國立臺灣大學生物資源暨農學院生物機電工程學系

碩士論文

Department of Biomechatronics Engineering
College of Bioresources and Agriculture
National Taiwan University

Master's Thesis

建立臺灣人泛參考基因組提升短序列回貼及 KIR 分型正確性

Constructing Taiwanese Reference Pangenome (TW-graph) to Improve Read Mapping Rates and KIR typing

邱顯鈞

Hsien-Chun Chiu

指導教授：陳倩瑜 博士

Advisor: Chien-Yu Chen, Ph.D.

中華民國 113 年 7 月

July 2024

doi:10.6342/NTU202401388

誌謝

首先，非常感謝指導教授陳倩瑜老師，入學前的晤談、開學前的選課建議、碩一結束的回顧及未來展望，至碩二確定題目之後，每回與老師的開會都有當頭棒喝的感覺，得到許多不同的見解建議與未來方向的提點。

另外，還有兩周一次的醫學院會議及參與的學長及老師們，特別是後來的口委，非常感謝每次報告完給的建議，而在口試方面更是給予許多建議，包含陳沛隆 醫師、楊雅倩 老師、許書睿 老師、許家郎 老師，以及其他與會成員。接著就是實驗室夥伴，首先最重要的就是同屆的亭堅、青勁、冠穎、承軒以及天洞，特別是亭堅，由於題目及其中所用到工具的關係，亭堅是在實驗過程中幫忙最多的，另外冠穎也有不少的互相幫忙的經驗，也是不可多得的實驗好夥伴，青勁則是經驗較豐富，並且常常會有領先大家不同的見解及經驗，也是幫忙了我不少，承軒及天洞則是在修課以及常駐在實驗室互相督促的好夥伴。其他還有弘曄學長，交接給我們最重要的工具，還有伯豪學長、毓聰學長、以及名翔學長，在碩一的時候也給了我們不少的建議與幫助，還有東祈學長也有解惑許多較艱深細節的問題。接著就是優秀的大專生及後來的研究助理，特別是鴻昇以及渤翰，會一起在實驗室奮鬥，並且給出許多實質上的意見，而且相處相當融洽。

還有許多其他實驗室的學長姐學弟妹跟同學，也都幫了相當多的忙，在此萬分感謝，謝謝你們。

我也從對生物資訊領域懵懵懂懂，至今學會了如何串接工具，並且需要時刻注意伺服器的運算情況、軟體需要的環境等等，都是實驗中需要注意的細節。再次感謝倩瑜老師及所有在我研究路上遇到的貴人，包含學長姐，同儕們等等。

doi:10.6342/NTU202401388

中文摘要

由於目前主流的個人定序技術，常為大量短序列，而短序列通常已經遺失了位置資訊，因此需要藉由對比參考基因組來將所有短序列回貼至參考基因組上。現有常見的參考基因組為 hg19 或 hg38，但該基因組取自少數個體，又缺少東亞人參與其中，因此對於臺灣人而言做為參考基因組可能會因個體差異而有所偏差。本研究使用了臺灣人體生物資料庫 (Taiwan Biobank) 的資料，所採用的基因變異集合為先前本實驗團隊所得，其中包含許多臺灣人特有的變異點位，本研究用以建立臺灣人泛參考基因組 (TW-graph)，以提升短序列回貼之品質及未來應用之準確性。過去的研究中多採用圖基因組方法來建立泛參考基因組，常用工具中又以 HISAT2 為最熱門者，因此本研究使用 bcftools 篩選變異點位以及 HISAT2 這個基於圖基因組概念的演算法，實現本研究欲建立臺灣人泛參考基因組的目標，以及建立做為對照組的 hg38 圖參考基因組 (意即不加入任何變異點位) 及全球泛參考基因組 (即為加入全球千人基因組點位計劃的變異點位資料) 共兩個版本的對照參考基因組。建立泛參考基因組後，再將臺灣人的短序列資料回貼並做後續分析，使用回貼率來做初步結果判讀及比較。本研究使用七個 Taiwan Biobank 的以及四個非 Taiwan Biobank 的短序列資料，觀察臺灣泛參考基因組對比上述的其他兩者參考基因組而言，回貼率有顯著提升，十一個樣本的總體回貼率對比 hg38-graph 有提升約1%的趨勢，對比 1000G-graph 有提升約0.9%的趨勢。本研究進而將所建立的臺灣人泛參考基因組，應用於KIR基因家族的等位基因分型，在採用的HPRC之44 個樣本中，唯一回貼短序列數 (Unique mapped reads) 的數據顯示，BWA-linear 表現會優於 1000G-graph 再優於 hg38-graph；而和先前一樣的十一個臺灣人樣本中，TW-graph 在 KIR 區域中的唯一回貼短序列數明顯比其他三個對照組多，雖然此刻缺乏正確答案作為評量之參考，仍期待未來有更多

doi:10.6342/NTU202401388

的實驗數據來探討回貼序列數的增加，是否助於提升 KIR 等位基因分型之準確性。

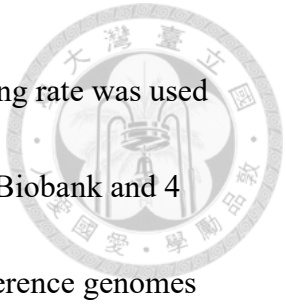
關鍵字：臺灣泛參考基因組、短序列回貼、臺灣人體生物資料庫、KIR、單核苷酸多態性

doi:10.6342/NTU202401388

英文摘要

Due to the current mainstream personal sequencing technology often producing large numbers of short sequences that have lost positional information, it is necessary to align these short sequences to a reference genome for comparison. The commonly used reference genomes are hg19 or hg38, but these genomes are derived from a small number of individuals and lack East Asian participation. Therefore, for Taiwanese people, using these as reference genomes may lead to biases due to population differences. This study uses data from the Taiwan Biobank, which contains variants unique to Taiwanese people. This data is used to construct a Taiwanese reference pangenome to improve the quality of short sequence alignment and future applications. Past research has often used graph-based genome methods to build reference pangenomes, with HISAT2 being the most popular. Therefore, this study uses bcftools to filter variants and HISAT2, an algorithm that helps to implement the concept of graph genomes, to construct a Taiwanese reference pangenome. For comparison, this study also created two versions of reference genomes: an hg38-graph reference genome (without adding any variant positions, hg38-graph) and a global reference pangenome (incorporating variant data from the global 1000 Genomes Project, 1000G-graph). After establishing the reference pangenomes, this study aligned Taiwanese short read data and

doi:10.6342/NTU202401388



performed subsequent analyses such as KIR allele typing. The mapping rate was used for result interpretation and comparison. This study adopts 7 Taiwan Biobank and 4 non-Taiwan Biobank short reads data. Compared to the other two reference genomes mentioned above, the Taiwanese reference pangenome shows a significant improvement in mapping rate. The overall mapping rate for 11 samples shows an improvement trend of about 1% compared to hg38-graph and about 0.9% compared to 1000G-graph. Regarding KIR typing, among the 44 samples from HPRC, the data on uniquely mapped reads shows that BWA-linear performs better than 1000G-graph, which in turn performs better than hg38-graph. For the eleven Taiwanese samples, TW-graph shows significantly more uniquely mapped reads in the KIR region compared to the other three control groups.

Keywords: Taiwanese reference pangenome, Short read mapping, Taiwan Biobank, KIR, SNP

doi:10.6342/NTU202401388

目　　次

doi:10.6342/NTU202401388

doi:10.6342/NTU202401388

圖　　次

==========

doi:10.6342/NTU202401388

表　　次

doi:10.6342/NTU202401388

doi:10.6342/NTU202401388

第一章 前言

1.1 背景介紹

在現代的次世代基因定序 (Next Generation Sequencing, NGS) 中，全基因組定序 (Whole Genome Sequencing, WGS) 是一項關鍵技術。為實現低成本之 WGS 技術，目前仰賴短序列 (short read)。這些短序列是由基因體中的 DNA 片段製成，並且通常具有數十至數百個鹼基對的長度，亦為目前主流的定序方法。然而，短序列回貼時存在一些挑戰。因此，為了提高短序列回貼的正確性，研究人員開始導入泛基因組的概念，期能更好地處理個體差異和變異點。這對於基因體學研究和醫學應用領域都具有重大意義，因為它有助於更全面地理解和利用個體的基因組資訊。

隨著個人全基因定序資料普及，科學家們可以收集到更多人的基因組資料，將這些資料整合在一起，並正從傳統的單倍型線性參考序列轉向以圖資料結構儲存的泛基因體，以更好地應對人類基因組的多樣性。從前我們只靠少數人的基因組來建立參考基因組 (reference genome)，參考基因組就像一個地圖，它可以幫助我們定位人類基因組上的特定位置。傳統線性基因組呈現了最常見的參考基因型，但在短序列回貼時，必須謹慎處理與參考基因組不匹配的情況。泛基因組 (pangenome) (Miga & Wang, 2021)就像一個更新的、更準確的地圖事先包含了觀察到的變異點，為短序列回貼提供了更多選擇，也幫助我們更好地理解人類基因組，且單倍型的線性基因組無法表現人體群體多樣性 (Sirén et al., 2021)，尤其在複雜區域或大區段變異的情況下，正確地將短序列回貼到參考基因組上面臨挑戰，例如：免疫基因區域。而表達泛基因組的其中一種方式是利用圖形的演算法結構，我們稱之為圖基因組 (graph genome)。

一般而言，在序列回貼之後，後續常見的分析有變異點偵測 (variant calling)、特定基因區域，如：HLA、KIR 及 TCR 等等基因，以上皆為與免疫相關並且複雜性高，臨床亦相當重要的基因，對其進行等位基因分型 (allele typing)，及拷貝數 (Copy number) 偵測。

doi:10.6342/NTU202401388

本研究的創新之處在於，它將臺灣人特有的常見和罕見遺傳變異納入圖基因組中，能幫助我們更好地理解臺灣人的疾病風險和其他健康狀況。而主要挑戰之一是如何整合臺灣人特有的遺傳變異和公開資料庫中已知的遺傳變異，如：IPD-KIR (Robinson et al., 2013) 2.10.0 (2020 年釋出之版本)等等，整合後除了原先對於臺灣人之短序列回貼影響分析外，對其後續 KIR 區域分型結果影響分析亦相當重要。另一個挑戰是如何開發高效的分析工具來分析這個龐大的數據集。此外，圖基因組將更有效幫助臨床如何運用低成本之短序列定序資料。

本研究會使用到線性回貼工具 BWA、圖基因組回貼工具 HISAT2、以及 samtools, bcftools 等等。在高通量基因組定序分析中，samtools 及 bcftools 是兩個重要且經常使用到之生物資訊工具。它們在處理和分析基因組數據方面具有高度的靈活性和效率，為基因組學研究提供了強有力的支持。

首先，samtools 是一個用於處理 SAM (Sequence Alignment Map) 檔和 BAM(Binary Alignment Map) 檔格式的工具。SAM 和 BAM 是基因組比對結果的標準格式，其中 SAM 是人類能看懂的純文字檔，而 BAM 是其二進位壓縮格式。samtools 的主要功能包括排序 (sort)、索引 (index)、格式轉換 (view)、抽取部分已回貼短序列 (view) 和回貼短序列計數 (flagstats) 等等。samtools 能夠對 BAM 文件進行快速排序和索引，這對於後續結果分析至關重要甚至不可或缺。並且可以將 SAM 格式轉換為 BAM 格式，從而節省存儲空間並提高處理速度。此外，samtools 還支持從 BAM 文件中提取特定區域的已回貼短序列，這在變異點檢測和基因組研究中舉足輕重。例如，在 KIR 基因區域的研究中，使用 samtools 提取該區域的已回貼短序列，以便進行下游分型及分析。總之，samtools 在基因組數據處理中不可或缺。

再者，bcftools 則是一個專門用於處理和分析 VCF (Variant Call Format) 文件的工具。VCF 文件是變異檢測結果的標準格式，記錄了基因組中的單核苷酸多態性 (SNPs)、插入或缺失 (Indels) 等變異資訊。bcftools 的主要功能包括變異點位過濾 (view, filter)、統計分析(stats)、格式轉換 (view) 和合併 (concat) 等等。

doi:10.6342/NTU202401388

bcftools 可以對 VCF 文件中的變異點位進行篩選，根據研究人員的標準篩選出其所需的變異。例如：通過設置不同等位基因頻率 (allele frequency) 閾值，bcftools 可以篩選出不同等位基因頻率之變異點。並且 bcftools 還支持對 VCF 統計分析，如：計算每個變異點之等位基因頻率等等，這對於基因組變異研究和相關研究具重要意義。此外，bcftools 還能夠將 VCF 格式轉換為其他格式，或將多個 VCF 文件合併為一個文件，以利後續分析。

1.2 研究目的

本研究旨在建立臺灣人的泛參考基因組 (TW-graph)，以提高臺灣人 WGS 等級的短序列回貼率，並使用 Graph-KIR (Lin et al., 2023) 作為工具，以分析臺灣人 KIR 分型，並且評估如何篩選欲加入之變異點位對於以上分析有所幫助。本實驗室先前完成了臺灣人的全基因定序數據分析 (Wu et al., 2021)，建立了高品質的變異點資料庫 TaiwanGenomes，其資料來源於 TaiWan BioBank (TWB)，包含了近六千萬個變異點。本研究第一步將從這個龐大的資料集中選擇局部資料，建立泛參考基因組，以平衡正確性和效能。在後續資料分析，我們將擴展 TW-graph 的應用，觀察 HISAT2 的定序數據分析。最後，我們將前述的變異點位篩選出 KIR 區域並且加至原先之 KIR-graph 中，並擴展到包括免疫基因等等相關分型結果，望能解決複雜基因的分析困難，尤其免疫相關等疾病相關的預測提供更精確的工具。

doi:10.6342/NTU202401388

第二章　文獻探討

2.1 泛基因組

目前人類常用的參考基因組基本上為線性的 hg19 或 hg38，並不能完全代表人類的基因甚至做為標準。而該參考基因為人類基因體計畫 (Human Genome Project, HGP) 中，少數幾位志願參加計畫者之基因序列組合而成。有研究闡明，目前之人類參考基因組，有 70%來自同一個體樣本，而且可能其中的一些點位為罕見的變異點 (rare variant) (Ballouz et al., 2019)，因此將其做為標準的參考基因組時，可能會有參考偏見 (reference bias)，影響後續的回貼正確率，如某基因片段難以回貼或是回貼數量及深度較其他地方少且淺等等。其中更可能因為不同地區的群體會有不同的基因的差異，因此各個地區的泛基因組參考基準也會不同。

原人類基因組 (GRCh37, hg19) 參考序列的個體樣本來源並且其中反映了一名男性捐贈者的 DNA 及其地區來源，並且該參考基因組多樣性不足，多數來自歐洲血統的人 (Miga & Wang, 2021)。基因組在各地之間有較顯著的差異。其中不論是洲際間，同洲不同地區之間，甚至是同地區都有不同的共同的基因組比例 (Auton et al., 2015)，因此泛基因組在當中就相當重要，需因應各地不同的基因組狀況，加入不同之變異點位以做為泛參考基因組。

2.2 線性基因組

短序列 (short read) 是目前常用的次世代定序 (Next-Generation Sequencing, NGS) 技術，其優勢為快速、高效並且低成本的一種 NGS 技術，並且 NGS 本身就比前代定序技術有著更高通量的特性，並且在診斷領域具有廣泛的應用。在此類定序中，基因在定序之前會經被打碎分成各小片段 (通常為 50 到 100 個鹼基對)。

而在序列皆為短序列的情況下，欲得知該個體之基因序列，則需要將短序列碎片回貼至參考基因組上，而在圖基因組演算法大量使用且基因組資料尚未充足時，就只能回貼至線性的參考基因組上，而此時就會需要用到線性

doi:10.6342/NTU202401388

的回貼工具，如：BWA-mem (Li, 2013)、Bowtie2 (Langmead & Salzberg, 2012) 等。如此一來，將短序列回貼後，即可得到該個體之基因組，亦可稱之為全基因體定序 (Whole Genome Sequence, WGS)。

2.3 圖基因組

可能在某些情況，如：大片段變異 (structural variation) 或是序列的插入或刪除 (indels) 等等，若是使用線性參考基因組都會影響短序列回貼的精確，因此可能影響後續的分析，如：變異點偵測 (variant calling) 以及後續的與基因體資料庫對比的變異點註釋分析 (variant annotation)、以及特定基因區域 (如：KIR) 之等位基因分型 (allele typing)。

在過去，基因組研究主要使用線性基因組，但隨著短序列定序技術的發展，以及圖基因組 (genome graph 或 graph genome) 演算法的開發 (Garrison et al., 2018; Kim et al., 2019; Rakocevic et al., 2019)，越來越多研究開始採用圖參考基因組 (Paten et al., 2017)。圖參考基因組可以包含所有已知的遺傳變異，因此具有更高的準確性和可靠性，如圖 1 所示，可以發現圖基因組會較線性基因組多了變異點位的資訊。

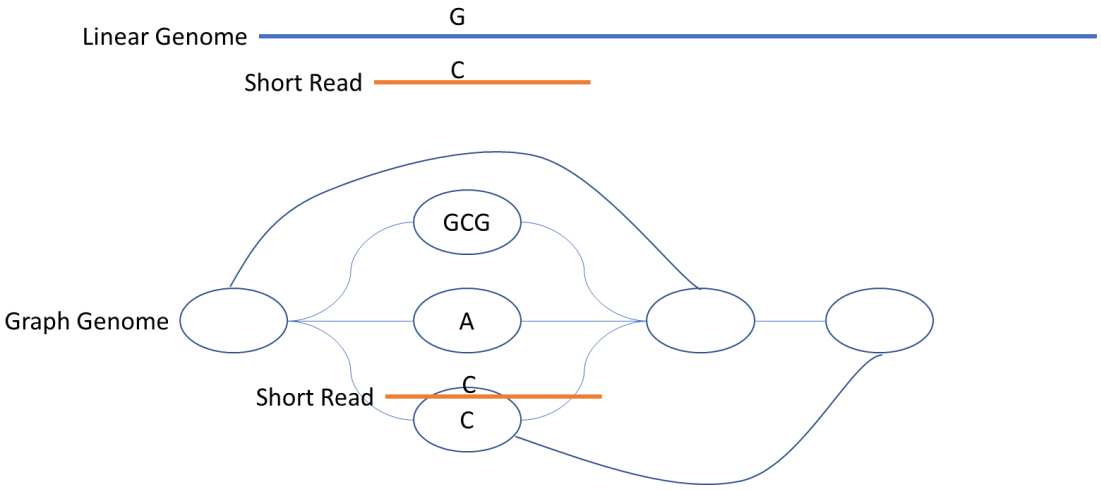


圖 1. 線性基因組與圖基因組之概念圖比較。橘色線條為短序列，藍色部份為參考基因組

doi:10.6342/NTU202401388

現有的生資短序列比對工具 (aligner or mapper)，線性基因組常見的有：Bowtie2 (Langmead & Salzberg, 2012)、BWA-mem (Li, 2013)，而圖基因組常見工具有：HISAT2 (Kim et al., 2019)、vg (Garrison et al., 2018)、vg giraffe (Sirén et al., 2021)，其中 HISAT2 在比對敏感度 (alignment sensitivity) 的表現為 99%，而 BWA-mem 和 Bowtie2 的表現皆為 97%上下，同為圖基因組比對工具的 vg 表現為 98%，可以看出 HISAT2 表現為最精準者，而在計算資源方面，不論是記憶體使用方面，抑或是執行速度的部分，都和線性工具不相上下，執行速度甚至是比線性工具更快。

2.4 圖基因組回貼工具 HISAT2

其論文中有多項比較：常見工具之比對敏感度 (包含所有模擬序列之比較，或是僅包含單核苷酸多態性 (Single-Nucleotide Polymorphism, SNP) 之序列)、執行速度 (單位為 number of pairs processed per second, PPS) 以及記憶體用量之比較 (Kim et al., 2019)。而 HISAT2 比對敏感度準確、執行速度快速，並且節省計算資源，其關鍵在於演算法採用以圖為基礎 (Graph-based) 的階層結構圖 FM 指數 (Hierarchical Graph FM-index, HGFM)，其中 FM 指數在電腦科學領域中是基於塊排序壓縮 (Burrows-Wheeler Transform, BWT) 之技術，因為 BWT 可應用於數據壓縮的特性，使得 FM 指數允許壓縮輸入文本，同時保有快速查詢子字串 (substring) 的特性。以上種種使得 HISAT2 採用的關鍵演算法 HGFM 能在速度、計算資源以及準確度對於其他工具取得相當的優勢。

而 HISAT2 正是使用此技術，來將原先記錄變異點位之 vcf 格式轉換成其建立參考基因組之格式 snp 檔及 haplotype 檔，並且使用前述技術將其儲存成 HGFM 的索引 (index) 方式，產出的 8 個 ht2 檔即為利用該技術建成的參考基因組，一旦該基因組建立後即可回貼短序列，並且可以根據不同情況使用不同的選項 (option) 或是其中的回貼分數計算方式來決定最後產出的回貼格式：序列比對地圖 (Sequence Alignment Map, SAM)，其中記錄的回貼位置

doi:10.6342/NTU202401388

就會根據前面下的不同指令選項而有所不同。

並且在其中有與其他比對工具比較，HISAT2 共有四個不同的設定，與 BWA-mem、Bowtie2、以及 vg 都各自兩個版本做比較，其佔有的優勢除了運算資源較同為圖基因組的 vg 快許多並且記憶體亦節省許多外，最重要的是其回貼靈敏度 (alignement sensitivity) 較兩者線性者優，因此其強調的是回貼的短序列是否為正確貼回，而非只是能回貼但是回貼正確率或品質並不高。

2.5 線性基因組回貼工具 BWA

BWA (Burrows-Wheeler Aligner) 是一種高效率的序列比對工具，廣泛應用於基因組研究中。其主要功能是將短序列 (reads) 回貼至參考基因組，以便後續的基因組分析。BWA 利用了 Burrows-Wheeler 變換 (Burrows-Wheeler transform, BWT) 和後綴陣列 (suffix array) 等資料結構，使得比對過程既快速又節省記憶體的使用。

BWA 的主要用途包括但不限於以下幾點：BWA 特別適合處理高通量測序技術 (如：Illumina) 產生的短序列樣本 (short reads)，並且比對結果可用於後續的變異檢測，如單核苷酸多態性 (Single Nucleotide Polymorphism, SNP)和序列上的插入或刪除 (insertion/deletion, Indel) 分析。此外，BWA 在全基因組重組、外顯子 (exon) 偵測和 RNA 定序 (RNA-seq) 的比對中也有廣泛應用。

其演算法特點如下：BWA 採用 BWT 和後綴陣列這兩大資料結構，具有以下特點：能夠在較短時間內完成大規模數據的比對任務，因此處理高通量測序數據較為高效。此外，得益於 BWT 和後綴陣列，BWA 在處理大規模基因組數據時較節省記憶體。BWA 亦提供多種比對模式，包括 BWA-backtrack、BWA-SW 和 BWA-MEM，滿足不同應用需求。其中，BWA-MEM 是目前最常用的一種模式，特別適合處理短序列 (short reads) 及長序列樣本 (long reads) 和有較高錯誤率的數據。

doi:10.6342/NTU202401388

BWA 因其高效性、記憶體較為節省並且支援多模式，成為基因組研究中不可或缺的工具。不同應用場景下，研究者需根據具體需求選擇最適合的比對工具，以獲得最佳的分析結果以利後續分析。在此研究中亦探討其對於臺灣人短序列回貼及 KIR 分型與 HISAT2 之結果差異分析。

2.6 圖基因組 KIR 分型工具 Graph-KIR

本研究使用圖基因組演算法為基礎的 KIR 分型工具，Graph-KIR (Lin et al., 2023) 做為 KIR 分型結果並分析之。其使用 Graph-KIR 工具進行 KIR 基因拷貝數估計和等位基因分型。首先，從 IPD-KIR 資料庫 (2020 年釋出之 2.10.0) 獲取 32 個多重序列比對 (Multiple sequence alignment, MSA)，將其合併成 15 個 MSA，並重建 MSA。接著，將 MSA 轉換為 HISAT2 專屬圖參考基因檔。

原先流程為：短序列 (格式：fastq) 經由 BWA 對照參考基因組 (hg38) 進行比對，而在本研究中於此處將建立的 TW-graph 及兩個對照組 hg38-graph 以及 1000G-graph 三個參考基因組加入與 BWA 做比較，並使用 samtools (Danecek et al., 2021) 提取 KIR 區域的短序列，擷取的 KIR 區域短序列隨後透過 HISAT2 比對到前述 15 個 MSA 轉換而成的 HISAT2 專屬圖基因組。最後，Graph-KIR 對每個等位基因分型 (allele typing) 並估算拷貝數 (Copy number)。

doi:10.6342/NTU202401388

第三章　材料與方法

3.1 臺灣人體生物資料庫 (TaiWan Biobank, TWB)

旨在收集並保存臺灣人口的遺傳和健康數據。臺灣人體生物資料庫的建立旨在促進基因組學研究，揭示與疾病相關的遺傳變異，並最終提升公共健康水準和醫療品質。其數據來源包括大量的臺灣志願者，涵蓋了多種族群和不同地區的人口樣本。這些樣本的收集過程包括詳細的健康檢查和生活習慣調查，並通過現代基因組學技術對基因組進行測序和分析。特別是，TWB 中的資料包含豐富的基因變異資訊，這對於識別臺灣人口中特有的遺傳點位具有重要意義。透過這些遺傳點位資訊，研究人員能夠深入探討臺灣人群中常見疾病的遺傳基礎。

本研究為針對全基因組定序 (Whole Genome Sequencing, WGS) 做分析，所使用之資料集為臺灣人體生物資料庫之資料，(TaiWan Biobank 申請案：TWBR10411-03 (項目名稱：Constructing Pharmacogenomics Testing Platform and Discovering Causal Genes of Abnormal Drug Responses；主持人：陳沛隆醫師，共同主持人：陳倩瑜博士))，該檔案經由基因組分析工具包 (Genome Analysis ToolKit, GATK) (DePristo et al., 2011; McKenna et al., 2010) 的途徑而產生，詳細流程記錄於 (Wu et al., 2021)，整體結果是為 TaiwanGenomes 資料庫 (https://genomes.tw/#/)。

該資料集 (TWB) 記錄了 1492 位臺灣人 WGS 變異點資料庫，此資料共 61,424,216 個變異點位，用以建立泛參考基因組 (TW-graph)，各類之變異點數量如圖 2 所示。

doi:10.6342/NTU202401388

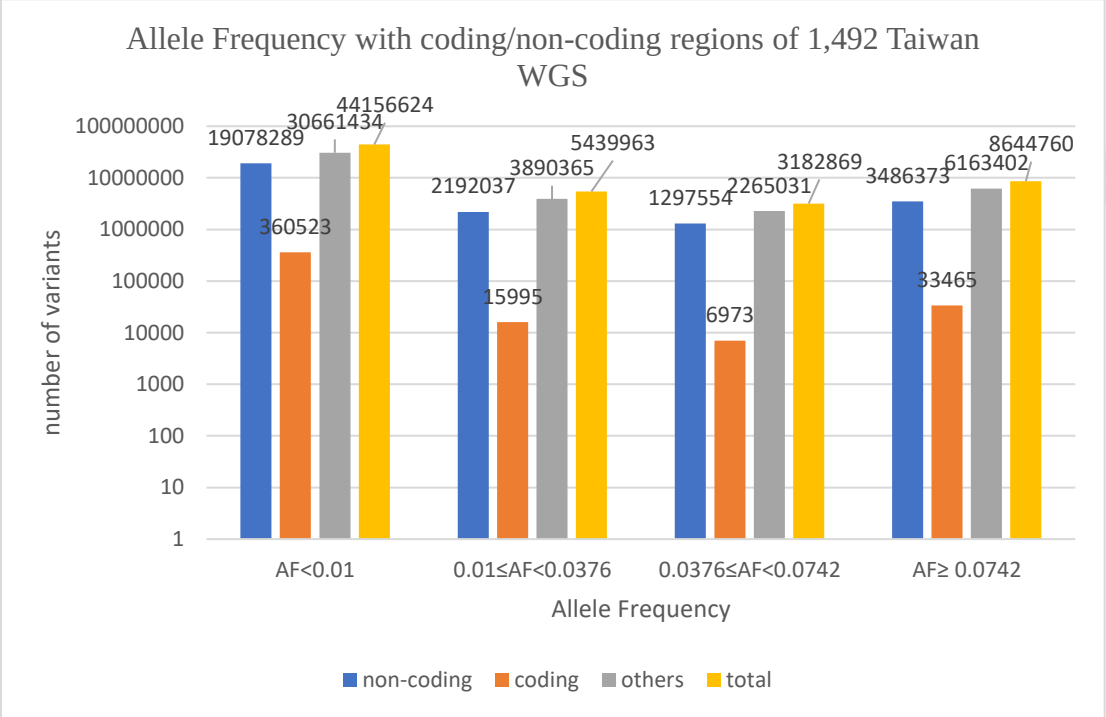


圖 2. 使用專用於 VCF 格式之工具 bcftools 其中的 csq 功能 (Danecek & McCarthy, 2017) 統計出位於編碼區以及非編碼區的統計圖 (經過對數刻度縮放)。其分類編碼及非編碼區的標準則來自另個工具 Ensembl 之 Variant Effect Predictor, VEP (McLaren et al., 2016)，其中灰色者代表 bcftools csq 指令判別不出為編碼區或非編碼區者。可以發現編碼區變異點相較之下少了非編碼區變異點非常多。

由圖 2 可以發現，其實 TWB 僅有少部分的點位是位於編碼區 (coding region)，其數量差距為非編碼區的數百甚至是數千分之一倍，意即多數變異點位為非編碼區的點位，這也是後續能納入篩選考量的部份。而該資料集變異點位密度如表 1，包含全部密度、其他基因區域密度 (除 HLA 及 KIR 以外者)、HLA 以及 KIR 四個區域的密度，可以發現 HLA 以及 KIR 區域確實是較其他區域複雜不少 (密度較密)。

doi:10.6342/NTU202401388

表 1. TWB vcf 各區域密度統計 (總長為 3,217,346,917 鹼基對 (b.p.)，HLA 區域為 4,970,458 b.p.，KIR 區域為 1,058,685 b.p.)

Different regions	Total	AF ≥ 0.01	AF ≥ 0.0376	AF ≥ 0.0742
Total	55,701,082 variants (57.76 b.p./variant)	13,641,766 variants (235.85 b.p./variant)	10,188,728 variants (315.78 b.p./variant)	8,652,777 variants (371.83 b.p./variant)
non-HLAKIR	55,547,870 variants (57.81 b.p./variant)	13,559,797 variants (236.83 b.p./variant)	10,122,598 variants (317.24 b.p./variant)	8,594,993 variants (373.63 b.p./variant)
HLA	153,212 variants (32.44 b.p./variant)	81,969 variants (60.64 b.p./variant)	66,130 variants (75.16 b.p./variant)	57,784 variants (86.02 b.p./variant)
KIR	34,348 variants (30.82 b.p./variant)	11,296 variants (93.72 b.p./variant)	8,194 variants (129.20 b.p./variant)	6,623 variants (159.85 b.p./variant)

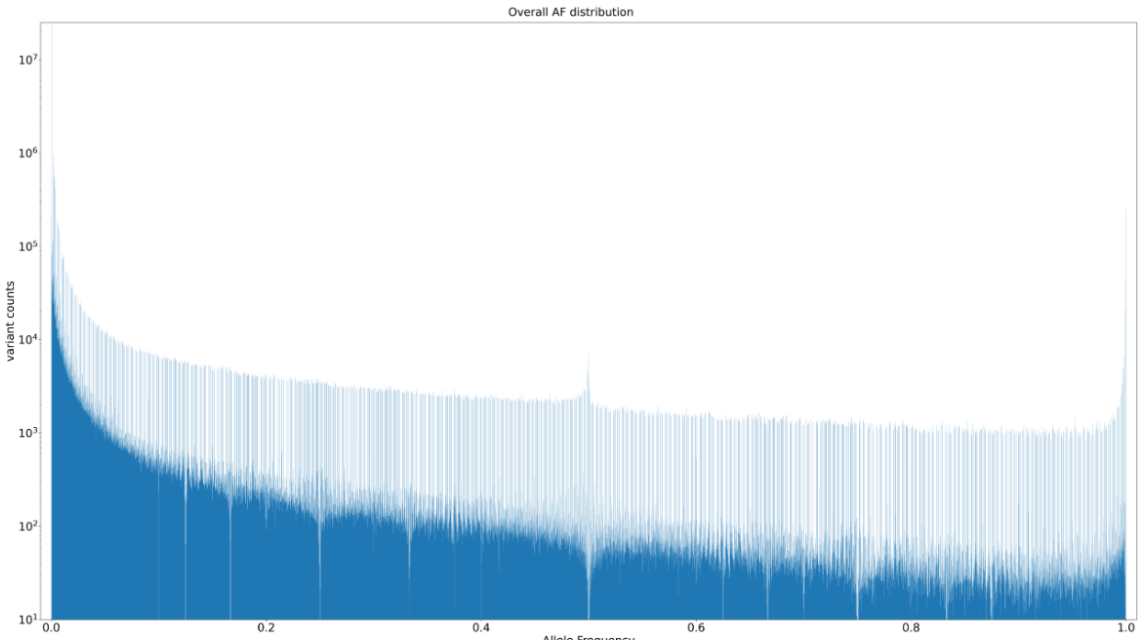


圖 3.變異點之等位基因頻率分布圖。可以發現大多數為較罕見的變異點位，其平均數為 0.061、中位數為 0.001，橫軸為 Allele Frequency，縱軸為 variants counts (log scale)

綜合圖 2 以及圖 3 可以發現，TWB 的變異點位除了多數為非編碼區的變異之外，也有大多數為罕見的變異，因此對於等位基因頻率的篩選是必然需要的。

doi:10.6342/NTU202401388

3.2 人類泛參考基因體聯盟

人類泛參考基因體聯盟 (Human Pangenome Reference Consortium, HPRC) 是一個致力於建立和維護人類泛參考基因組的國際合作組織。此聯盟的目標是通過收集和分析來自不同族群和地理區域的人類基因組數據，創建一個包含更全面的基因變異和結構的參考基因組 (Liao et al., 2023)，以取代現有的單一參考基因組 (如 hg19 和 hg38)。這一舉措旨在更好地反映全球人類遺傳多樣性，並提高基因組研究和應用的精確性。

值得注意的是，HPRC 在 KIR（Killer-cell Immunoglobulin-like Receptor）基因區域的分型結果顯示，使用泛參考基因組的短序列回貼效果顯著優於傳統的單一參考基因組。KIR 基因區域以其高度多樣性和複雜性著稱，對免疫系統功能有重要影響。HPRC 通過結合多個個體的基因數據，成功地提升了 KIR 區域的回貼準確性，從而提供了更為詳細和準確的基因型資訊，這對於研究人類免疫系統和相關疾病具有重大意義。本研究中使用到其中的 44 個短序列樣本及其 KIR 分型結果來分析部分實驗結果。

3.3 建立臺灣人泛參考圖基因組及其他對照組

此階段目標為建立臺灣人泛參考圖基因組 (TW-graph) 及其線性對照組、無加入變異點圖基因組以及全球泛參考基因組 (BWA-linear, hg38-graph, and 1000G-graph)。

首先執行 hisat2_extract_snps_haplotypes_VCF.py，需要 hg38 之 DNA 序列 (格式：.fasta) 做為骨架以及前述之 TWB 資料，格式為變異點偵測格式 (Variant Calling Format, vcf)，接著會產生兩個檔案為 HISAT2 特有記錄點位的格式為 snp 檔以及 haplotype 檔，接著用 HISAT2 的 Linux 指令 hisat2-build，並且一樣的 hg38 之 DNA 序列做為骨架建出我們所需之 TW-graph，此處的圖基因組即為 HISAT2 會產生的 8 個後綴為 ht2 的 HISAT2 專屬圖參考基因檔，有了這 8 個檔案 (泛參考圖基因組)，即可開始對雙端短序列 (Pair-end reads, 檔案格式：.fastq) 回貼。

doi:10.6342/NTU202401388

而為了比較結果，本研究亦準備了不同版本之參考基因組做為對照組，除了經典的線性參考基因組，以 hg38 做為骨架的 BWA-linear，hg38 基因組 (hg38-graph，無加入任何變異點位或單核苷酸多態性之資訊) 以及加入千人基因組計畫 (1000 Genome project, 1000G) 常見單核苷酸多態性 (Common SNPs) 資訊之泛參考圖基因組 (1000G-graph)，為等位基因頻率 (Allele frequency) 至少 1%，其數目為 15,313,604 筆，其中建立 1000G-graph 流程類似 TW-graph，抽取 snp 檔以及 haplotype 檔之程式改為 hisat2_extract_snps_haplotypes_UCSC.py。

其中針對 TW-graph 會有不同的篩選變異點位的標準，如：等位基因頻率大於等於 0.0376 或是大於等於 0.0742，是因為在先前研究 (Pritt et al., 2018) 表明這些是表現較好之等位基因頻率閾值，亦針對免疫相關基因 HLA 及 KIR 等等做特別處理，將位於兩基因所有變異點位留下，其他部份則同前述進行兩個等位基因頻率閾值篩選，如圖 4 所示。

doi:10.6342/NTU202401388

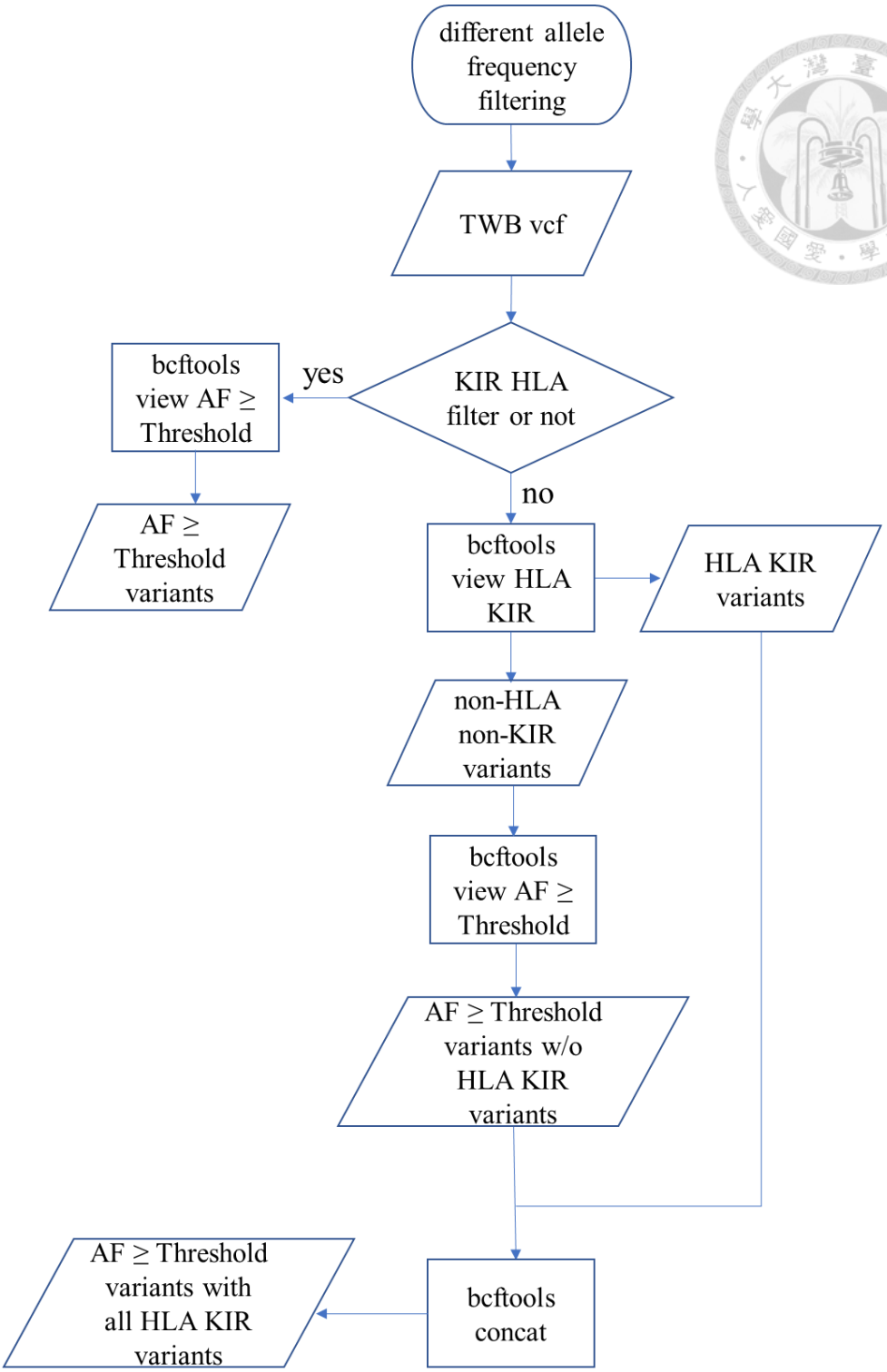


圖 4. 不同等位基因頻率及對特定區域如：HLA 及 KIR 等等進行篩選之流程圖。

Threshold = 0.01 (1%) or 0.0376 (3.76%) or 0.0742 (7.42%)

doi:10.6342/NTU202401388

3.4 短序列回貼比較

準備好前述之各版本參考基因組後即可開始執行回貼，首先將欲回貼之短序列檔案，本研究目前使用之檔案為 TWB 計畫中的七個以及四個非 TWB 共十一個雙端短序列檔案 (pair-end short read, 格式：fastq)。接著使用 Linux 指令 hisat2 來進行回貼，輸出之檔案可藉由參數調整，除了常見之回貼檔案格式，序列比對地圖 (Sequence Alignment Map，格式: sam) 外，HISAT2 可產出回貼概要 (alignment summary) 的純文字檔案，BWA-linear 則使用 samtools 工具裡的 flagstat 指令來觀察其回貼率，將兩者用以初步分析比較不同版本之參考基因組的回貼率 (mapping rate)。HISAT2 建立不同圖基因組、回貼以及最後比較流程圖如圖 5 所示。

doi:10.6342/NTU202401388

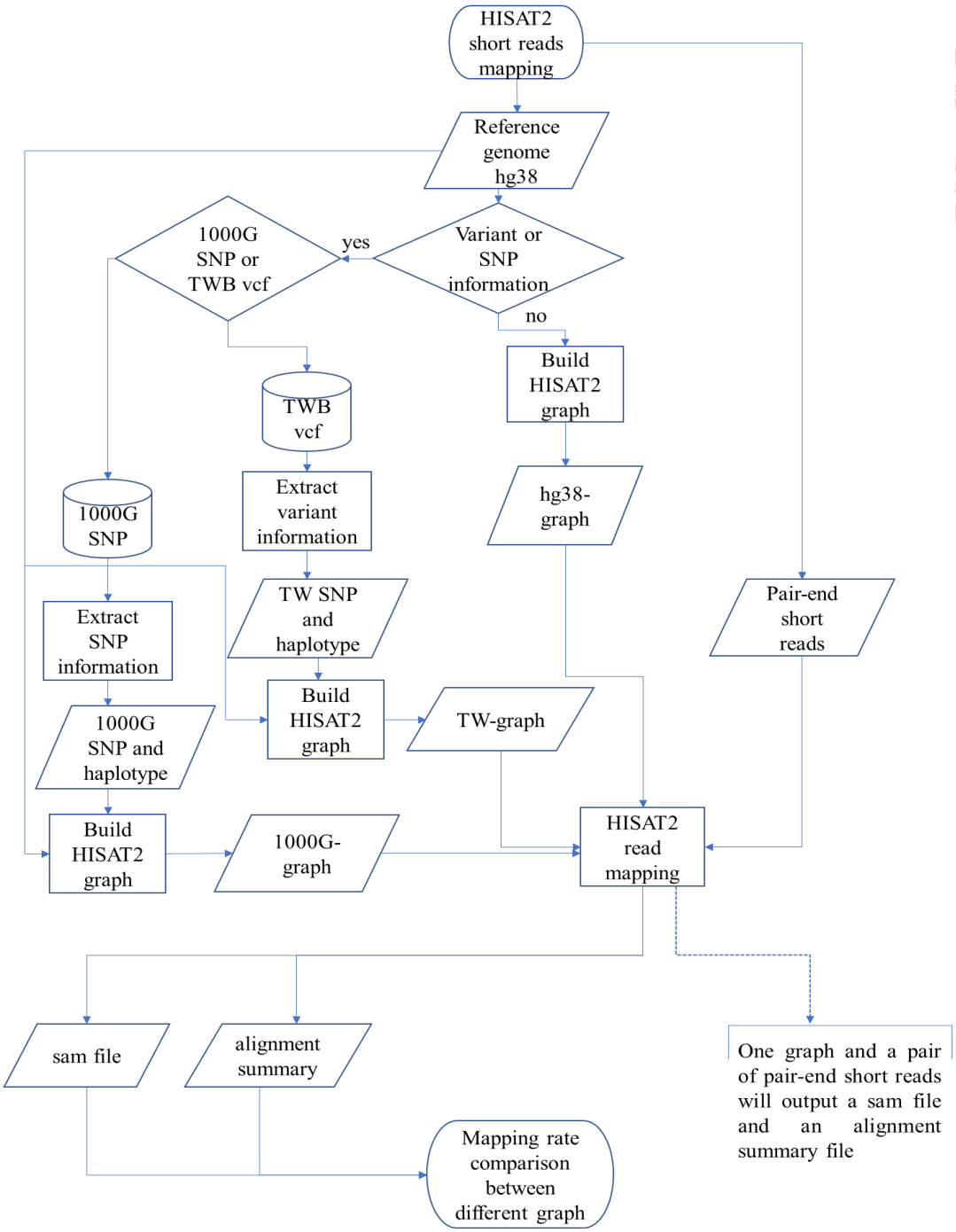


圖 5. 整體利用 HISAT2 來建立參考基因組之 WGS 短序列回貼流程圖

doi:10.6342/NTU202401388

3.5 Graph-KIR 整合臺灣人變異點位

將前述過的 TWB vcf，先利用 hg38 座標位置將 KIR 區域的變異點位篩出，再利用前述的方法將 vcf 格式轉換成 snp 檔和 haplotype 檔，接著使用 BLAST 工具將 hg38 與 Graph-KIR 之十五個 DNA 骨架序列進行比對，並且將位置欄位 (pos column) 轉換後與 Graph-KIR 一致，並且使用此新版本的 Graph-KIR 進行第二階段回貼及 KIR 分型與原版 Graph-KIR 進行結果比較，其流程如圖 6。

doi:10.6342/NTU202401388

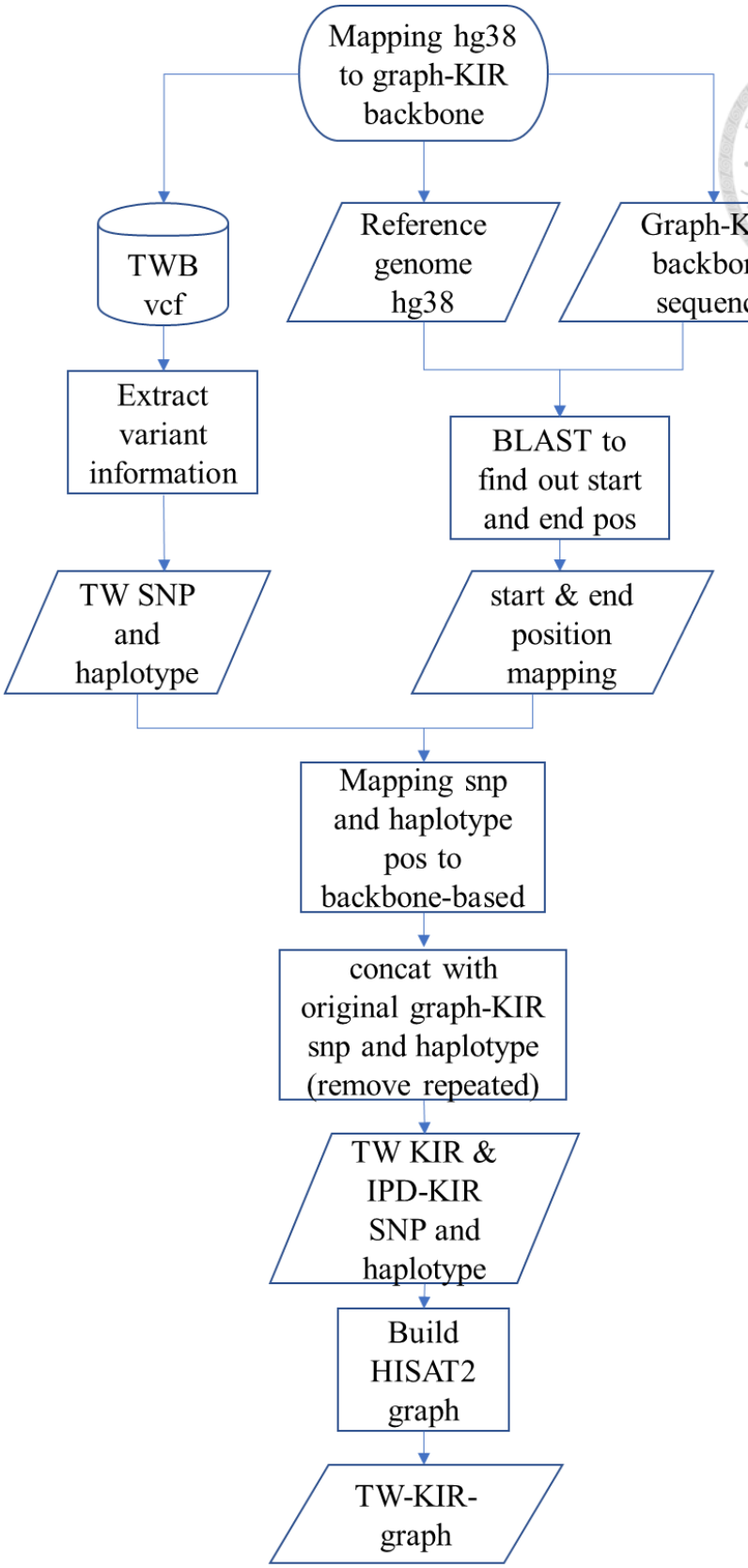


圖 6. 將臺灣人 KIR 區域變異點位與 KIR-graph 對齊整合並重建成 TW-KIR-graph 之流程圖

doi:10.6342/NTU202401388

3.6 短序列回貼對 KIR 分型結果影響分析

在先前得到 sam 檔後進行以下後處理 (postprocessing)，利用 samtools 中的 sort 指令及 view 指令將其壓縮轉成二進位格式的 bam 檔 (Binary Alignment Map)，有了經過這些後處理的 bam 檔後就可以將其作為 Graph-KIR 之輸入，並對其唯一回貼短序列數 (Unique mapped reads, 意即將有重複回貼至不同位置之短序列去除後的序列數目) 以及分型結果進行分析。

doi:10.6342/NTU202401388

第四章　結果與討論

本章節將涵蓋不同面向的結果比較，順序如下：首先 4.1 小節為經過 3.5 小節 (圖 6) 的處理後，TWB 的 KIR 區域的變異點位，與原先 Graph-KIR 工具中的 KIR-graph 及其原始 IPD-KIR 資料庫的重複及數量統計、後續則是針對回貼率的比較 (4.2 小節)、HPRC 的 KIR 分型的分析 (有標準答案) (4.3 小節)、臺灣人的短序列回貼至 KIR 區域概況、KIR 分型分析 (無標準答案) (4.4 小節)、在 KIR-graph 內部整合進 TWB KIR 區域變異點的先後結果差異 (4.5 小節)，以及最後針對各小節的討論 (4.6 小節)。

4.1 原始資料基本統計

由於 TWB 的 KIR 區域的變異點位與 Graph-KIR 工具中原始的 KIR-graph 所使用的 fasta 參考序列不同，因此經過處理之後，會發現 TWB 的 KIR 變異點，會與部份 KIR-graph 的變異點位 (IPD-KIR 之基因頻率 ＞0.1 者) 重複，亦會與 IPD-KIR 所記錄的變異點位重複，而 TWB 本身 KIR 的變異數目與 KIR-graph、IPD-KIR 兩者重複的數目如表 2 表 3 所述。

doi:10.6342/NTU202401388

表 2. TWB vcf KIR 區域點位數目與建立 KIR-graph 之變異數及兩者重複數量

Graph-KIR backbones	TWB variants	KIR-graph variants	overlapped
KIR2DL2	1255	33	0
KIR2DL3	1256	64	6
KIR2DL4	307	83	2
KIR2DL5	274	95	0
KIR2DP1	951	118	18
KIR2DS2	1237	21	0
KIR2DS3	1275	48	2
KIR2DS4	1314	115	3
KIR2DS5	1281	19	0
KIR3DL1	930	154	2
KIR3DL2	723	72	2
KIR3DL3	478	140	7
KIR3DP1	335	18	11
KIR3DS1	950	13	0
KIR2DL1S1	1263	403	28
total	13,829	1,396	81

doi:10.6342/NTU202401388

表 3. IPD-KIR 之全變異數、IPD-KIR 與 KIR-graph 兩者差集的變異數以及 TWB vcf 與差集的變異重複數量 (IPD-KIR 之出現頻率 >0.1 者才會拿去建立 KIR-graph，意即 KIR-graph variants 是 IPD-KIR snp 之子集合)

Graph-KIR backbones	IPD-KIR variants	IPD-KIR freq ≤ 0.1 variants	overlapped
KIR2DL2	85	52	2
KIR2DL3	230	166	27
KIR2DL4	217	134	11
KIR2DL5	226	131	1
KIR2DP1	212	94	6
KIR2DS2	111	90	3
KIR2DS3	137	89	2
KIR2DS4	291	176	3
KIR2DS5	257	238	7
KIR3DL1	490	336	10
KIR3DL2	308	236	8
KIR3DL3	341	201	23
KIR3DP1	117	99	23
KIR3DS1	58	45	1
KIR2DL1S1	645	242	4
total:	3,725	2,329	131

而在扣除所有重複後 (表 2 表 3 之 overlapped 欄位)，會剩下 13617 個 snp，若是無刪除變異則剩下 13052 個 snp，再來需再扣除在 hg38 上被記錄為 snp 但與 Graph-KIR backbone 之序列相同者共重複 2956 個鹼基，因此屬於臺灣人特有的 (無記錄於 IPD-KIR 上的 snp) KIR 區域點位則是 10661 個 snp，而沒有刪除變異 (w/o_deletion) 的則是 10096 個點位。

doi:10.6342/NTU202401388

4.2 WGS 短序列回貼率

除 BWA-linear 者為由 HISAT2 工具所產生檔案之一，回貼概要 (alignment summary)，並且與 BWA-linear 利用 samtools flagstats 產生之回貼結果做比較，記錄回貼率之情形，即為該樣本所有短序列中有多少比例成功回貼至參考基因組上，目前統計樣本數為七個 TWB 的樣本 (NGS1_20170303H, NGS1_20170305G, NGS1_20170312A, NGS1_20170312B, NGS1_20170601D, NGS1_20170602H, NGS1_20170603A) 以及四個非 TWB 樣本 (NV0027-01_S8_L001, NV0027-01_S8_L002, NV0027-01_S8_L003, NV0027-01_S8_L004)，原因為 TWB 樣本可能會因為 TW-graph 裡擁有 TWB 的變異點位資訊進而使 TWB 的短序列 (fastq) 更容易貼至 TW-graph 裡，因此需要非 TWB 的短序列樣本來評估，不論 TWB 短序列樣本與否，加入 TWB 變異點位都能成功使臺灣人的短序列能成功回貼至 TW-graph 上。表 4 表 5 分別為 TWB 樣本及非 TWB 樣本的回貼率。

doi:10.6342/NTU202401388

表 4. TWB 短序列樣本回貼率 (mapping rate) (由上至下分別是 BWA-linear, hg38-graph, 1000G-graph, TW-graph, TW-graph_AF ≥ 0.01, TW-graph_AF ≥ 0.0376, TW-graph_AF ≥ 0.0742, TW-graph_AF ≥ 0.01_all_HLAKIR, TW-graph_AF ≥ 0.0376_all_HLAKIR, TW-graph_AF ≥ 0.0742_all_HLAKIR，短序列樣本順序同前文所述)

Mapping rate	TWB-1	TWB-2	TWB-3	TWB-4	TWB-5	TWB-6	TWB-7
linear	**99.76%**	**99.80%**	**99.88%**	**99.90%**	**99.87%**	**99.88%**	**99.88%**
Graph1	95.96%	96.07%	95.88%	96.47%	96.63%	96.54%	96.34%
Graph2	96.09%	96.22%	96.05%	96.62%	96.76%	96.67%	96.48%
Graph3	96.66%	96.78%	96.67%	97.13%	97.35%	97.34%	97.19%
Graph4	96.68%	96.78%	96.67%	97.13%	97.35%	97.34%	97.20%
Graph5	96.71%	96.81%	96.69%	97.16%	97.37%	97.35%	97.22%
Graph6	96.70%	96.81%	96.69%	97.15%	97.36%	97.35%	97.22%
Graph7	96.68%	96.78%	96.67%	97.13%	97.35%	97.34%	97.20%
Graph8	96.71%	96.81%	96.69%	97.16%	97.37%	97.35%	97.22%
Graph9	96.70%	96.81%	96.69%	97.15%	97.36%	97.35%	97.22%

doi:10.6342/NTU202401388

表 5. 非 TWB 短序列樣本回貼率，由上至下順序同表 4，短序列樣本順序同前文所述

Mapping rate	Non-TWB-1	Non-TWB-2	Non-TWB-3	Non-TWB-4
linear	**99.89%**	**99.89%**	**99.89%**	**99.86%**
Graph1	95.64%	95.61%	95.47%	93.93%
Graph2	95.76%	95.73%	95.60%	94.09%
Graph3	N/A	96.56%	96.43%	94.87%
Graph4	96.66%	96.61%	96.48%	94.91%
Graph5	96.67%	96.63%	96.49%	94.92%
Graph6	96.66%	96.61%	96.47%	94.90%
Graph7	96.66%	96.61%	96.48%	94.91%
Graph8	96.67%	96.63%	96.49%	94.92%
Graph9	96.66%	96.61%	96.47%	94.90%

由表 4 及表 5 兩表可以看出，表現最好的都是 BWA-linear，再者是 TW-graph 接著是 Global-graph，末者為 Global-linear。

doi:10.6342/NTU202401388

4.3 不同參考基因組之 HPRC 短序列樣本之 KIR 分型結果差異

由於目前現在臺灣人的短序列並無 KIR 分型結果作為標準答案，因此先以有分型結果能作為標準答案的 HPRC 短序列樣本以及作為其參考基因組有幫助的基因組：1000G-graph，以及作為比較對照組的另外兩個參考基因組：hg38-graph 以及 BWA-linear 做分型結果比較，首先會先觀察 KIR 區域的唯一回貼短序列數 (表 6)，接著再比較三者之間的分型敏感度 (表 7) (Sensitivity, FN 意即少分型時會使 sensitivity 降低，而 FP 即多分型時則不會)。

表 6. HPRC 短序列樣本於 KIR 區域之唯一回貼短序列數，括號內小數為該行 Unique mapped reads 扣除 BWA 的 Unique mapped reads 之後除以 BWA 的 Unique mapped reads 之比值

Reference genome	Unique mapped reads
BWA-linear	**27,529.91**
hg38-graph	26,753.14 (-0.028)
1000G-graph	26,863.59 (-0.024)

表 7. HPRC 短序列樣本之 KIR 分型結果

Reference genome	Sensitivity	FN	FP
BWA-linear	**0.979**	**17**	**16**
hg38-graph	0.956	36	11
1000G-graph	0.959	34	10

上表可以發現，BWA-linear 表現較 1000G-graph 好又比 hg38-graph 更好，因此在臺灣人短序列樣本 (含 TWB 以及非 TWB 短序列樣本) 中，會先以 BWA-linear 作為標準來比較。

doi:10.6342/NTU202401388

4.4 不同參考基因組之 TWB 短序列樣本之 KIR 分型結果差異

本小節欲探討於第一階段加入臺灣人的變異點對後續 KIR 區域的分型是否有影響，因此一樣會先觀察在 KIR 區域的平均唯一回貼短序列數做為回貼好壞的初步評判標準 (表 8 至表 11，各樣本數的唯一回貼短序列數則記錄在附錄表 1 及附錄表 2)，接著會以 4.3 小節表現最好的 BWA-linear 做為標準來比較不同參考基因組的 KIR 分型結果，此部份無標準答案。

表 8. TWB 短序列樣本於 KIR 區域之唯一回貼短序列數，括號內小數為該行 Unique mapped reads 扣除 BWA 的 Unique mapped reads 之後除以 BWA 的 Unique mapped reads 之比值

Different reference genome	The number of unique mapped reads (mean)	Standard deviation
BWA-linear	26,225.43	3,110.99
hg38-graph	25,346.29 (-0.034)	2,756.69
1000G-graph	25,472.29 (-0.029)	2,805.97
TW-graph	26,307.43 (0.003)	3,144.23
TW-graph_AF ≥ 0.01	26,310.29 (0.003)	3,139.97
TW-graph_AF ≥ 0.0376	26,316.00 (0.003)	3,144.01
TW-graph_AF ≥ 0.0742	26,320.00 (0.004)	3,141.19
TW-graph_AF ≥ 0.01_all_HLAKIR	26,315.14 (0.003)	3,135.96
TW-graph_AF ≥ 0.0376_all_HLAKIR	26,320.00 (0.004)	3,144.24
TW-graph_AF ≥ 0.0742_all_HLAKIR	26,316.57 (0.003)	3,146.06

doi:10.6342/NTU202401388

表 9. 非TWB短序列樣本於KIR區域之唯一回貼短序列數，括號內小數為該行Unique mapped reads 扣除 BWA 的 Unique mapped reads 之後除以 BWA 的 Unique mapped reads 之比值

Different reference genome	The number of unique mapped reads (mean)	Standard deviation
BWA-linear	5,553.5	213.3137
hg38-graph	5,538 (-0.003)	220.1772
1000G-graph	5,541 (-0.002)	215.6873
TW-graph	5,548.67 (-0.001)	247.5282
TW-graph_AF ≥ 0.01	5,572.5 (0.003)	210.1018
TW-graph_AF ≥ 0.0376	5,579.5 (0.005)	212.0536
TW-graph_AF ≥ 0.0742	5,583.5 (0.005)	206.9269
TW-graph_AF ≥ 0.01_all_HLAKIR	5,577.5 (0.004)	217.1699
TW-graph_AF ≥ 0.0376_all_HLAKIR	5,579.5 (0.005)	214.2726
TW-graph_AF ≥ 0.0742_all_HLAKIR	5,578.5 (0.005)	211.8933

表 10. TWB 短序列樣本之 KIR 分型結果 (L: 較 BWA-linear 少分型出的等位基因，M: 較 BWA-linear 多分型出的等位基因)

Result vs BWA (TWB samples)	concordance rate	L	M
hg38-graph	0.948 (128/135)	5	0
1000G-graph	0.948 (128/135)	5	0
TW-graph	0.985 (128/135)	0	0
TW-graph_AF ≥ 0.01	0.985 (133/135)	0	0
TW-graph_AF ≥ 0.0376	0.985 (133/135)	0	0
TW-graph_AF ≥ 0.0742	0.985 (133/135)	0	0
TW-graph_AF ≥ 0.01_all_HLAKIR	0.985 (133/135)	0	0
TW-graph_AF ≥ 0.0376_all_HLAKIR	0.985 (133/135)	0	0
TW-graph_AF ≥ 0.0742_all_HLAKIR	0.985 (133/135)	0	0

doi:10.6342/NTU202401388

表 11. 非 TWB 短序列樣本之 KIR 分型結果 (L: 較 BWA-linear 少分型出的等位基因，M: 較 BWA-linear 多分型出的等位基因)

Result vs BWA (non-TWB samples)	concordance rate	L	M
hg38-graph	0.944 (68/72)	0	1
1000G-graph	0.944 (68/72)	0	0
TW-graph (NV0027-01_S8_L001 = N/A)	0.981 (53/54)	0	0
TW-graph_AF ≥ 0.01	0.986 (71/72)	0	1
TW-graph_AF ≥ 0.0376	0.972 (70/72)	0	0
TW-graph_AF ≥ 0.0742	0.986 (71/72)	0	0
TW-graph_AF ≥ 0.01_all_HLAKIR	1.000 (72/72)	0	0
TW-graph_AF ≥ 0.0376_all_HLAKIR	0.972 (70/72)	0	1
TW-graph_AF ≥ 0.0742_all_HLAKIR	0.986 (71/72)	0	0

由表 8 至表 11 可以發現有加入臺灣人變異點位 (七個 TW-graph) 相較加入全球常見點位 (1000G-graph) 以及沒加點位者 (hg38-graph)，確實是有幫助，與 BWA-linear 結果相距不大。

doi:10.6342/NTU202401388

4.5 相同參考基因組中不同 KIR-graph (有無整合臺灣人 KIR 區域變異點位) 之比較

由於在 4.4 小節的結論中得知，TW-graph 整體雖較 1000G-graph 以及 hg38-graph 好，但是與 BWA-linear 的分型結果差異並不大，因此會需要在後續第二階段 KIR-graph 中也加入屬於臺灣人的變異點，將其加入原先 IPD-KIR 資料庫所沒有的 snp 檔及 haplotype 檔 (前四小節的比較都是在第一階段 WGS 回貼時所建立的圖基因組時有納入變異點)，在此小節會比較有無加入臺灣人點位進 KIR-graph 是否會影響唯一回貼短序列數 (表 12 至表 15，各樣本的唯一回貼短序列數則記錄於附錄表 3 至附錄表 6) 及 KIR 分型結果 (表 16 至表 19)。而四個非 TWB 樣本實則同一人，因此在分型結果上亦有將四者合併後的結果對比統計，記錄在附錄表 7 及附錄表 8。

表 12. 整合臺灣人 KIR 區域變異點位 (TW-KIR-graph) 對於 TWB 短序列樣本之唯一回貼短序列數，括號內小數為該行 Unique mapped reads 扣除 BWA 的 Unique mapped reads 之後除以 BWA 的 Unique mapped reads 之比值

Different reference genome	The number of unique mapped reads (mean)	Standard deviation
BWA-linear	26,218.86	3,114.64
hg38-graph	25,337.71 (-0.034)	2,757.19
1000G-graph	25,463.71 (-0.029)	2,807.33
TW-graph	26,300.86 (0.003)	3,148.02
TW-graph_AF ≥ 0.01	26,303.71 (0.003)	3,143.77
TW-graph_AF ≥ 0.0376	26,309.43 (0.003)	3,147.85
TW-graph_AF ≥ 0.0742	26,313.43 (0.004)	3,144.98
TW-graph_AF ≥ 0.01_all_HLAKIR	26,308.57 (0.003)	3,139.78
TW-graph_AF ≥ 0.0376_all_HLAKIR	26,313.43 (0.004)	3,148.04
TW-graph_AF ≥ 0.0742_all_HLAKIR	26,310 (0.003)	3,149.83

doi:10.6342/NTU202401388

表 13. 整合臺灣人 KIR 區域移除刪除變異點的其他臺灣人變異點位 (TW-KIR-graph_w/o_deletion) 對於 TWB 短序列樣本之唯一回貼短序列數，括號內小數為該行 Unique mapped reads 扣除 BWA 的 Unique mapped reads 之後除以 BWA 的 Unique mapped reads 之比值

Different reference genome	The number of unique mapped reads (mean)	Standard deviation
BWA-linear	26,238	3,107.14
hg38-graph	25,356 (-0.034)	2,749.85
1000G-graph	25,483.71 (-0.029)	2,800.25
TW-graph	26,320 (0.003)	3,140.52
TW-graph_AF ≥ 0.01	26,322.86 (0.003)	3,136.28
TW-graph_AF ≥ 0.0376	26,328.57 (0.003)	3,140.35
TW-graph_AF ≥ 0.0742	26,332.57 (0.004)	3,137.48
TW-graph_AF ≥ 0.01_all_HLAKIR	26,327.71 (0.003)	3,132.29
TW-graph_AF ≥ 0.0376_all_HLAKIR	26,332.57 (0.004)	3,140.54
TW-graph_AF ≥ 0.0742_all_HLAKIR	26,329.14 (0.003)	3,142.33

doi:10.6342/NTU202401388

表 14. 整合臺灣人 KIR 區域變異點位 (TW-KIR-graph) 對於非 TWB 短序列樣本之唯一回貼短序列數，括號內小數為該行 Unique mapped reads 扣除 BWA 的 Unique mapped reads 之後除以 BWA 的 Unique mapped reads 之比值

Different reference genome	The number of unique mapped reads (mean)	Standard deviation
BWA-linear	5,553.5	214.55
hg38-graph	5,538 (-0.003)	221.06
1000G-graph	5,541.5 (-0.002)	217.10
TW-graph	5,553.33 (-0.00003)	251.04
TW-graph_AF ≥ 0.01	5,572.5 (0.003)	211.28
TW-graph_AF ≥ 0.0376	5,580.5 (0.005)	213.84
TW-graph_AF ≥ 0.0742	5,584 (0.005)	208.61
TW-graph_AF ≥ 0.01_all_HLAKIR	5,577.5 (0.004)	218.41
TW-graph_AF ≥ 0.0376_all_HLAKIR	5,580 (0.005)	215.91
TW-graph_AF ≥ 0.0742_all_HLAKIR	5,579.5 (0.005)	213.80

doi:10.6342/NTU202401388

表 15. 整合臺灣人 KIR 區域除刪除變異點位外的其他臺灣人變異點位 (TW-KIR-graph_w/o_deletion) 對於非 TWB 短序列樣本之唯一回貼短序列數，括號內小數為該行 Unique mapped reads 扣除 BWA 的 Unique mapped reads 之後除以 BWA 的 Unique mapped reads 之比值

Different reference genome	The number of unique mapped reads (mean)	Standard deviation
BWA-linear	5,556	212.90
hg38-graph	5,541 (-0.003)	218.51
1000G-graph	5,544 (-0.002)	215.45
TW-graph	5,557.33 (0.0002)	249.56
TW-graph_AF ≥ 0.01	5,575 (0.003)	209.60
TW-graph_AF ≥ 0.0376	5,583 (0.005)	212.18
TW-graph_AF ≥ 0.0742	5,586.5 (0.005)	207.05
TW-graph_AF ≥ 0.01_all_HLAKIR	5,580 (0.004)	216.73
TW-graph_AF ≥ 0.0376_all_HLAKIR	5,582.5 (0.005)	214.27
TW-graph_AF ≥ 0.0742_all_HLAKIR	5,581.5 (0.005)	211.88

表 12 與表 8 相比可以發現，整體唯一回貼短序列數是下降的，表 13 與表 8 相比唯一回貼短序列數是整體上升的。而表 14 與表 9 相比，整體唯一回貼短序列數差異不大，而表 15 與表 9 比較會發現，唯一回貼短序列數微幅上升一些。

而表 16 至表 19 則是針對不同的第一階段的參考基因組以及第二階段 TW-KIR-graph 與 TW-KIR-graph_w/o_deletion 與原本的 KIR-graph 進行分型結果的比較，因無標準答案並且十個參考基因組都會有結果上的變化，因此十個參考基因組皆有與原 KIR-graph 進行結果上的比較。

doi:10.6342/NTU202401388

表 16. 整合臺灣人 KIR 區域變異點位 (TW-KIR-graph) 對於 TWB 短序列樣本 KIR 分型結果對比 (都與相同的參考基因組比較，如：BWA-linear 與 BWA-linear_TW-KIR-graph 比、hg38-graph 與 hg38-graph_TW-KIR-graph 比，以此類推) (L: 較 KIR-graph 少分型出的等位基因，M: 較原版 KIR-graph 多分型出的等位基因)

Result vs KIR-graph (TWB samples)	concordance rate	L	M
BWA-linear	0.830 (112/135)	0	0
hg38-graph	0.838 (109/130)	0	0
1000G-graph	0.831 (108/130)	0	0
TW-graph	0.837 (113/135)	0	0
TW-graph_AF ≥ 0.01	0.837 (113/135)	0	0
TW-graph_AF ≥ 0.0376	0.837 (113/135)	0	0
TW-graph_AF ≥ 0.0742	0.837 (113/135)	0	0
TW-graph_AF ≥ 0.01_all_HLAKIR	0.837 (113/135)	0	0
TW-graph_AF ≥ 0.0376_all_HLAKIR	0.837 (113/135)	0	0
TW-graph_AF ≥ 0.0742_all_HLAKIR	0.837 (113/135)	0	0

doi:10.6342/NTU202401388

表 17. 整合臺灣人 KIR 區域除刪除變異點位外的其他臺灣人變異點位 (TW-KIR-graph_w/o_deletion) 對於 TWB 短序列樣本 KIR 分型結果對比 (都與相同的參考基因組比較，如：BWA-linear 與 BWA-linear_TW-KIR-graph_w/o_deletion 比、hg38-graph 與 hg38-graph_TW-KIR-graph_w/o_deletion 比，以此類推) (L: 較 KIR-graph 少分型出的等位基因，M: 較 KIR-graph 多分型出的等位基因)

Result vs KIR-graph (TWB samples)	Concordance rate	L	M
BWA-linear	0.844 (114/135)	0	0
hg38-graph	0.869 (113/130)	0	0
1000G-graph	0.846 (110/130)	0	0
TW-graph	0.885 (115/135)	0	0
TW-graph_AF ≥ 0.01	0.859 (116/135)	0	0
TW-graph_AF ≥ 0.0376	0.859 (116/135)	0	0
TW-graph_AF ≥ 0.0742	0.859 (116/135)	0	0
TW-graph_AF ≥ 0.01_all_HLAKIR	0.859 (116/135)	0	0
TW-graph_AF ≥ 0.0376_all_HLAKIR	0.859 (116/135)	0	0
TW-graph_AF ≥ 0.0742_all_HLAKIR	0.859 (116/135)	0	0

doi:10.6342/NTU202401388

表 18. 整合臺灣人 KIR 區域變異點位 (TW-KIR-graph) 對於非 TWB 短序列樣本 KIR 分型結果對比 (都與相同的參考基因組比較，如：BWA-linear 與 BWA-linear_TW-KIR-graph 比、hg38-graph 與 hg38-graph_TW-KIR-graph 比，以此類推) (L: 較 KIR-graph 少分型出的等位基因，M: 較 KIR-graph 多分型出的等位基因)

Result vs KIR-graph (non-TWB samples)	Concordance rate	L	M
BWA-linear	0.764 (55/72)	0	0
hg38-graph	0.753 (55/73)	0	0
1000G-graph	0.750 (54/72)	0	0
TW-graph	0.778 (42/54)	0	0
TW-graph_AF ≥ 0.01	0.767 (56/73)	0	0
TW-graph_AF ≥ 0.0376	0.764 (55/72)	0	1
TW-graph_AF ≥ 0.0742	0.778 (56/72)	0	1
TW-graph_AF ≥ 0.01_all_HLAKIR	0.764 (55/72)	0	0
TW-graph_AF ≥ 0.0376_all_HLAKIR	0.781 (57/73)	0	0
TW-graph_AF ≥ 0.0742_all_HLAKIR	0.778 (56/72)	0	1

doi:10.6342/NTU202401388

表 19. 整合臺灣人 KIR 區域除刪除變異點位外的其他臺灣人變異點位 (TW-KIR-graph_w/o_deletion) 對於非 TWB 短序列樣本 KIR 分型結果對比 (都與相同的參考基因組比較，如：BWA-linear 與 BWA-linear_TW-KIR-graph_w/o_deletion 比、hg38-graph 與 hg38-graph_TW-KIR-graph_w/o_deletion 比，以此類推) (L: 較 KIR-graph 少分型出的等位基因，M: 較 KIR-graph 多分型出的等位基因)

Result vs KIR-graph (non-TWB samples)	Concordance rate	L	M
BWA-linear	0.792 (57/72)	0	0
hg38-graph	0.740 (54/73)	0	0
1000G-graph	0.750 (55/73)	0	0
TW-graph	0.815 (44/54)	0	0
TW-graph_AF ≥ 0.01	0.767 (56/73)	0	0
TW-graph_AF ≥ 0.0376	0.792 (57/72)	0	0
TW-graph_AF ≥ 0.0742	0.778 (56/72)	0	0
TW-graph_AF ≥ 0.01_all_HLAKIR	0.778 (56/72)	0	0
TW-graph_AF ≥ 0.0376_all_HLAKIR	0.767 (56/73)	0	0
TW-graph_AF ≥ 0.0742_all_HLAKIR	0.792 (57/72)	0	0

可以從表 16 至表 19 發現，加入臺灣人 KIR 區域之變異點位對於後續第二階段不同的 KIR-graph 的分型結果有相當程度的影響，許多分型結果都產生了改變，並且發現少數 TW-graph 有較 KIR-graph 多分型出等位基因 (表 18)。

doi:10.6342/NTU202401388

4.6 討論

4.6.1 WGS 短序列回貼率

在 0 中可以發現，泛參考圖基因組的表現較無加入變異點圖基因組來的好，而對於臺灣人短序列回貼而言，加入臺灣人之變異點位資料又會較加入全球點位者佳，因此該結果可以做為對於針對臺灣人短序列回貼泛參考基因組具優勢之初步證據。但是由於 BWA-linear 在其中表現又是最好的，因此在沒有短序列之正確解答情況下，無法確定是否如同文獻回顧 2.4 中所言，真的為 HISAT2 回貼之品質就較好但回貼率略微低於 BWA-linear，亦或是 BWA-linear 回貼率較高，並且錯誤率也不高，因此對於後續分析較有幫助。

4.6.2 不同參考基因組之 HPRC 短序列樣本之 KIR 分型結果差異

而在 4.3 中可以發現 BWA-linear 在唯一回貼短序列數以及分型結果正確性都是表現最佳的，除後續分析無答案之臺灣人短序列 KIR 分型結果之三個對照組 (BWA-linear, hg38-graph, 1000G-graph) 中以 BWA-linear 作為結果標準外，還有一個觀察重點為：為何加入全球常見單核苷酸多態性之資訊對於 HPRC 短序列樣本的 KIR 區域之唯一回貼短序列數仍舊較 BWA-linear 少，甚至在 KIR 分型結果的正確率也較 BWA-linear 低，也許與 HISAT2 的指令選項 --no-spliced-alignment 有關，由於 HISAT2 論文中的 Alignment sensitivity 的結果與其他工具的比較時，其選項是有--no-spliced-alignment 的，因此可能會影響分型結果及唯一短序列回貼數等後續結果，這都是值得探討的部分。

4.6.3 不同參考基因組之 TWB 短序列樣本之 KIR 分型結果差異

而在臺灣人短序列樣本的結果中可以發現，唯一回貼短序列數基本上加入臺灣人變異點位的七個版本的參考基因組，都會比三個對照組多，以此可以側面說明，對於臺灣人短序列樣本而言，加入臺灣人變異點位者，也許是真的能使 KIR 區域的基因較正確回貼至正確位置，即便原始的 WGS

doi:10.6342/NTU202401388

回貼率較 BWA-linear 較低，但在 KIR 區域的唯一回貼短序列數卻是較多的，正文表中是平均數量，而各樣本的短序列回貼數在附錄中呈現。

再者就是觀察到 KIR 分型結果的部分，可以觀察到與 hg38-graph 以及 1000G-graph 相比，其結果與 BWA-linear 較為相近，並且較無拷貝數上的差異，但是在沒有標準答案並且 BWA-linear 於 HPRC 表現較其於兩者良好的情況下，對於有無必要使用 TW-graph 等等參考泛基因組於 KIR 區域分型就顯得十分重要，因為結果差異性並不大，但是對於個別樣本可能又會有不同的分型結果甚至是拷貝數變異預測，因此在 KIR 分型中，是否需要在 WGS 回貼中加入臺灣人變異對位亦為重要的討論觀察重點。

4.6.4 相同參考基因組中不同 KIR-graph (有無整合臺灣人 KIR 區域變異點位) 之比較

由目前的結果發現，在第二階段 KIR-graph 整合進臺灣人之 KIR 變異點位，對於後續結果具相當大區別，唯一回貼短序列數的結果來看，TW-KIR-graph 的結果會較原版的差，可能和加入的變異點位太多有關，後續可以考慮將內含子也移除，而 TW-KIR-graph_w/o_deletion 的結果會較好，推測其原因可能為將 hg38 利用 BLAST 對齊 Graph-KIR 所使用之序列時，其刪除點位的座標可能會使後續建立圖基因組時有影響，正文表中是平均數量，而各樣本的短序列回貼數在附錄中呈現。而從分型結果來看，其結果甚至比 4.4 中 TW-graph 與 KIR-graph 之 BWA-linear 還大，因此可以確定對於後續分型結果是有相當程度的影響的，但對於無答案的臺灣人短序列樣本來說的話，暫時無法確認，該影響是好抑或是導致更偏差，此也為待觀察之重點。

4.6.5 其他討論及本研究之限制

而在本研究中，亦有一項相當重要之觀察重點在以上各部分實驗中，就是對於不同的等位基因頻率的閾值所篩選出的變異點位，對於不論是短序列回貼正確率的影響，或是更下游的 KIR 分型結果及拷貝數變異，以目前

doi:10.6342/NTU202401388

的實驗結果來看，七個版本的 TW-graph 差異相距並不大，並且亦看不出規律性，等位基因頻率越大表現不一定越好，而沒有篩選過，將全部臺灣人變異點位都加入者，也並非是表現最好者，因此等位基因頻率的影響尚不明確。

並且從唯一回貼短序列數及其標準差的觀察中發現，其實 TW-graph 對比 BWA-linear 優勢不甚明顯，因此若是能找出是否多貼的那些短序列真的與 TW-graph 的 vcf 記錄的位置相同，更能說明 TW-graph 對後續分析是有幫助的。

此外，針對本研究尚有一些限制及其討論，首先：由於本研究仍進行中，整體以圖基因組演算法為基礎的工具僅有 HISAT2 一個，而該工具可能對於重複性高的序列處理能力不是甚好。再者，HISAT2 本身也有許多不同的指令選項可能需要做調整，如先前提到的：--no-spliced-alignment 等等，可能有些是適合 WGS 定序的，而有些是適合 RNA-seq 的，並且根據不同的使用情境，會有不同偵測敏感度的需求，其指令亦有不同，間隔扣分 (gap penalty) 機制可能也是導致第一階段 WGS 回貼時的回貼率，在 BWA-linear 與其它使用 HISAT2 的差異的主要原因。第三：雖然前面提到等位基因頻率對於後續短序列回貼亦或是 KIR 分型及拷貝數結果差異並不大，但是在建立 HISAT2 專屬圖基因組時有發現一些警告訊息 (Warning messages)，如圖 7 所示，其訊息可以得知建立基因組時可能會有圖過大的問題，而在最大的圖基因組 (TW-graph) 中，甚至會需要使用--large-index 選項才能成功建立參考基因組，並且在後續回貼過程中會使速度變慢，並且在實驗過程中發現，NV0027-01_S8_L001 樣本會無法回貼，將四個非 TWB 樣本 (實際上為同一人，因此合併後與 TWB 樣本結果比較才更合理) 合併後，更是七個 TW-graph 都會有記憶體 (共 1007 G) 不足而無法成功回貼完成的問題，因此篩選變異點位是重要且必要的過程之一，而如何篩選是相當重要的議題，後續研究該要參考 2021 年研究 (Chen et al., 2021)

doi:10.6342/NTU202401388

的標準 AF ≥ 10% (該研究使用 vg 工具)。

```
Warning: a local graph exploded (offset: 0, length: 14819)
Warning: a local graph exploded (offset: 14819, length: 14814)
Warning: a local graph exploded (offset: 29633, length: 11222)
Warning: a local graph exploded (offset: 40855, length: 10100)
Warning: a local graph exploded (offset: 50955, length: 13133)
Warning: a local graph exploded (offset: 64088, length: 14583)
Warning: a local graph exploded (offset: 78671, length: 15106)
Warning: a local graph exploded (offset: 93777, length: 16135)
Warning: a local graph exploded (offset: 109912, length: 15271)
Warning: a local graph exploded (offset: 125183, length: 14575)
Warning: a local graph exploded (offset: 139758, length: 17044)
```

圖 7. hisat2-build 指令，若是參考圖基因組太大時，會出現的警告訊息

因此對於整體流程而言，可能需要加入不同的以圖基因組為基礎之演算法的其他工具，如：vg, vg giraffe, GraphAligner (Rautiainen & Marschall, 2020), PanGenie (Ebler et al., 2022), deBGA (Liu et al., 2016), ODGI (Guarracino et al., 2022), 以及 BrownieAligner (Heydari et al., 2018) 等等，除了 HISAT2 本身的限制性之外，加入不同圖基因組互相比較，亦為較客觀公平之作法，並且從中找出較適合臺灣人，或是較適合 WGS 短序列之回貼工具，再來就是整合臺灣人 KIR 區域點位進入 KIR-graph 的方法也是一個也許有改善的空間，可能兩者序列比對的起點不對，導致加入 KIR-graph 之後會有不同的分型問題等等，因此方法可能需要有一定程度的調整，對於 KIR 部分的分析才會更加完善。

doi:10.6342/NTU202401388

第五章　結論

本研究目的為建立臺灣人泛參考基因組，而在不同版本基因組之建立及後續回貼時，其執行時間於各版本間差異並不大，並且 HISAT2 會略快於 BWA。

而在臺灣人短序列回貼率部分，撇除 BWA-linear 較所有 HISAT2 之回貼率高外 (可能因計分方式或是演算法導致)，HISAT2 眾版本中，七個 TW-graph 會較 hg38-graph 以及 1000G-graph 的回貼率略微高，而 BWA-linear 回貼率較高可能如 4.6 所言，可能正確率較低也不一定，而在 KIR 區域的唯一回貼短序列數可以間接說明，在臺灣人短序列之 KIR 區域回貼是有所幫助的。

而從 4.4 以後可以發現以下結論：於後續 KIR 分型時，雖然臺灣人短序列皆無 KIR 區域分型標準答案，但除了第一階段 WGS 回貼時建立的 TW-graph 外，於第二階段 KIR-graph 整合進臺灣人 KIR 區域變異點位對結果有相當程度影響，使許多短序列樣本中的許多分型結果都與 KIR-graph 有所差異，可能和加入的變異點位內含子遠多過外顯子有關。而在以上兩個主要結果之下，亦有做出不同程度篩選臺灣人變異點位的方式，目前是以等位基因頻率以 0.01、0.0376 與 0.0742 作為三個閾值，並且對 HLA 及 KIR 區域有特別留下全部變異點的版本，共七個版本的臺灣泛參考基因組，從中發現目前等位基因頻率對於後續結果影響差異並不明確，包含前述之 WGS 回貼以及 KIR 分型結果，除全無篩選版本者 (TW-graph) ，會因為變異點位太多，使建立專屬圖基因組檔較慢且需要以選項 large-index 建立外，亦會有樣本對於該參考基因組回貼時，造成記憶體不足的情形，影響後續結果之外，其餘六者 (TW-graph_AF ≥ 0.01, TW-graph_AF ≥ 0.0376, TW-graph_AF ≥ 0.0742, TW-graph_AF ≥ 0.01_all_HLAKIR, TW-graph_AF ≥ 0.0376_all_HLAKIR, TW-graph_AF ≥ 0.0742_all_HLAKIR) 可挑一組作為代表性評估，甚至需要更嚴格的標準 AF ≥ 10% 建立新的 TW-graph。

而未來望可以補足的部分有以下幾點：首先，若是有長讀取序列組裝 (long read assemblies) 做為回貼及 KIR 分型參考答案為佳。再者，在臺灣人整合 KIR 區域的點位上可以用不同的方法常識改善，如：從一開始就將臺灣人的序列加入

doi:10.6342/NTU202401388

MSA 比對，後續建立含臺灣人點位的 KIR-graph 時應會較本文中的方法合理且正確。第三，加入不同的圖基因組回貼工具比較，較為客觀且公平。第四，在第一階段短序列回貼時將--no-spliced-alignment 選項以及間隔扣分 (gap penalty) 等選項加入，以符合 HISAT2 及 Graph-KIR 論文中結果生成的參數選項。第五，利用 bcftools 的 norm 來標準化 TWB 的 vcf 資料，亦有可能影響後續 Graph-KIR 靠左對齊的座標及分型結果。第六，在整合臺灣人 KIR 區域變異點位時目前是無等位基因頻率篩選閾值的，也許可以考慮原先篩選 IPD-KIR 的方式，以頻率做篩選標準 (雖頻率於兩者意義不同，IPD-KIR 是 snp 於某 KIR 分型之出現頻率，而 TWB vcf 則是臺灣人族群之等位基因頻率)，或是其他篩選方式對於 TW-KIR-graph 的建立也會有些許影響。第七，因為四個非 TWB 樣本應該要是合併之後深度才和 TWB 樣本一樣為 30X，因此後續應將四檔案合併後再進行結果比較較為合理，其基本統計亦記錄在附錄中。最後，對於等位基因頻率的評估需再審慎思考，以及是否有其他篩選變異點位之方式。

doi:10.6342/NTU202401388

參考文獻

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., . . . National Eye Institute, N. I. H. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68-74.

Ballouz, S., Dobin, A., & Gillis, J. A. (2019). Is it time to change the reference genome? *Genome Biology*, *20*(1), 159.

Chen, N.-C., Solomon, B., Mun, T., Iyer, S., & Langmead, B. (2021). Reference flow: reducing reference bias using multiple population genomes. *Genome Biology*, *22*(1), 8.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2).

Danecek, P., & McCarthy, S. A. (2017). BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics*, *33*(13), 2037-2039.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J.,

doi:10.6342/NTU202401388

Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491-498.

Ebler, J., Ebert, P., Clarke, W. E., Rausch, T., Audano, P. A., Houwaart, T., Mao, Y., Korbel, J. O., Eichler, E. E., Zody, M. C., Dilthey, A. T., & Marschall, T. (2022). Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. *Nature Genetics*, *54*(4), 518-525.

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*(9), 875-879.

Guarracino, A., Heumos, S., Nahnsen, S., Prins, P., & Garrison, E. (2022). ODGI: understanding pangenome graphs. *Bioinformatics*, *38*(13), 3319-3326.

Heydari, M., Miclotte, G., Van de Peer, Y., & Fostier, J. (2018). BrownieAligner: accurate alignment of Illumina sequencing data to de Bruijn graphs. *BMC Bioinformatics*, *19*(1), 311.

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature*

doi:10.6342/NTU202401388

Biotechnology, *37*(8), 907-915.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357-359.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997. Retrieved March 01, 2013

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., . . . Paten, B. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312-324.

Lin, H.-Y., Chuang, H.-W., Hung, T.-K., Wang, T.-J., Lin, C.-J., Hsu, J. S., Hsu, C.-L., Yang, Y.-C., Chen, P.-L., & Chen, C.-Y. (2023). Graph-KIR: Graph-based KIR Copy Number Estimation and Allele Calling Using Short-read Sequencing Data. *bioRxiv*, 2023.2011.2029.568665.

Liu, B., Guo, H., Brudno, M., & Wang, Y. (2016). deBGA: read alignment with de Bruijn graph-based seed and extension. *Bioinformatics*, *32*(21), 3224-3232.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., & Daly, M. (2010). The Genome Analysis

doi:10.6342/NTU202401388

Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, *20*(9), 1297-1303.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122.

Miga, K. H., & Wang, T. (2021). The Need for a Human Pangenome Reference Sequence. *Annu Rev Genomics Hum Genet*, *22*, 81-102.

Paten, B., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Res*, *27*(5), 665-676.

Pritt, J., Chen, N.-C., & Langmead, B. (2018). FORGe: prioritizing variants for graph genomes. *Genome Biology*, *19*(1), 220.

Rakocevic, G., Semenyuk, V., Lee, W.-P., Spencer, J., Browning, J., Johnson, I. J., Arsenijevic, V., Nadj, J., Ghose, K., Suciu, M. C., Ji, S.-G., Demir, G., Li, L., Toptaş, B. Ç., Dolgoborodov, A., Pollex, B., Spulber, I., Glotova, I., Kómár, P., . . . Kural, D. (2019). Fast and accurate genomic analyses using genome graphs. *Nature Genetics*, *51*(2), 354-362.

Rautiainen, M., & Marschall, T. (2020). GraphAligner: rapid and versatile sequence-

doi:10.6342/NTU202401388

to-graph alignment. *Genome Biology*, *21*(1), 253.

Robinson, J., Halliwell, J. A., McWilliam, H., Lopez, R., & Marsh, S. G. (2013). IPD--the Immuno Polymorphism Database. *Nucleic Acids Res*, *41*(Database issue), D1234-1240.

Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J. A., Hickey, G., Chang, P.-C., Carroll, A., Gupta, N., Gabriel, S., Blackwell, T. W., Ratan, A., Taylor, K. D., Rich, S. S., Rotter, J. I., Haussler, D., Garrison, E., & Paten, B. (2021). Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, *374*(6574), abg8871.

Wu, D.-C., Hsu, J. S.-J., Chen, C.-Y., Shih, S.-H., Liu, J.-F., Tsai, Y.-C., Lee, T.-L., Chen, W.-A., Tseng, Y.-H., Lo, Y.-C., Lin, H.-Y., Chen, Y.-C., Chen, J.-Y., Chang, D. T.-H., Guo, W.-H., Mao, H.-H., & Chen, P.-L. (2021). Complete genomic profiles of 1,496 Taiwanese reveal curated medical insights. *medRxiv*, 2021.2012.2023.21268291.

doi:10.6342/NTU202401388

附錄

由於本文中的唯一回貼短序列數皆為平均數，因此在附錄中特別補充各樣本的唯一回貼短序列做為參考，以下表 1 表 2 為第一階段的不同基因組之 KIR 區域唯一回貼短序列數，表 3 至表 6 為第二階段 TW-KIR-graph 以及 TW-KIR-graph_w/o_deletion 與第一階段不同參考基因組之 KIR 區域唯一回貼短序列數，而最後兩表則是將四個非 TWB 樣本合併之後的分型結果比較。

附錄表 1. 內文中採用平均唯一回貼短序列數 (Unique mapped reads, UMR)，恐對數據解讀判讀有誤，因此附上各樣本對各基因組的數據，此為 TWB 樣本 (由上至下分別是 BWA-linear, hg38-graph, 1000G-graph, TW-graph, TW-graph_AF ≥ 0.01, TW-graph_AF ≥ 0.0376, TW-graph_AF ≥ 0.0742, TW-graph_AF ≥ 0.01_all_HLAKIR, TW-graph_AF ≥ 0.0376_all_HLAKIR, TW-graph_AF ≥ 0.0742_all_HLAKIR，短序列樣本順序: NGS1_20170303H, NGS1_20170305G, NGS1_20170312A, NGS1_20170312B, NGS1_20170601D, NGS1_20170602H, NGS1_20170603A)

UMR	TWB-1	TWB-2	TWB-3	TWB-4	TWB-5	TWB-6	TWB-7
linear	24,006	24,544	29,526	22,698	24,908	25,778	32,118
Graph1	22,892	23,380	28,344	22,698	24,904	24,694	30,512
Graph2	23,030	23,530	28,516	22,692	24,906	24,894	30,738
Graph3	24,126	24,608	29,648	22,712	24,918	25,886	32,254
Graph4	24,132	24,616	29,654	22,720	24,916	25,890	32,244
Graph5	24,134	24,630	29,650	22,712	24,914	25,910	32,262
Graph6	24,126	24,634	29,654	22,728	24,938	25,898	32,262
Graph7	24,140	24,624	29,652	22,728	24,920	25,900	32,242
Graph8	24,130	24,632	29,662	22,722	24,928	25,902	32,264
Graph9	24,120	24,626	29,658	22,722	24,934	25,888	32,268

doi:10.6342/NTU202401388

附錄表 2. 內文中採用平均唯一回貼短序列數 (Unique mapped reads, UMR)，恐對數據解讀判讀有誤，因此附上各樣本對各基因組的數據，此為非 TWB 樣本 (由上至下分別是 BWA-linear, hg38-graph, 1000G-graph, TW-graph, TW-graph_AF ≥ 0.01, TW-graph_AF ≥ 0.0376, TW-graph_AF ≥ 0.0742, TW-graph_AF ≥ 0.01_all_HLAKIR, TW-graph_AF ≥ 0.0376_all_HLAKIR, TW-graph_AF ≥ 0.0742_all_HLAKIR，短序列樣本順序: NV0027-01_S8_L001, NV0027-01_S8_L002, NV0027-01_S8_L003, NV0027-01_S8_L004)

UMR	Non-TWB-1	Non-TWB-2	Non-TWB-3	Non-TWB-4
linear	5,620	5,728	5,676	5,190
Graph1	5,612	5,716	5,662	5,162
Graph2	5,606	5,720	5,664	5,174
Graph3	N/A	5,750	5,696	5,200
Graph4	5,642	5,744	5,690	5,214
Graph5	5,648	5,754	5,698	5,218
Graph6	5,638	5,750	5,714	5,232
Graph7	5,648	5,762	5,692	5,208
Graph8	5,642	5,764	5,696	5,216
Graph9	5,640	5,752	5,704	5,218

由上兩表可以發現其結論與正文中的 4.4 小節一樣，七個 TW-graph 較 BWA-linear 好並且比 1000G-graph 優，最後才是 hg38-graph。而下四表是第二階段加入臺灣人 KIR 區域變異點位的 TW-KIR-graph 以及 TW-KIR-graph_w/o_deletion 的唯一回貼短序列數。

doi:10.6342/NTU202401388

附錄表 3. 內文中採用平均唯一回貼短序列數 (Unique mapped reads, UMR)，恐對數據解讀判讀有誤，因此附上各樣本對各基因組的數據，此為 Phase 2: TW-KIR-graph 對於 TWB 樣本的數據 (由上至下分別是 BWA-linear, hg38-graph, 1000G-graph, TW-graph, TW-graph_AF ≥ 0.01, TW-graph_AF ≥ 0.0376, TW-graph_AF ≥ 0.0742, TW-graph_AF ≥ 0.01_all_HLAKIR, TW-graph_AF ≥ 0.0376_all_HLAKIR, TW-graph_AF ≥ 0.0742_all_HLAKIR，短序列樣本順序: NGS1_20170303H, NGS1_20170305G, NGS1_20170312A, NGS1_20170312B, NGS1_20170601D, NGS1_20170602H, NGS1_20170603A)

UMR	TWB-1	TWB-2	TWB-3	TWB-4	TWB-5	TWB-6	TWB-7
linear	23,998	24,534	29,504	22,712	24,846	25,804	32,134
Graph1	22,884	23,368	28,318	22,710	24,842	24,718	30,524
Graph2	23,020	23,516	28,492	22,706	24,844	24,918	30,750
Graph3	24,118	24,598	29,626	22,726	24,856	25,912	32,270
Graph4	24,124	24,606	29,632	22,734	24,854	25,916	32,260
Graph5	24,126	24,620	29,628	22,726	24,852	25,936	32,278
Graph6	24,118	24,624	29,632	22,742	24,876	25,924	32,278
Graph7	24,132	24,614	29,630	22,742	24,858	25,926	32,258
Graph8	24,122	24,622	29,640	22,736	24,866	25,928	32,280
Graph9	24,112	24,616	29,636	22,736	24,872	25,914	32,284

doi:10.6342/NTU202401388

附錄表 4. 內文中採用平均唯一回貼短序列數 (Unique mapped reads, UMR)，恐對數據解讀判讀有誤，因此附上各樣本對各基因組的數據，此為 Phase 2: TW-KIR-graph_w/o_deletion 對於 TWB 樣本的數據 (由上至下分別是 BWA-linear, hg38-graph, 1000G-graph, TW-graph, TW-graph_AF ≥ 0.01, TW-graph_AF ≥ 0.0376, TW-graph_AF ≥ 0.0742, TW-graph_AF ≥ 0.01_all_HLAKIR, TW-graph_AF ≥ 0.0376_all_HLAKIR, TW-graph_AF ≥ 0.0742_all_HLAKIR，短序列樣本順序: NGS1_20170303H, NGS1_20170305G, NGS1_20170312A, NGS1_20170312B, NGS1_20170601D, NGS1_20170602H, NGS1_20170603A)

UMR	TWB-1	TWB-2	TWB-3	TWB-4	TWB-5	TWB-6	TWB-7
linear	24,022	24,556	29,526	22,734	24,872	25,826	32,130
Graph1	22,906	23,390	28,340	22,730	24,868	24,740	30,518
Graph2	23,044	23,540	28,514	22,728	24,870	24,942	30,748
Graph3	24,142	24,620	29,648	22,748	24,882	25,934	32,266
Graph4	24,148	24,628	29,654	22,756	24,880	25,938	32,256
Graph5	24,150	24,642	29,650	22,748	24,878	25,958	32,274
Graph6	24,142	24,646	29,654	22,764	24,902	25,946	32,274
Graph7	24,156	24,636	29,652	22,764	24,884	25,948	32,254
Graph8	24,146	24,644	29,662	22,758	24,892	25,950	32,276
Graph9	24,136	24,638	29,658	22,758	24,898	25,936	32,280

doi:10.6342/NTU202401388

附錄表 5. 內文中採用平均唯一回貼短序列數 (Unique mapped reads, UMR)，恐對數據解讀判讀有誤，因此附上各樣本對各基因組的數據，此為 Phase 2: TW-KIR-graph 對於非 TWB 樣本的數據 (由上至下分別是 BWA-linear, hg38-graph, 1000G-graph, TW-graph, TW-graph_AF ≥ 0.01, TW-graph_AF ≥ 0.0376, TW-graph_AF ≥ 0.0742, TW-graph_AF ≥ 0.01_all_HLAKIR, TW-graph_AF ≥ 0.0376_all_HLAKIR, TW-graph_AF ≥ 0.0742_all_HLAKIR，短序列樣本順序: NV0027-01_S8_L001, NV0027-01_S8_L002, NV0027-01_S8_L003, NV0027-01_S8_L004)

UMR	Non-TWB-1	Non-TWB-2	Non-TWB-3	Non-TWB-4
linear	5,610	5,736	5,678	5,190
Graph1	5,604	5,722	5,664	5,162
Graph2	5,598	5,728	5,666	5,174
Graph3	N/A	5,760	5,700	5,200
Graph4	5,632	5,752	5,692	5,214
Graph5	5,638	5,762	5,704	5,218
Graph6	5,628	5,758	5,718	5,232
Graph7	5,638	5,770	5,694	5,208
Graph8	5,632	5,772	5,700	5,216
Graph9	5,630	5,760	5,710	5,218

doi:10.6342/NTU202401388

附錄表 6. 內文中採用平均唯一回貼短序列數 (Unique mapped reads, UMR)，恐對數據解讀判讀有誤，因此附上各樣本對各基因組的數據，此為 Phase 2: TW-KIR-graph_w/o_deletion 對於非 TWB 樣本的數據 (由上至下分別是 BWA-linear, hg38-graph, 1000G-graph, TW-graph, TW-graph_AF ≥ 0.01, TW-graph_AF ≥ 0.0376, TW-graph_AF ≥ 0.0742, TW-graph_AF ≥ 0.01_all_HLAKIR, TW-graph_AF ≥ 0.0376_all_HLAKIR, TW-graph_AF ≥ 0.0742_all_HLAKIR，短序列樣本順序: NV0027-01_S8_L001, NV0027-01_S8_L002, NV0027-01_S8_L003, NV0027-01_S8_L004)

UMR	Non-TWB-1	Non-TWB-2	Non-TWB-3	Non-TWB-4
linear	5,608	5,738	5,682	5,196
Graph1	5,602	5,724	5,668	5,170
Graph2	5,596	5,730	5,670	5,180
Graph3	N/A	5,762	5,704	5,206
Graph4	5,630	5,754	5,696	5,220
Graph5	5,636	5,764	5,708	5,224
Graph6	5,626	5,760	5,722	5,238
Graph7	5,636	5,772	5,698	5,214
Graph8	5,630	5,774	5,704	5,222
Graph9	5,628	5,762	5,712	5,224

而以上附錄表 1 至附錄表 6 也與 4.5 小節的結論相似，在 TWB 樣本中，TW-KIR-graph_w/o_deletion > KIR-graph > TW-KIR-graph，而在非 TWB 樣本中則是相距都不大，TW-KIR-graph 以及 TW-KIR-graph_w/o_deletion 會略為比 KIR-graph 多一些。

doi:10.6342/NTU202401388

由於四個非 TWB 樣本其實是同一人，因此在後續分型結果分析時，將四個檔案合併後，用原本的方法回貼、以及最後一步的分型，其分型結果比較統計如下：

附錄表 7. 整合臺灣人 KIR 區域變異點位 (TW-KIR-graph) 對於非 TWB 短序列樣本 (四檔案合併後) KIR 分型結果對比 (都與相同的參考基因組比較，如：BWA-linear_KIR-graph 與 BWA-linear_TW-KIR-graph 比、hg38-graph_KIR-graph 與 hg38-graph_TW-KIR-graph 比，以此類推) (L: 較 KIR-graph 少分型出的等位基因，M: 較 KIR-graph 多分型出的等位基因)

Result vs KIR-graph (non-TWB samples)	Concordance rate	L	M
BWA-linear	0.778 (14/18)	0	0
hg38-graph	0.889 (16/18)	0	0
1000G-graph	0.833 (15/18)	0	0
TW-graph	0.833 (15/18)	0	0
TW-graph_AF ≥ 0.01	0.833 (15/18)	0	0
TW-graph_AF ≥ 0.0376	0.833 (15/18)	0	0
TW-graph_AF ≥ 0.0742	0.833 (15/18)	0	0
TW-graph_AF ≥ 0.01_all_HLAKIR	0.833 (15/18)	0	0
TW-graph_AF ≥ 0.0376_all_HLAKIR	0.833 (15/18)	0	0
TW-graph_AF ≥ 0.0742_all_HLAKIR	0.833 (15/18)	0	0

doi:10.6342/NTU202401388

附錄表 8. 整合臺灣人 KIR 區域變異點位 (TW-KIR-graph_w/o_deletion) 對於非 TWB 短序列樣本 (四檔案合併後) KIR 分型結果對比 (都與相同的參考基因組比較，如：BWA-linear_KIR-graph 與 BWA-linear_TW-KIR-graph_w/o_deletion 比、hg38-graph 與 hg38-graph_TW-KIR-graph_w/o-deletion 比，以此類推) (L: 較 KIR-graph 少分型出的等位基因，M: 較 KIR-graph 多分型出的等位基因)

Result vs KIR-graph (non-TWB samples)	Concordance rate	L	M
BWA-linear	0.778 (14/18)	0	0
hg38-graph	0.778 (14/18)	0	0
1000G-graph	0.722 (13/18)	0	0
TW-graph	0.833 (15/18)	0	0
TW-graph_AF ≥ 0.01	0.944 (17/18)	0	0
TW-graph_AF ≥ 0.0376	0.944 (17/18)	0	0
TW-graph_AF ≥ 0.0742	0.944 (17/18)	0	0
TW-graph_AF ≥ 0.01_all_HLAKIR	0.944 (17/18)	0	0
TW-graph_AF ≥ 0.0376_all_HLAKIR	0.944 (17/18)	0	0
TW-graph_AF ≥ 0.0742_all_HLAKIR	0.944 (17/18)	0	0

最後兩表可以發現一樣是 TW-graph 會比 1000G-graph 以及 hg38-graph 與 BWA-linear 較接近，並且 TW-KIR-graph 帶來的變化會比 TW-KIR-graph_w/o_deletion 更為劇烈，以上與正文 4.5 小節中的結果一致，但是在第二階段 TW-KIR-graph 以及 TW-KIR-graph_w/o_deletion 一樣會使得其分型結果與原先不同，但因合併後的樣本數只剩一個，因此可能需要更多樣本來說明，加入臺灣人 KIR 變異點會使分型結果有所改變。

doi:10.6342/NTU202401388