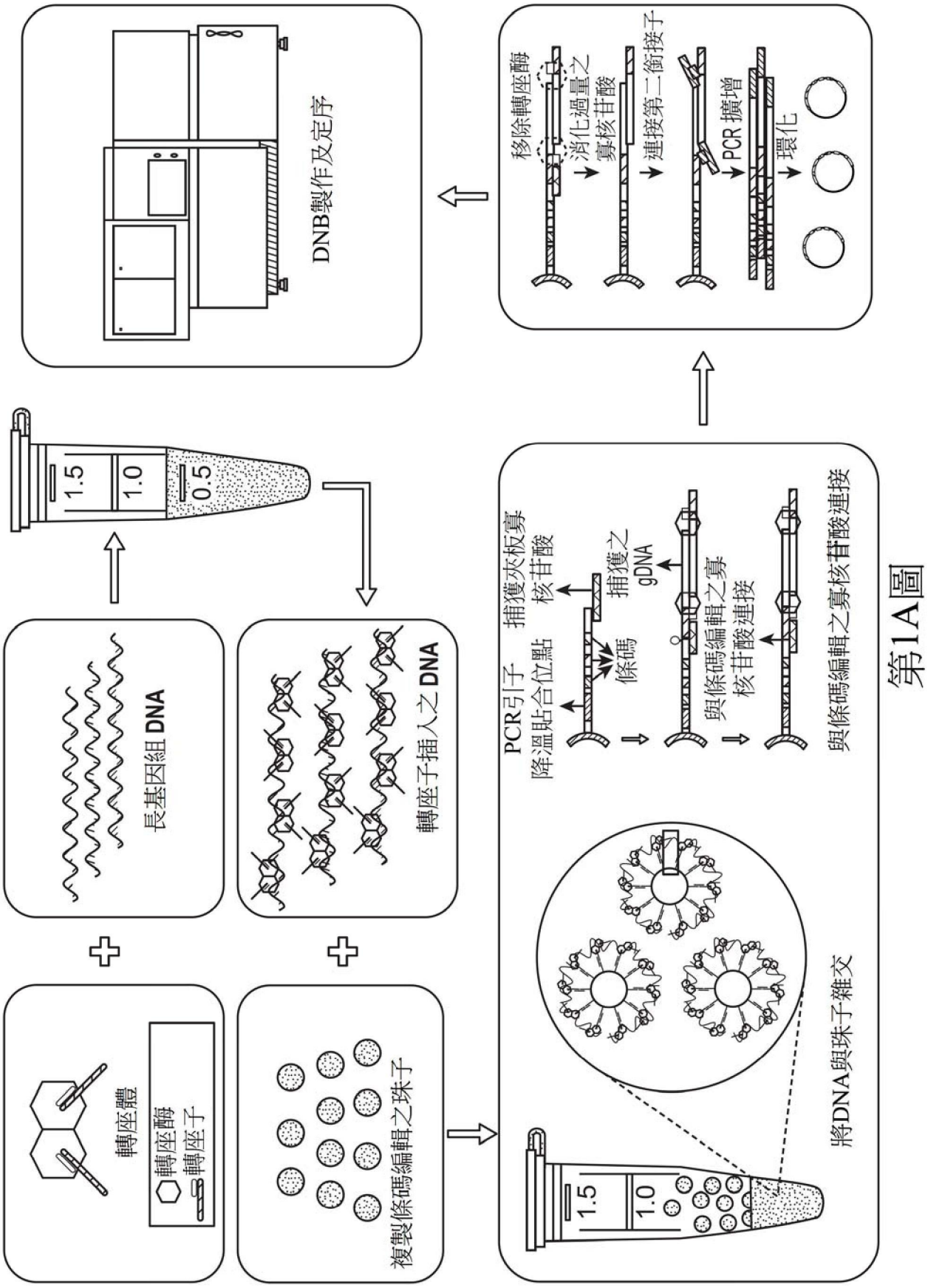
【第40項】 如請求項1所述之方法，其中該珠子群體總共包含至少100,000個不同標籤序列。

【第41項】 如請求項40所述之方法，其中該珠子群體

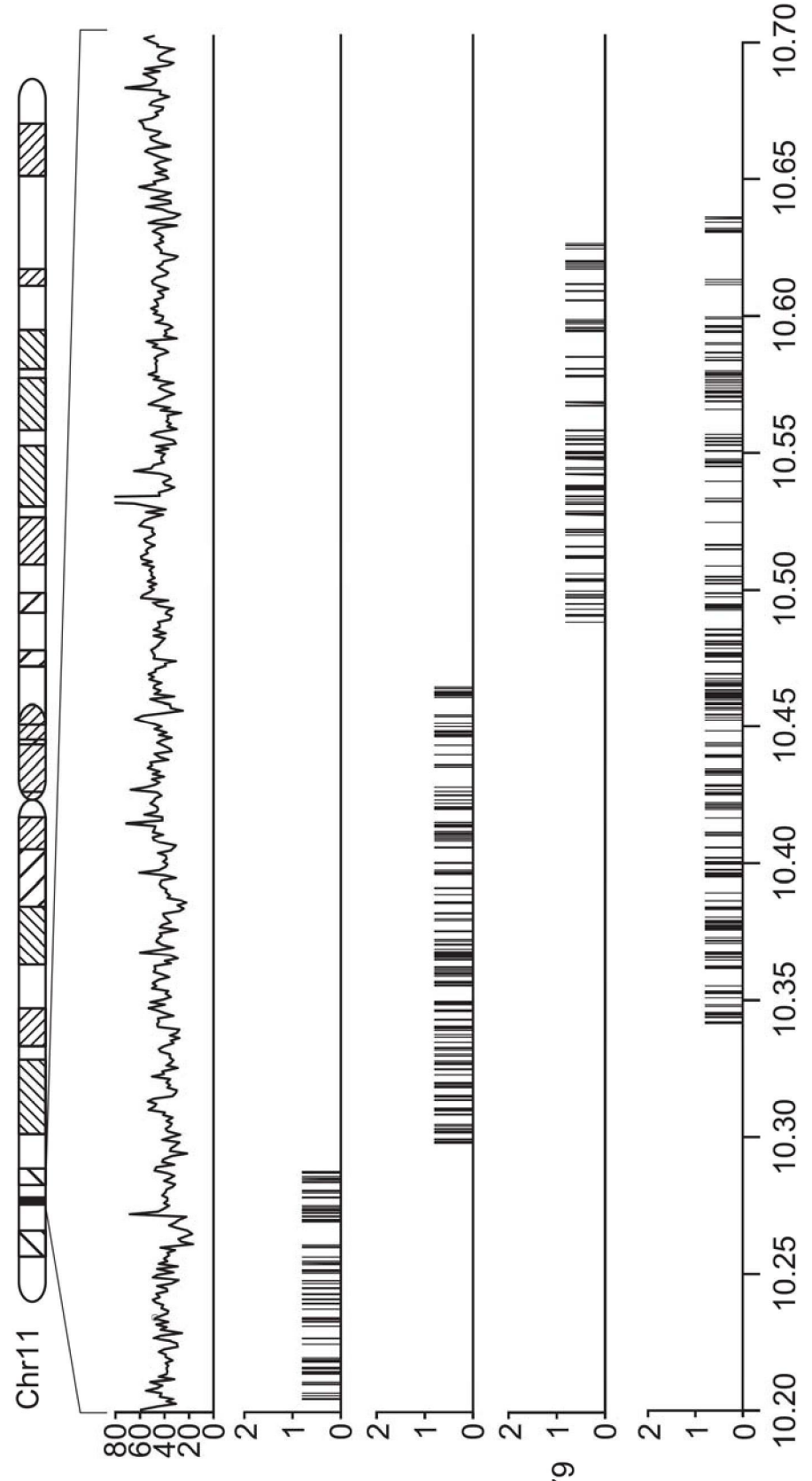
總共包含至少1百萬個不同標籤序列。

【第42項】 如請求項5所述之方法，其中該等亞片段之尺寸在200至2,000 bp之長度之範圍內。

【發明圖式】


轉座體
轉座酶
轉座子
長基因組 DNA
轉座子插入之 DNA
複製條碼編輯之珠子
1.5
1.0
0.5
PCR引子
降溫貼合位點
捕獲夾板寡
核苷酸
條碼
捕獲之
gDNA
與條碼編輯之寡
核苷酸連接
與條碼編輯之寡核苷酸連接
將DNA與珠子雜交
移除轉座酶
消化過量之
寡核苷酸
連接第二銜接子
PCR 擴增
環化
DNB製作及定序

第1A圖


Chr11
覆蓋率
80
60
40
20
0
條碼 #56
條碼 #2345
條碼 #144579
條碼 #n
2
1
0
10.20
10.25
10.30
10.35
10.40
10.45
10.50
10.55
10.60
10.65
10.70
位置(Mb)

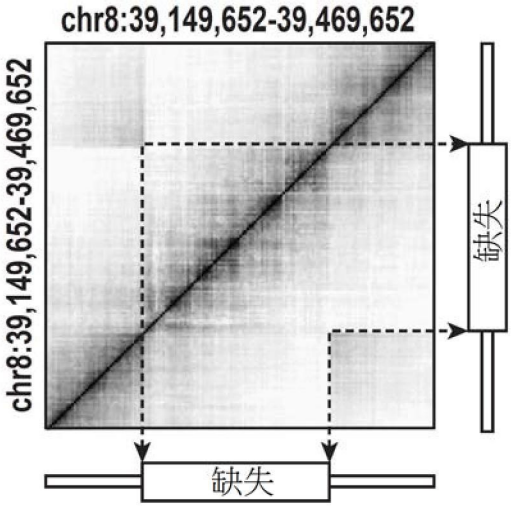
第1B圖

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
每個條碼之片段

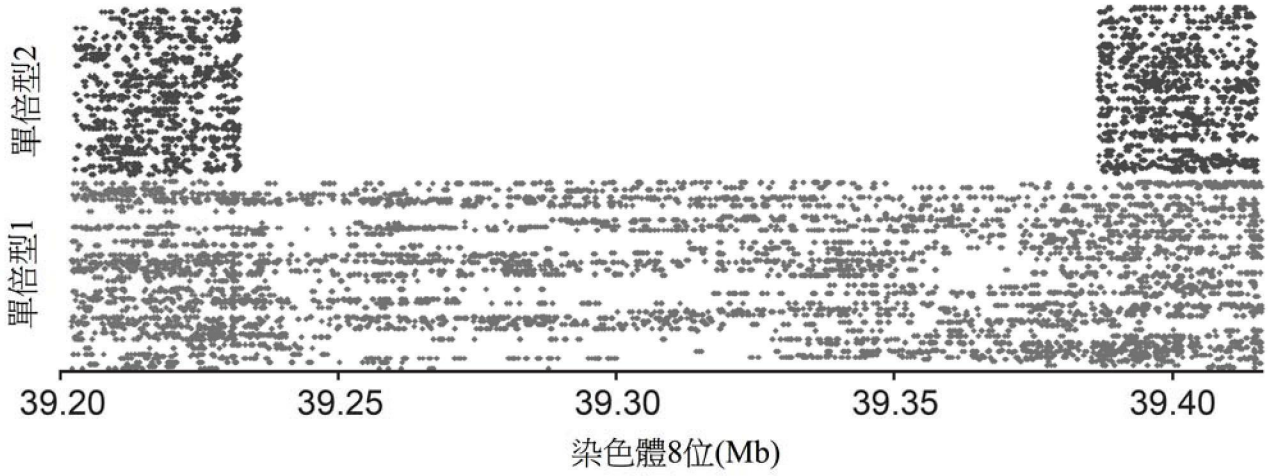
第1C圖

0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.00 0.04 0.08 0.12 0.16 0.20 0.24 0.28 0.32 0.36 0.40 0.44 0.48 0.52
覆蓋之長片段之分率

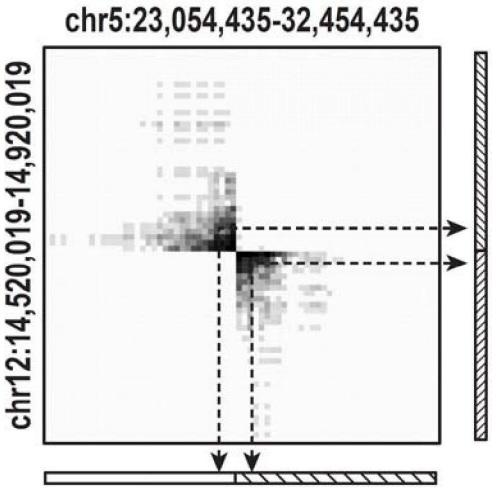
第1D圖



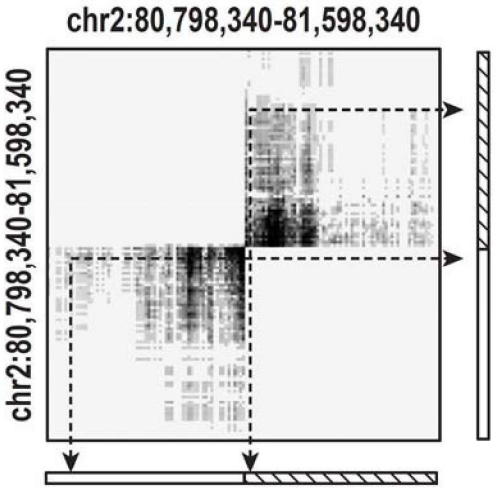
第2A圖



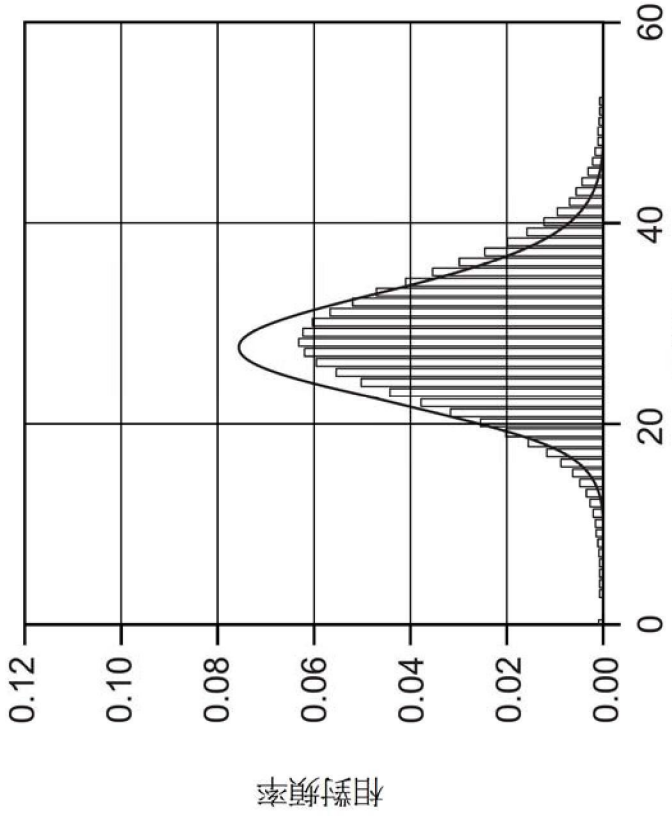
第2B圖



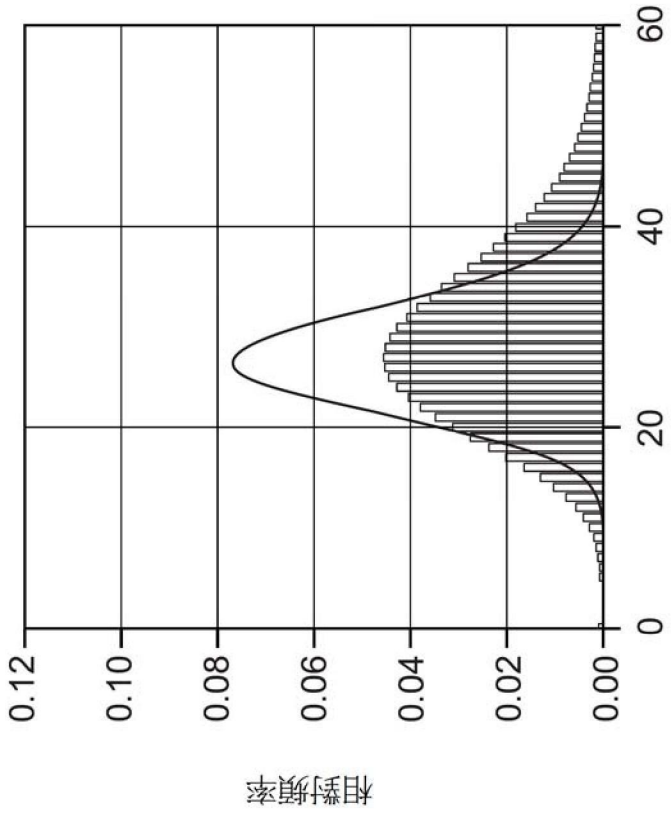
第2C圖



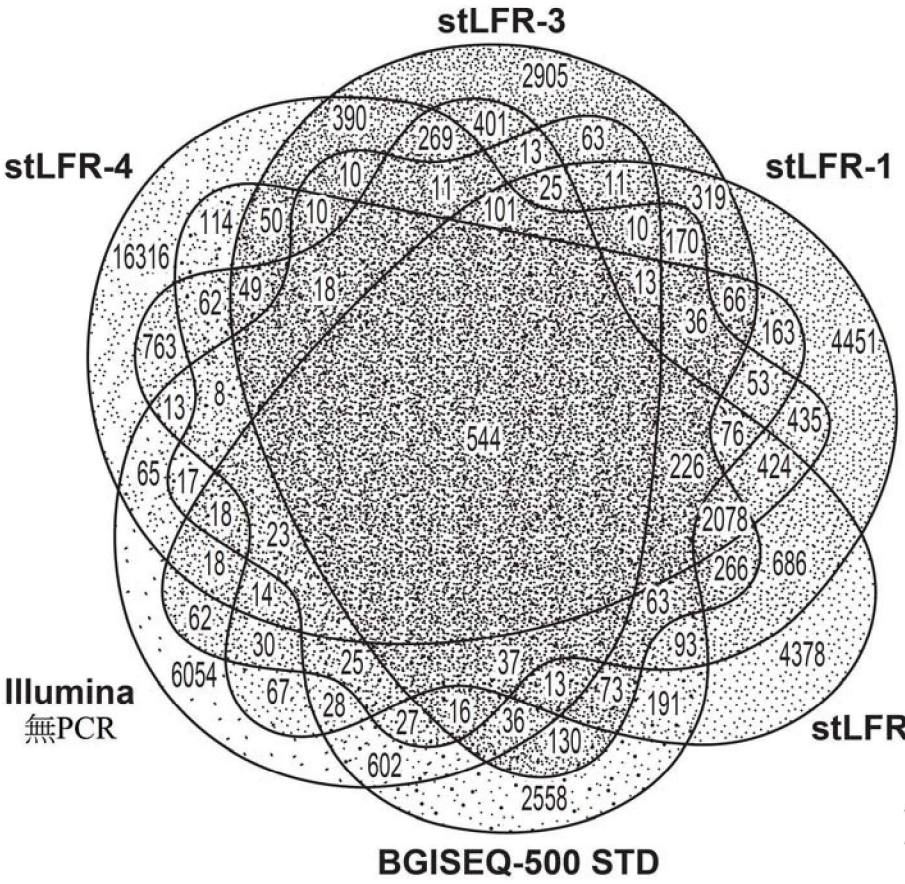
第2D圖


0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
覆蓋率
相對頻率

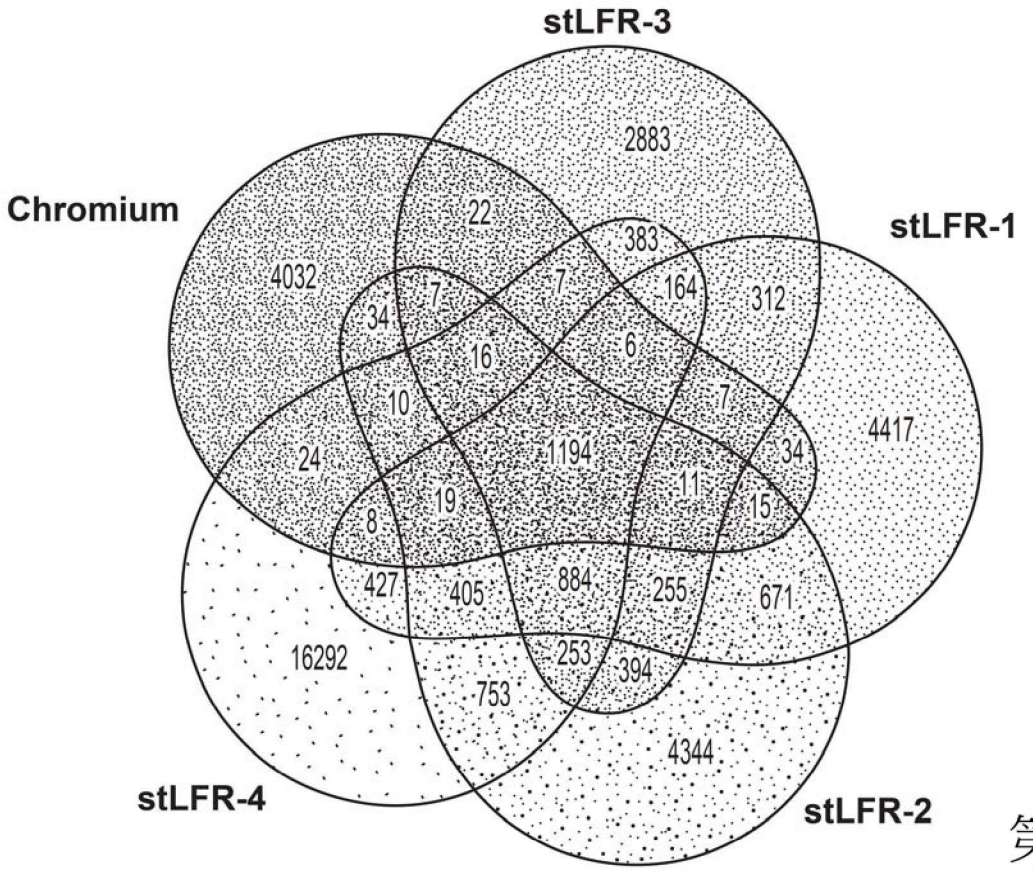
第3B圖


0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
20
40
60
覆蓋率
相對頻率

第3A圖

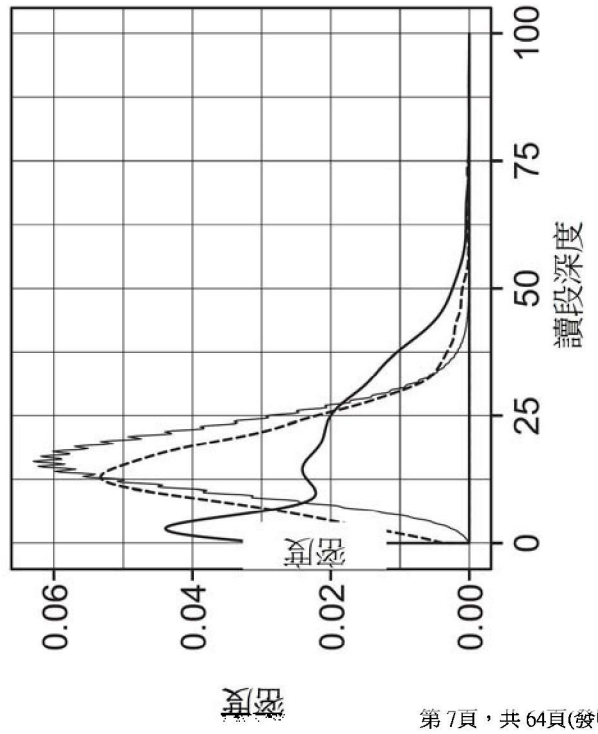

stLFR-3
stLFR-4
stLFR-1
Illumina
無PCR
stLFR-2
BGISEQ-500 STD
2905
390
269
401
63
13
10
11
25
11
319
101
10
170
114
50
10
16316
13
66
62
49
18
36
163
4451
763
53
8
435
544
76
13
226
424
65
17
2078
18
23
266
686
18
14
63
62
30
93
4378
25
37
6054
67
13
73
191
28
27
16
36
130
602
2558

第4A圖

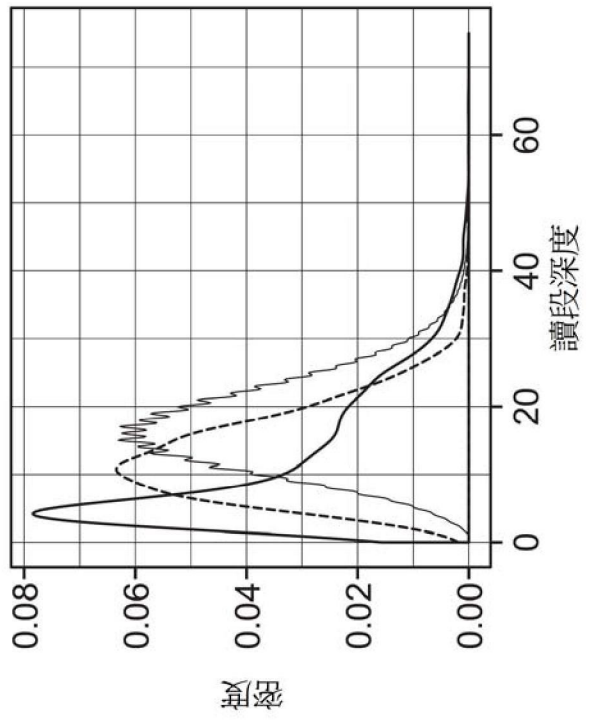

stLFR-3
Chromium
stLFR-1
stLFR-4
stLFR-2
2883
22
383
4032
7
7
164
312
34
6
16
10
7
4417
24
1194
34
11
19
15
8
427
405
884
255
671
16292
253
394
753
4344

第4B圖

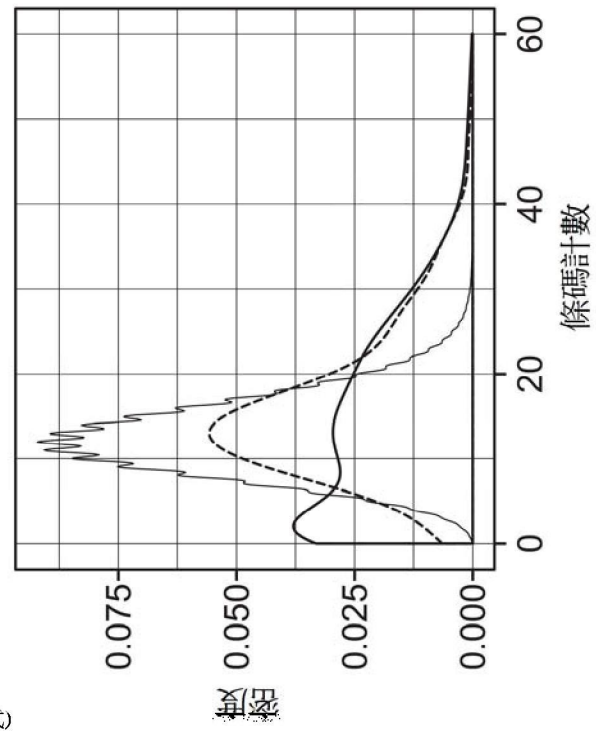
第5A圖



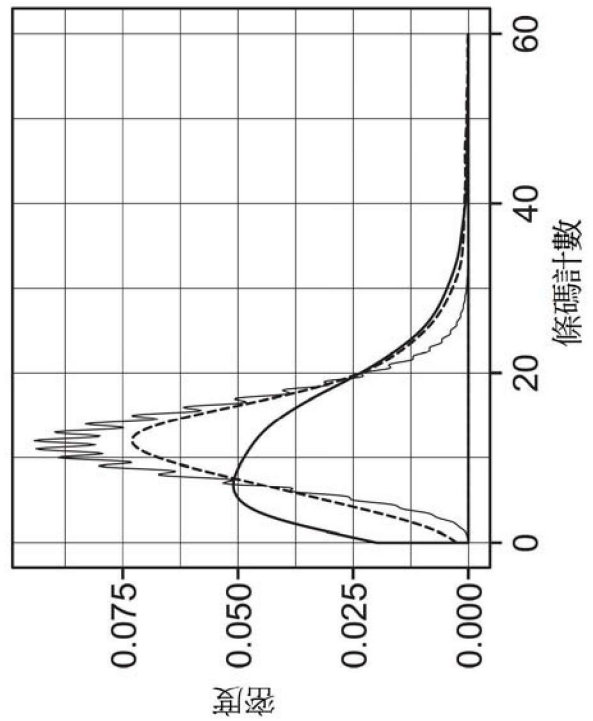
第5B圖

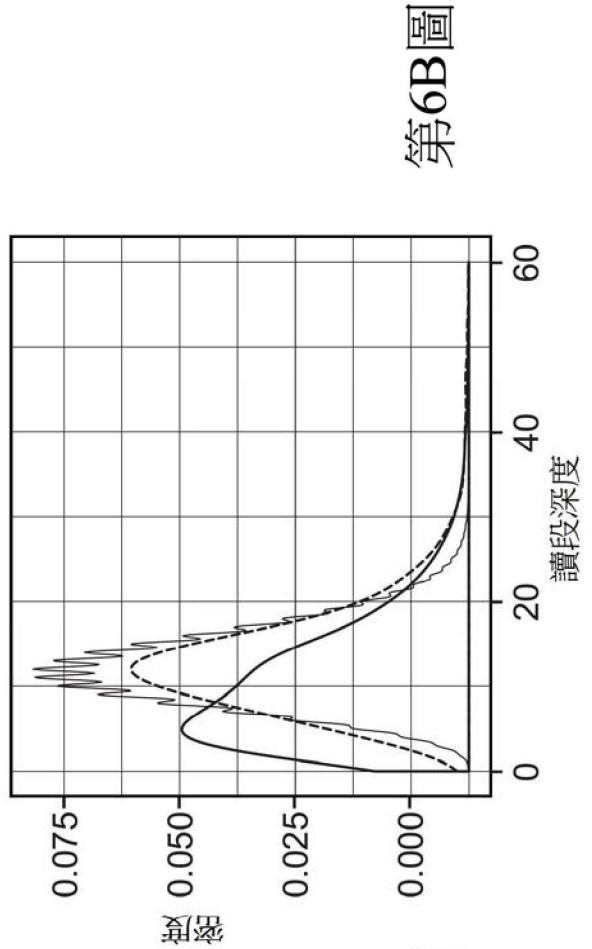


第5C圖

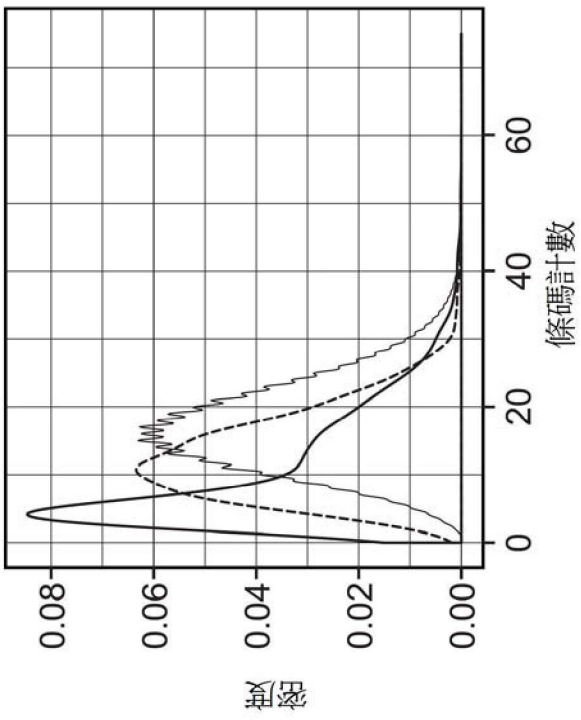


第5D圖

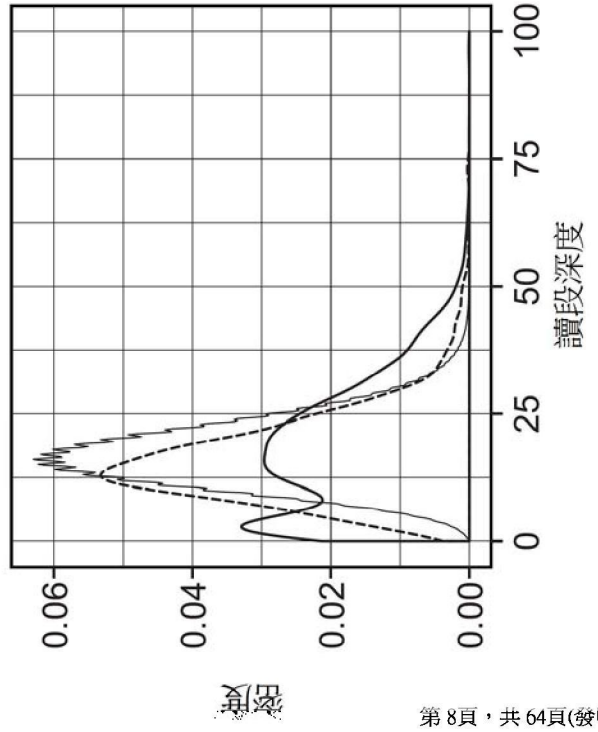




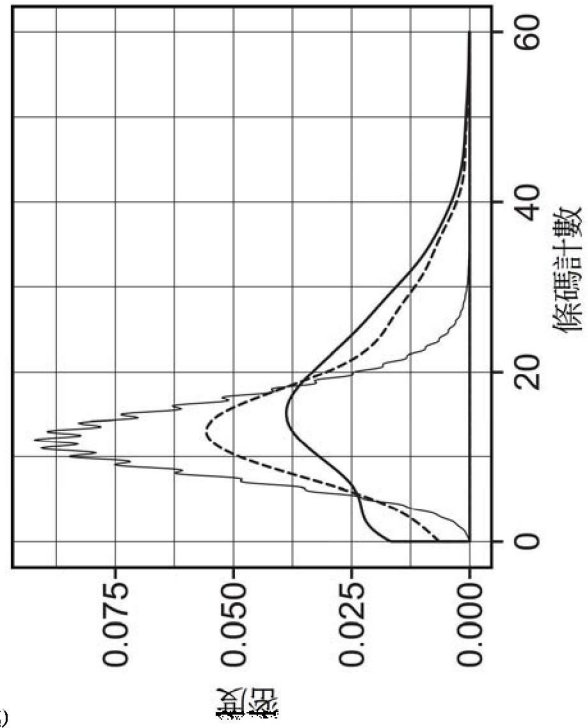
第6B圖



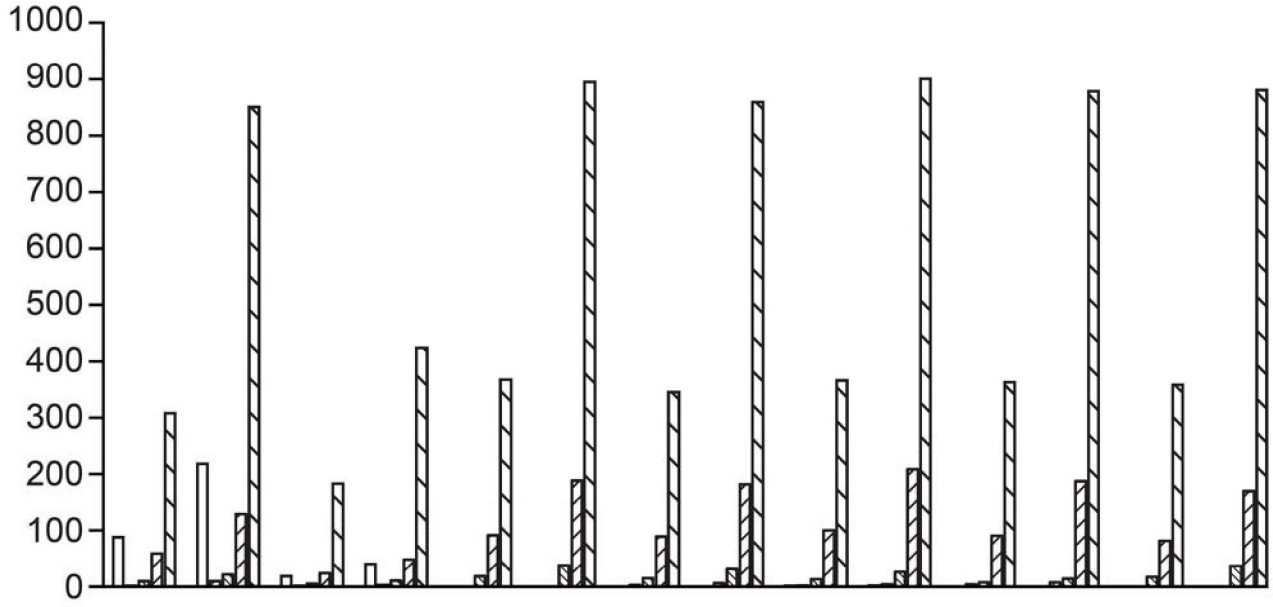
第6D圖



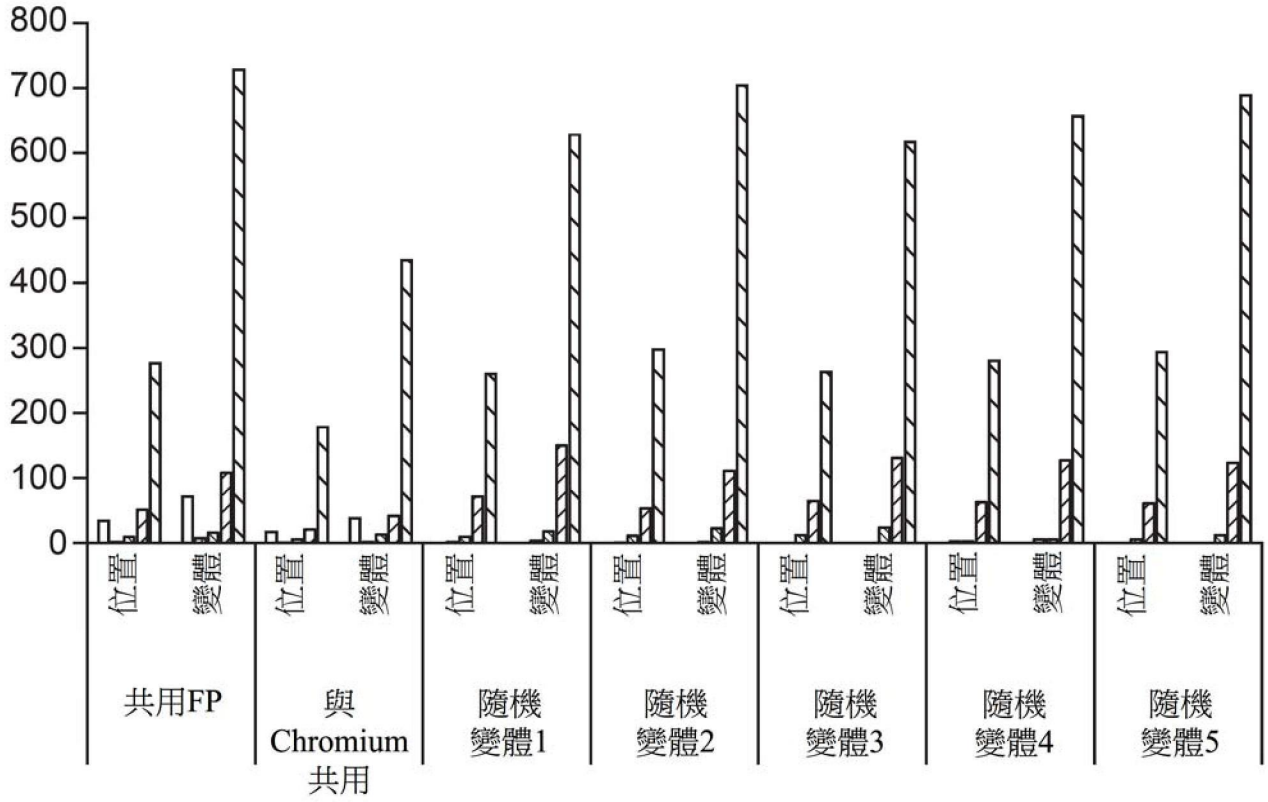
第6A圖



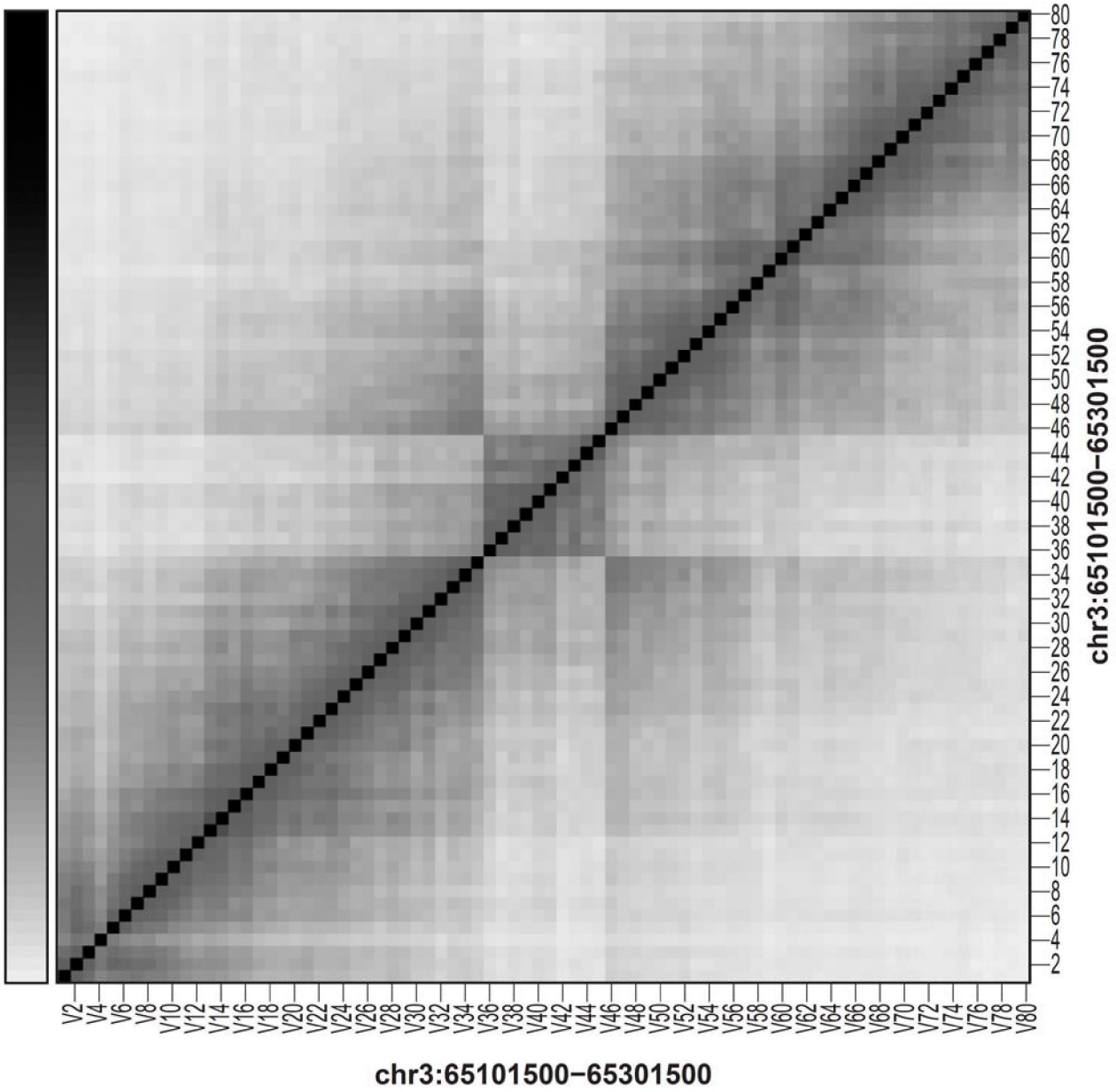
第6C圖


1000
900
800
700
600
500
400
300
200
100
0

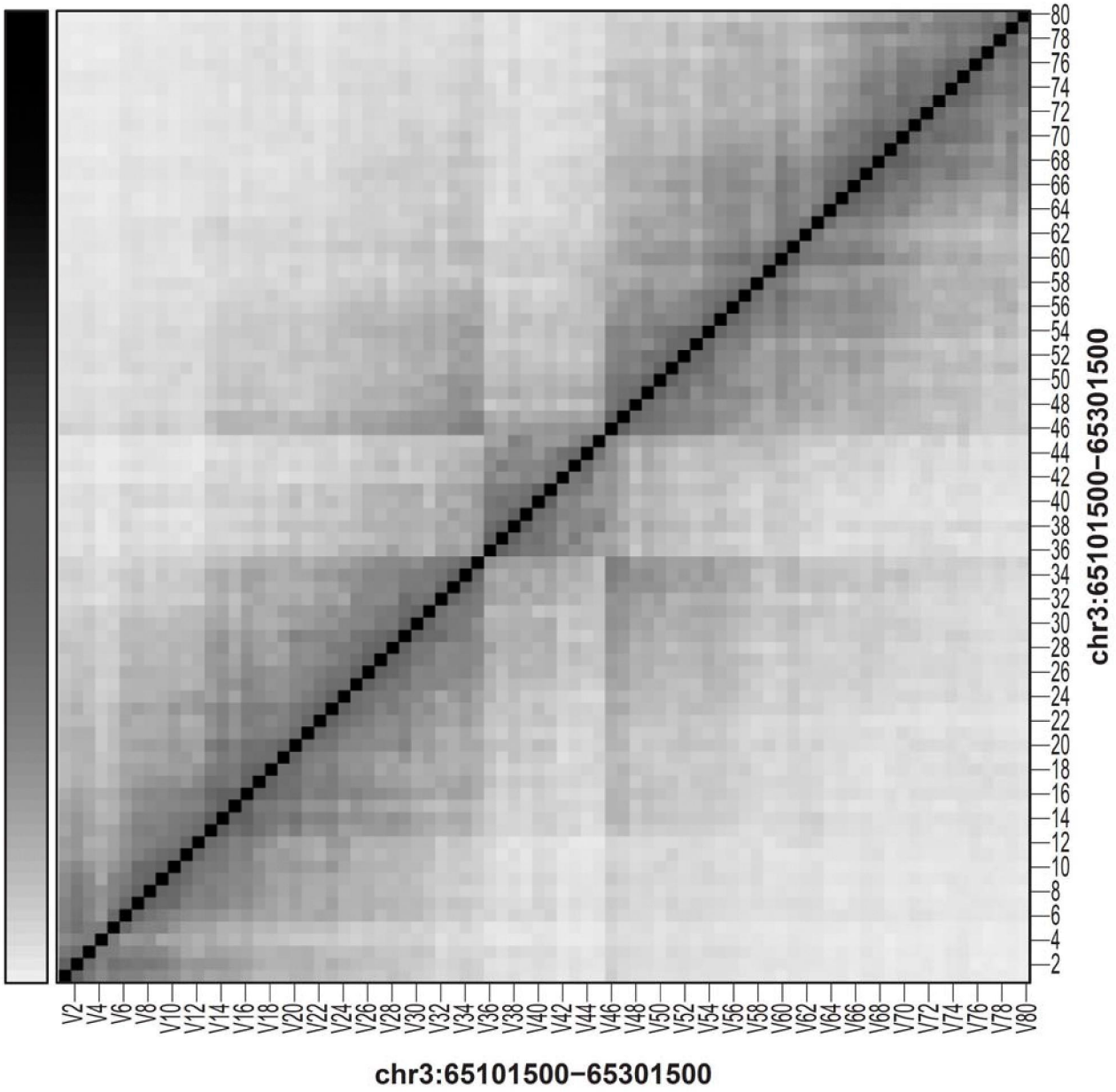
第7A圖


800
700
600
500
400
300
200
100
0
位置
變體
共用FP
與
Chromium
共用
隨機
變體1
隨機
變體2
隨機
變體3
隨機
變體4
隨機
變體5

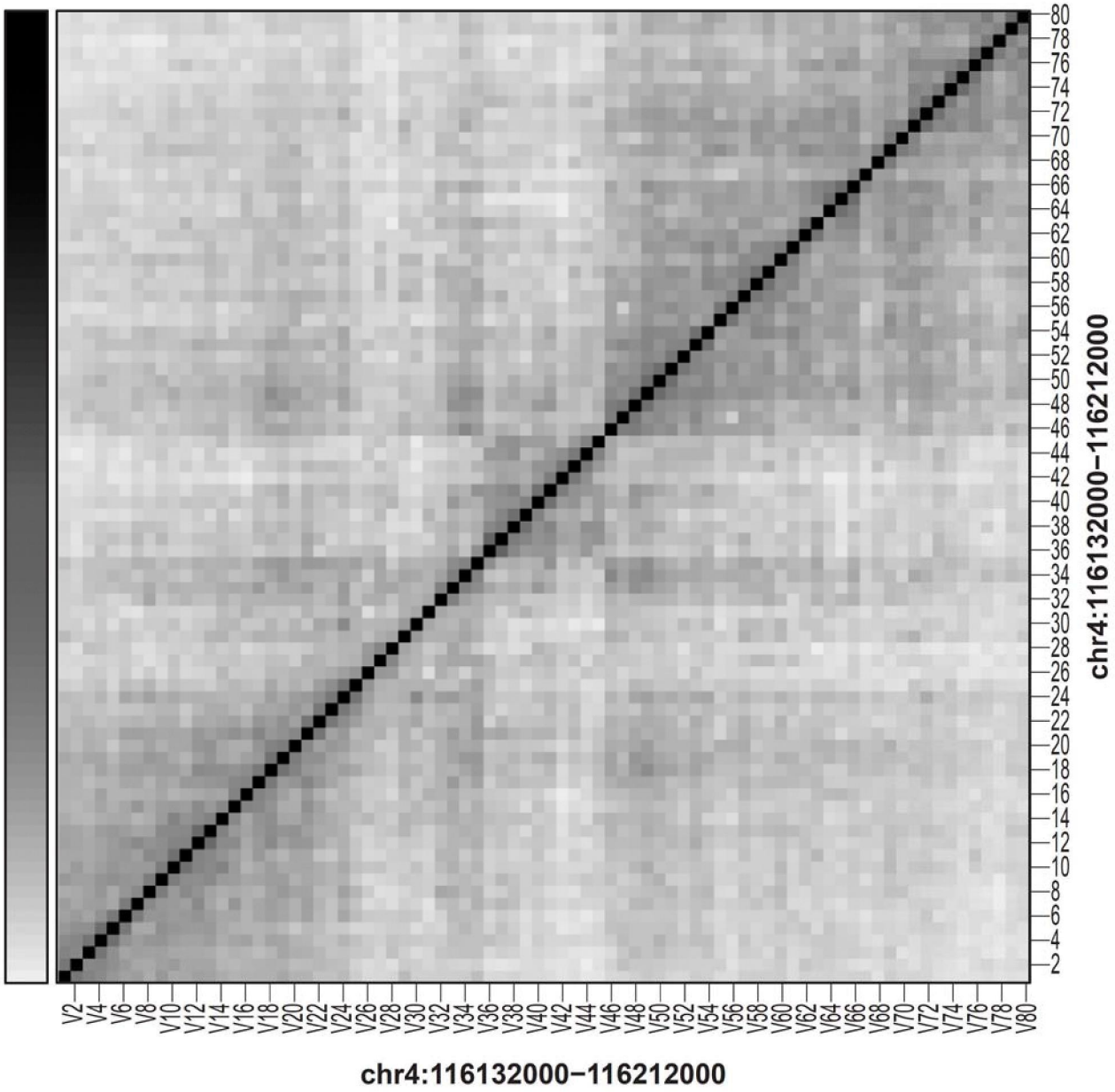
第7B圖



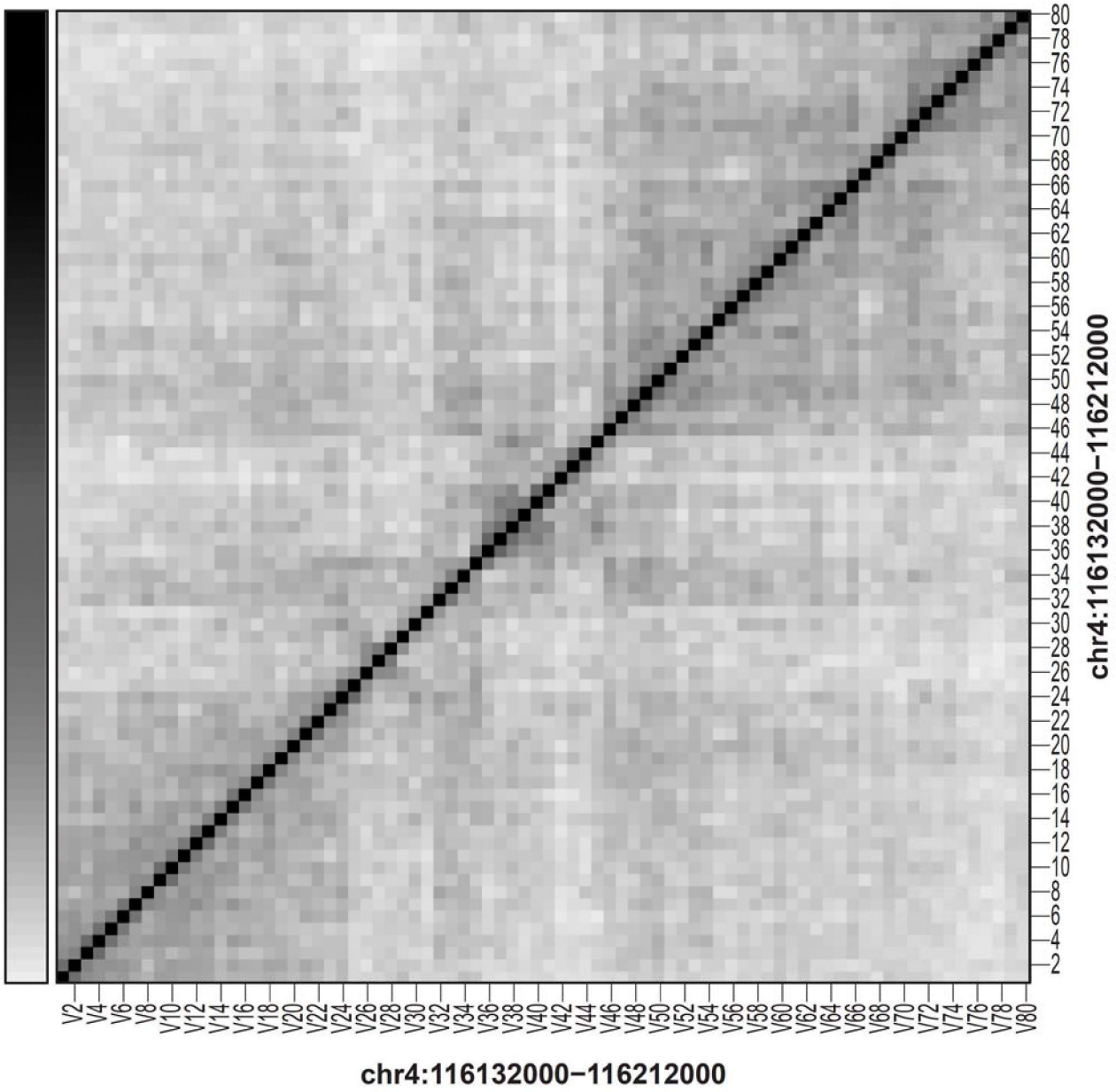
第8A圖



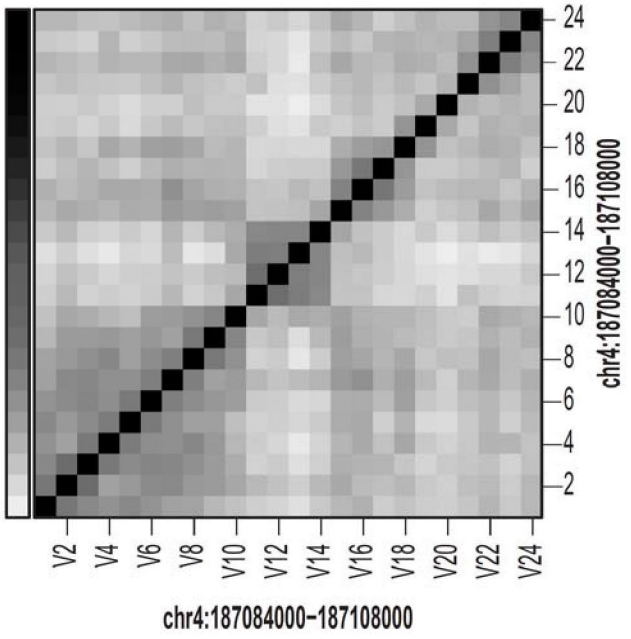
第8B圖



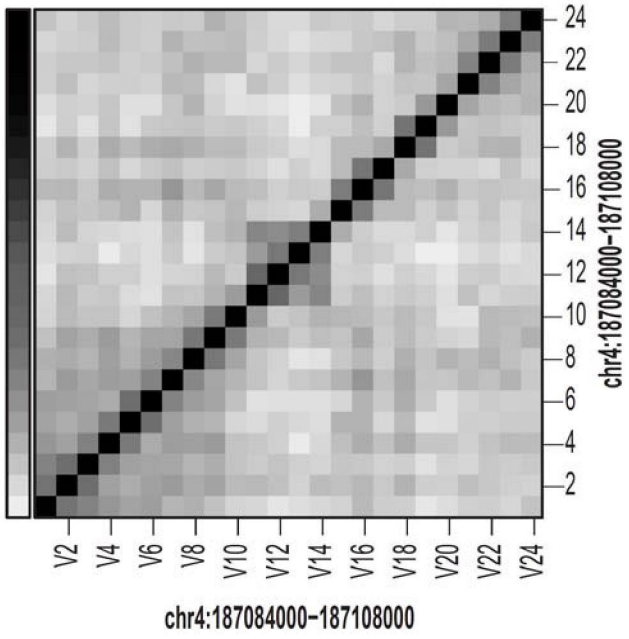
第8C圖



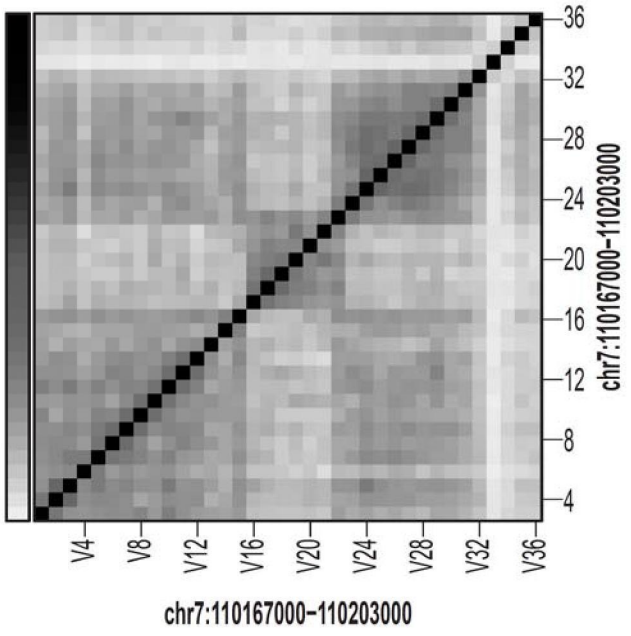
第8D圖



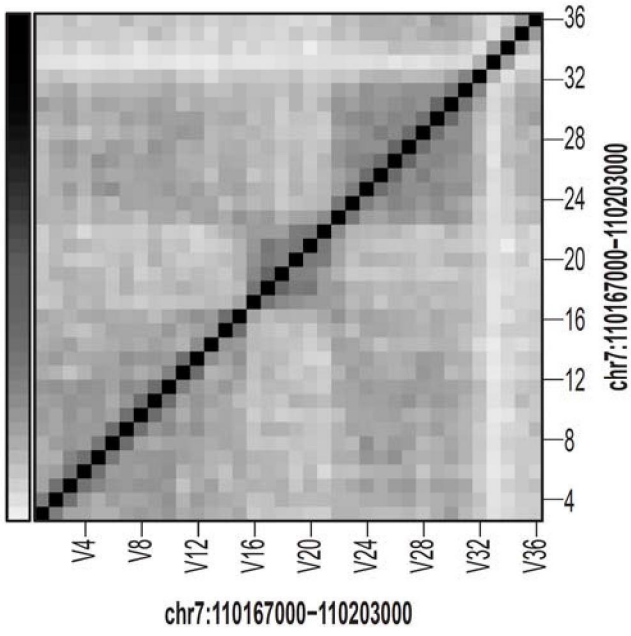
第8E圖



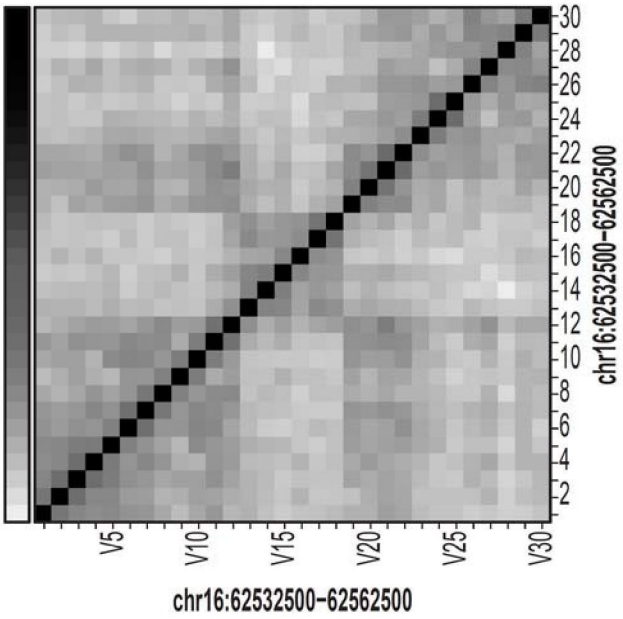
第8F圖



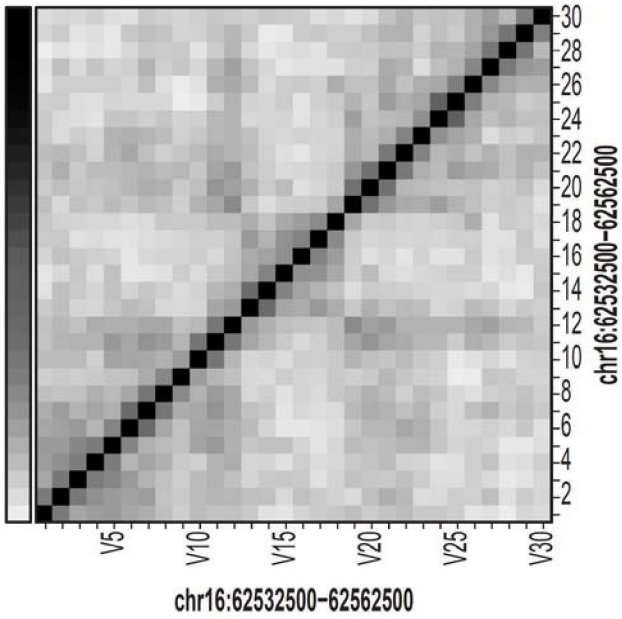
第8G圖



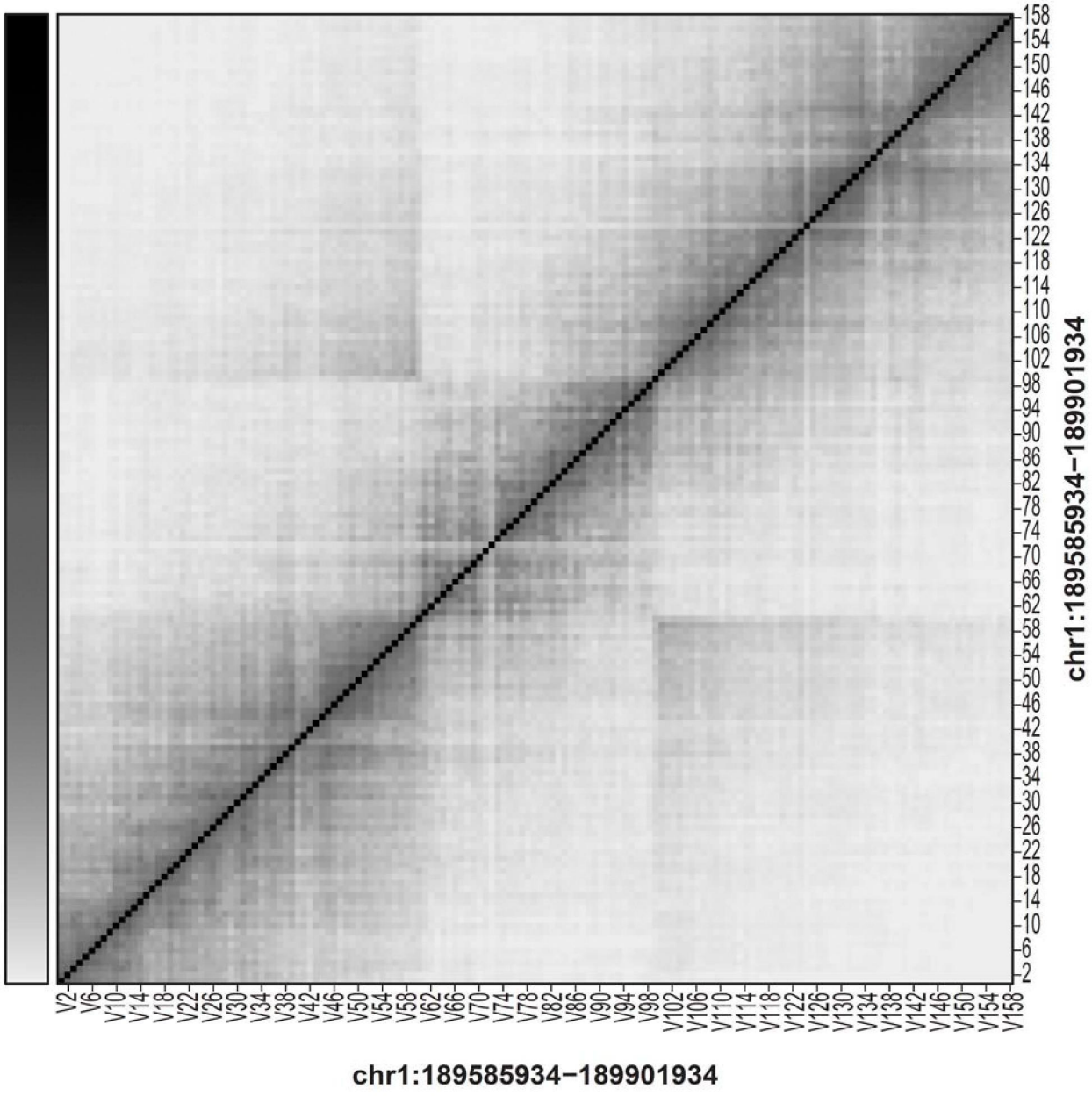
第8H圖


chr16:62532500-62562500
chr16:62532500-62562500
V5
V10
V15
V20
V25
V30
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30

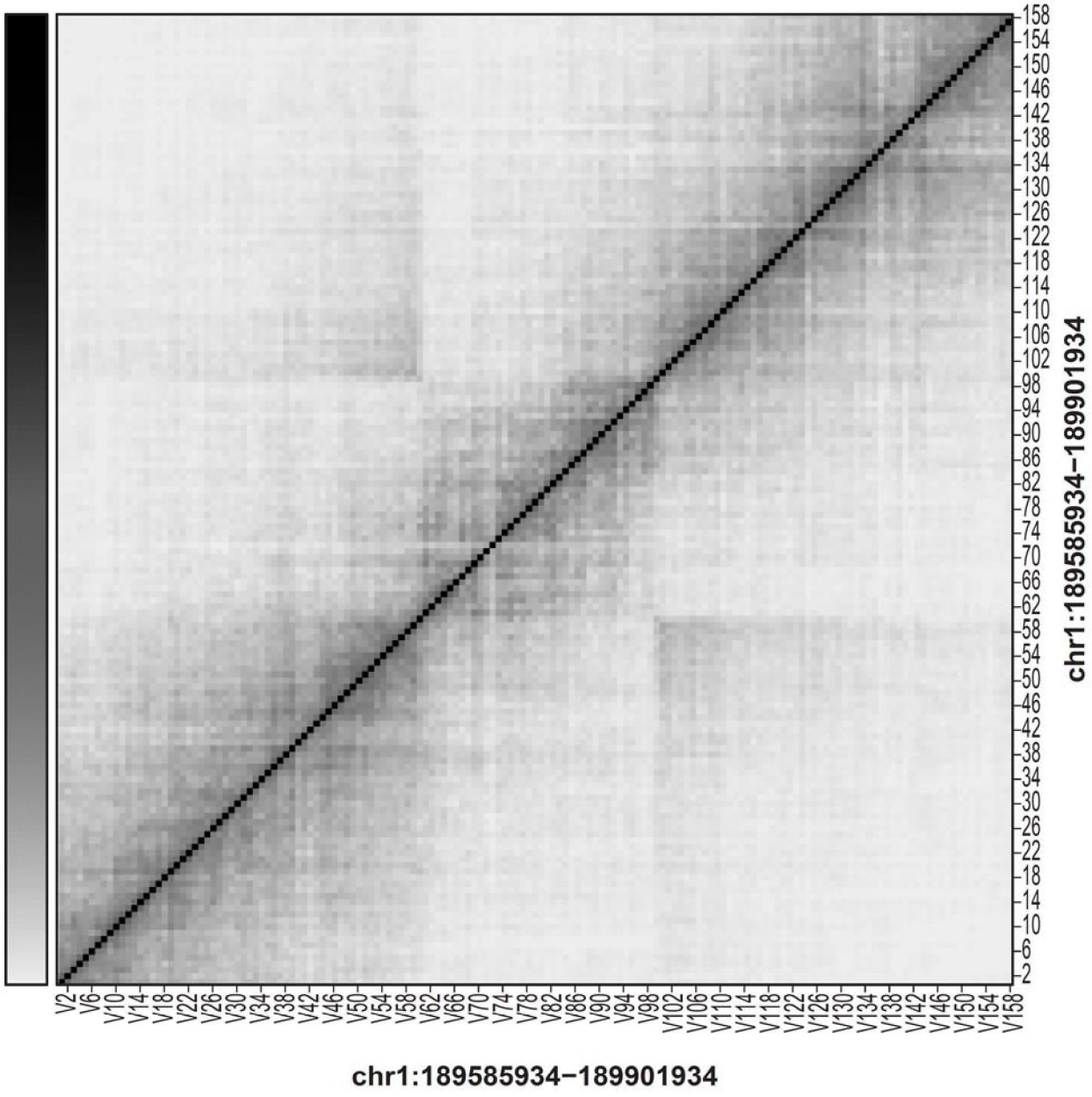
第8I圖


chr16:62532500-62562500
chr16:62532500-62562500
V5
V10
V15
V20
V25
V30
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30

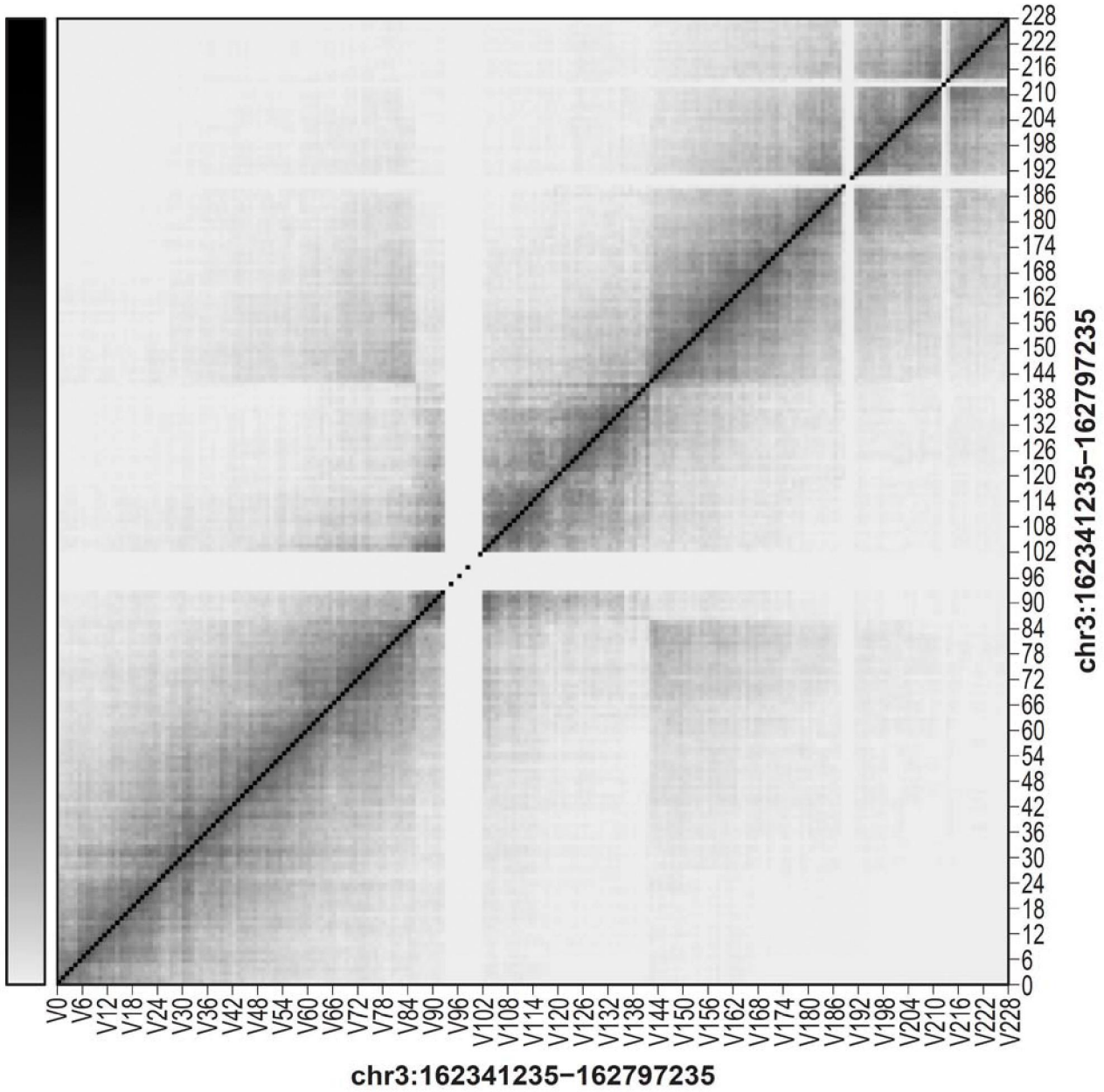
第8J圖



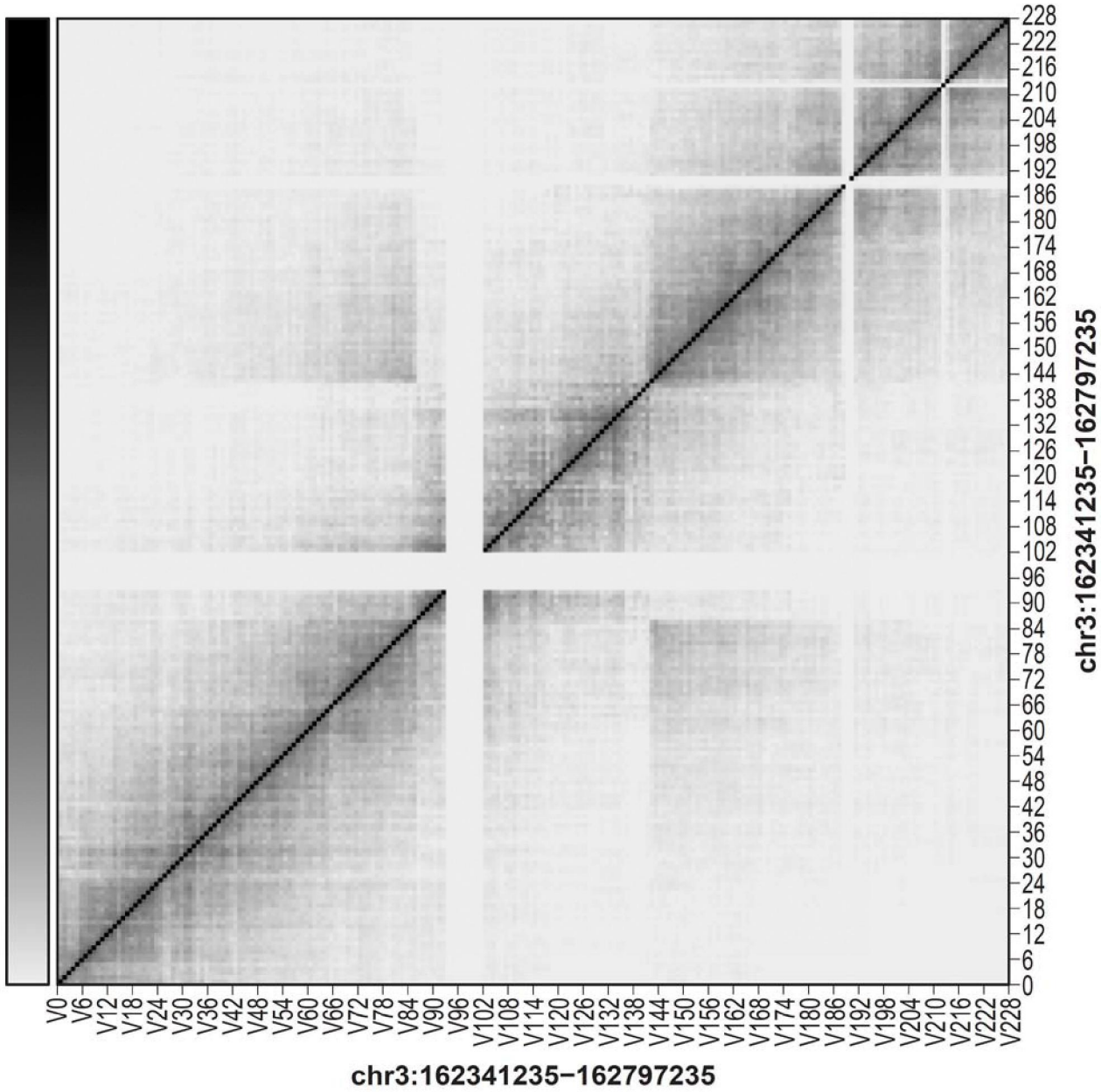
第8K圖



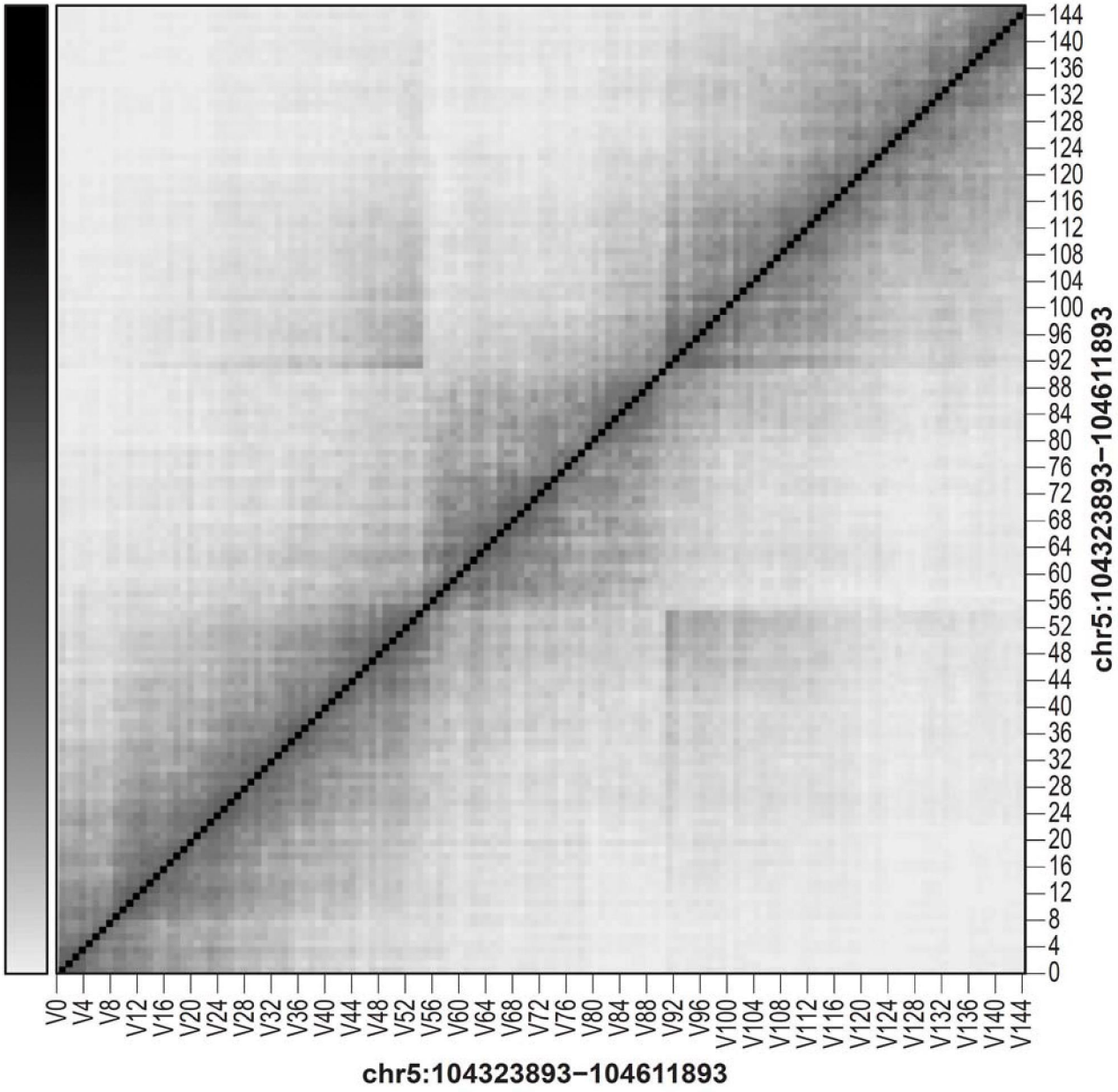
第8L圖



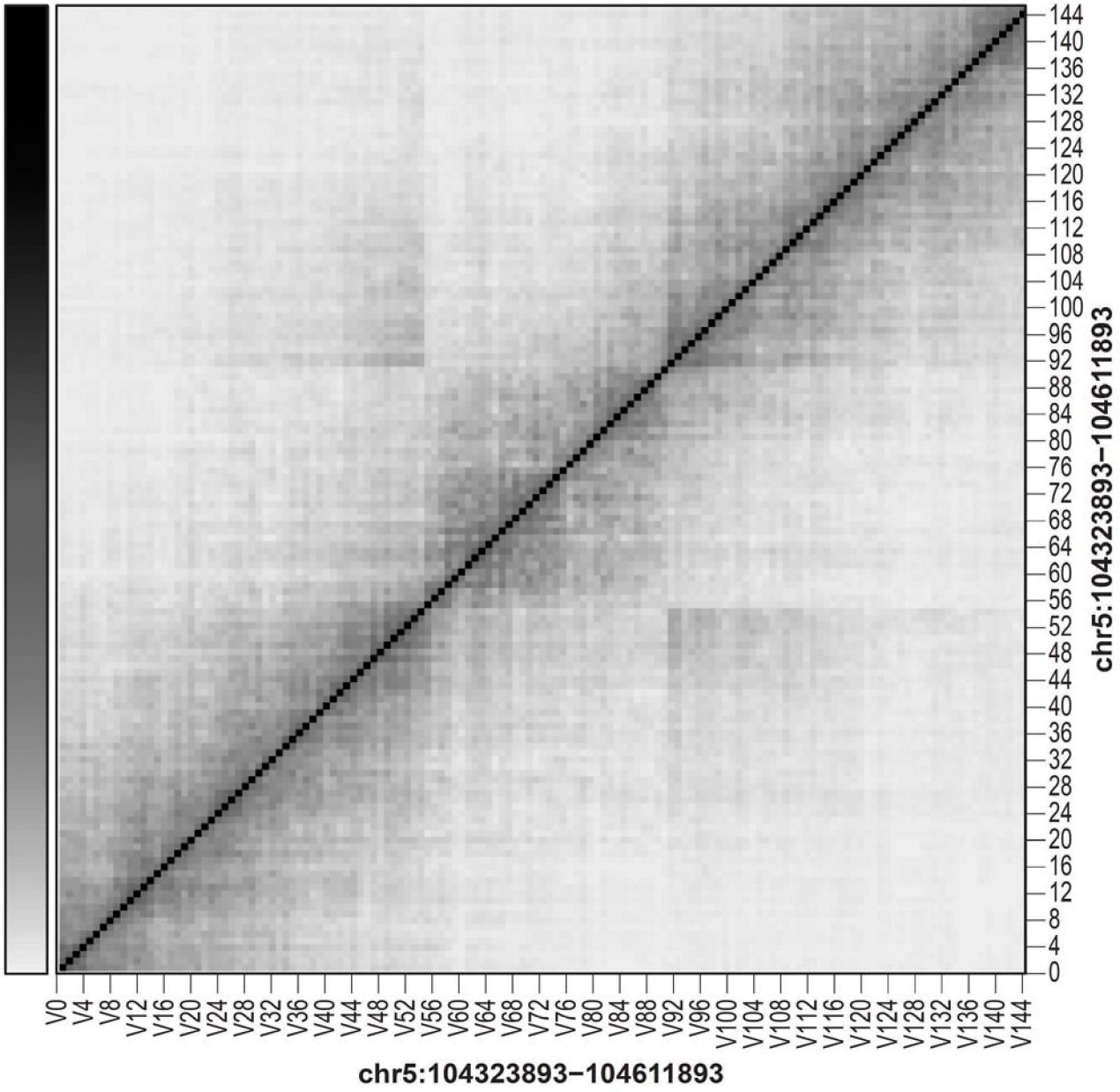
第8M圖



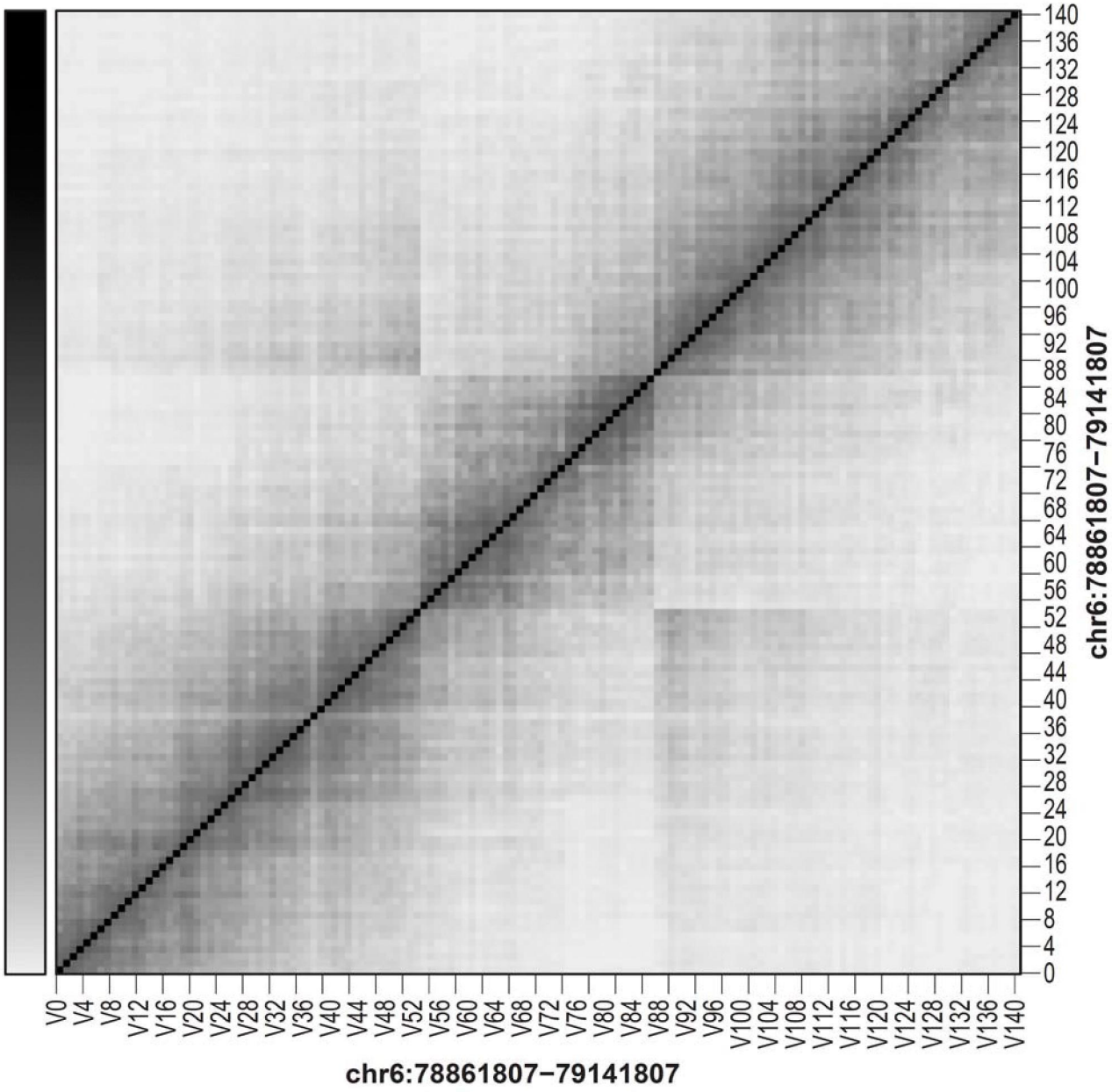
第8N圖



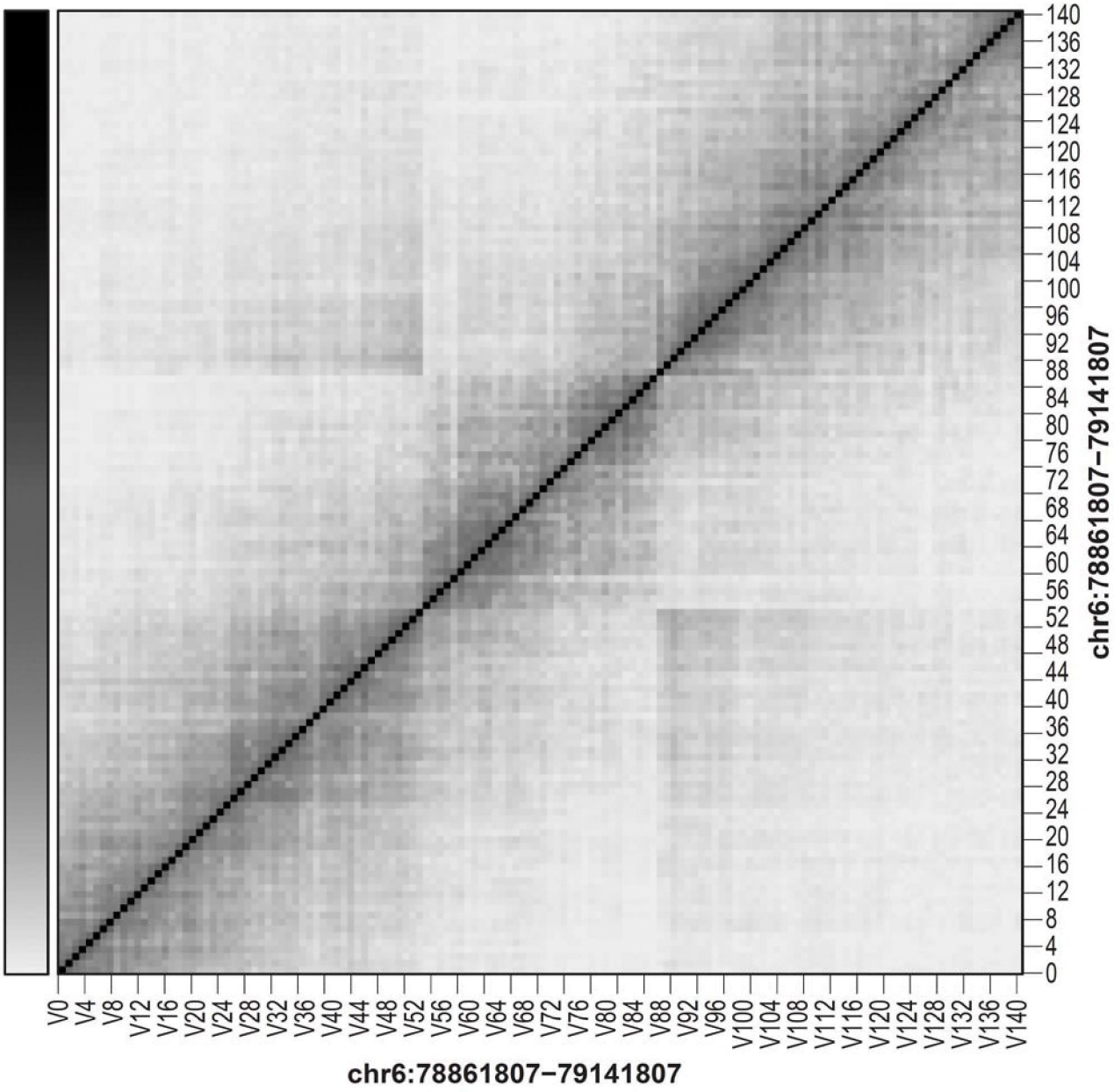
第8O圖



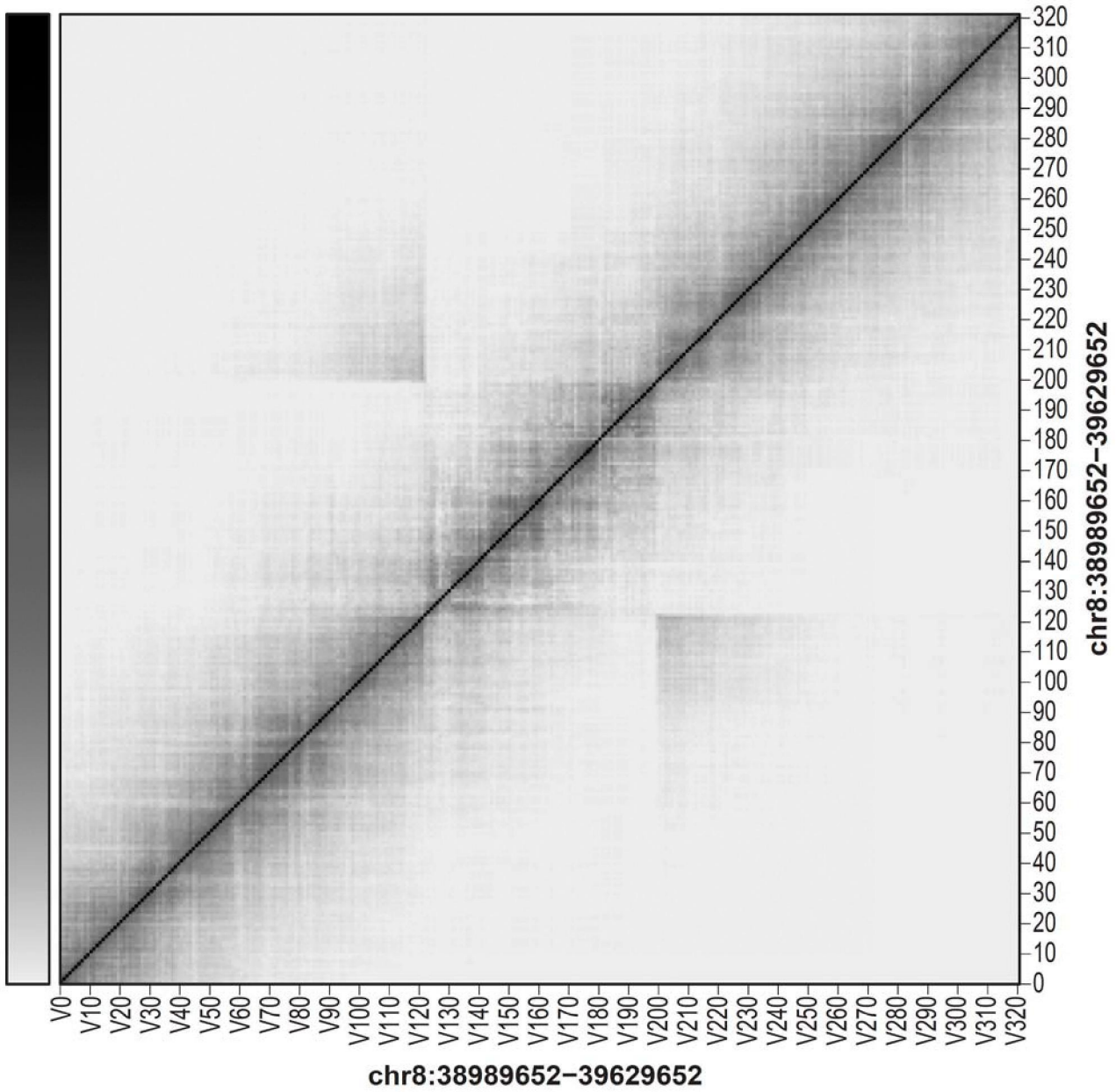
第8P圖



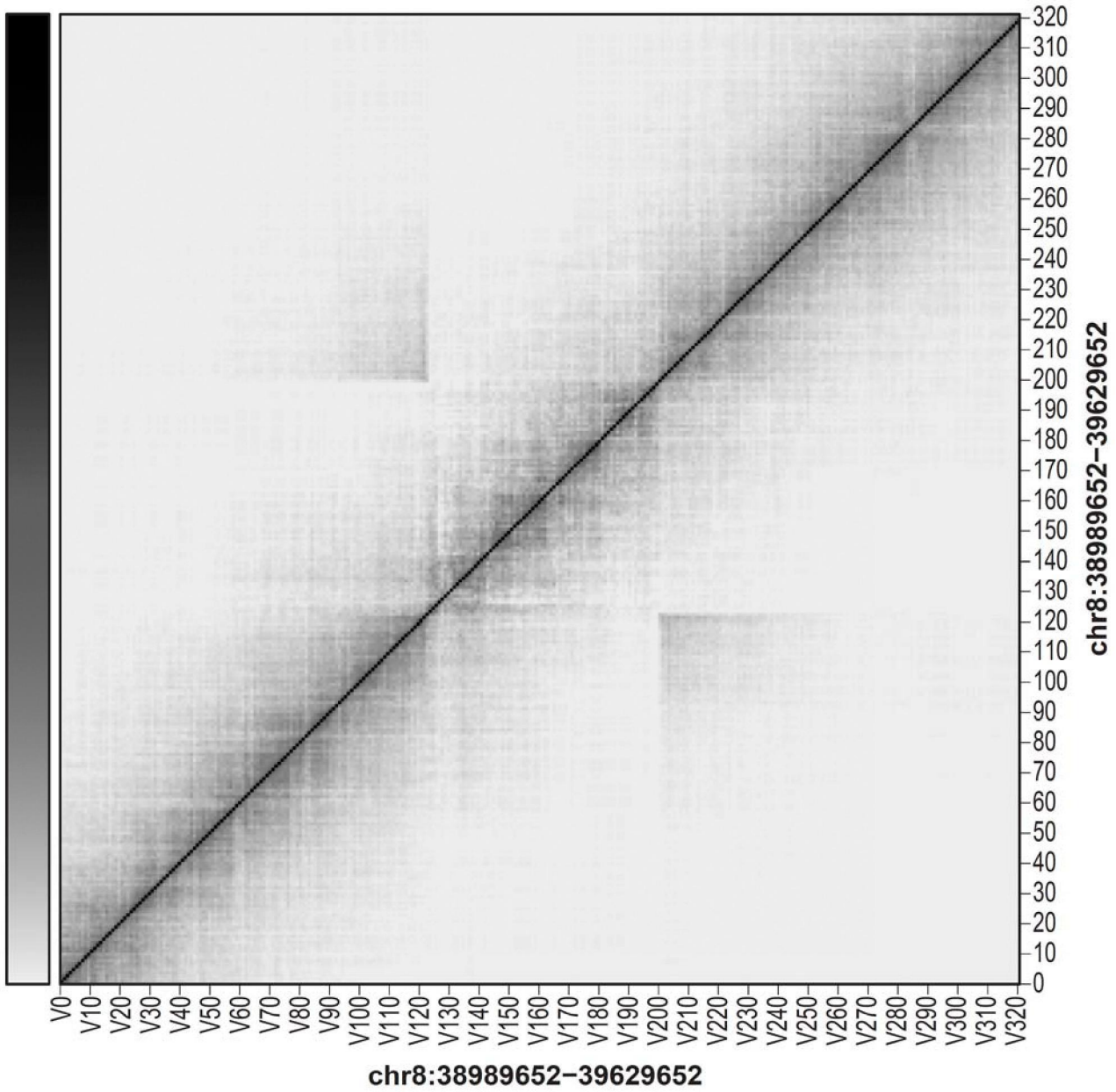
第8Q圖



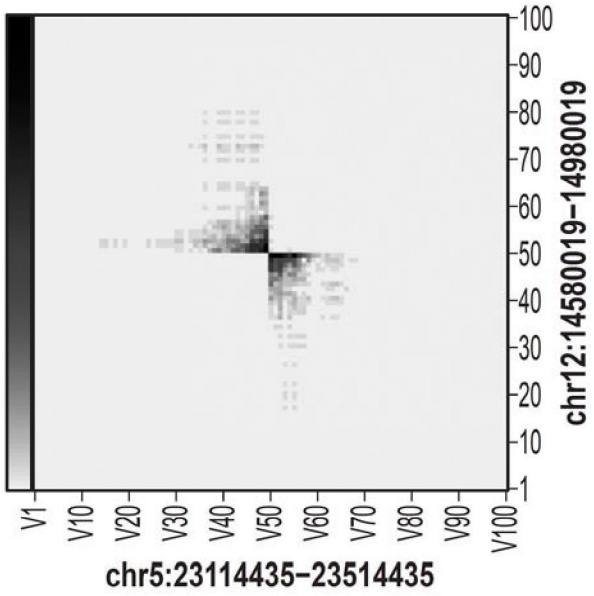
第8R圖



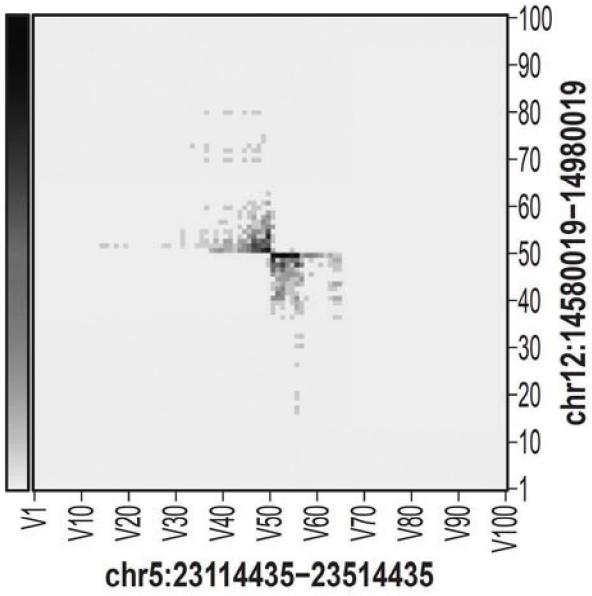
第8S圖



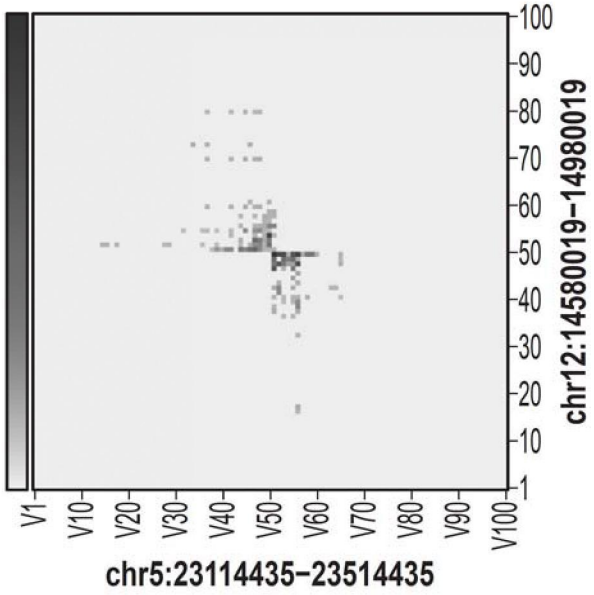
第8T圖



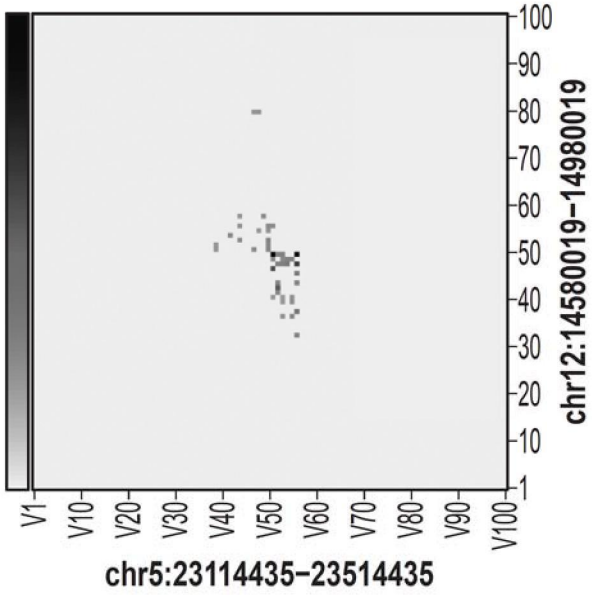
第9A圖



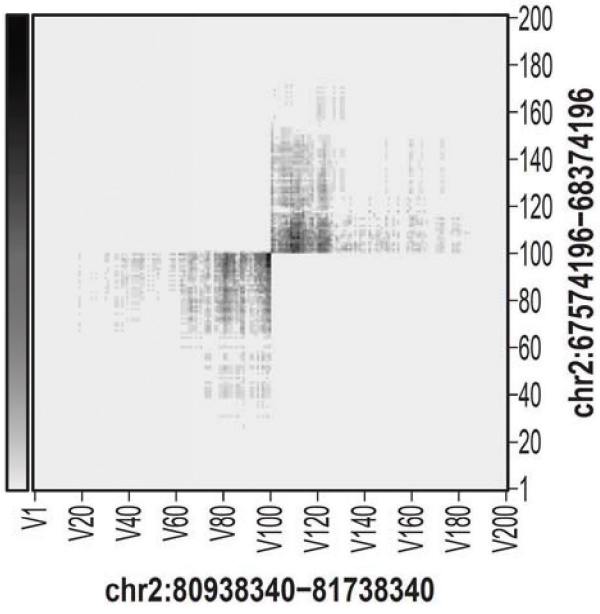
第9B圖



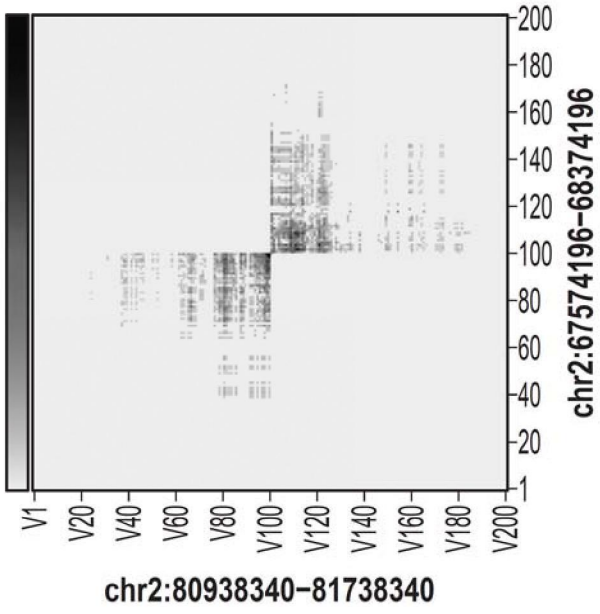
第9C圖



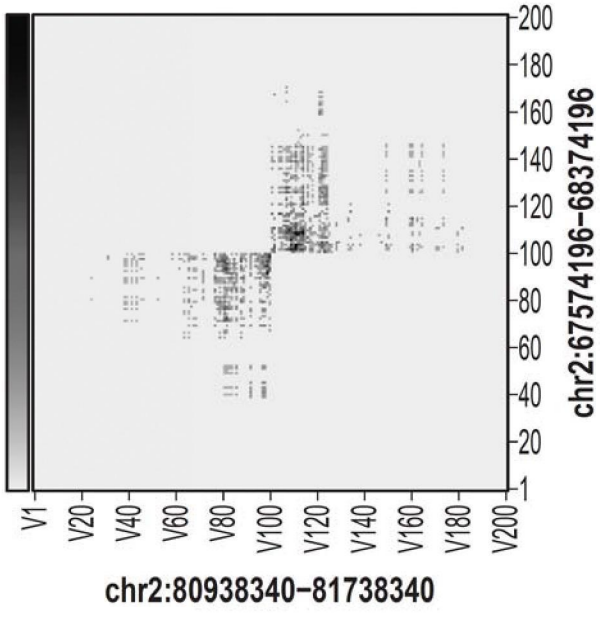
第9D圖



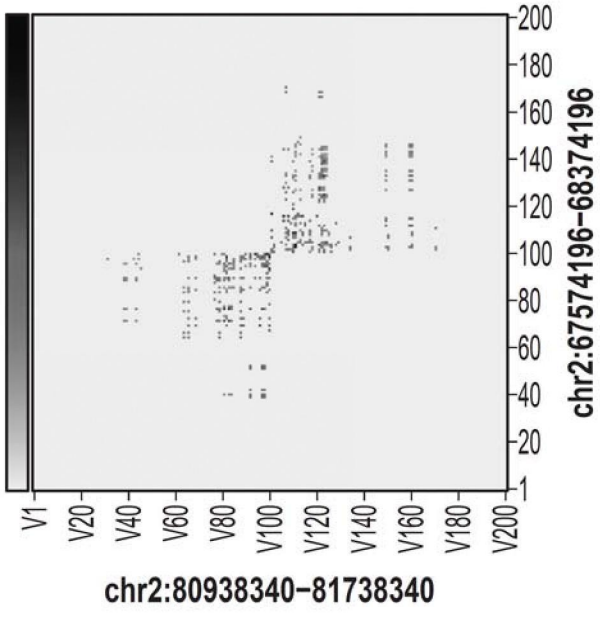
第9E圖



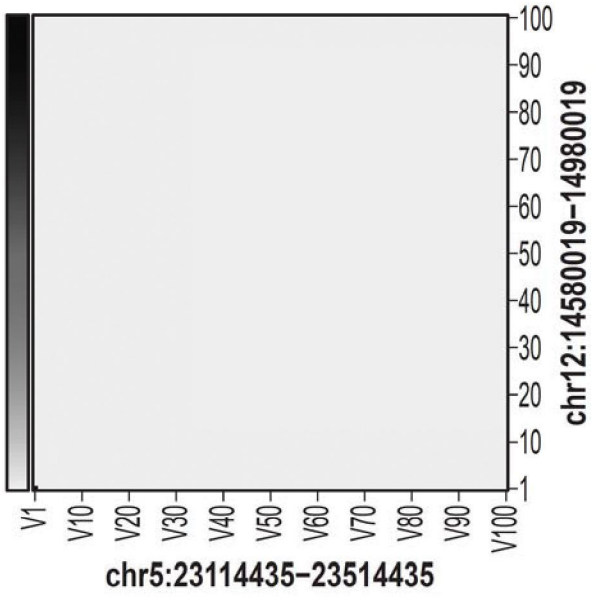
第9F圖



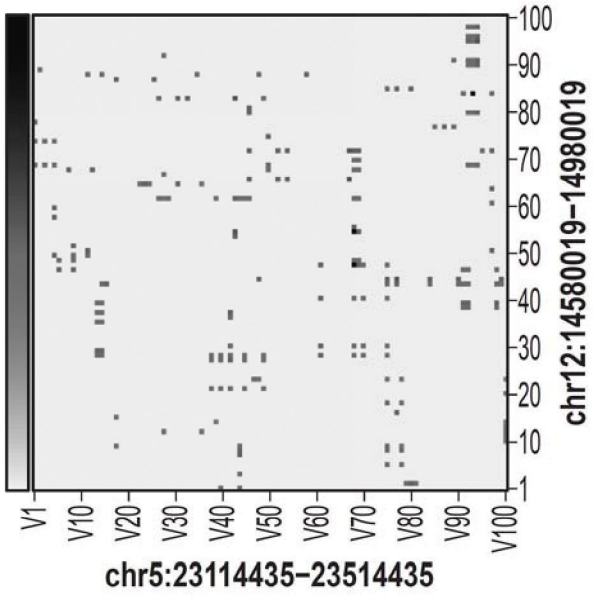
第9G圖



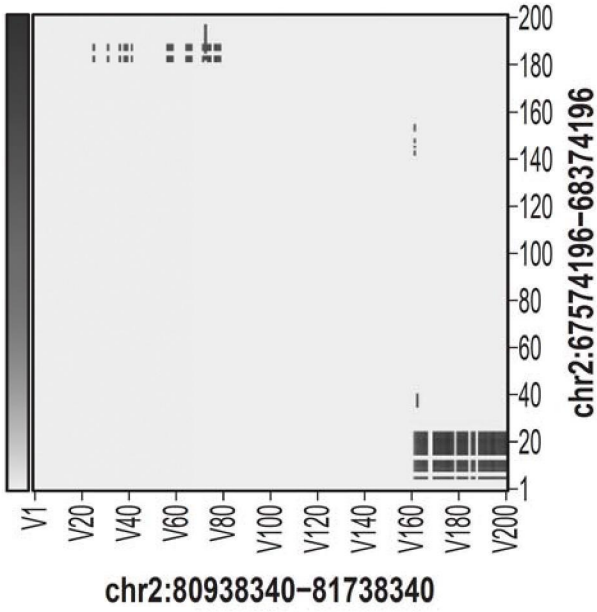
第9H圖



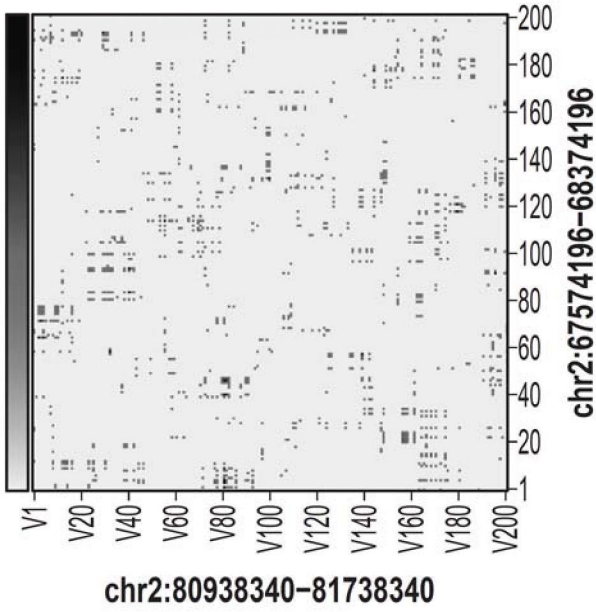
第9I圖



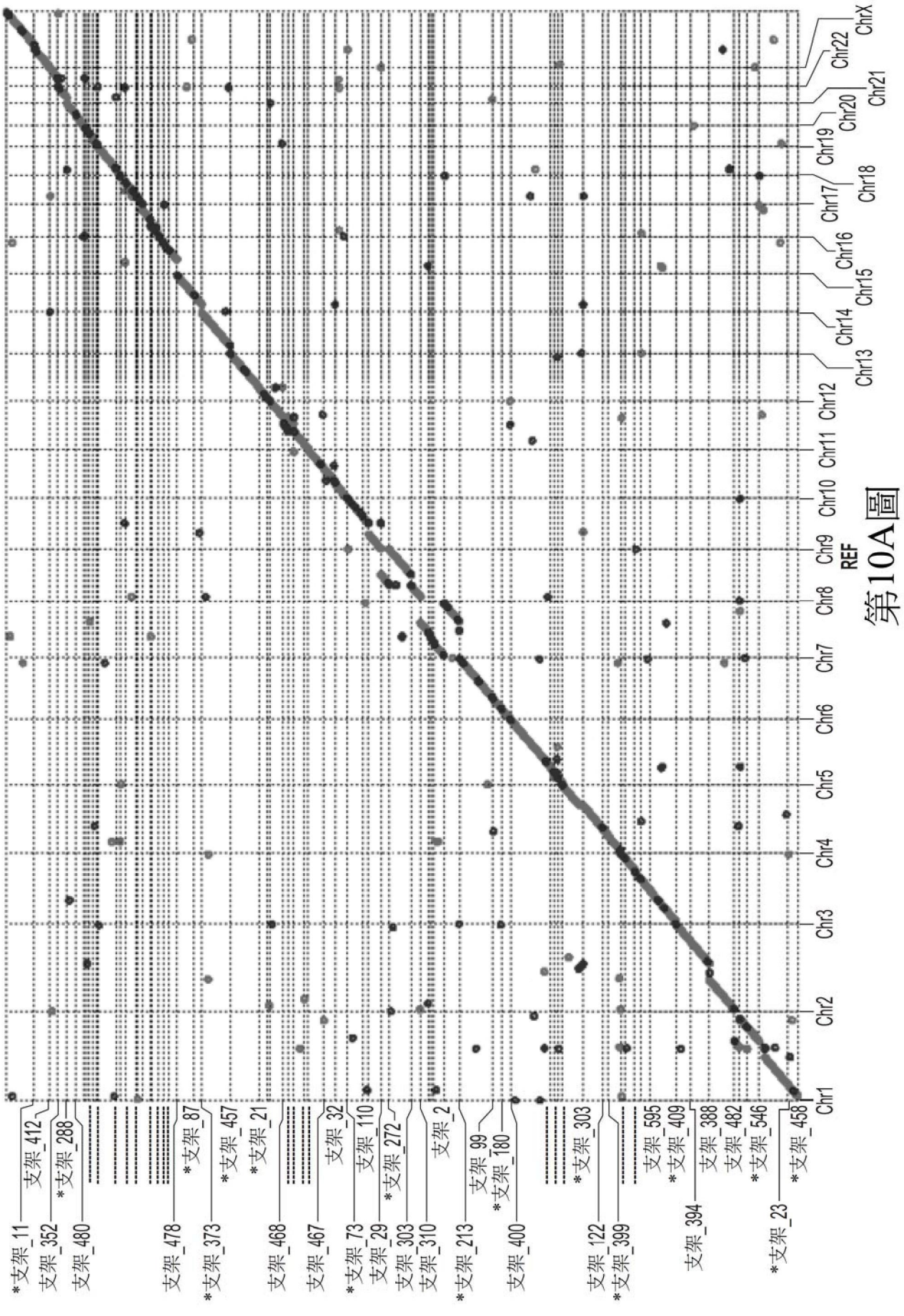
第9J圖



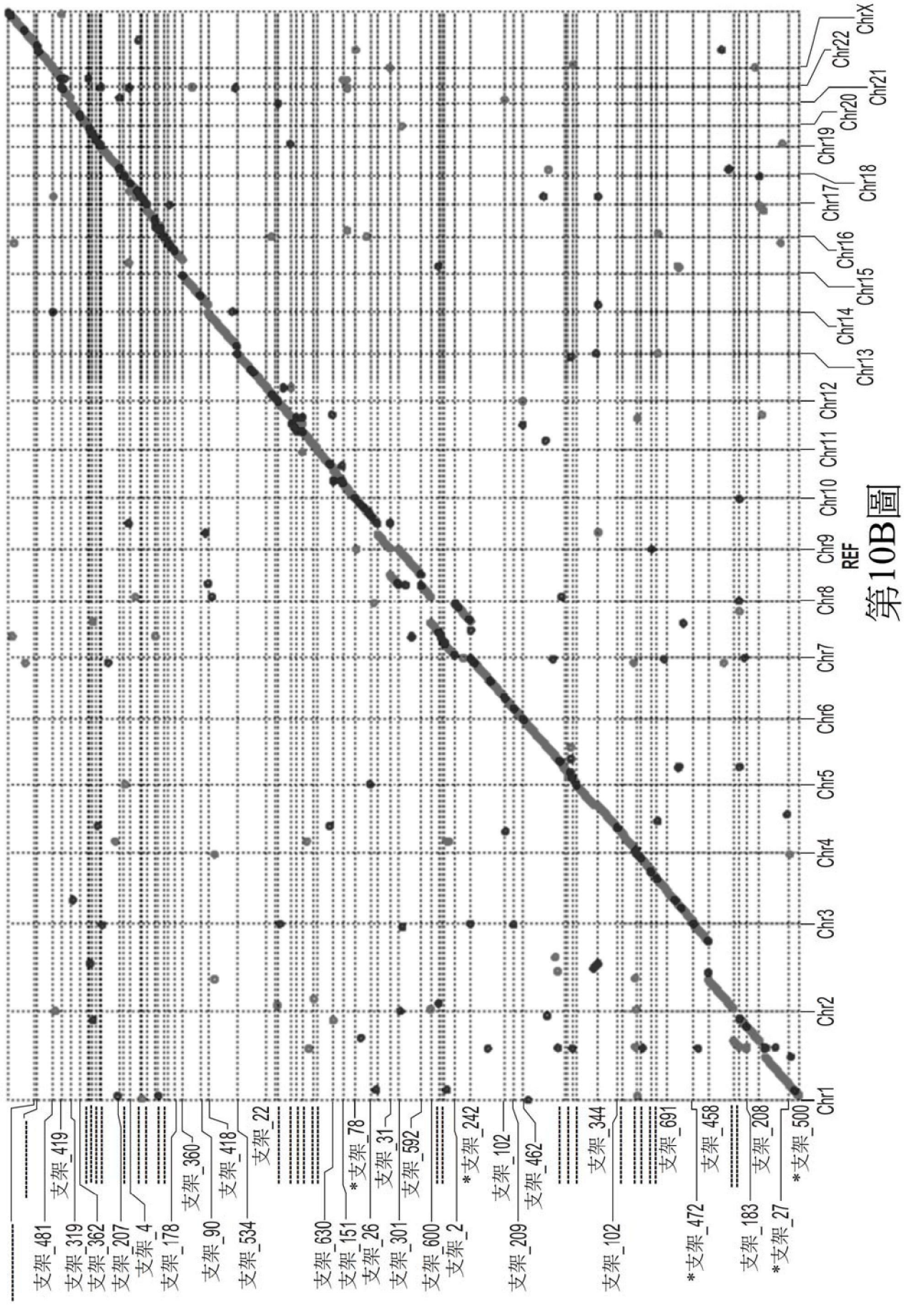
第9K圖



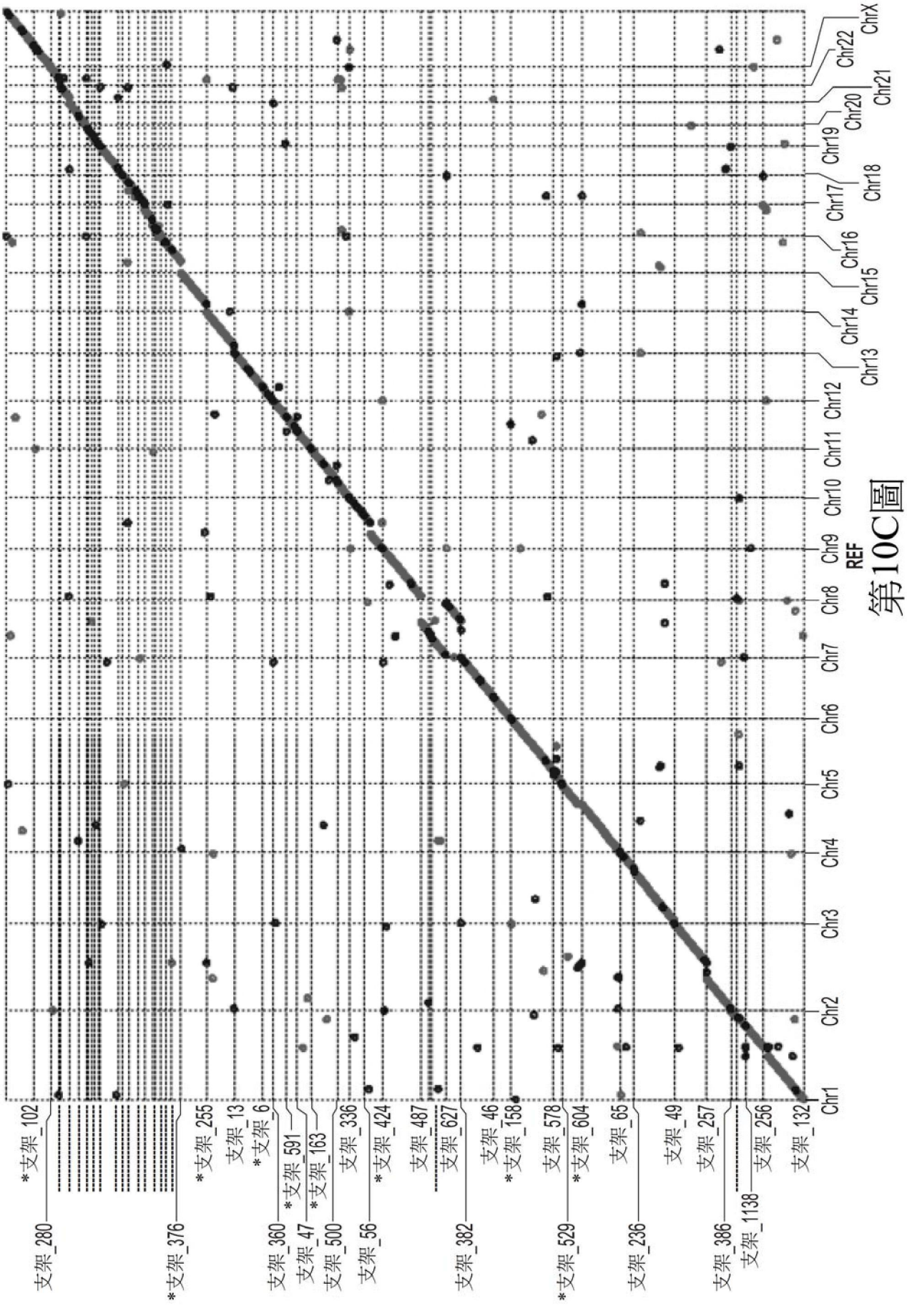
第9L圖


*支架_11
支架_412
支架_352
*支架_288
支架_480
支架_478
*支架_87
*支架_373
*支架_457
*支架_21
支架_468
支架_467
支架_32
*支架_73
支架_110
支架_29
*支架_272
支架_303
支架_310
*支架_213
支架_2
支架_99
*支架_180
支架_400
*支架_303
支架_122
*支架_399
支架_595
*支架_409
支架_394
支架_388
支架_482
*支架_546
*支架_23
*支架_458
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
REF
QRY

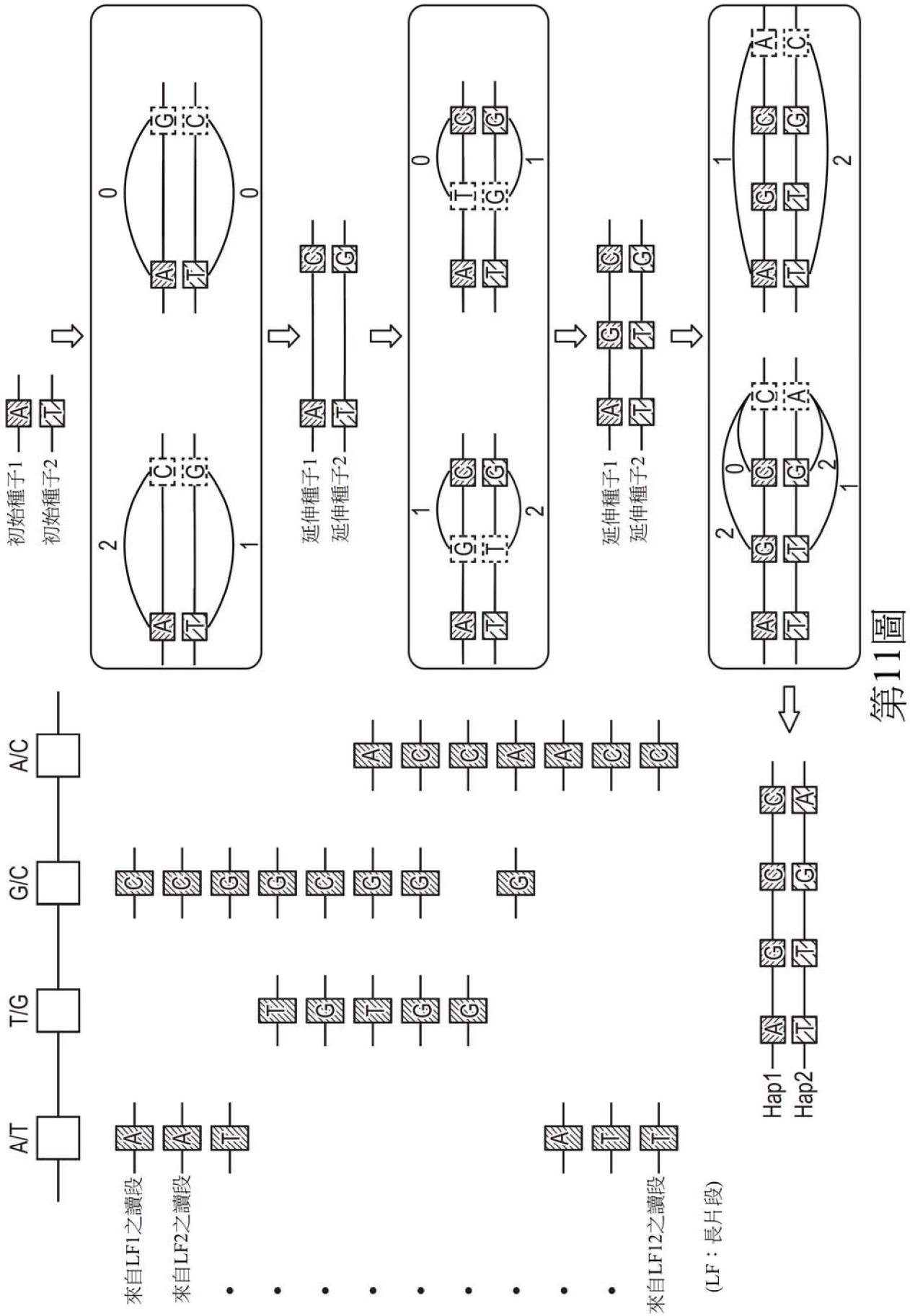
第10A圖


支架_481
支架_419
支架_319
支架_362
支架_207
支架_4
支架_178
支架_360
支架_90
支架_418
支架_534
支架_22
支架_630
支架_151
*支架_78
支架_26
支架_31
支架_301
支架_592
支架_600
支架_2
*支架_242
支架_102
支架_209
支架_462
支架_344
支架_102
支架_691
*支架_472
支架_458
支架_183
支架_208
*支架_27
*支架_500
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
REF
ORY

第10B圖


*支架_102
支架_280
*支架_376
*支架_255
支架_13
*支架_6
支架_360
*支架_591
支架_47
*支架_163
支架_500
支架_336
支架_56
*支架_424
支架_487
支架_627
支架_382
支架_46
*支架_158
支架_578
*支架_604
*支架_529
支架_65
支架_236
支架_49
支架_257
支架_386
支架_1138
支架_256
支架_132
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
Chr9
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
ChrX
REF
QRY

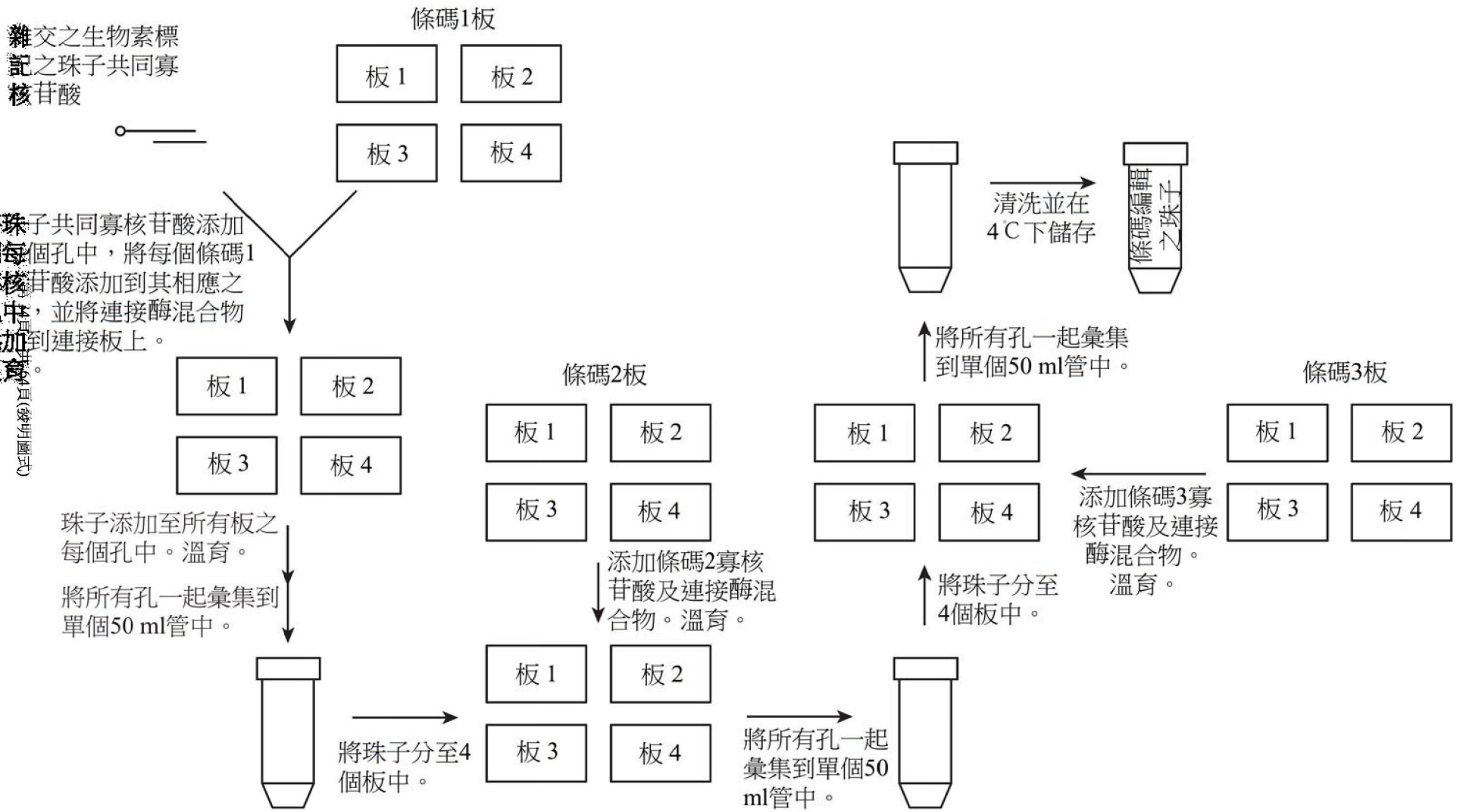
第10C圖


初始種子1
初始種子2
延伸種子1
延伸種子2
延伸種子1
延伸種子2
A/T
T/G
G/C
A/C
來自LF1之讀段
來自LF2之讀段
來自LF12之讀段
(LF：長片段)
Hap1
Hap2

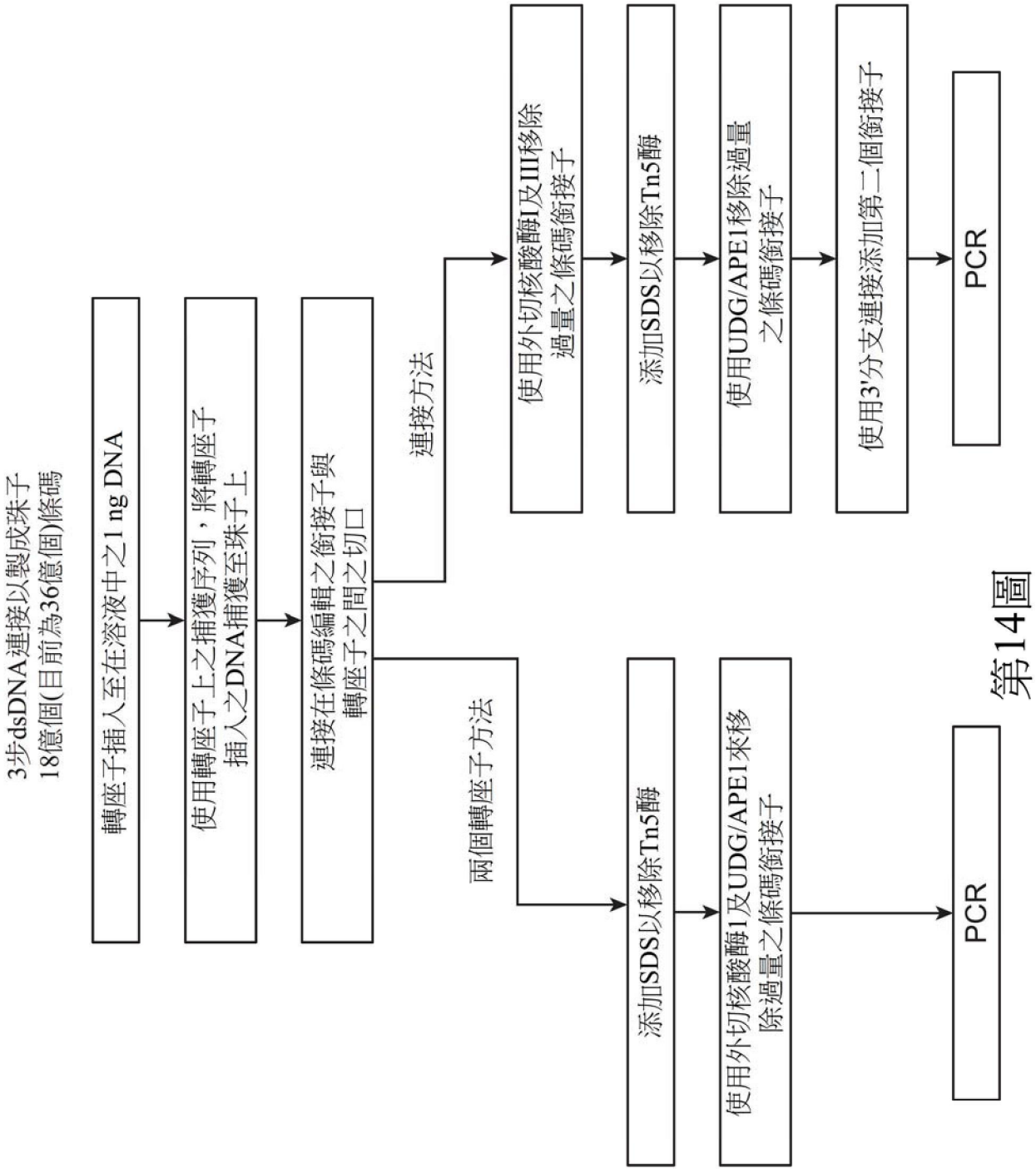
第11圖


/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTG CTCTGG [條碼] TCTGCG
GGGCATCGGTACAGCAAGACGAGACC
將第一個條碼連接到共同銜接子
/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTGCTCTGG [條碼] TCTGCG
GGGCATCGGTACAGCAAGACGAGACC [條碼]
與鏈黴抗生物素蛋白塗覆之珠子結合
/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTGCTCTGG [條碼] TCTGCG
GGGCATCGGTACAGCAAGACGAGACC [條碼]
將第二個條碼連接到共同銜接子
/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTGCTCTGG [條碼] TCTGCG [條碼]
GGGCATCGGTACAGCAAGACGAGACC [條碼] AGACGC [條碼] GGAAGG
將第三個條碼連接到共同銜接子
/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTGCTCTGG [條碼] TCTGCG [條碼] CCTTCC [條碼] TAGCATGGACTATGG
GGGCATCGGTACAGCAAGACGAGACC [條碼] AGACGC [條碼] GGAAGG [條碼] ATCGTACCTGATACC
變性以移除底部鏈
/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTGCTCTGG [條碼] TCTGCG [條碼] CCTTCC [條碼] TAGCATGGACTATGG
雜交橋寡核苷酸
/52-Bio/AAAAAAAACCCGTAGCCATGTCGTTCTGCTCTGG [條碼] TCTGCG [條碼] CCTTCC [條碼] TAGCATGGACTATGG
AUCGUACCUGAUACCGCUAGGAACCACUAGUACAGCAGUCACG

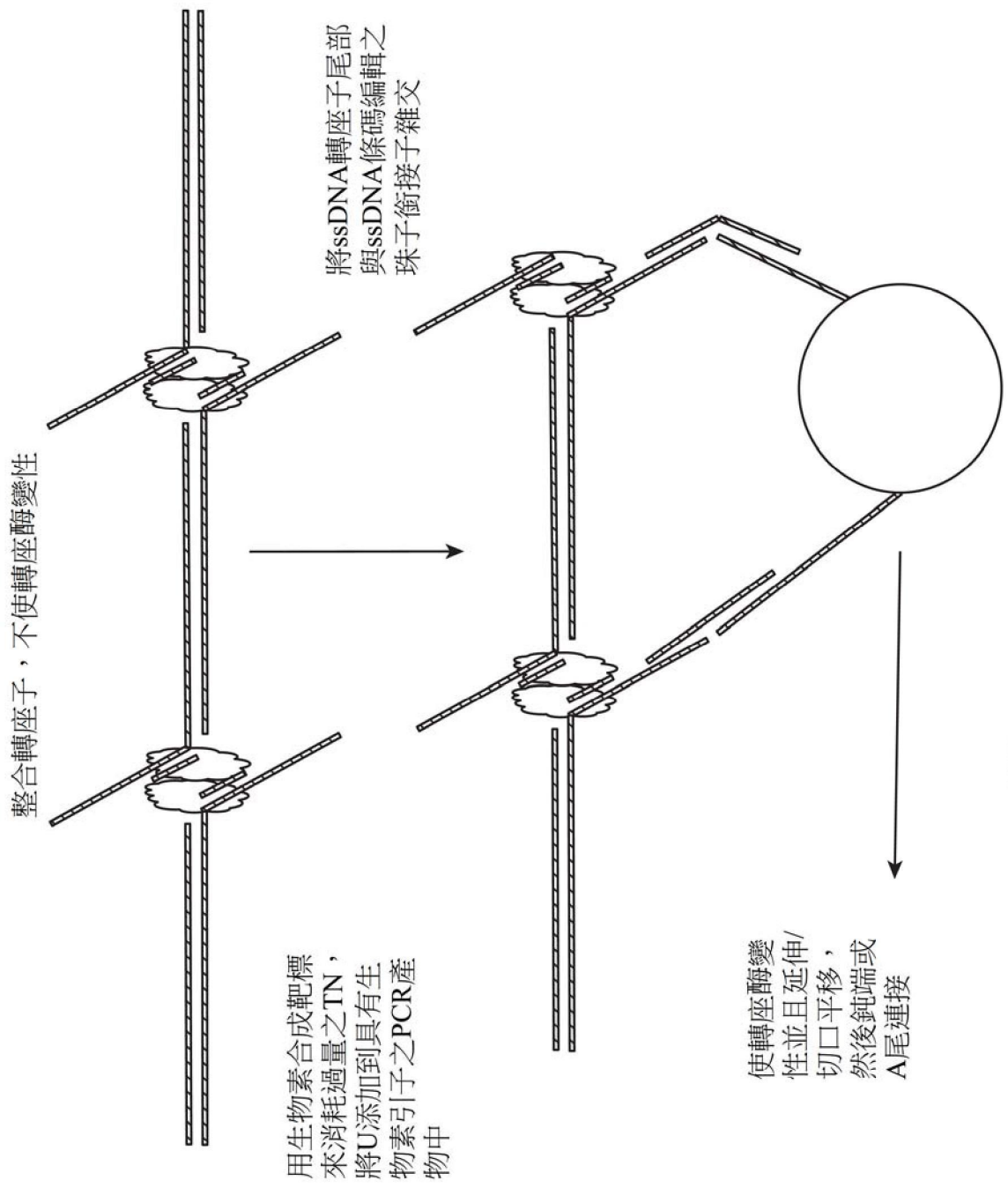
第12圖


條碼1板
雜交之生物素標記之珠子共同寡核苷酸
板 1
板 2
板 3
板 4
珠子共同寡核苷酸添加每個孔中，將每個條碼1寡核苷酸添加到其相應之孔中，並將連接酶混合物添加到連接板上。
溫育。
珠子添加至所有板之每個孔中。溫育。
將所有孔一起彙集到單個50 ml管中。
將珠子分至4個板中。
條碼2板
添加條碼2寡核苷酸及連接酶混合物。溫育。
將所有孔一起彙集到單個50 ml管中。
將珠子分至4個板中。
條碼3板
添加條碼3寡核苷酸及連接酶混合物。溫育。
將所有孔一起彙集到單個50 ml管中。
清洗並在4℃下儲存
條碼編輯之珠子

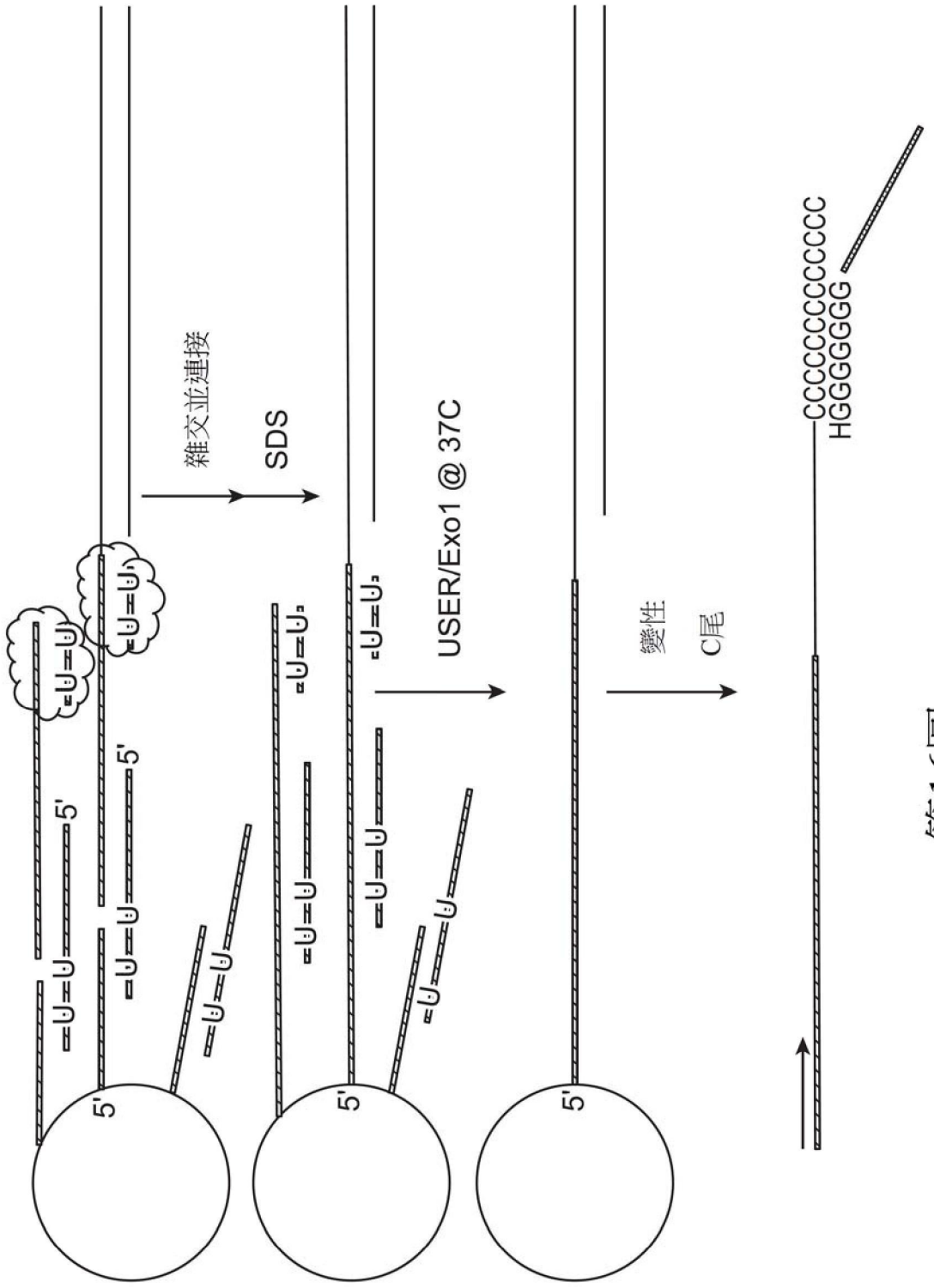
第13圖


3步dsDNA連接以製成珠子
18億個(目前為36億個)條碼
轉座子插入至在溶液中之1 ng DNA
使用轉座子上之捕獲序列，將轉座子插入之DNA捕獲至珠子上
連接在條碼編輯之銜接子與轉座子之間之切口
連接方法
使用外切核酸酶I及III移除過量之條碼銜接子
添加SDS以移除Tn5酶
使用UDG/APE1移除過量之條碼銜接子
使用3'分支連接添加第二個銜接子
PCR
兩個轉座子方法
添加SDS以移除Tn5酶
使用外切核酸酶1及UDG/APE1來移除過量之條碼銜接子
PCR

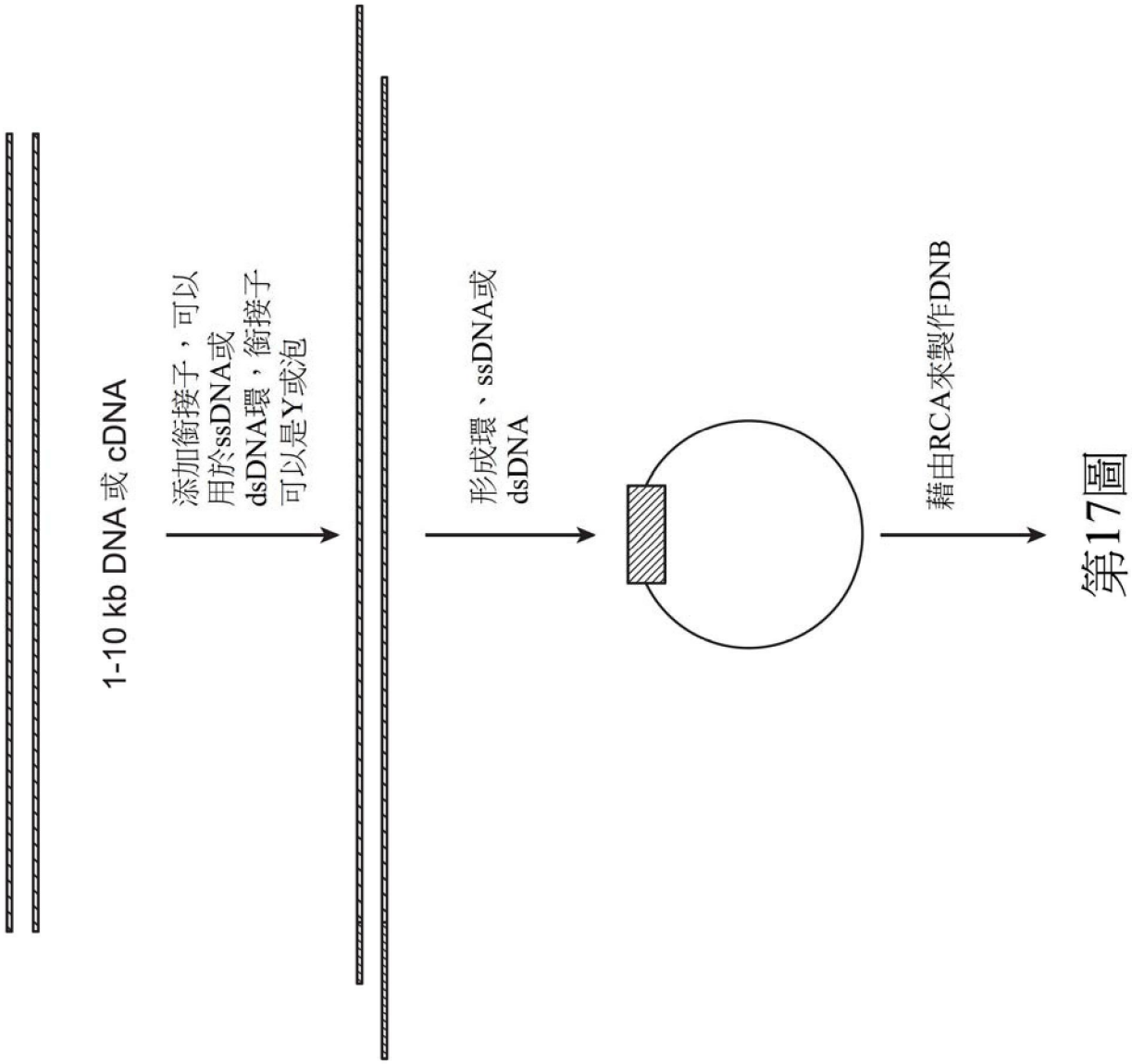
第14圖


整合轉座子，不使轉座酶變性
將ssDNA轉座子尾部與ssDNA條碼編輯之珠子銜接子雜交
用生物素合成靶標來消耗過量之TN，將U添加到具有生物素引子之PCR產物中
使轉座酶變性並且延伸/切口平移，然後鈍端或A尾連接

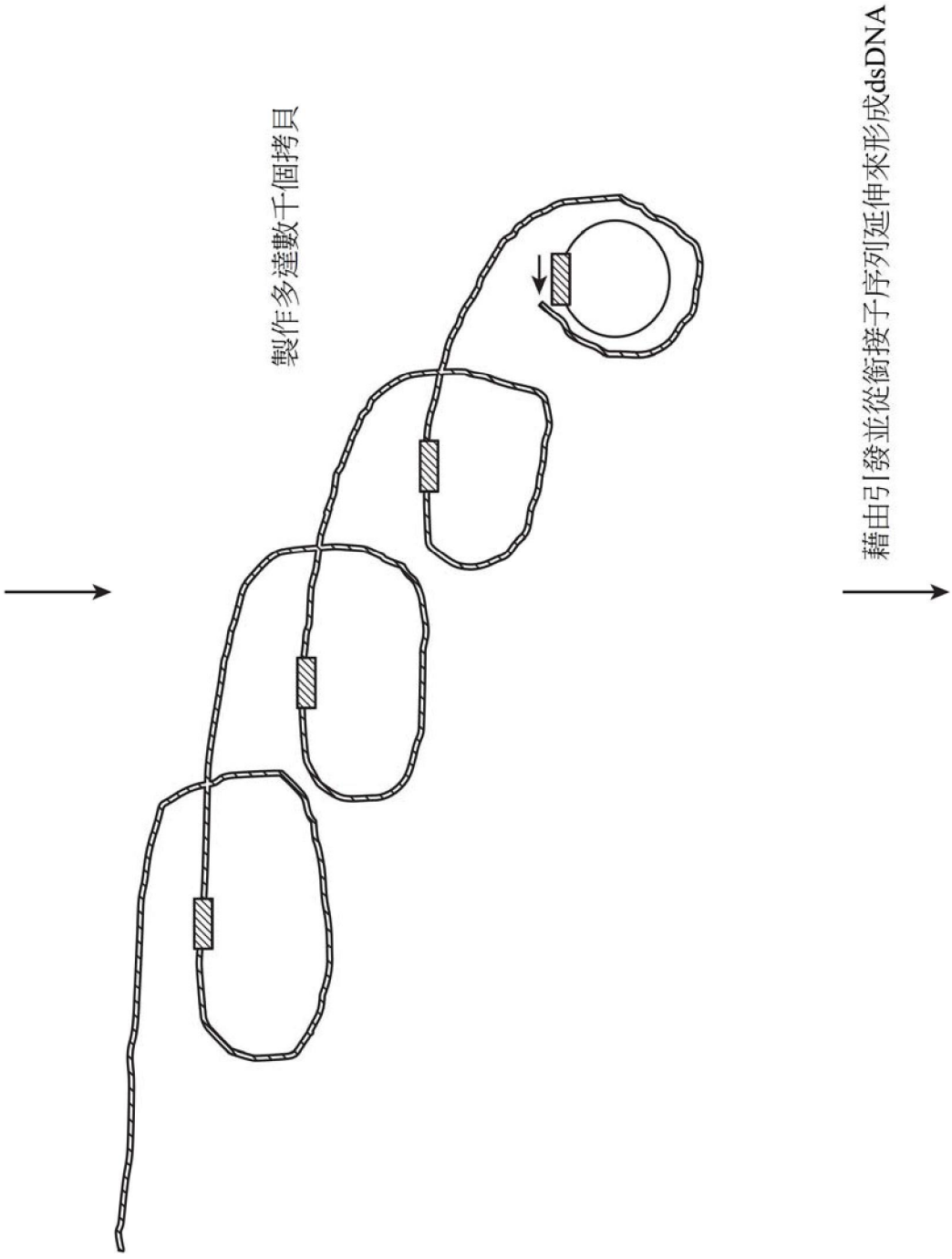
第15圖



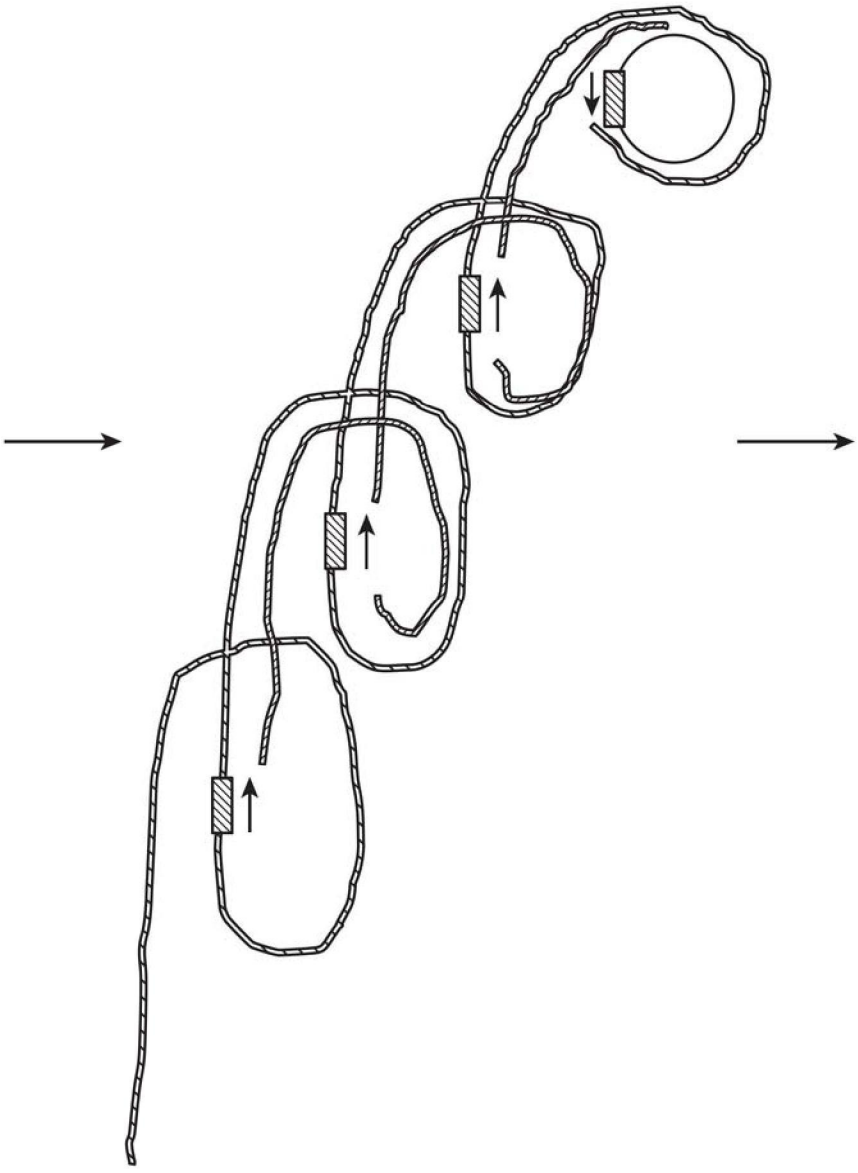
第16圖


1-10 kb DNA 或 cDNA
添加銜接子，可以
用於ssDNA或
dsDNA環，銜接子
可以是Y或泡
形成環、ssDNA或
dsDNA
藉由RCA來製作DNB

第17圖

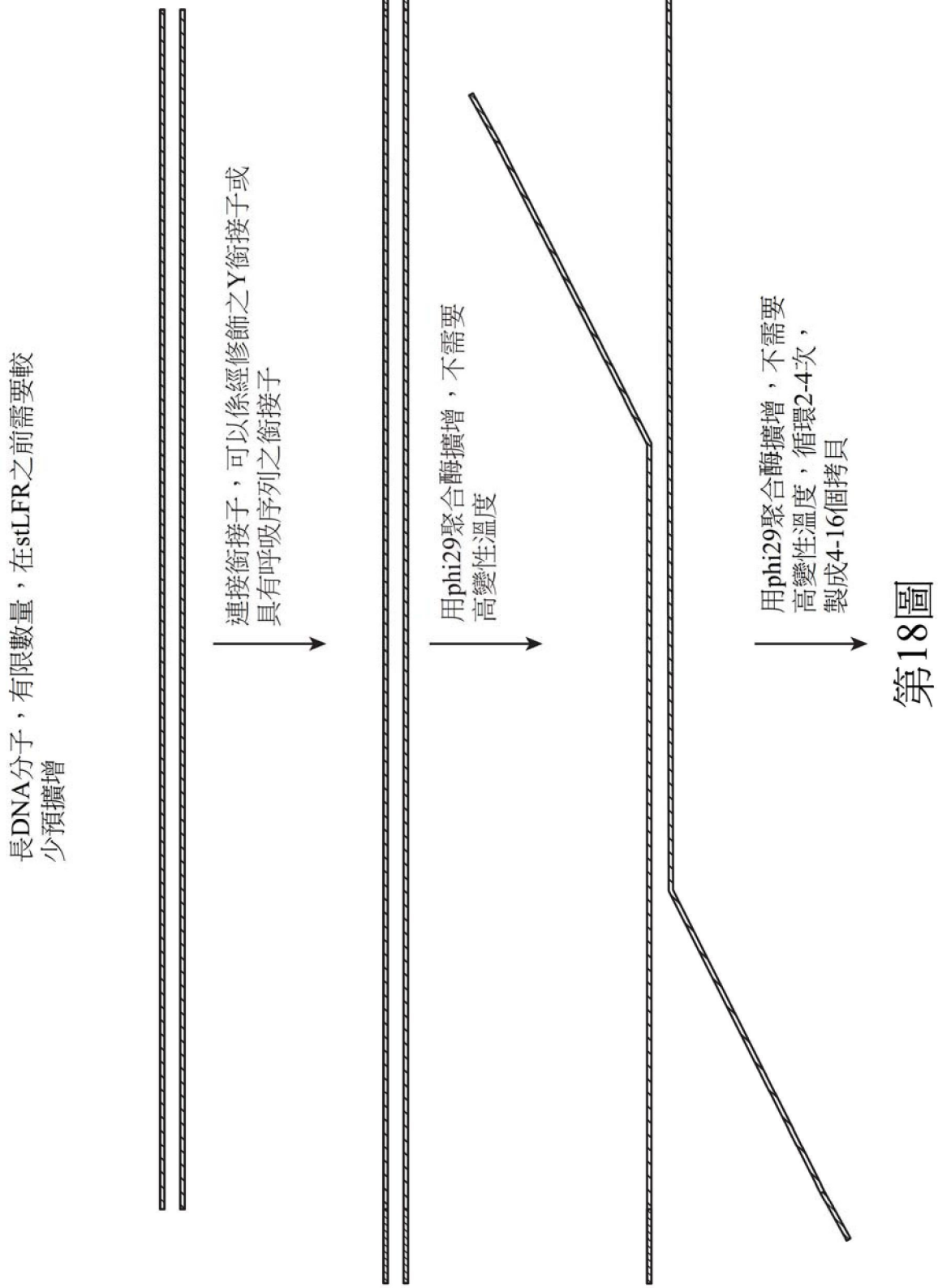


第17圖 (續1)

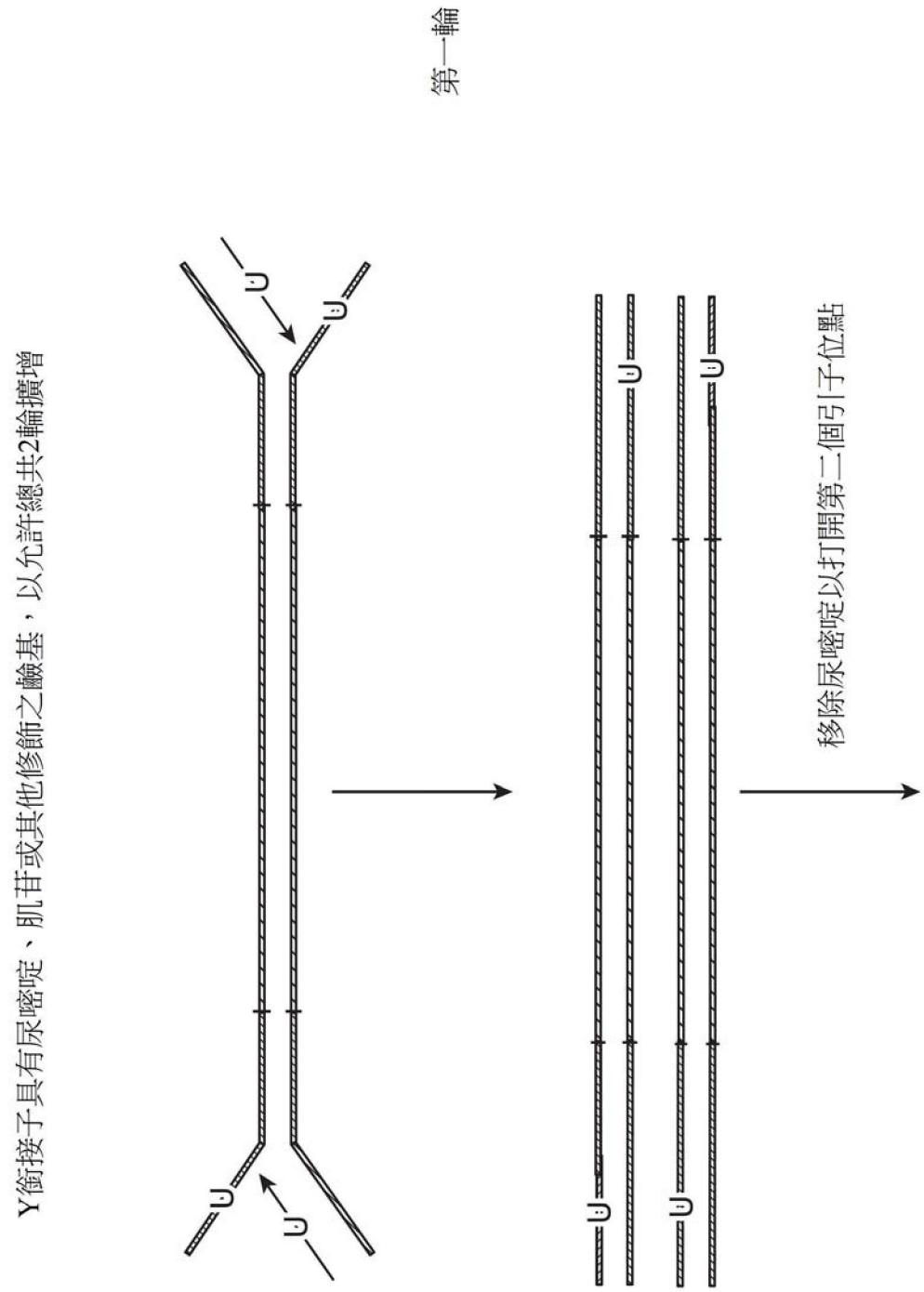


雙鏈DNB現在可以被轉座子插入並被stLFR珠子捕獲，可能產生原始分子之數千個拷貝並且全部在相同之DNA鏈上，此使得能夠藉由stLFR定序來高度覆蓋原始分子

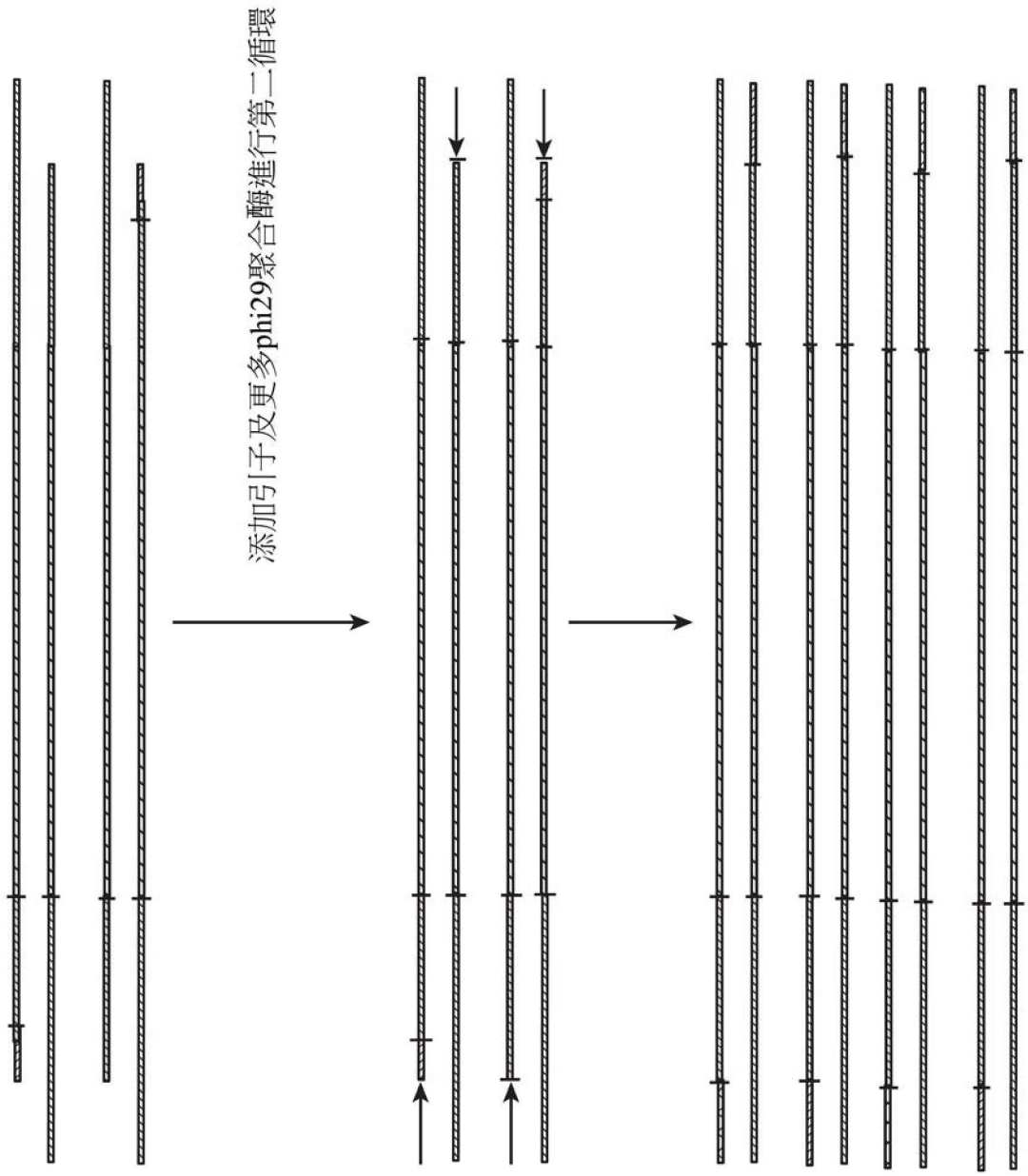
第17圖 (續2)


長DNA分子，有限數量，在stLFR之前需要較少預擴增
連接銜接子，可以係經修飾之Y銜接子或具有呼吸序列之銜接子
用phi29聚合酶擴增，不需要高變性溫度
用phi29聚合酶擴增，不需要高變性溫度，循環2-4次，製成4-16個拷貝

第18圖



第18圖 (續1)

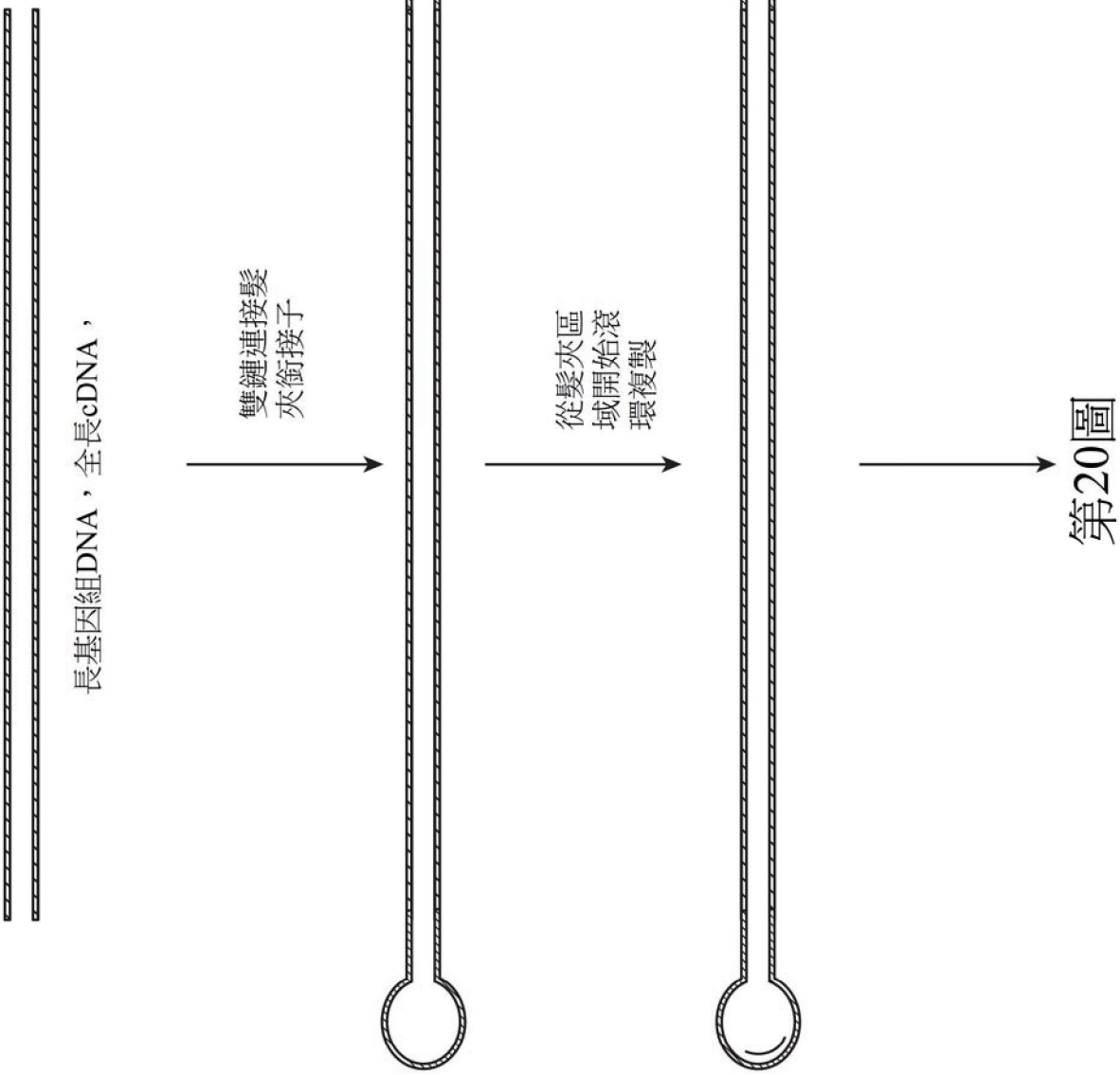


第18圖 (續2)

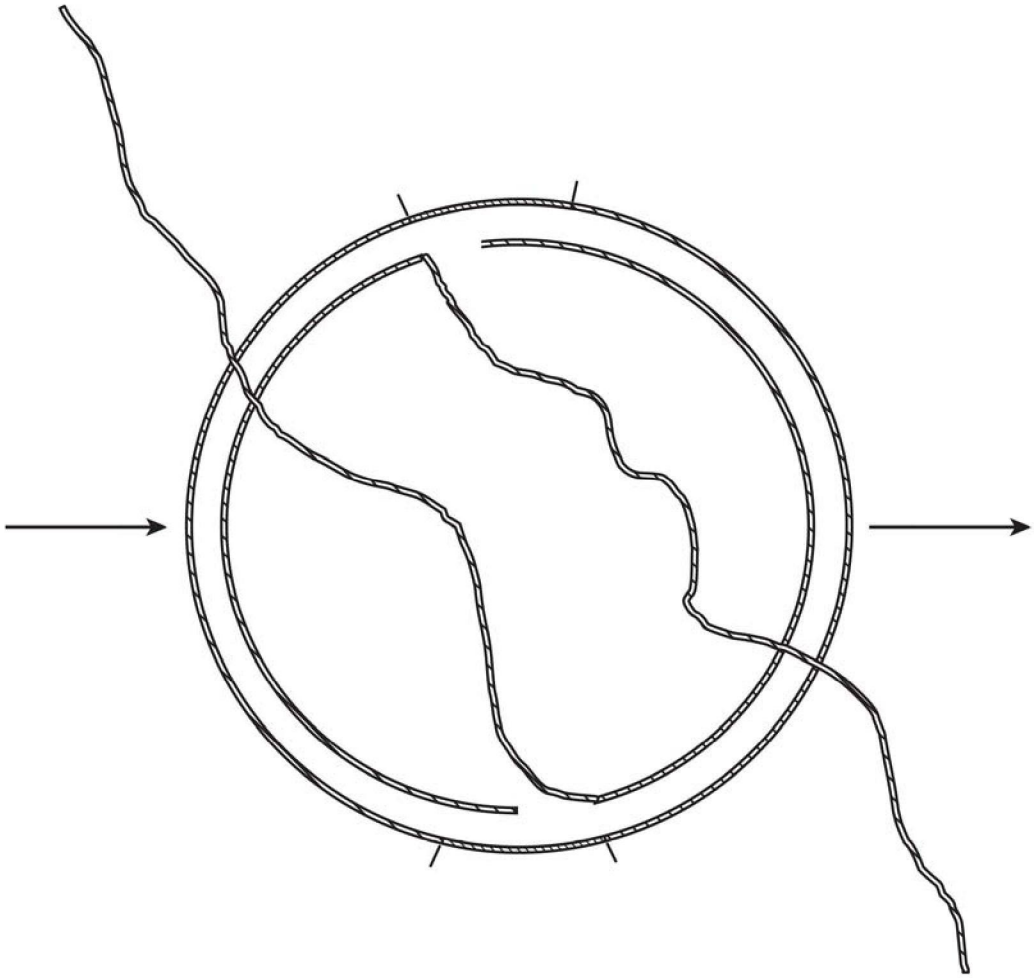
3'連接銜接子
覆蓋珠子
低濃度之隨機切口酶、中等濃度之klenow及高濃度之連接酶。珠子及DNA在stLFR濃度下。當產生切口並藉由Klenow來打開至間隙時，立即連接並將長片段鎖定在珠子上。允許產生切口並且打開更多間隙以使更多銜接子連接到間隙中。

第19圖

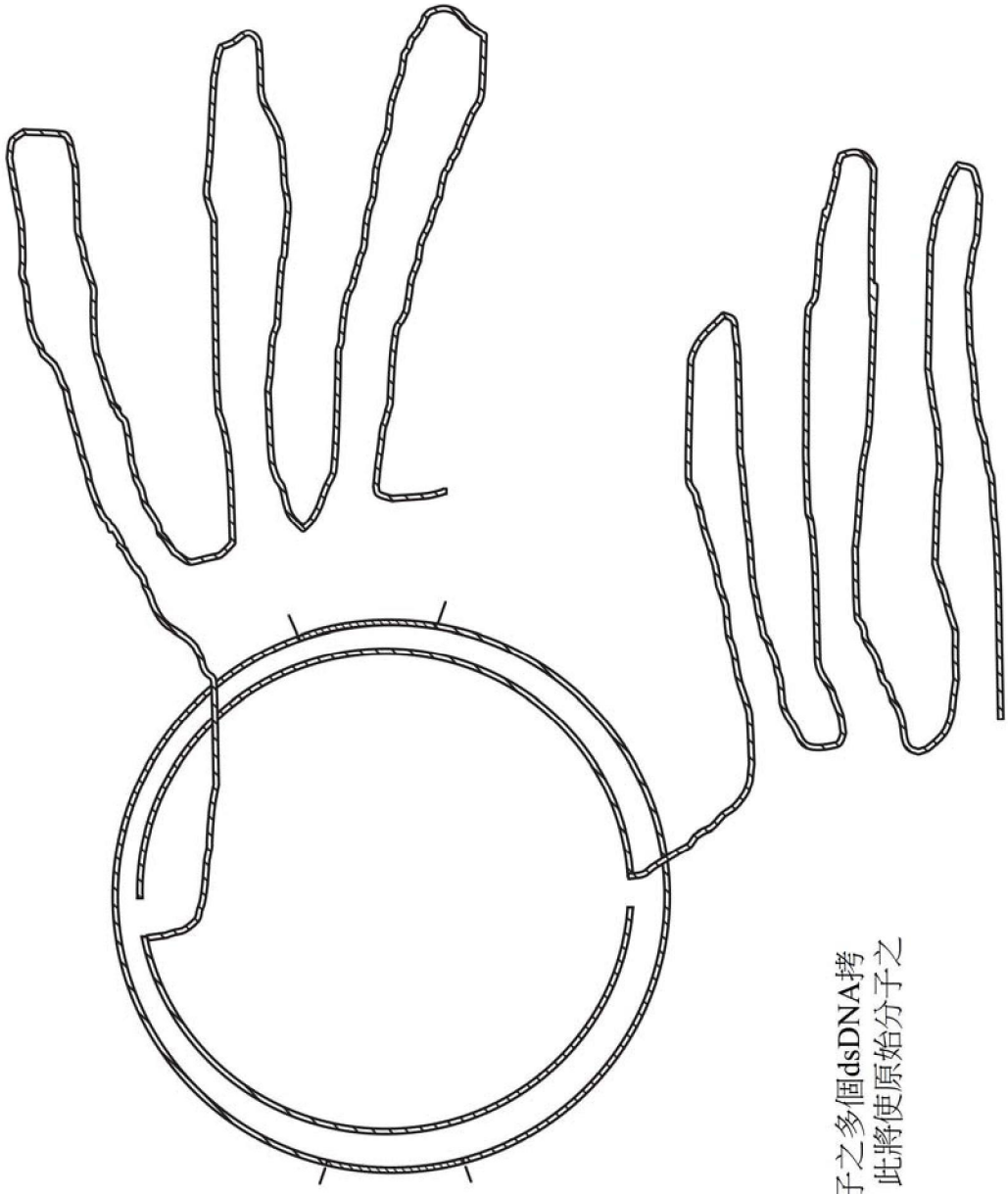
第19圖 (續1)



第20圖

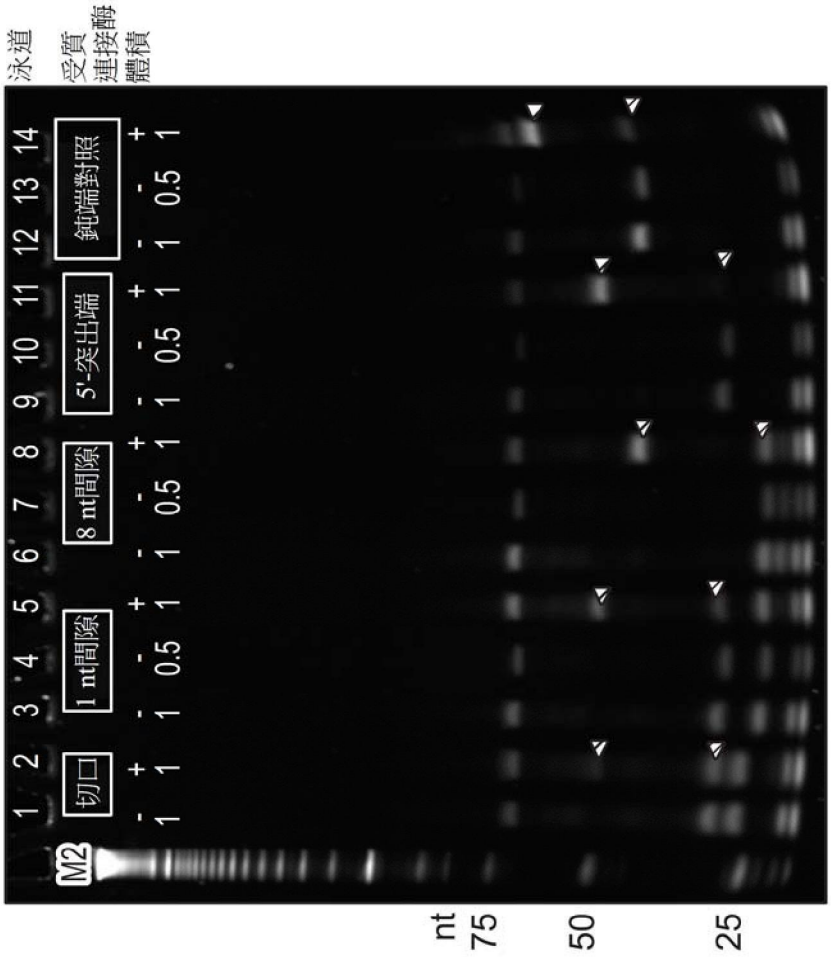


第20圖 (續1)

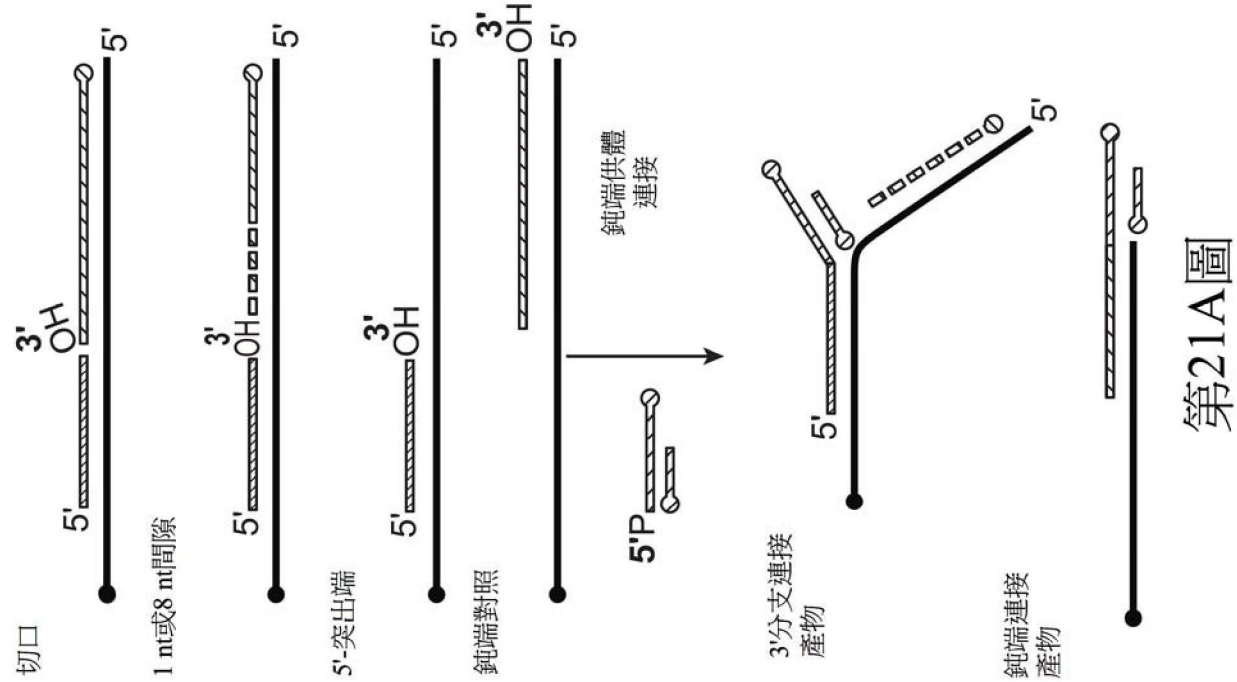


多聯體中之原始分子之多個dsDNA拷貝準備進入stLFR，此將使原始分子之覆蓋率更高

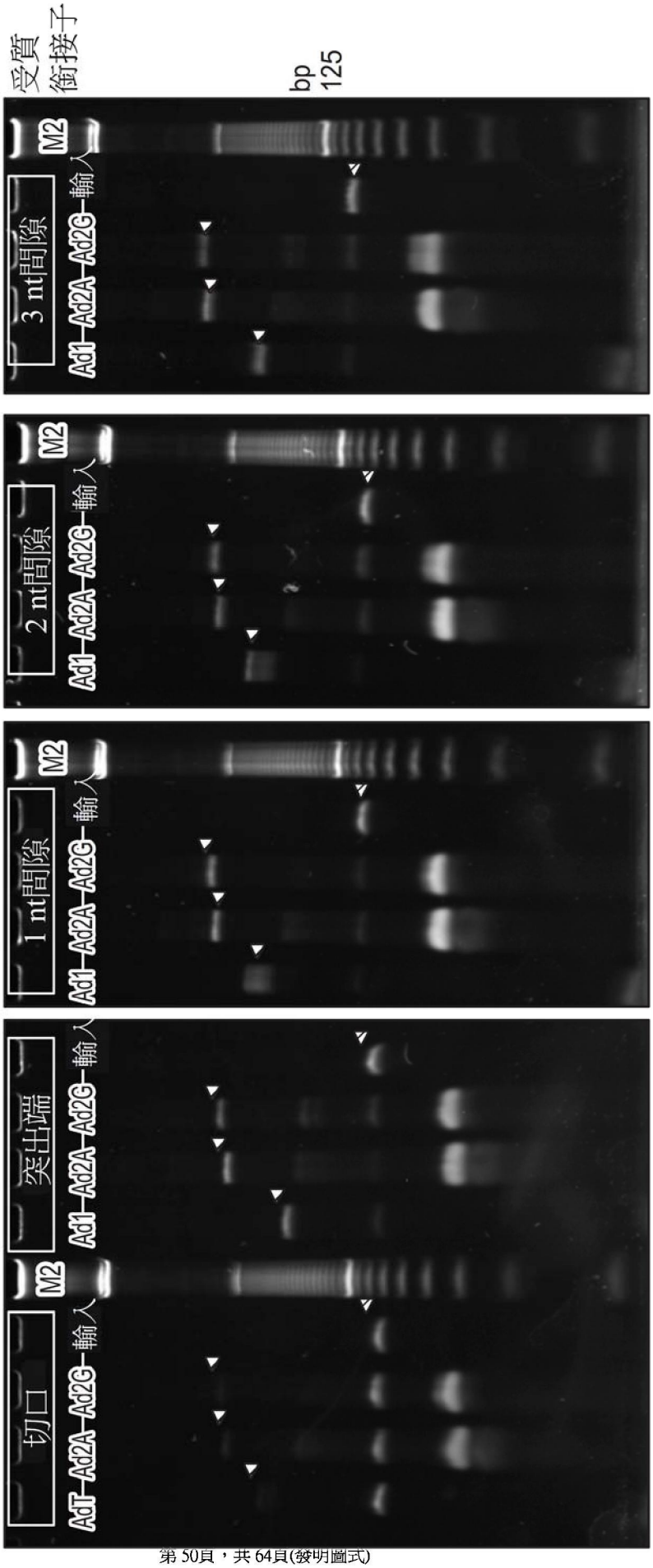
第20圖 (續2)


泳道
受質
連接酶
體積
1 2 3 4 5 6 7 8 9 10 11 12 13 14
M2
切口
1 nt間隙
8 nt間隙
5'-突出端
鈍端對照
- + - + - + - + - + - + - +
1 1 1 0.5 1 1 0.5 1 1 0.5 1 1 0.5 1
nt
75
50
25

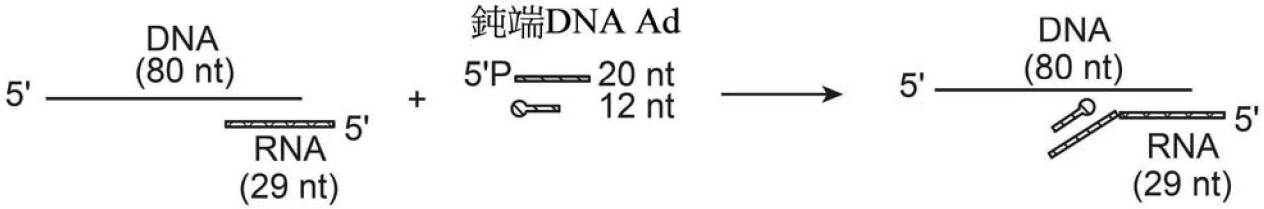
第21B圖


切口
1 nt或8 nt間隙
5'-突出端
鈍端對照
鈍端供體
連接
5'P
3'分支連接
產物
鈍端連接
產物
3'
OH
5'

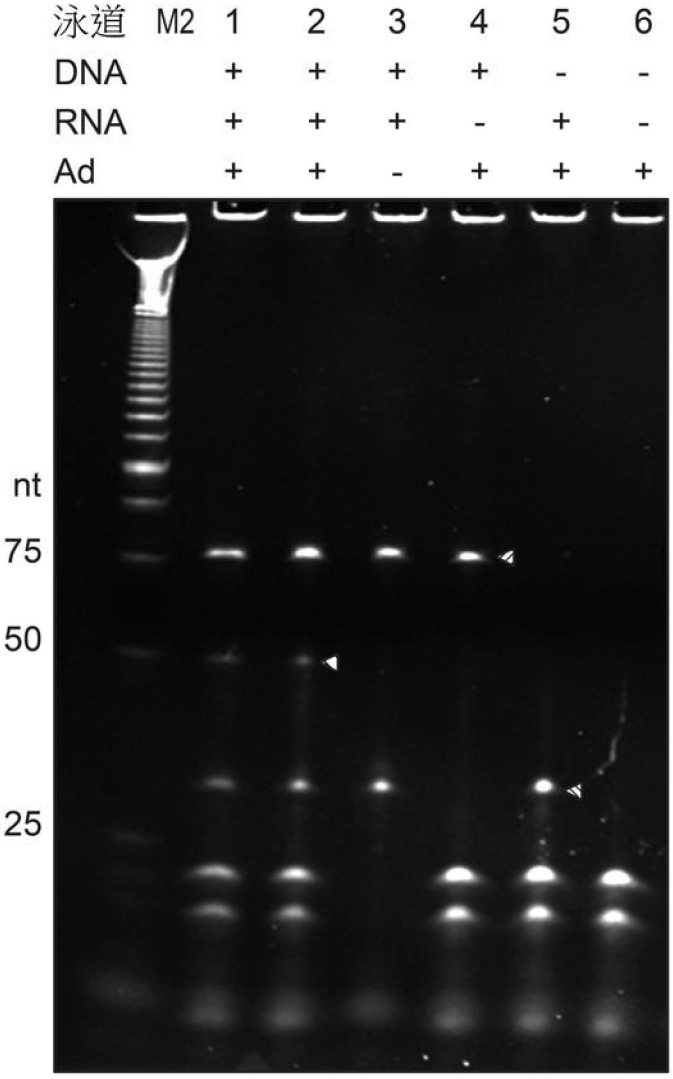
第21A圖



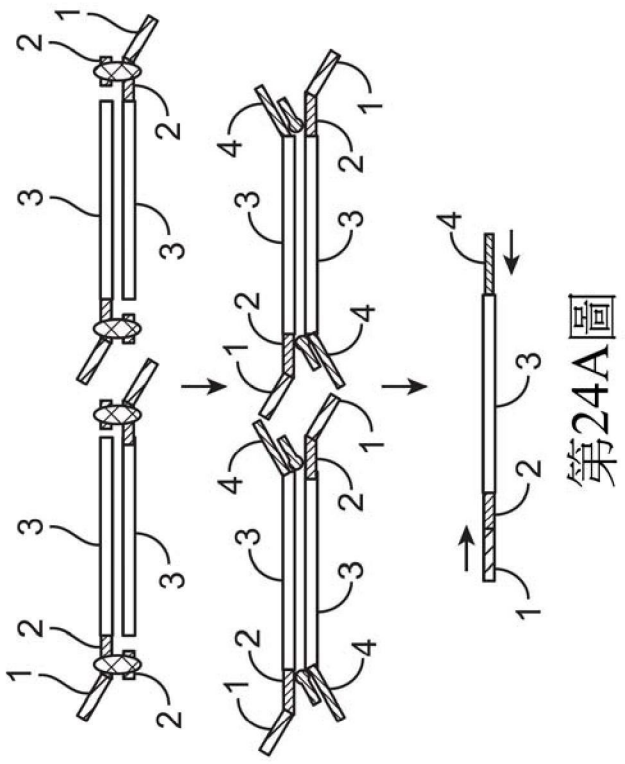
第22A圖　第22B圖　第22C圖　第22D圖


DNA
(80 nt)
5'
5'
RNA
(29 nt)
+
鈍端DNA Ad
5'P
20 nt
12 nt
DNA
(80 nt)
5'
5'
RNA
(29 nt)

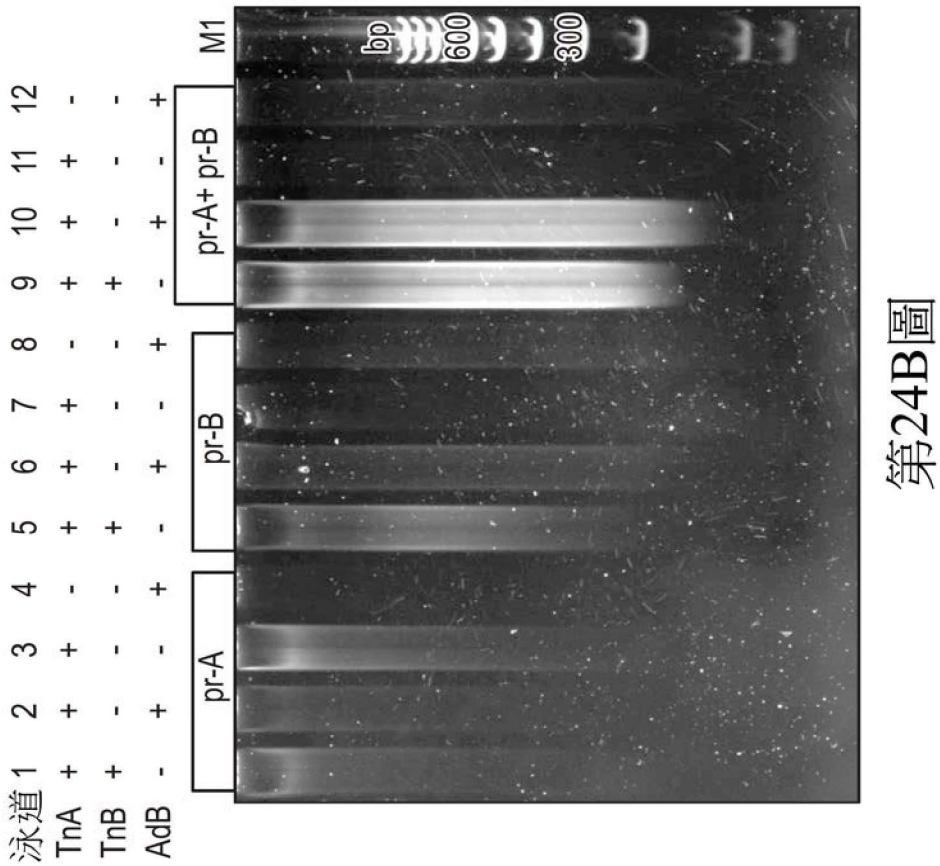
第23A圖


泳道 M2 1 2 3 4 5 6
DNA + + + + - -
RNA + + + - + -
Ad + + - + + +
nt
75
50
25

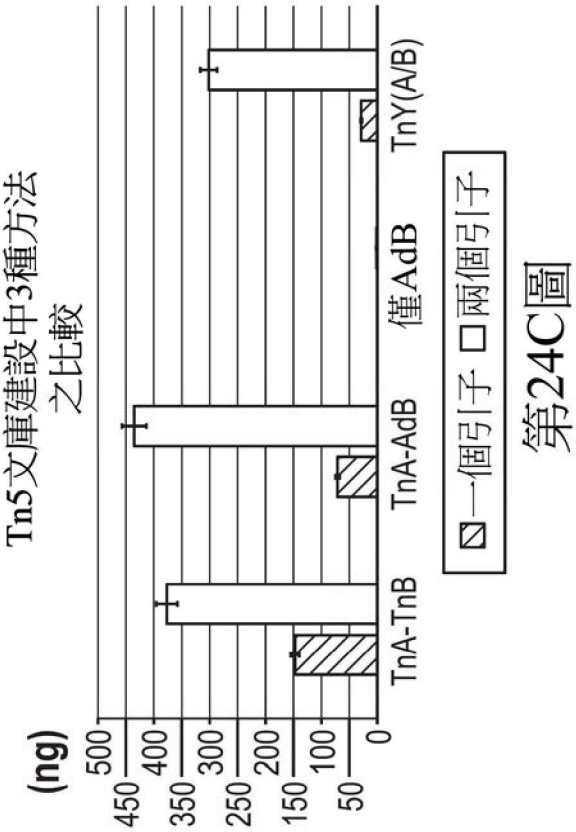
第23B圖



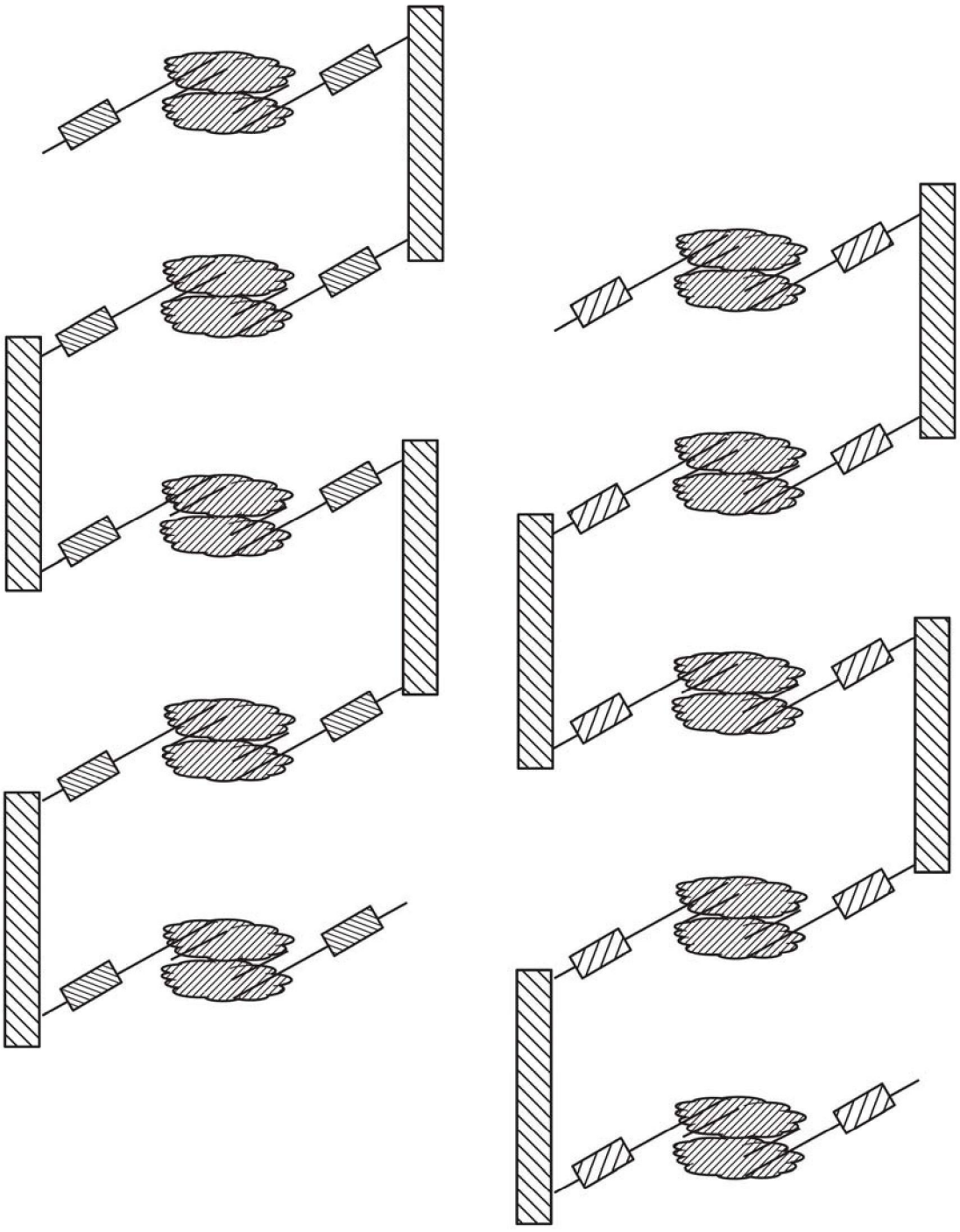
第24A圖



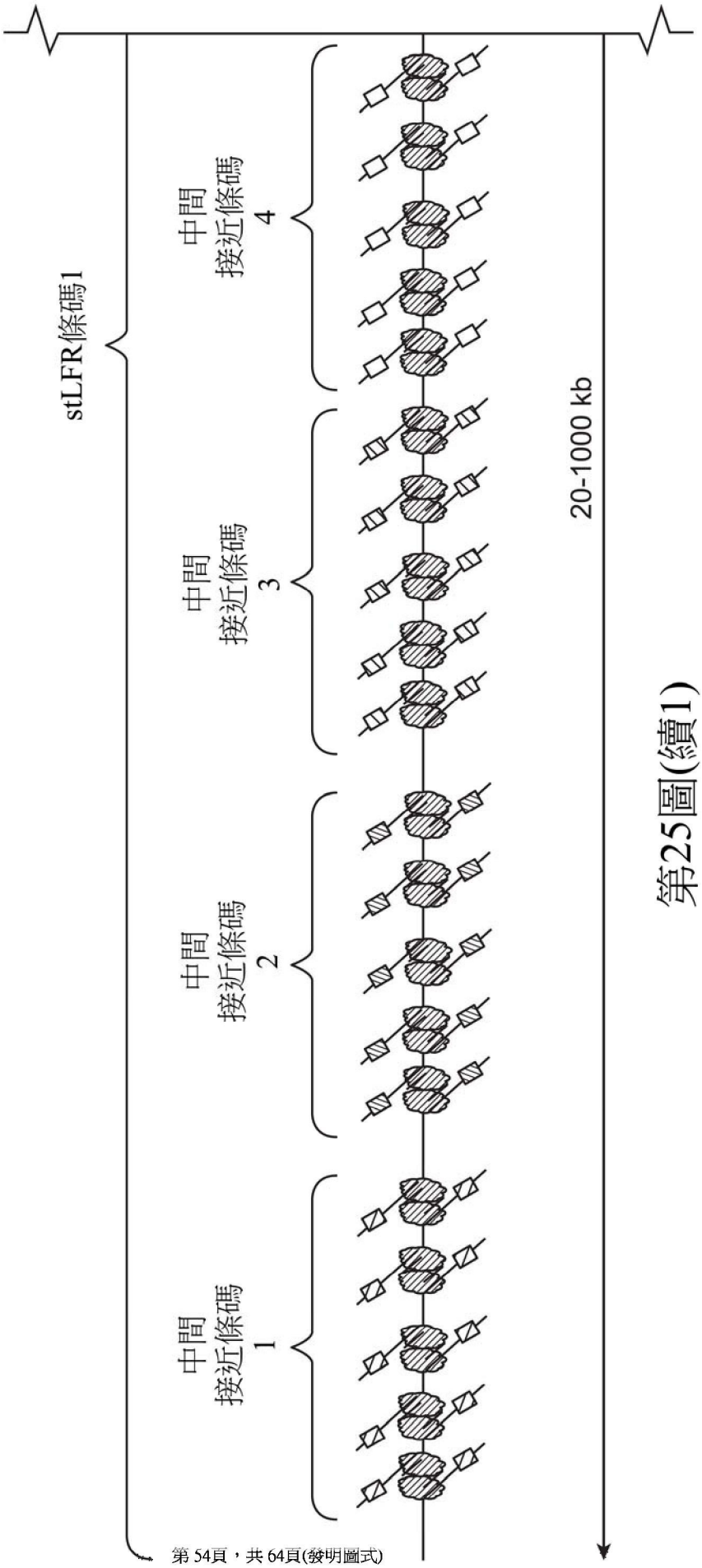
第24B圖



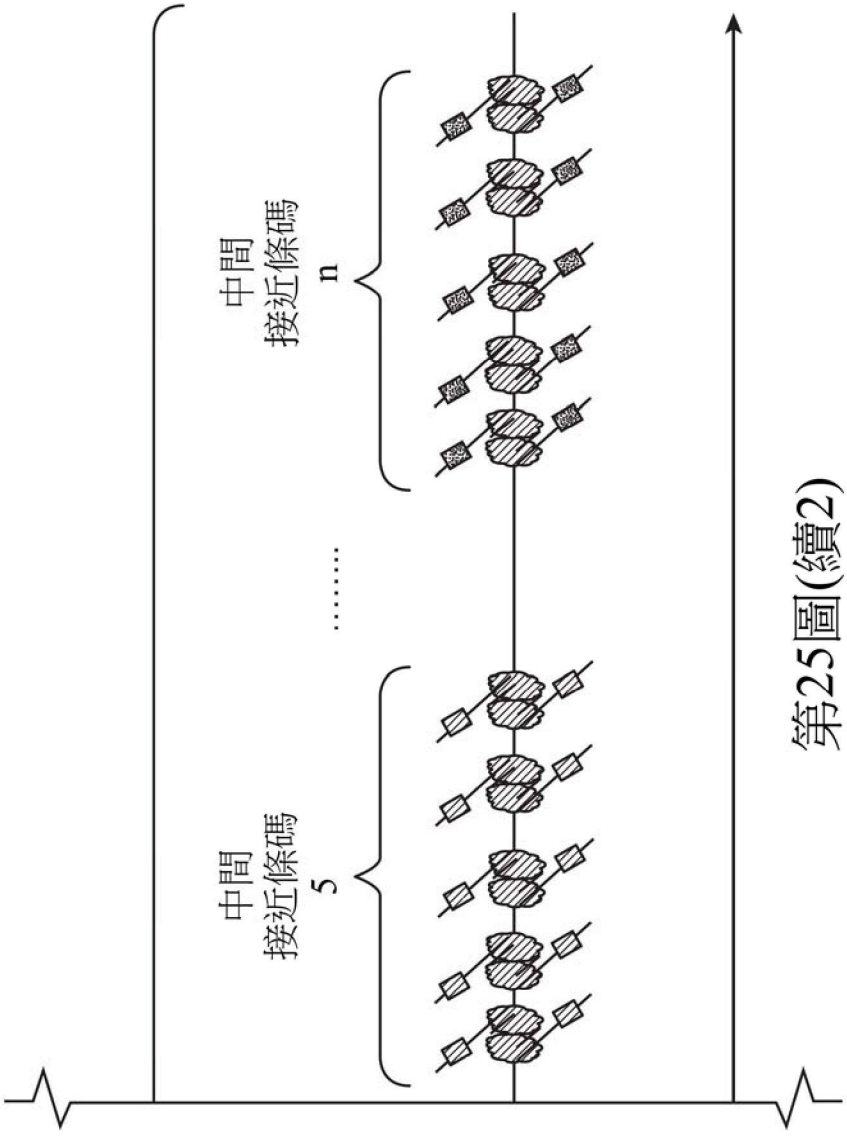
第24C圖



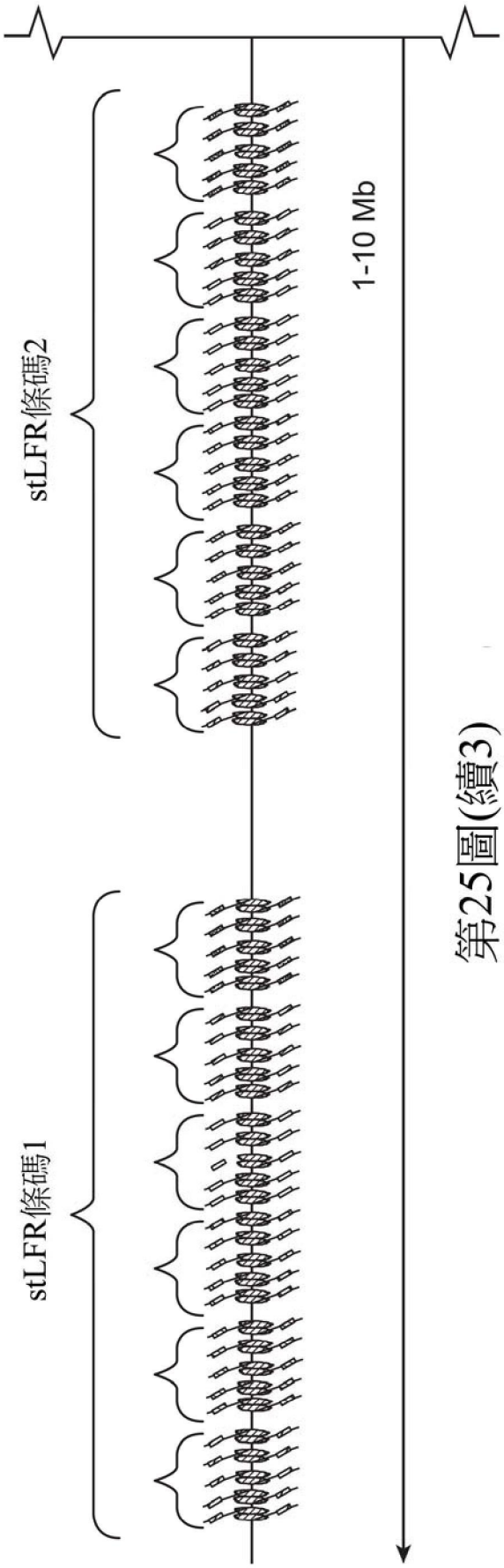
第25圖



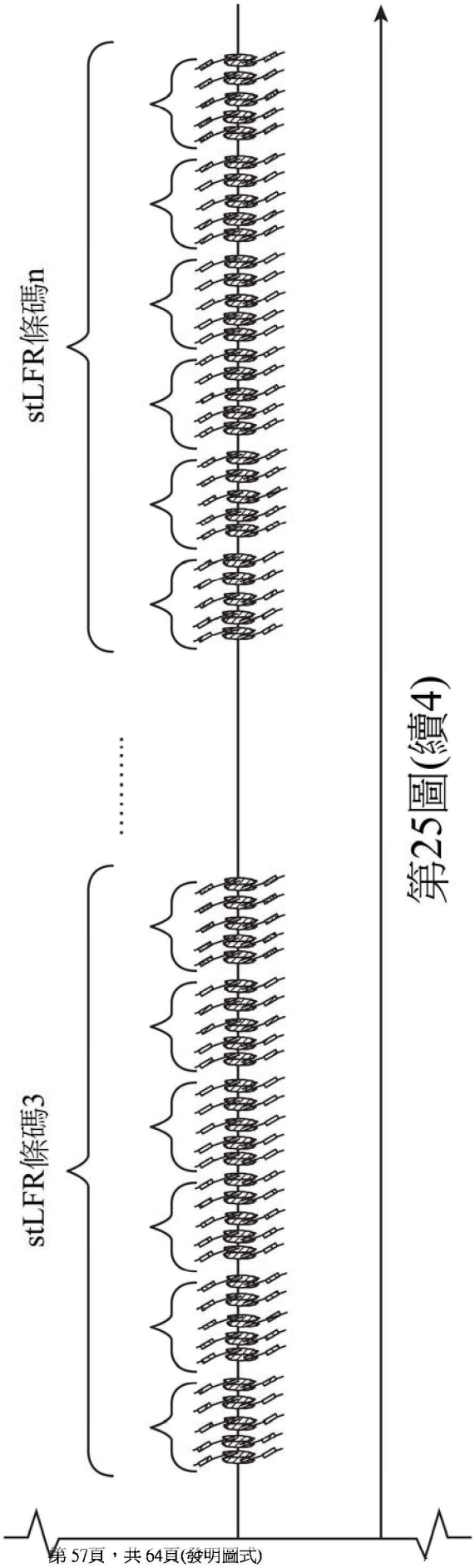
第25圖(續1)



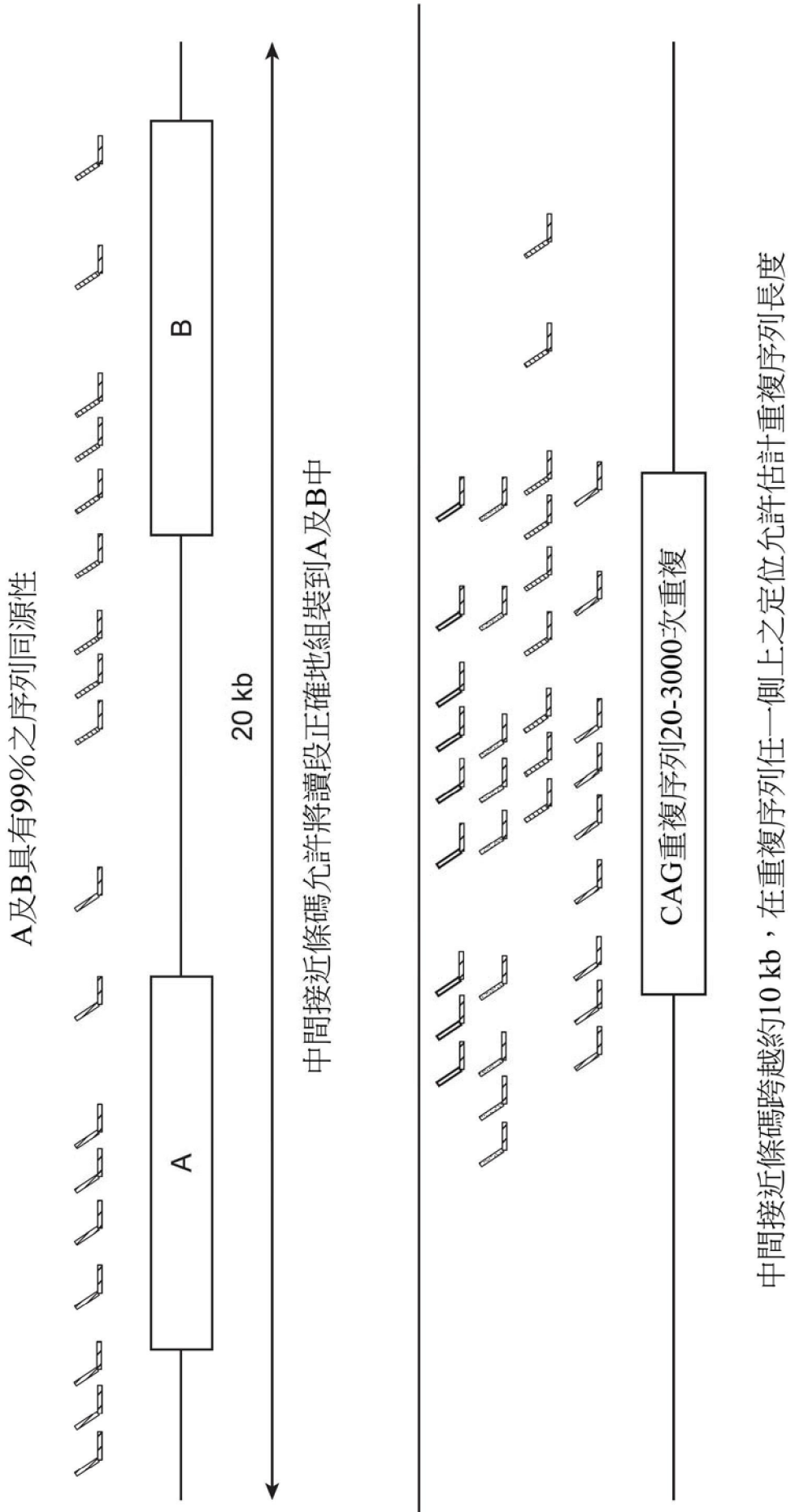
第25圖(續2)



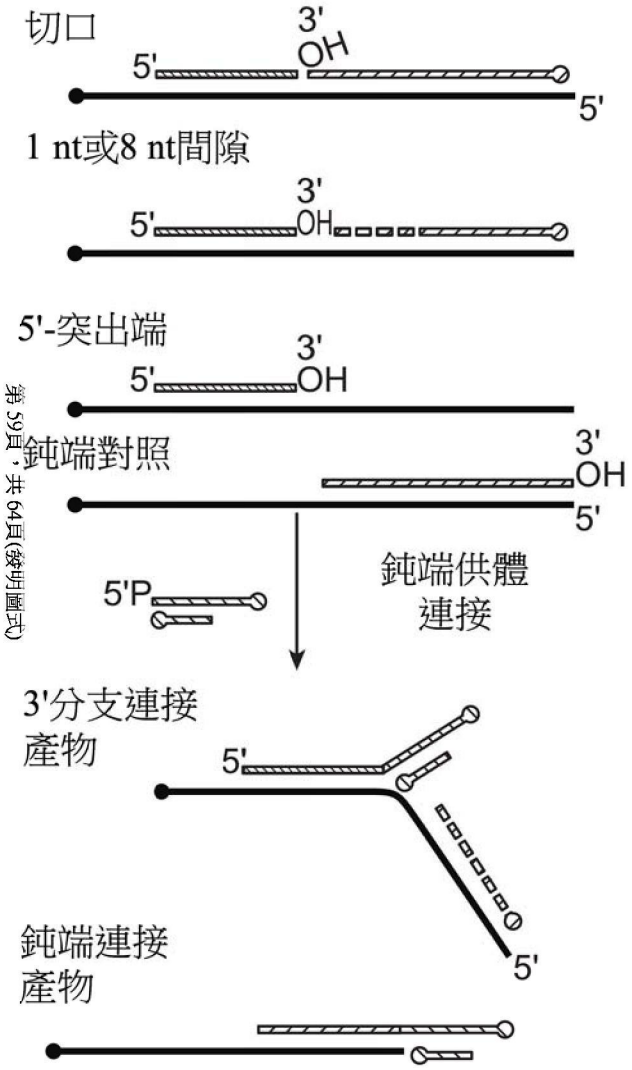
第25圖(續3)



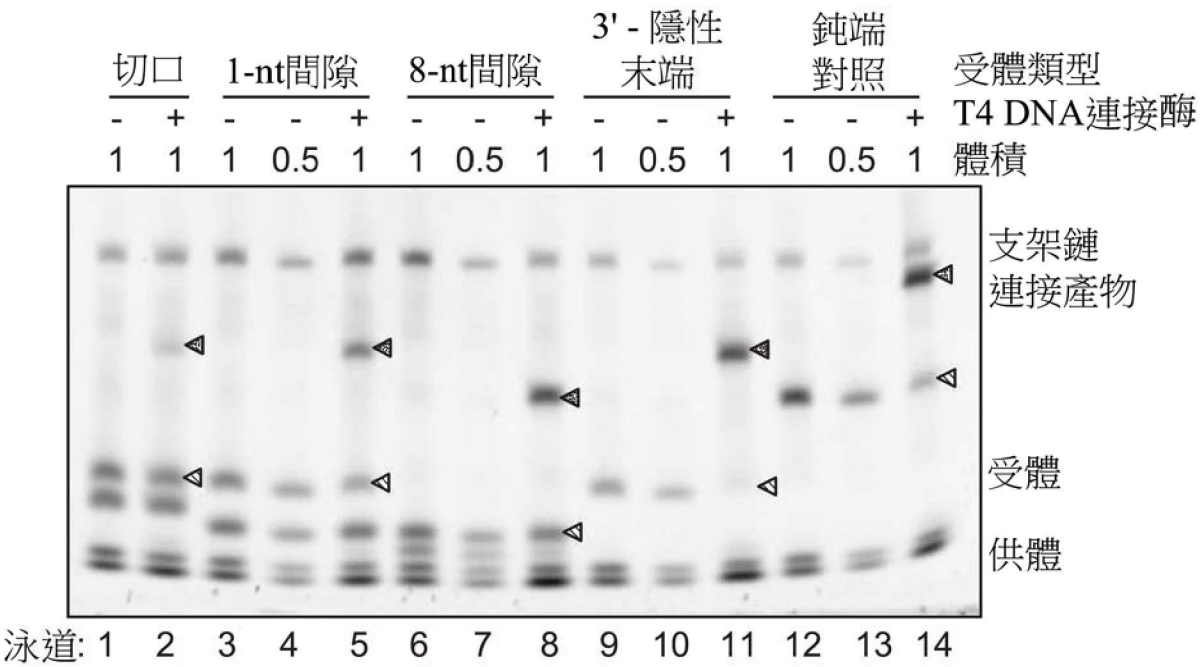
第25圖(續4)


A及B具有99%之序列同源性
A
B
20 kb
中間接近條碼允許將讀段正確地組裝到A及B中
CAG重複序列20-3000次重複
中間接近條碼跨越約10 kb，在重複序列任一側上之定位允許估計重複序列長度

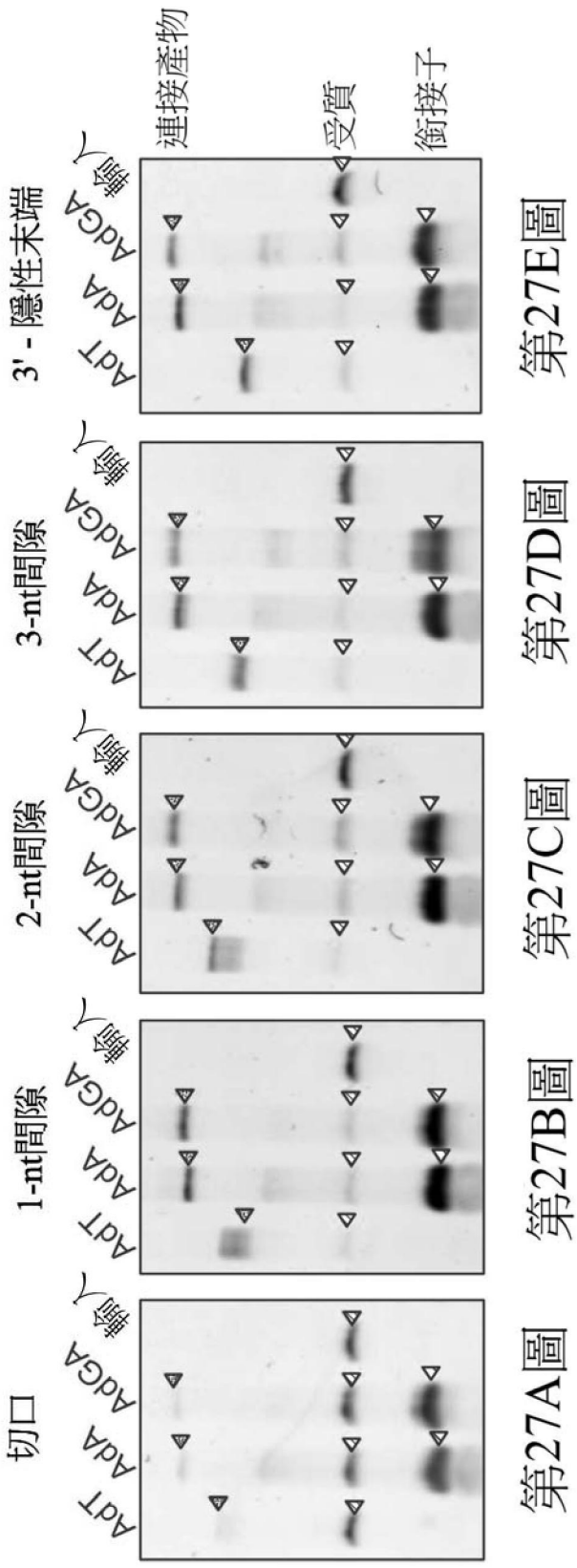
第25圖(續5)


切口
3'
OH
5'
5'
1 nt或8 nt間隙
3'
5'
OH
5'-突出端
3'
5'
OH
鈍端對照
3'
OH
5'
5'P
鈍端供體
連接
3'分支連接
產物
5'
5'
鈍端連接
產物

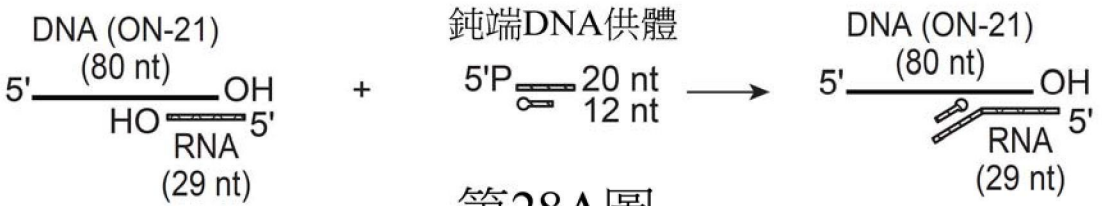
第26A圖


切口
1-nt間隙
8-nt間隙
3' - 隱性
末端
鈍端
對照
受體類型
T4 DNA連接酶
體積
- + - - + - - + - - + - - +
1 1 1 0.5 1 1 0.5 1 1 0.5 1 1 0.5 1
支架鏈
連接產物
受體
供體
泳道: 1 2 3 4 5 6 7 8 9 10 11 12 13 14

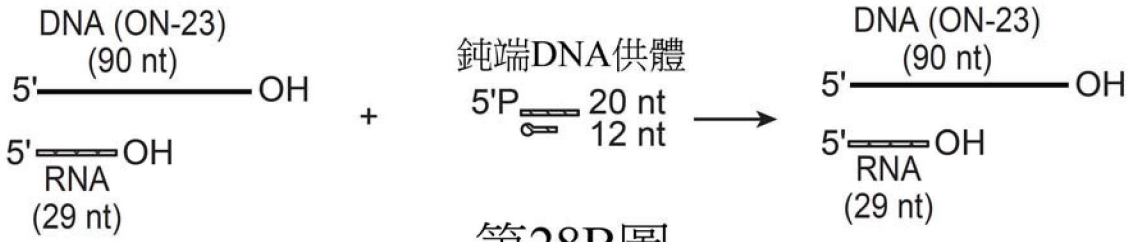
第26B圖



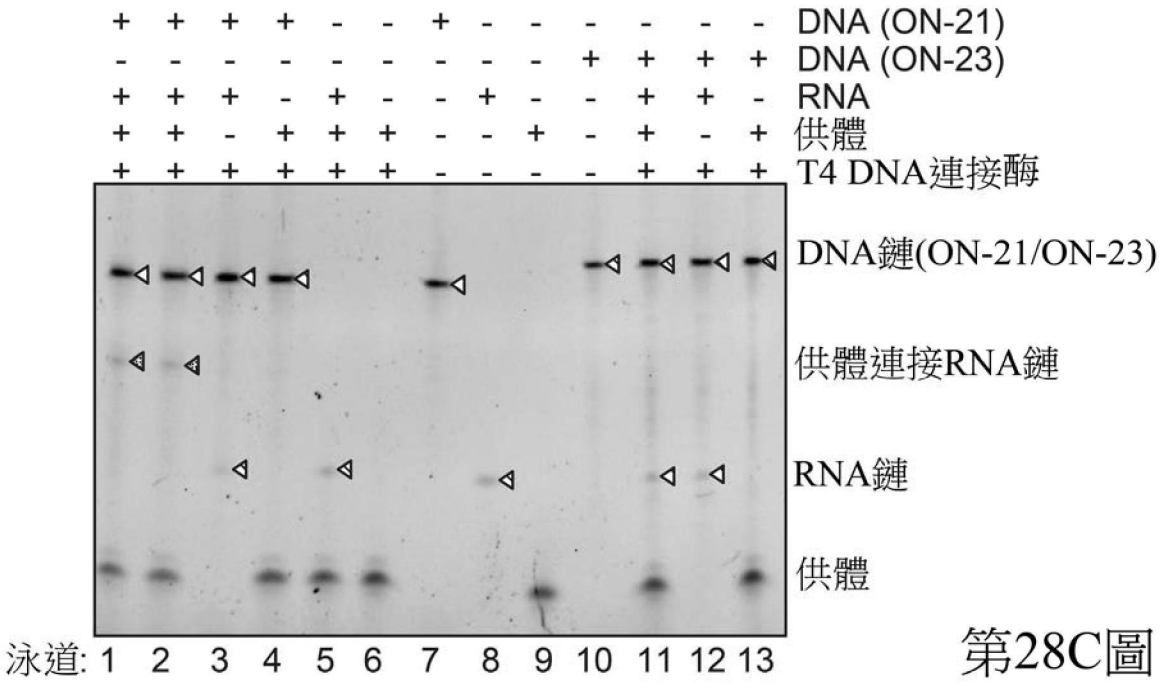
第27A圖 第27B圖 第27C圖 第27D圖 第27E圖



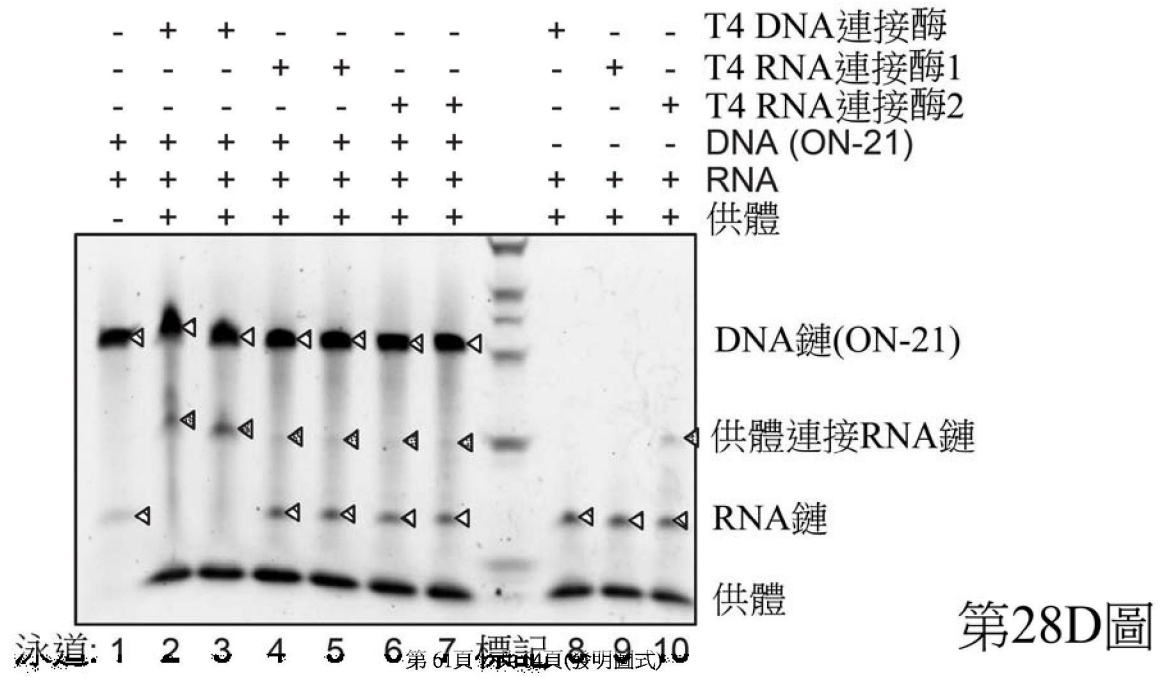
第28A圖



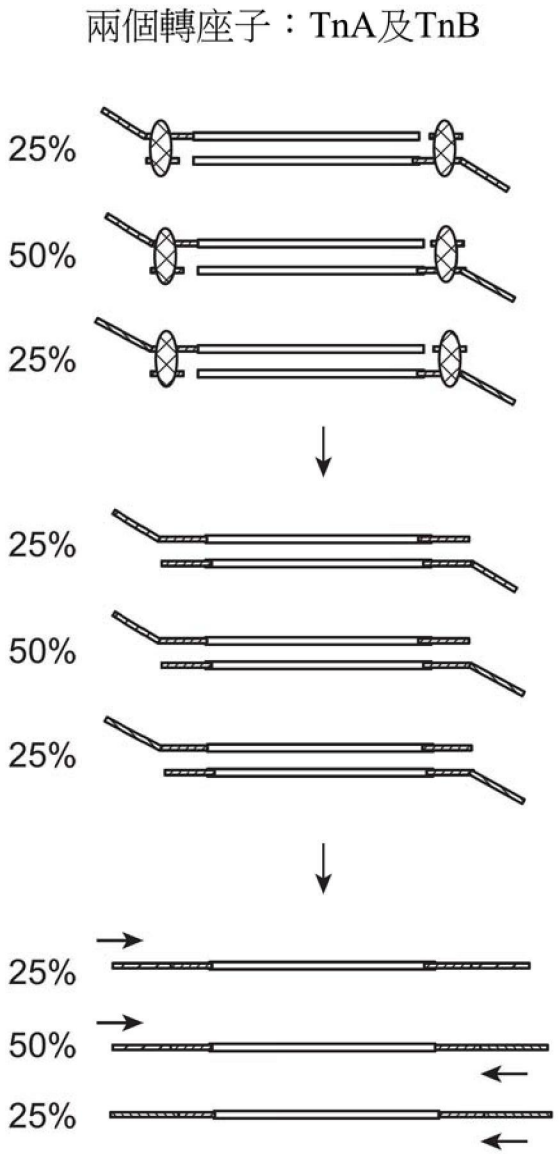
第28B圖



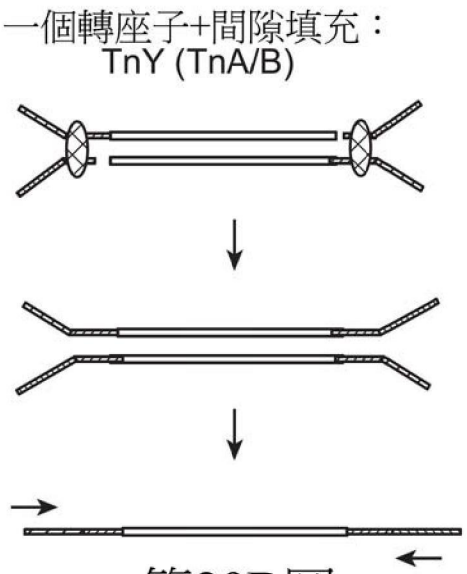
第28C圖



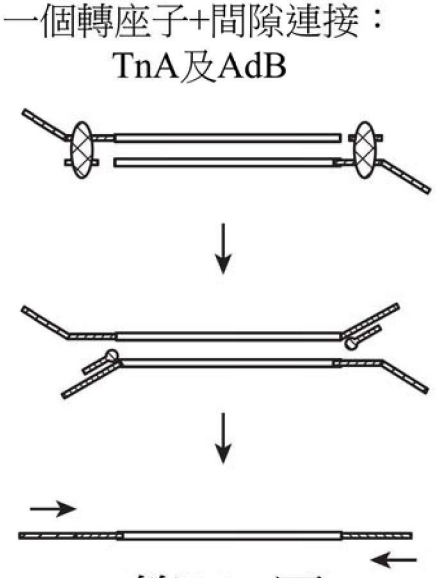
第28D圖



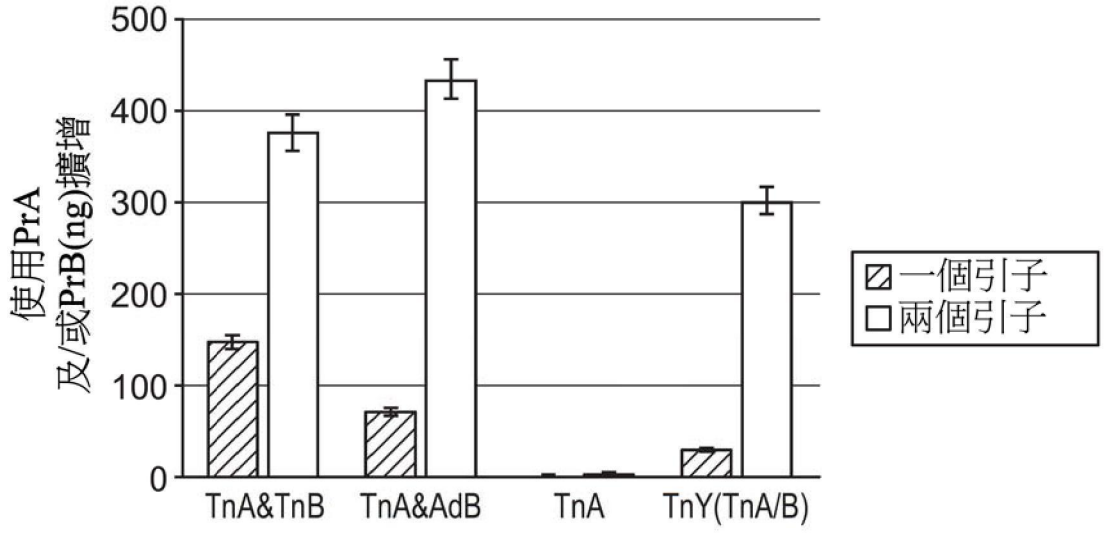
第29A圖



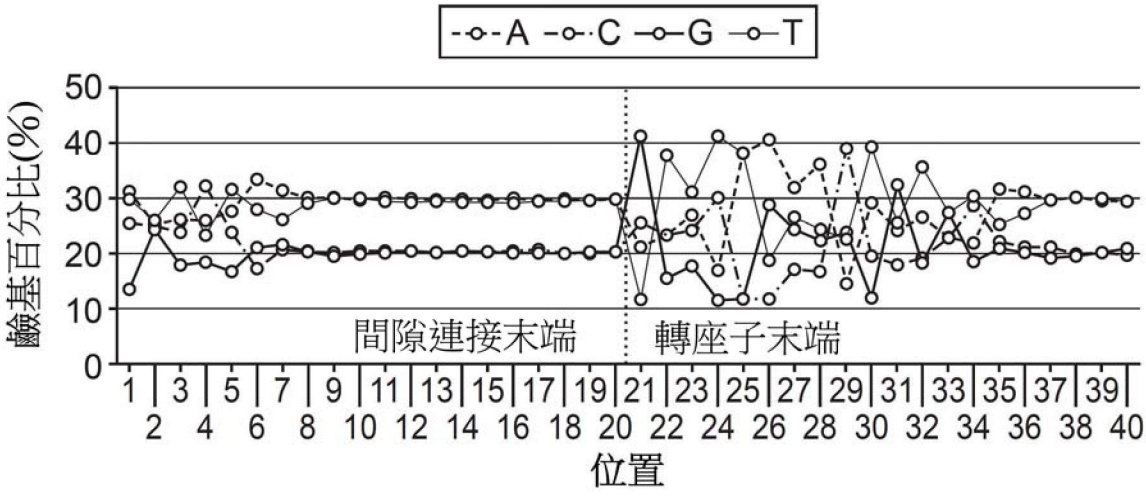
第29B圖



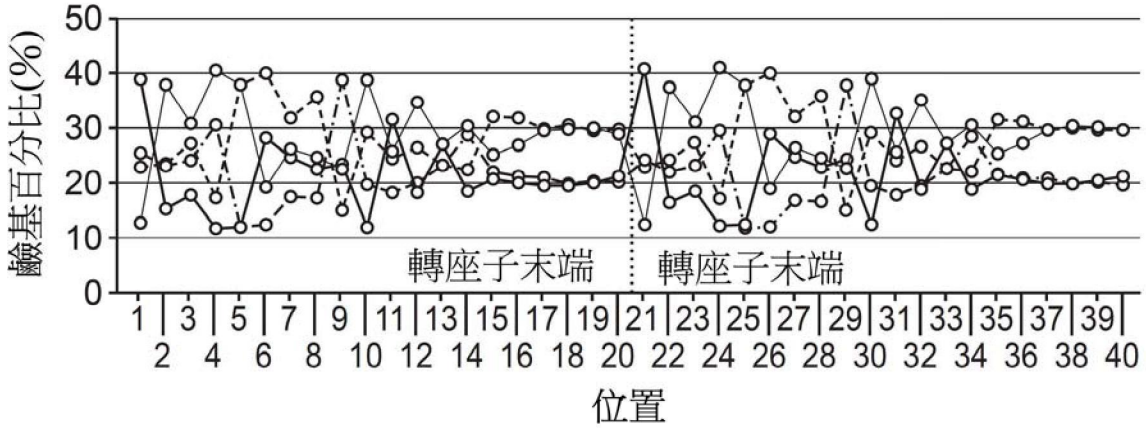
第29C圖



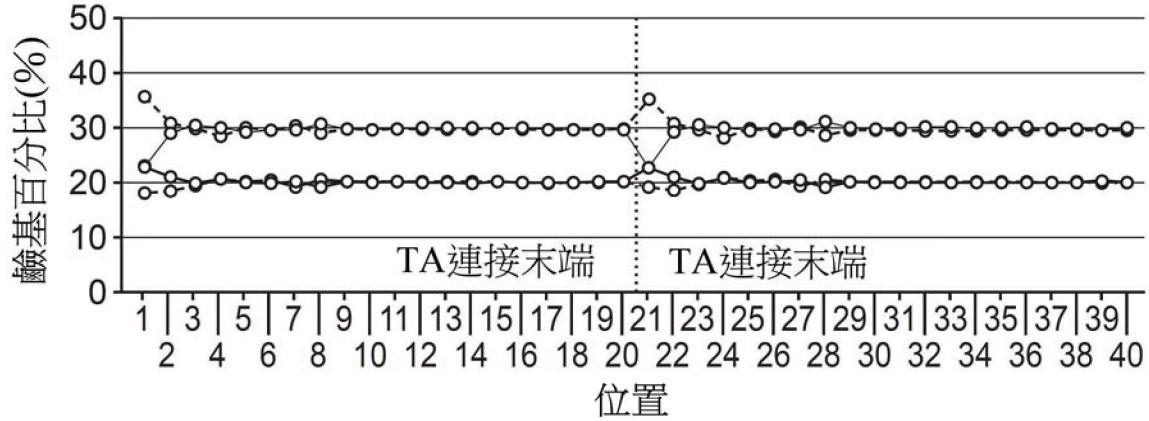
第29D圖



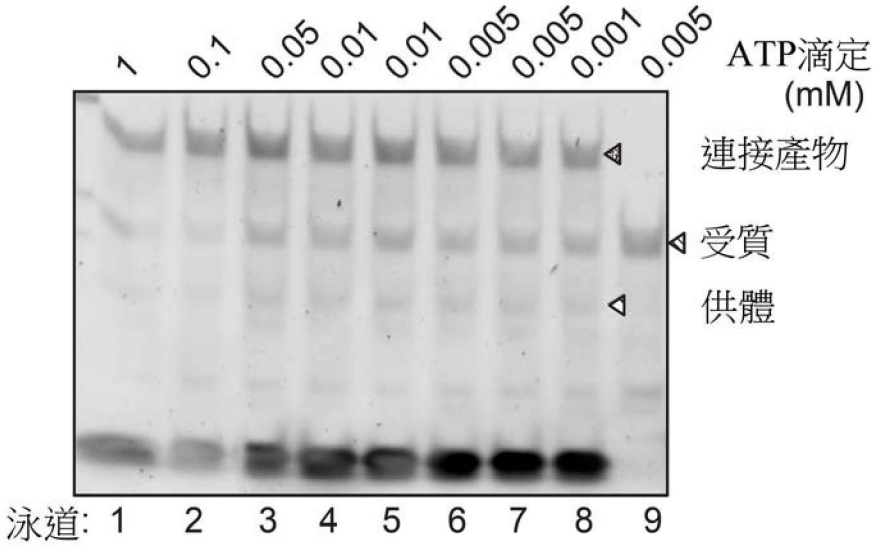
第30A圖



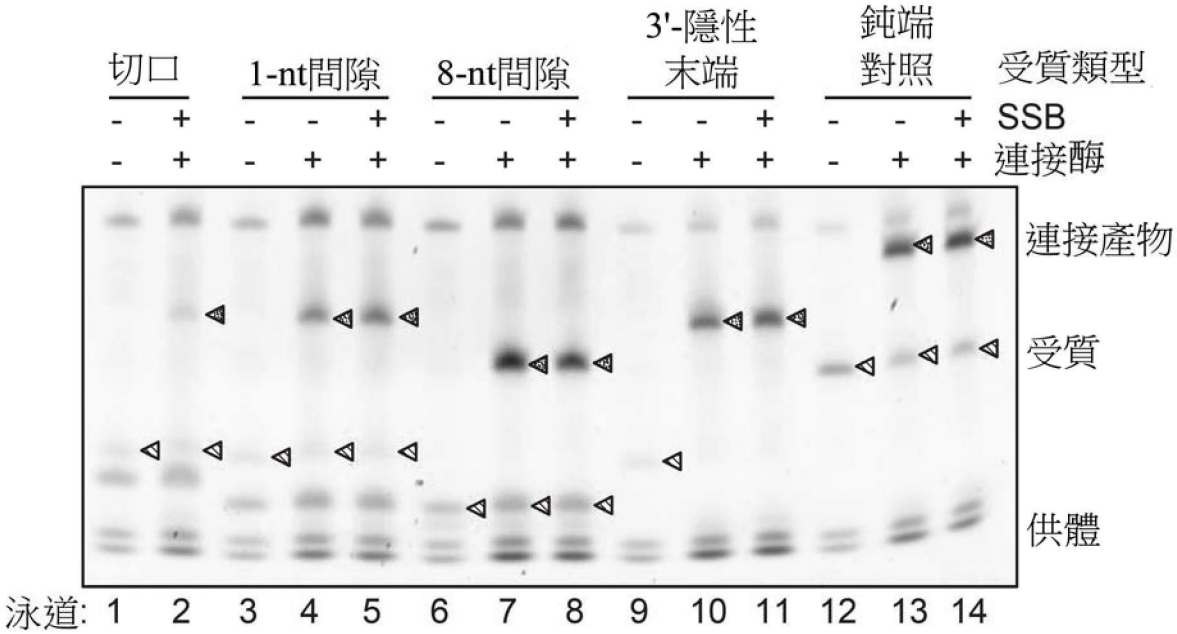
第30B圖



第30C圖


1
0.1
0.05
0.01
0.01
0.005
0.005
0.001
0.005
ATP滴定
(mM)
連接產物
受質
供體
泳道: 1 2 3 4 5 6 7 8 9

第31A圖


切口
1-nt間隙
8-nt間隙
3'-隱性
末端
鈍端
對照
受質類型
SSB
連接酶
連接產物
受質
供體
泳道: 1 2 3 4 5 6 7 8 9 10 11 12 13 14

第31B圖