

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

Graph-KIR 在真實資料上的評估
及在台灣族群之應用

Evaluation of Graph-KIR on Real Data and its Application
to Taiwanese Population

王亭堅

Ting-Jian Wang

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



致謝



回首望去，畢業的時光來得如此迅速，彷彿昨日才剛踏進台大開始碩班生活。在這兩年，我衷心感謝身邊的同學顯鈞、青勁、冠穎和承軒，你們的優秀和支持讓這段旅程更加豐富，能與你們共事是我的榮幸，從你們身上學到了許多。感謝實驗室的學長姐名翔、柏豪、毓聰、如秀和弘曄，你們走在前面，卻不忘提攜和關心我們，熱心的和我們分享經驗並給予建議，讓我們更有前進的動力。特別感謝倩瑜老師的指導。自從加入實驗室後，老師引領我們參與各種實驗和計畫，提供同儕之間合作和學習的機會，並推動我們與外界交流，獲得更豐富的學術經驗。研究上，老師的建議和鼓勵對我來說彌足珍貴，這兩年在老師身邊學到了很多，非常感謝老師。剛加入 c4Lab 時，理工科出生的我在生物資訊領域文獻閱讀上遇到不小的困難，好在碩一下時上了陳沛隆醫師、陳倩瑜老師和許書睿老師精心開設的 NGS 課程，讓我有機會補齊基因和病理相關的知識，縮短認知的差距，對所學開始有更深入的了解，萬分感謝老師們。感謝考研時帶我參觀台大宿舍的敏彰，以及一路上給了我許多建議的政彤和駿榮，是你們的啟發和鼓勵讓我在 Gap year 期間找到了攻讀碩士的動力。最後，感謝一直陪伴在我身邊的家人和羽雯，你們的信任和支持是我前進最大的動力。

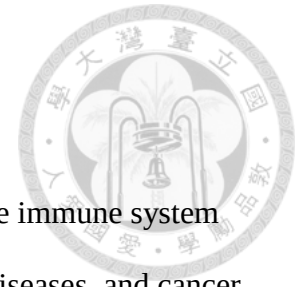
中文摘要



殺手細胞免疫球蛋白樣受體（Killer Immunoglobulin-like Receptors，KIR）基因在免疫系統中扮演關鍵角色，與器官移植、自體免疫疾病及癌症密切相關。KIR 基因表現高度多態性，這使得透過次世代高通量短序列進行個體等位基因型識別充滿挑戰。實驗室先前開發的 Graph-KIR 在生成資料與真實資料的雙重驗證中證實為目前表現最佳的次世代定序（Next-Generation Sequencing，NGS）高通量短序列 KIR 等位基因分型工具，優於其主要競爭對手 PING。然而，Graph-KIR 在真實資料與生成資料的表現存在落差。儘管聚焦在真實資料中非新穎等位基因的部分，其表現仍然不及生成資料。為了釐清 Graph-KIR 在真實資料中遭遇的挑戰，本研究深入分析先前 Graph-KIR 在人類全基因組參考聯盟（Human Pangenome Reference Consortium，HPRC）中發生的錯誤預測案例。HPRC 分析結果顯示，Graph-KIR 在缺失變異區域的短序列回貼表現不佳，缺失變異的忽略或錯誤識別，使正確答案無法被輸出。此外，拷貝數錯誤預測案例的分析顯示了工具對合併基因拷貝數預測的困難。本研究從常見錯誤案例探討算法優化的可能性，並進行了初步的更新與校正，最新版本的 Graph-KIR 在真實資料上表現顯著改善。優化後的算法被運用於 Taiwan Biobank 台灣人 1492 樣本的 KIR 基因分型，測試其在大規模樣本群的實用性。初步結果中，少量樣本因模型缺乏對拷貝數預測深度區間的限制，預測出異常的拷貝數分布。模型經過調整，將二倍體的參考資訊納入適配過程，提升 Graph-KIR 拷貝數預測的自動性與可靠性，並輸出相對合理的結果。最後，本研究統計了 1492 台灣人的全基因定序資料的等位基因分型結果，計算個體中常見的拷貝數分布以及族群中高頻率出現的等位基因型。本研究首次報導的台灣人高精度 7 碼 KIR 等位基因統計，為未來族群 KIR 研究及個人化精準醫療的發展奠定了重要基礎。

關鍵字：殺手細胞免疫球蛋白樣受體、次世代定序、拷貝數、等位基因、台灣人

英文摘要



Killer Immunoglobulin-like Receptors(KIR) play a crucial role in the immune system and are closely associated with organ transplantation, autoimmune diseases, and cancer. The high polymorphism of KIR gene family makes individual genotyping through next-generation sequencing(NGS) high-throughput short-read a challenging task. The previously developed Graph-KIR has been proven to be the best-performing NGS KIR genotyping tool to date, outperforming its main competitor, PING, in both simulated and real data validation. However, there is a performance gap between simulated and real data for Graph-KIR, with its performance on non-novel alleles in real data still falling short of its performance on simulated data. To elucidate the challenges faced by Graph-KIR in real data, this study conducted an in-depth analysis of the previous erroneous prediction cases of Graph-KIR on Human Pangenome Reference Consortium(HPRC) real data. The HPRC analysis results indicate that Graph-KIR performs poorly in aligning short reads in regions with deletion variants, leading to the incorrect or missed identification of deletion variants and the inability to output correct results. Additionally, the analysis of copy number error prediction cases revealed difficulties in predicting merged gene copy numbers. This study explored the potential for algorithm optimization from common error cases and performed preliminary updates and corrections, resulting in significant performance improvements of the latest version of Graph-KIR on real data. The optimized algorithm was applied to KIR gene typing of 1492 samples from the Taiwan Biobank to verify its practicality on a large sample set. Preliminary results showed a small number of samples with abnormal copy number distributions due to the model's lack of constraints on copy number prediction depth intervals. Diploid reference information was incorporated into the model to limit the

fitting interval to 2 copy numbers, enhancing the automaticity and reliability of Graph-KIR copy number prediction. Finally, this study summarized the 7-digit allele typing results of the whole genome sequencing data of 1492 Taiwanese individuals, calculating the common copy number distributions in individuals and the high-frequency allele types in the population, providing important reference data for subsequent KIR studies in the Taiwanese population.

Keywords: Killer Immunoglobulin-like Receptors, Next generation sequencing, Copy number, Allele, Taiwanese

目次



致謝	i
中文摘要	ii
英文摘要	iii
目次	v
圖次	viii
表次	ix
第一章 前言	1
1.1 KIR	1
1.2 KIR 基因多樣性	1
1.3 NGS 高通量短序列 KIR 基因分型	2
1.4 研究目標與實驗設計	4
第二章 文獻回顧	5
2.1 T1K	5
2.2 PING	5
2.3 Graph-KIR	6
2.3.1 拷貝數預測模型與預測策略	7
2.3.2 等位基因分型算法	7
2.4 KIR 族群等位基因型統計	8



第三章 研究方法.....	9
3.1 HPRC 資料集.....	9
3.2 TBB 資料集	10
3.3 等位基因分型算法檢驗.....	10
3.4 等位基因分型算法更新.....	11
3.4.1 缺失變異分數調整	11
3.4.2 同型合子等位基因分型算法	11
3.5 拷貝數預測模型更新	12
第四章 研究結果與討論	13
4.1 HPRC 真實資料拷貝數預測錯誤案例分析	13
4.2 HPRC 非新穎變異等位基因錯誤預測案例分析	14
4.2.1 案例解析： <i>KIR3DP1</i> 同分預測	15
4.2.2 案例解析：同型合子預測偏差	17
4.3 分型算法更新 HPRC 驗證表現	19
4.4 Graph-KIR 在台灣人族群上的實作與分析.....	21
4.4.1 拷貝數異常預測與模型缺陷校正結果.....	21
4.4.2 台灣人拷貝數分布統計	24
4.4.3 台灣人等位基因型統計	25

4.5 未來工作.....	27
第五章 結論.....	28
參考文獻.....	29
附錄 A.....	31
附錄 B.....	34
附錄 C.....	36



圖次



圖 1 KIR 等位基因命名法	2
圖 2 <i>KIR3DP1</i> 基因前段長片段缺失區域無序列回貼	16
圖 3 <i>KIR3DP1</i> 基因前段長片段缺失區域 WGS 回貼狀況	16
圖 4 同型合子被預測為異型合子案例在缺失變異區域的回貼狀況 ..	19
圖 5 TBB 1492 人 <i>KIR3DL3</i> 拷貝數分布，	22
圖 6 二倍體基因平均深度與 WGS 回貼平均深度的關聯性	23
圖 7 不正常拷貝數預測結果範例。樣本 ID: NGS1_20170109C	23
圖 8 校正後拷貝數預測結果範例。樣本 ID: NGS1_20170109C	23
圖 9 參考二倍體基因平均深度與 KIR 區域基因深度之關聯性	24
圖 10 台灣 1492 人前 10 常見個體 KIR 基因拷貝數分布	25
圖 11 台灣 1492 人各 KIR 基因最常見的等位基因型，等位基因頻 率大於 5% 的等位基因顯示於圖上	26

表次



表 1 SKIRT 新穎變異標註符號.....	10
表 2 HPRC 44 樣本各基因拷貝測錯誤預測案例統計	14
表 3 非新穎等位基因型錯誤預測案例和實際檢驗案例數目統計	15
表 4 HPRC 44 樣本允許 <i>KIR3DP1</i> 同分預測之驗證準確度	17
表 5 Graph-KIR 各算法更新版本於 HPRC 44 樣本之表現.....	20

第一章 前言

1.1 KIR

殺手細胞免疫球蛋白樣受體（Killer Immunoglobulin-like Receptors，KIR）是一類位於自然殺手細胞及少數 T 細胞表面的受體，這類受體通過與人類白細胞抗原（Human Leukocyte Antigen，HLA）的互動，來激起或抑制免疫細胞的殺傷程序。KIR 基因具備多態性且在個體間存在差異，這些差異與器官移植的免疫排斥、多種癌症的發展及自體免疫疾病的進程密切相關。精確識別 KIR 等位基因型對於深入理解其不同疾病中的作用極為關鍵。這一過程將加速相關基礎研究的進展，並有望優化臨床應用，提供針對性更強的治療策略。

1.2 KIR 基因多樣性

KIR 基因家族位於人類第 19 號染色體 q13.4 的區域，序列長度介於 70 kbp 到 270 kbp 之間。本研究所探討的 KIR 多樣性，大致分為以下三個層級：

（1）基因層級多樣性；（2）拷貝數變異（Copy Number Variation，CNV）多樣性；以及（3）等位基因型多樣性。首先，KIR 家族由 17 種 KIR 基因所組成，各基因長度介於 5 kbp 到 13 kbp，分別為 8 個激活基因（*KIR2DS1*、*KIR2DS2*、*KIR2DS3*、*KIR2DS4*、*KIR2DS5*、*KIR3DS1* 和 *KIR2DL4*）、7 個抑制基因（*KIR2DL1*、*KIR2DL2*、*KIR2DL3*、*KIR2DL5A*、*KIR2DL5B*、*KIR3DL1*、*KIR3DL2* 和 *KIR3DL3*）和 2 個假基因（*KIR2DP1* 和 *KIR3DP1*）。再來，各 KIR 基因存在拷貝數變異的可能，拷貝數變異通常介於 0 到 4 拷貝數，其中拷貝數為 0 到 2 的情形最常見，拷貝數為 3 較少發生，拷貝數為 4 十分罕見，各 KIR 基因拷貝數變異發生的機率也不同，如 *KIR3DL3* 在不同樣本中拷貝數通常都為 2，不易發生拷貝數變異。最後，同 KIR 基因上的序列變異，為等位基因型層級的多樣性。不同的等位基因命名如圖 1 所示，KIR 等位基因由 7 碼標註，前 3 碼的差異，表示在 KIR 基因編碼區域（Coding region）存在影響轉錄

的胺基酸序列的非同義變異 (Nonsynonymous variant)。中間 2 碼則用於區分編碼區域區域存在變異，但不影響轉錄胺基酸序列的同義變異 (Synonymous variant)。而最後 2 碼則用於標註非編碼區 (Noncoding Region) 存在序列差異。當前已被發現的等位基因型紀錄於 IPD-KIR 資料庫 (Robinson et al., 2012)，該資料庫提供等位基因序列的多序列比對資訊。KIR 的等位基因資料庫目前並不完備，仍有許多新穎的 KIR 等位基因型尚未被發現和記錄。最新版本為 IPD-KIR-2.13 於 2023 年 12 月釋出，共紀錄 2219 個 KIR 等位基因型。儘管 KIR 基因呈現高度的多樣性，但是 KIR 基因的基因結構十分相似，不同的基因間有相似的外顯子 (exon) 和內涵子 (intron) 結構。除此之外，各 KIR 基因之間及 KIR 等位基因型之間序列亦有相當高的一致性，先前研究也透過對 KIR 等位基因型的序列比對來顯現基因間序列的相似度 (Marin et al., 2021)。

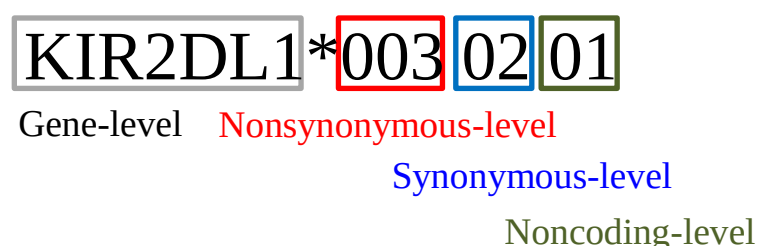


圖 1 KIR 等位基因命名法

1.3 NGS 高通量短序列 KIR 基因分型

隨著次世代測序 (Next-Generation Sequencing, NGS) 技術的發展，短序列分析已成為基因體學研究中強大的工具，尤其是由於其低成本和高通量的特性。這些優勢使得 NGS 技術更具普及性，有潛力為廣泛的用戶提供服務，從而促進個性化醫學的進展。

要從 NGS 短序列拼湊出個體 KIR 基因型的全貌並不容易，KIR 高度多態且高度相似的特性，造成傳統線性短序列回貼方法的不適用。短序列在 KIR



基因之間會交互回貼，源自於個體的 KIR 序列無法從短序列回貼結果識別。由於 KIR 區域序列變異組合的複雜性，工具得透過與傳統變異呼叫流程不同的機制，搭配對各個 KIR 基因拷貝數變異的預測，開發算法確認樣本存在的等位基因型。此外，由於資料庫中只片面記載了已確認存在的等位基因型，而真實樣本中仍可能存在未知的新穎等位基因，所以在 KIR 等位基因分型時，工具必須在新穎變異存在的情況下，以最相似的已知等位基因型標註新穎等位基因型，甚至報導存在的新穎變異，讓使用者看出樣本 KIR 基因型的實際樣貌。

實驗室團隊先前開發的 NGS 高通量短序列 KIR 等位基因分型工具 Graph-KIR (Hong-Ye Lin et al., 2023)，以圖形段序列回貼策略產生的精確回貼結果為基礎，搭配提出的拷貝數預測及等位基因分型策略，有效克服了傳統線性回貼參考序列資訊不充足所造成的偏差，對樣本 KIR 等位基因型做出更加準確的預測。該工具經生成資料及真實資料雙重驗證，為當前 NGS 短序列 KIR 基因分型領域的領先解法，優於當前其他同樣以 NGS 短序列為輸入資料的工具，如 PING (Marin et al., 2021) 和 T1K (Li song et al., 2023)。

Graph-KIR 在不存在新穎等位基因的生成資料上，有完美的拷貝數預測結果，並達到九成以上的等位基因分型準確度。然而，在真實資料上，Graph-KIR 拷貝數預測仍存在錯誤案例，且等位基因分型的準確度與生成資料相比下降了不少。就算只觀察真實資料中非新穎等位基因的預測狀況，仍與生成資料的表現有約兩成的落差，這代表面對真實世界樣本的變化性，工具等位基因型仍存在改進的空間。

1.4 研究目標與實驗設計

為提升 Graph-KIR 對真實世界樣本 KIR 基因識別的準確度，並增加其在醫療和後續臨床分析中的應用價值，本研究對先前 Graph-KIR 在 HPRC 真實資料的驗證結果進行更加深入的探討。本實驗首先將分析 Graph-KIR 在拷貝數預測中出現的錯誤，並分析其在真實資料上非新穎等位基因分型表現不如生成資料的原因。釐清工具在真實資料上遇到的挑戰，是進步的重要依據，本研究也期望針對所發現的問題，對當前算法提出修正和更新，加強 Graph-KIR 對真實資料分型結果的可靠性。最後，本研究使用該工具對台灣基因體資料庫（Taiwan Biobank, TWBB）（Feng et al., 2022）中 1492 名台灣人樣本資料進行 KIR 等位基因分型，測試 Graph-KIR 在大規模樣本群的實用性，並對台灣族群的 7 碼 KIR 等位基因分型結果進行統計分析，期望這些數據能為未來台灣族群 KIR 基因型相關研究提供支持。

第二章 文獻回顧

本章節分為四部分，前三部分介紹當前主要的 NGS 高通量短序列 KIR 等位基因分型工具，分別為 T1K、PING 和 Graph-KIR，最後一部分介紹族群 KIR 等位基因分型之研究。

2.1 T1K

T1K (Li Song et al., 2022) 以類似轉錄組定序 (RNA sequencing, RNA-Seq) 計算豐度的方式，使用最大期望算法 (Expectation-maximization algorithm) 計算樣本短序列與各 KIR 等位基因間的豐度，最終為每個 KIR 基因輸出兩個等位基因型。該工具優勢為可接受 RNA-Seq、全基因定序 (Whole genome sequencing, WGS) 和全外顯子測序 (Whole exon sequencing, WES) 資料對樣本 KIR 和 HLA 等位基因分型。此外在等位基因分型過程中，T1K 有效利用了大多數工具排除的在相似 KIR 基因間多重回貼的短序列。不過儘管 T1K 針對高度多態基因提出了新穎的等位基因分型流程，該工具在 KIR 分型中對各基因只輸出兩個等位基因的設定並不符合 KIR 拷貝數可能大於二的情況，KIR 拷貝數的預測成為該工具對 KIR 等位基因分型表現的侷限。對於 KIR 序列的分型，雖然工具有提供 5 碼和 7 碼的分型選項，但 T1K 在期刊中僅強調了其在 3 碼蛋白質序列層級分型的成果，未提出數據支持工具對於處理高精度完整 KIR 序列的能力。

2.2 PING

PING (Pushing Immunogenetics to the Next Generation) (Marin et al., 2021) 如其名期許將免疫基因體研究推廣至次世代基因定序資料上，該研究開發了一個直觀的分析流程，以 NGS 短序列作為輸入資料，進行 KIR 拷貝數預測及高精度的基因分型。PING 的特色是使用了多重參考序列和特色虛擬探針。多重參考序列由各個 KIR 基因篩選 5 個等位基因型構成，經過驗證，此方法能有效降低



原先單一參考序列造成短序列無法正確映射的問題。特色虛擬探針應用於偵測各個 KIR 基因的拷貝數，以及用於偵測一些因短序列錯貼而難以識別其存在性的目標基因型。探針的長度為 32 bp，用來抓取涵蓋與其完全一致內容的讀取序列，在拷貝數預測階段，各個 KIR 基因探針抓取的短序列數目，會經由 *KIR3DL3* 探針抓取的短序列數目正規化，來得到深度比，最後整樣本群中各個 KIR 基因的深度比，以人工的方式區分拷貝數變異。

在 2023 年 PING 的研究單位發表了將此流程運用於 WGS 的成果，以下簡稱此版工具為 PING-wgs (Marin et al., 2023)。在 1366 個生成資料中，PING-wgs exon 區域的基因分型可以達到 95% 的精確度，而面對由 1000 Genomes Project 提供的 215 個歐洲人的真實資料，平均每個 KIR 基因中，無法定出 exon 區域的基因型接近或低於所有等位基因總數的 5%，代表面對存在新穎變異的真實資料，PING-wgs 也無法達到如生成資料這般優秀的結果。

PING-wgs 是目前除了 Graph-KIR 外，對 WGS 資料達到最高 5 碼精確度等位基因分型的工具，不過該工具仍存在許多缺點：(1) 該工具無法偵測非編碼區的序列資訊，進行 7 碼精度的等位基因分型；(2) 在拷貝數預測算法中，PING 將 *KIR3DL3* 視為二倍體去正規化其他基因的深度數值，但仍有研究顯示 *KIR3DL3* 存在結構變異的風險，使得該作法並非完全可靠；(3) 需要人工來識別拷貝數變異；(4) PING 的 KIR 參考序列及虛擬探針，無法隨著 IPD-KIR 的版本更新而更新，而是停留在該工具開發時的版本 (IPD-KIR v2.7.1)。

2.3 Graph-KIR

Graph-KIR (Hong-Ye Lin et al., 2023) 利用圖形回貼算法 (Daehwan Kim et al., 2019) 來應對 KIR 基因間過度相似的問題，使短序列回貼準確度上升，並基於更可靠的深度資訊與短序列上的變異資訊，開發拷貝數預測模型和等位基因分型算法。



2.3.1 拷貝數預測模型與預測策略

Graph-KIR 拷貝數預測模型由多個常態分佈組成，該模型模擬了個體中 KIR 相同拷貝數基因短序列回貼深度的聚集及不同拷貝數深度的線性分佈特性。模型參數由拷貝數分布的線性關西及拷貝數的發生機率來設定。工具透過調整拷貝數為 2 常態分佈平均數的深度位置，來匹配模型與樣本 KIR 基因深度，最終當模型與樣本的 KIR 基因深度擬合後，拷貝數便由深度隸屬度最高的拷貝數常態分佈來確定。模型中未使用 *KIR3DL3* 作為 2 拷貝數深度的參考依據，避免了 *KIR3DL3* 拷貝數變異存在造成正規化偏差的可能性，但也使得模型缺乏二倍體深度參考資訊，僅依照模型擬合程度來預測拷貝數。

針對短序列在圖形回貼時仍容易交互回貼，導致無法獲得可靠回貼深度的兩個高度相似基因組合 *KIR2DL5A* 和 *KIR2DL5B*，以及 *KIR2DL1* 和 *KIR2DS1*，Graph-KIR 在圖形映射階段會先合併相似基因組的圖形參考索引，透過合併圖形上的總回貼深度判斷合併基因組的總拷貝數，合併基因組內個別的基因拷貝數再經由後續等位基因分型算法來識別。該策略能降低過度相似基因間短序列回貼的偏差，使拷貝數預測表現上升，但也使得拷貝數的預測高度依賴等位基因分型結果，當分型算法存在錯誤預測時，這些基因的拷貝數預測也會受到影響。

2.3.2 等位基因分型算法

Graph-KIR 由個體中個別 KIR 基因短序列上的變異資訊，來計算個別 KIR 基因中，同預測拷貝數大小的等位基因組合與短序列回貼結果的相似度分數。由於 IPD-KIR 並未提供等位基因的變異標註資訊，Graph-KIR 由等位基因型的多序列比對建構相對於共識序列的變異點資料庫，並在相似度計算時，偵測短序列在標註變異點位的序列與各等位基因型的相似度，每偵測到一個與對照等位基因型不相符的變異，在程式算法中稱其為「負變異」，就會使短序列與等位基因相似度降低 0.001。短序列與拷貝數大於等於 2 拷貝的等位基因組合的相似度，則



由組合中與短序列相似度分數較高的等位基因型決定。最後，所有短序列與對照等位基因組合相似度的總乘機取上取上以十為底數的對數，則為樣本在該 KIR 基因中特定等位基因組合的相似度分數，分數最高的等位基因組合會被視作預測結果輸出。此算法能有效的包容異型合子的狀況，使短序列只要與等位基因組合中的任一等位基因序列相似，就不會過度降低整體的相似度分數。然而，先前驗證結果也顯示存在輸出多組同相似度預測的情形，此算法鑑別短序列與等位基因組合匹配程度的效能有待進一步的檢驗。

2.4 KIR 族群等位基因型統計

先前對於短序列 KIR 基因分型的族群統計研究，通常僅限於 KIR 基因層級的分析或是對樣本單倍體基因組合的推測，缺乏更高解析度的等位基因層級分析與統計。PING 是少數在等位基因層級進行族群研究的團隊 (Leonardo M. Amorim et al., 2021)，研究對象為 2130 個無親緣關係的健康成年北美人，但由於工具限制，他們的分析僅限於 5 位數的 KIR 等位基因型，並且 PING 在真實資料驗證的表現也沒有 Graph-KIR 來得優秀。Graph-KIR 使本研究能夠突破基因層級的分析限制，得以在群體中研究 7 位數的等位基因型。

第三章 研究方法



本章節將分為五個部分，前兩部分介紹本研究所使用的驗證資料集和台灣族群資料集，第三部分介紹檢驗等位基因錯誤預測案例所使用的工具與方式，第四部份介紹更新的等位基因分型算法，第五部分則介紹更新的拷貝數預測模型。

3.1 HPRC 資料集

評估 Graph-KIR 的真實樣本集由人類全基因組參考聯盟 (Human Pangenome Reference Consortium, HPRC) (Wang et al., 2022) 提供。在總共 47 個樣本中，44 個在 NCBI 上有可取得的全基因定序雙端短序列 (WGS pair-end short reads) 樣本資料如下：HG002、HG00438、HG005、HG00621、HG00673、HG00733、HG00735、HG00741、HG01071、HG01106、HG01109、HG01175、HG01243、HG01258、HG01358、HG01361、HG01891、HG01928、HG01952、HG01978、HG02055、HG02080、HG02109、HG02145、HG02148、HG02257、HG02572、HG02622、HG02630、HG02717、HG02723、HG02818、HG02886、HG03098、HG03453、HG03486、HG03492、HG03516、HG03540、HG03579、NA18906、NA19240、NA20129、NA21309。驗證 Graph-KIR 在 HPRC 數據集基因分型表現的參考答案由 SKIRT (Tsung-Kai Hung et al., 2023) 提供，該研究將來自 IPD-KIR 數據庫的等位基因映射到高品質的基因分型單倍體上，這些單倍體是通過組裝 HiFi 長序列數據所生成，透過映射結果對樣本等位基因分型標註。已注釋的等位基因經過人工驗證後，在資料集中發現了許多新穎的等位基因，各種等位基因的標示如表 1 所示。

表 1 SKIRT 新穎變異標註符號

符號	意義	實例
#	用於識別在編碼區域中具有非同義變異、分段刪除或與另一基因融合的等位基因	KIR3DL3# (99.92% of 04101)
+	用於識別在編碼區域中具有同義變異的等位基因	KIR3DL3*035+, KIR3DL3*01406+
=	用於識別與 IPD-KIR 只記錄外顯子的等位基因匹配的等位基因	KIR2DS5*003=
\$	用於識別具有非編碼區變異的等位基因	KIR2DS4*00601\$

3.2 TBB 資料集

台灣基因體資料庫 (Taiwan Biobank, TWBB) (Feng et al., 2022) 由 101 年開始執行，至今已涵蓋了大於 150,000 多名：(1) 年齡介於 30 到 70 歲、(2) 具台灣國籍、(3) 未罹患癌症台灣人的身體檢測、檢體採集及健康相關問卷資訊。本研究利用 108 年申請計畫中 (TWBB application case: TWBR10411-03; Project: Constructing Pharmacogenomics Testing Platform and Discovering Causal Genes of Abnormal Drug Responses; Principal Investigator: Dr. Pei-Lung Chen) 取得的 1492 WGS pair-end short reads 資料，來驗證 Graph-KIR 在大規模樣本群的實用性，並對台灣族群基因分型結果做進一步的統計與分析。

3.3 等位基因分型算法檢驗

本研究專注於探討 HPRC 驗證結果中非新穎等位基因錯誤預測發生的原因。首先，本實驗使用由 Hong-Ye Lin 開發的 HLA 和 KIR 等位基因型多序列比對可視化工具 pyHLAMSA (來源：<https://github.com/linnil1/pyHLAMSA>)，協助找出錯誤預測等位基因和標註答案等位基因之間有差異的變異點位。接著透過 IGV (Integrative Genomics Viewer) (Helga Thorvaldsdóttir et al., 2013) 觀察在變異點位 Graph-KIR 的短序列實際回貼情況。最後，從程式碼中撈取等位基因分



形算法對分數計算的資訊，包括各等位基因或等位基因組合的相似度分數，以及每個短序列上相對對照等位基因型所偵測到的負變異。通過這些步驟，Graph-KIR 變異點偵測的正確性將被確認，以此查找標註答案等位基因相似分數較低無法被輸出的原因。本實驗並未詳細觀察全部的錯誤案例，而是對錯誤案例進行初步的分類，分類參考的指標包含基因拷貝數同型、同型異型合子和來源基因等，將狀況相近的案例歸類，並在同一類別中挑選幾個案例做深入的分析。

3.4 等位基因分型算法更新

本小節介紹 Graph-KIR 算法的修正與更新，包括兩個部分，第一部分介紹對於為缺失變異（Deletion variant）之負變異的分數調整方式與策略，第二部分介紹同型合子基因的識別與等位基因分型的細節。

3.4.1 缺失變異分數調整

本實驗調整了 Graph-KIR 偵測到缺失變異點位之負變異的扣分，相似度降低幅度由原本的 0.001 逐步減少，測試不同分數調整對 HPRC 驗證結果的影響，同時觀察分數調整對錯誤案例修正的比例，以此來決定最終的調整分數。

3.4.2 同型合子等位基因分型算法

本實驗觀察所有短序列有覆蓋到的變異點位，並判斷該點位是否為異型合子點位，只有深度大於 10 的變異點且存在多於一種變異的點位會被判斷，若深度次多的變異深度除以總深度大於拿基因拷貝數乘以 2 去除 1 的值，則被視為一個異型合子點位。唯有沒有識別出任何異型合子點位的基因，能被視作同型合子等位基因。針對同型合子等位基因，單一等位基因型與短序列的相似分數最高的等位基因型會直接被輸出為 Graph-KIR 的預測結果



3.5 拷貝數預測模型更新

本研究將二倍體深度資訊加入適配模型，來限制拷貝數為二的常態分佈區間，使模型適配於合理的深度位置。受到 Stargazer (Seung-Been Lee et al., 2019) 啟發，二倍體的深度資訊由 3 個基因提供，分別為 VDR、RYR1 和 EGFR，這 3 個基因在 Genomic Variant (MacDonald JR et al., 2014) 資料庫中顯示低全基因缺失或複製機率。在 Graph-KIR 完成 WGS 短序列回貼後，上述 3 個基因的回貼平均深度被截取，作為算法二倍體參考深度 μ_d ，並計算 3 基因回貼深度的標準差 σ_d ，作為二倍體深度的參考浮動區間。本實驗將拷貝數預測模型中拷貝為 2 常態分佈的平均值 μ_2 限制於近似二倍體深度的區間 $\mu_d \pm \sigma_d$ ，來與 KIR 基因深度資訊匹配。更新後的 Graph-KIR 拷貝數預測辦法亦可假設 KIR3DL3 拷貝數為 2 來預測樣本 KIR 拷貝數，此時 μ_d 由樣本 KIR3DL3 深度提供，使模型擬合後拷貝數為 2 的常態分佈與 KIR3DL3 深度值所在的聚落適配。

第四章 研究結果與討論



本論文的實驗結果分為五個部分，前兩部分為 Graph-KIR 在 HPRC 資料集上錯誤預測案例的分析結果，分別探討在拷貝數預測和等位基因分型兩階段所發現的問題，第三部分為等位基因分型算法更新後 HPRC 之驗證結果，第四部份為 Graph-KIR 應用在台灣族群上的成果，第五部分為當前研究的未來工作。

4.1 HPRC 真實資料拷貝數預測錯誤案例分析

在 HPRC 樣本拷貝數預測結果中，共有 33 個基因發生錯誤的拷貝數預測，其中 30 個錯誤預測來自 Graph-KIR 中的合併基因組 *KIR2DL5AB* 和 *KIR2DL1S1* (表 2)。Graph-KIR 在尚未利用等位基因分型算法區分合併等位基因組內個別基因的拷貝數前 (Before allele typing)，拷貝數預測模型能準確的由合併基因組的總深度預測合併基因組的總拷貝數。然而在等位基因分型後 (After allele typing)，分型算法無法完美區分合併基因組內個別基因的拷貝數，所有屬於合併基因組的錯誤預測皆發生在此階段。本研究發現 Graph-KIR 等位基因分型算法在計算相似度分數時，只要短序列上的變異與等位基因組合中其中一個等位基因相符，則認為該短序列與等位基因組相似而不會調降相似度。當拷貝數較多的時候，需要計算的相似度的等位基因型組合也較多，且短序列只要組合中其一等位基因型相似，就不會降低相似度分數，造成短序列與不同等位基因組合間輸出的相似度難以被鑑別。合併基因的總拷貝數通常較單獨基因的拷貝數多，上述問題在合併基因拷貝數的預測時被體現出來，造成了拷貝數變異預測的差異。

至於另外 3 個錯誤案例，2 個發生在不同樣本的 *KIR3DL2* 上，回貼結果顯示與答案標註拷貝數不成比例的回貼深度。其中一個案例的回貼深度為 37 被預測為 2 拷貝，且該深度與常態為 2 拷貝的 *KIR3DL3* 深度 44 相近，然而在答案標註中 *KIR3DL2* 的拷貝數為 1，*KIR3DL2* 短序列實際回貼情形是否正確有待進一步的確認 (附圖 A-1、附圖 A-2)。最後一個案例發生在 *KIR2DL2*，答案拷

貝數為 1，實際回貼深度為 8，被預測為 0 拷貝。完全缺失基因的回貼深度通常接近 0，但在本案例中，儘管有一定的回貼深度，卻仍會被預測為完全缺失（附圖 A-3）。對於完全缺失的狀況，縮小模型中拷貝數為 0 分布常態分佈的範圍或是以深度閾值來判斷，或許才是更恰當的做法。

表 2 HPRC 44 樣本各基因拷貝測錯誤預測案例統計

CN diff \ KIR gene	2DL1	2DS1	2DL5A	2DL5B	3DL2	2DL2	Total
Before allele typing	0	0	0	0	2	1	3
After allele typing	9	9	6	6	2	1	33

4.2 HPRC 非新穎變異等位基因錯誤預測案例分析

HPRC 非新穎等位基因錯誤預測的案例一共有 41 個，為提升驗證效率，本實驗並未檢驗所有案例，而是在不同的錯誤預測狀況中挑選數個案例解析，並歸納錯誤預測背後的原因。表 3 紀錄對案例檢驗的初步分類和數目，以及實際深入檢驗的案例數目，被歸類在假基因（Psuedo gene）和合併基因組（Merged gene）的案例不會被放在其他類別。經過對 13 個錯誤案例的詳細分析，本研究發現除了前述對合併基因等位基因分型的困難，Graph-KIR 於存在缺失變異區域的短序列回貼表現不佳是造成錯誤預測的主因。諸多案例中顯示，在標註答案等位基因型為缺失變異的區域，仍常存在序列回貼，導致答案等位基因的相似度分數下降，無法被作為答案輸出。此外，在缺失變異點位也經常發生短序列無法被回貼的狀況，特別是位於 KIR 等位基因中經常缺失的區域或是答案等位基因型長片段缺失區域，結果顯示回貼深度通常較淺，甚至無任何短序列回貼。該現象在一些案例中，連帶影響了缺失變異周圍的序列回貼與其他變異識別。實際個案短序列回貼情形可參考附錄 B。同樣是缺失變異造成的錯誤預測，以下兩小節將介紹本研究歸納的兩種對準確度影響最大的狀況，詳細呈現其產生錯誤預測的過

程，分別為 *KIR3DP1* 的多組等位基因產生同分預測的情形，以及同型合子易被判為異型合子的偏差。

表 3 非新穎等位基因型錯誤預測案例和實際檢驗案例數目統計

Classification	Number of cases	Number of Evaluated cases
Pseudo gene	16	3
Answer: Homozygous Predict: Heterozygous	6	2
Answer: Heterozygous Predict: another Heterozygous	5	2
Answer: Homozygous Predict: another Homozygous	4	1
Answer: Heterozygous Predict: Homozygous	1	1
CN = 1	4	2
Merged gene	5	2
Total	41	13

4.2.1 案例解析：*KIR3DP1* 同分預測

在非新穎變異等位基因型 41 個錯誤預測中，有 16 個案例來自假基因的錯誤預測，本研究發現了 11 個案例屬於 *KIR3DP1*004e*、*KIR3DP1*0020202* 和 *KIR3DP1*0020210* 之間的錯誤預測，並挑選了數個案例檢驗。Graph-KIR 對這 3 個等位基因計算出相同的相似度分數。IGV 回貼結果顯示，在圖形映射後 *KIR3DP1* 在前 2000 bp 區域幾乎沒有任何序列回貼，pyHLAMSA 顯示 *KIR3DP1* 在該區域經常存在大片段且不連續的缺失（圖 2）。由於該區域沒有

任何序列回貼，缺失片段中間的序列變異也無法被識別，導致上述 3 個需要靠此區域變異區分的等位基因型分數相同，最後 Graph-KIR 只能在多個同分的預測組合中輸出隨機的結果，驗證準確度因而下降。本研究也發現在全基因映射階段，*KIR3DP1* 在開頭區域深度本來就非常淺（圖 3），圖中白色短序列為低映射品質（Mapping quality）分數的短序列回貼，或許短序列無法回貼不僅僅是圖形映射算法的問題。



圖 2 *KIR3DP1* 基因前段長片段缺失區域無序列回貼

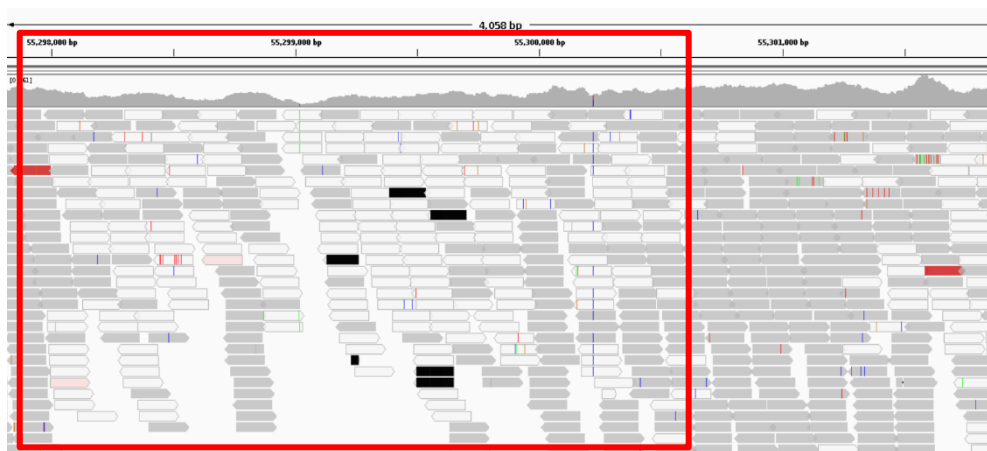


圖 3 *KIR3DP1* 基因前段長片段缺失區域 WGS 回貼狀況

若將上述同分預測納入 Graph-KIR 預測結果，只要同分預測中涵蓋答案標註的等位基因型就算正確預測，Graph-KIR 在 HPRC 數據集的表現如表 4 所示。加入同分預測後，Graph-KIR 分數明顯提升，特別在去除所有新穎等位基因的階段，7 碼準確度由 0.711 提升至 0.817。雖然同分預測的問題並未真正被解決，但接納同分預測後分數改變的幅度證實 *KIR3DP1* 同分狀況對工具表現的影響，也說明若能解決此問題，基因分型準確度改善的潛力。

表 4 HPRC 44 樣本允許 *KIR3DP1* 同分預測之驗證準確度

Filter	Methods	7-digits	5-digits	3-digit	Gene
-fusion	Number of alleles	248	713	825	825
	Graph-KIR-origin	0.706	0.739	0.749	0.979
	Graph-KIR-modified	0.770	0.750	0.760	0.979
-fusion -#	Number of alleles	237	651	740	740
	Graph-KIR-origin	0.713	0.745	0.766	0.978
	Graph-KIR-modified	0.781	0.756	0.776	0.978
-fusion -# -+ -=-	Number of alleles	222	516	548	548
	Graph-KIR-origin	0.721	0.814	0.865	0.978
	Graph-KIR-modified	0.793	0.826	0.876	0.987
-fusion -# -+ -=- \$	Number of alleles	142	143	145	145
	Graph-KIR-origin	0.711	0.853	0.897	0.979
	Graph-KIR-modified	0.817	0.888	0.924	0.979

Graph-KIR-origin: Graph-KIR 原始版本。

Graph-KIR-modified: Graph-KIR 允許 *KIR3DP1* 同分預測之版本。

4.2.2 案例解析：同型合子預測偏差

對於參考答案標註為同型合子的 KIR 基因，Graph-KIR 常會產生錯誤的異型合子預測結果。在非新穎變異錯誤案例中，有 6 個錯誤預測屬於此類情形，IGV 回貼結果顯示這些案例確實為同型合子，不該被預測為異型合子。不過因為標註答案的缺失變異區域存在與其不相符的回貼結果，導致答案同型合子等位基



因組合在 Graph-KIR 中的相似度分數降低。同時，特定異型合子等位基因組合因為涵蓋與缺失變異區域回貼狀況相符的序列，而不會在該區域被扣分，最終產生了相對高的相似度分數。以圖 4 樣本 HG00438 在 2 拷貝數的 *KIR2DL4* 的等位基因分型狀況為例，答案標註為 *KIR2DL4**0010201 的同型合子，而分型結果卻是 *KIR2DL4**0010201 加上 *KIR2DL4**0010308 的異型合子。圖上方為 pyHLAMSA 的多序列比對資訊，其中 *KIR2DL4**consensus 為 *KIR2DL4* 所有等位基因型多序列比對的共識序列，同時也是 Graph-KIR 在該基因圖形回貼的骨幹序列，符號 - 則表示序列共識序列相同，符號 * 代表相對共識序列為缺失序列。圖中答案等位基因組存在 34 bp 的缺失序列，但在該區域回貼結果中只有一個短序列偵測到了缺失變異且長度不一致，同時有多條短序列映射於答案標註的缺失變異區域，這導致最終標註答案同型合子基因組的相似度分數被降低為 -168.189 分，扣分皆來自此缺失變異區域，此區域外所有的變異點位短序列回貼結果皆與標註答案相符。而被做為答案的異型合子組合，則因為在此區域外與 *KIR2DL4**0010308 不相符的短序列多與 *KIR2DL4**0010201 相符，所以算法能將短序列視為來自 *KIR2DL4**0010201 不被扣分，加上回貼在圖中缺失變異區域的序列，可以被視為來自在該處無缺失變異的 *KIR2DL4**0010308，亦不會造成扣分，最終相似度分數為最高分的 -12.212 而被視為答案。

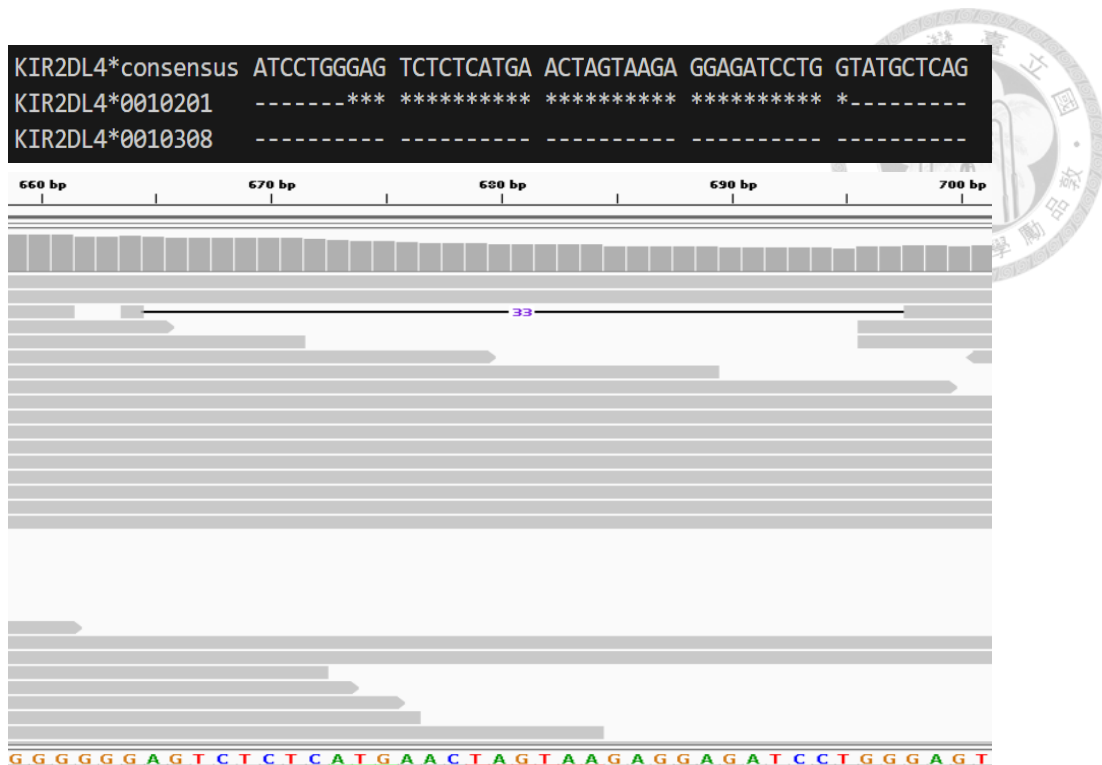


圖 4 同型合子被預測為異型合子案例在缺失變異區域的回貼狀況

4.3 分型算法更新 HPRC 驗證表現

為了降低上述問題對 Graph-KIR 預測準確度的影響，算法進行了改動與更新。改動後 HPRC 驗證準確度如表 5 所示，各版本皆允許假基因同分預測納入 Graph-KIR 預測結果，以此避免假基因同分狀況的分數浮動，影響工具算法改動前後效能的觀察。為了降低錯誤缺失變異偵測偏差，Graph-KIR 對照短序列上變異與等位基因型上的序列時，在缺失變異點位偵測到的負變異的乘積被調降。經過不同的分數調整，相似度乘機調降為 0.1 的時候 Graph-KIR 在真實資料上的表現能有最大的改善。然而缺失變異扣分降低對準確度的改善效果有限，無法真正消除錯誤缺失變異扣分的影響，也無法解決錯誤缺失判定情況下，算法青睞異型合子預測的偏差。本研究在原始算法中新增了對同型合子基因的識別及分型，更有效地提升了 Graph-KIR 在 HPRC 驗證集的表現。最後，Graph-KIR 合併上述兩種算法改動的版本在 HPRC 驗證集非新穎等位基因的部分有最佳的表現，7 碼準確度由 0.817 提升至 0.866。

表 5 Graph-KIR 各算法更新版本於 HPRC 44 樣本之表現

Filter	Methods	7-digits	5-digits	3-digit	Gene
-fusion	Number of alleles	248	713	825	825
	Graph-KIR	0.770	0.750	0.760	0.979
	Graph-KIR-dv-0.1	0.798	0.743	0.759	0.979
	Graph-KIR-Homo	0.806	0.767	0.767	0.979
	Graph-KIR-Homo-dv-0.1	0.835	0.764	0.770	0.979
-fusion -#	Number of alleles	237	651	740	740
	Graph-KIR	0.781	0.756	0.776	0.978
	Graph-KIR-dv-0.1	0.810	0.750	0.777	0.978
	Graph-KIR-Homo	0.819	0.777	0.786	0.978
	Graph-KIR-Homo-dv-0.1	0.848	0.773	0.789	0.978
-fusion -# -+ -=-	Number of alleles	222	516	548	548
	Graph-KIR	0.793	0.826	0.876	0.978
	Graph-KIR-dv-0.1	0.820	0.821	0.891	0.978
	Graph-KIR-Homo	0.833	0.853	0.891	0.978
	Graph-KIR-Homo-dv-0.1	0.860	0.851	0.889	0.978
-fusion -# -+ -=- \$	Number of alleles	142	143	145	145
	Graph-KIR	0.817	0.888	0.924	0.979
	Graph-KIR-dv-0.1	0.831	0.888	0.931	0.979
	Graph-KIR-Homo	0.850	0.916	0.938	0.979
	Graph-KIR-Homo-dv-0.1	0.866	0.923	0.945	0.979

表中版本皆經過允許同分預測之調整。

Graph-KIR: Graph-KIR 算法未調整版本。

Graph-KIR-dv-0.1: Graph-KIR 加入缺失變異分數調整版本。

Graph-KIR-Homo: Graph-KIR 加入同型合子基因識別分型版本。

Graph-KIR-Homo-dv-0.1: Graph-KIR 加入缺失變異分數調整和同型合子基因識別分型版本。



4.4 Graph-KIR 在台灣人族群上的實作與分析

4.4.1 拷貝數異常預測與模型缺陷校正結果

在 Graph-KIR 對 TWB 1492 樣本拷貝數的預測結果中，36 個樣本被發現存在 *KIR3DL3* 拷貝數變異，其中 34 個樣本被預測 4 拷貝，2 個樣本被預測為 1 拷貝（圖 5a）。然而，*KIR3DL3* 的拷貝數變異屬於罕見現象，因此在多個樣本觀察到 *KIR3DL3* 拷貝數變異並不正常。透過對不正常樣本中，拷貝數預測模型與 KIR 基因深度匹配狀況的觀察，發現模型與 KIR 深度聚落匹配的位置並不合理。如在樣本 NGS1_20170109C 中，在模型與深度聚落擬合後，*KIR3DL3* 和在同深度聚落的基因皆被判為 4 個拷貝（圖 7），而實際 KIR 發生 4 拷貝機率極低，所以此明顯為錯誤的預測。

為校正模型缺乏深度預測範圍限制的問題，本研究將樣本二倍體深度資訊加入預測模型，限制拷貝數為 2 的常態分佈的預測區間，避免模型僅因與 KIR 基因深度分布擬合分數較高而坐落於不正常的區間，產生不合理的拷貝數預測。為確認校正方法的可行性，本研究利用 TBB 1492 個樣本 WGS 回貼結果，統計了樣本二倍體參考區域深度與全基因回貼深度平均值的關聯性（圖 6），兩者呈現高度相關，相關係數為 0.862，這說明從局部區域取得二倍體深度資訊足以代表樣本基因整體的深度現象。

模型更新後算法被重新運用於 TBB 樣本拷貝數的預測。在新的結果中，原本 36 個存在 *KIR3DL3* 拷貝數變異的樣本皆被重新預測為正常的 2 拷貝（圖 5b）（圖 8），然而有新的 9 個樣本 *KIR3DL3* 拷貝數被預測為 1。由所有樣本參考二倍體深度與 *KIR3DL3* 75 分位深度及 KIR 所有基因 75 分位深度的平均之間的關聯性（圖 9），發現相對其他正常樣本，新的 9 個 *KIR3DL3* 拷貝數為 1 的樣本 *KIR3DL3* 的深度及 KIR 區域的深度都比參考二倍體深度來得淺，除此之外，也發現樣本名稱上標示了相近的日期，有可能為同批釋出之樣本，分別為

NGS2_20150106G、NGS2_20150106H、NGS2_20150107A、NGS2_20150107B、NGS2_20150107C、NGS2_20150107D、NGS2_20150110C、NGS2_20150110D、NGS2_20150208E，因此推測 KIR 區域深度較淺只發生在這批樣本上，屬於樣本本生的性質造成的結果。進一步觀察了拷貝數預測的結果（附錄C），發現儘管短序列回貼深度較淺，但仍明顯存在可被視為不同拷貝數的深度聚落，人工判斷 *KIR3DL3* 基因深度所在的聚落更可能拷貝數為 2，非拷貝數為 1。面對非工具缺陷造成的不正常拷貝數預測結果，本研究下參數使 Graph-KIR 假設 *KIR3DL3* 拷貝數為 2，同時調整常態分數模型標準差的參數，由預設的 0.8 降低至 0.6，使各拷貝數的常態分佈更集中，以此在 KIR 區域深度較淺狀況下更合理的預測拷貝數。參數調整重新對上述 9 個樣本進行預測後，所有 *KIR3DL3* 拷貝數皆被校正為 2（圖 5c）。算法的更新使 Graph-KIR 能自動化的輸出更佳的拷貝數預測結果，也提供了針對特殊案例處理的經驗，未來使用者亦可在拷貝數出現異常預測的情況下，根據實際狀況調整參數來重新預測和校正結果。

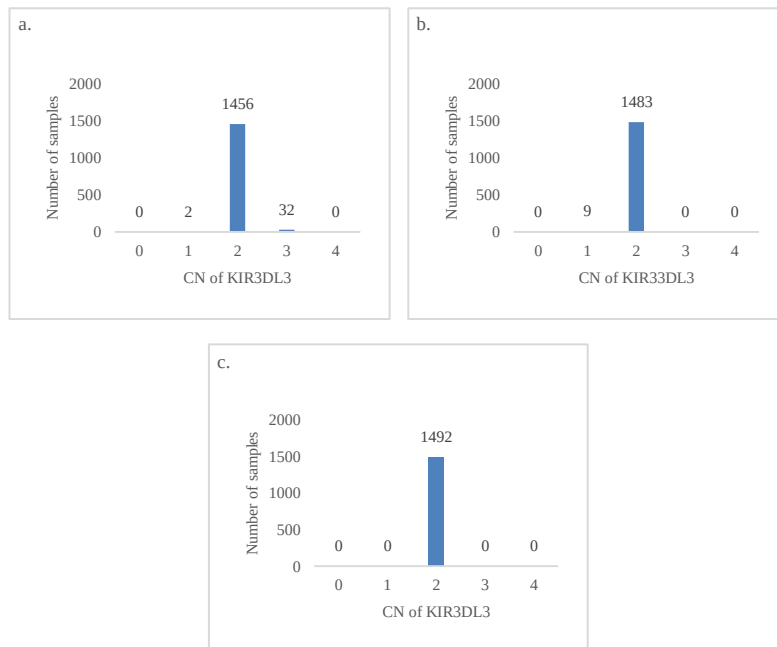


圖 5 TBB 1492 人 *KIR3DL3* 拷貝數分布，

(a) 模型校正前; (b) 模型校正後; (c) 特例校正後

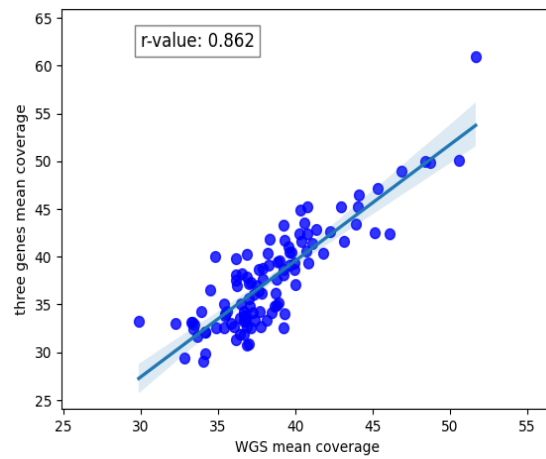


圖 6 二倍體基因平均深度與 WGS 回貼平均深度的關聯性

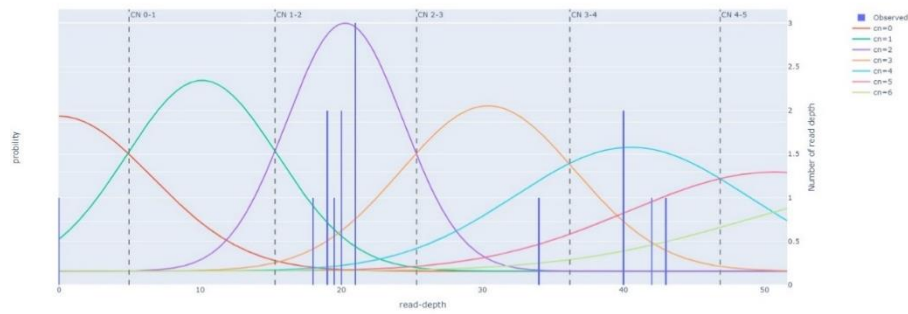


圖 7 不正常拷貝數預測結果範例。樣本 ID: NGS1_20170109C

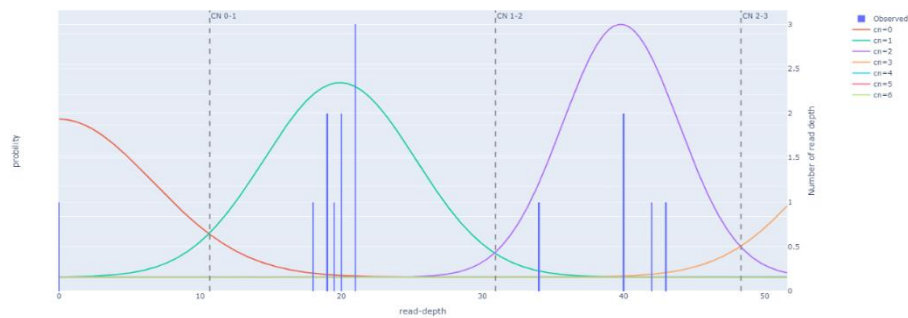


圖 8 校正後拷貝數預測結果範例。樣本 ID: NGS1_20170109C

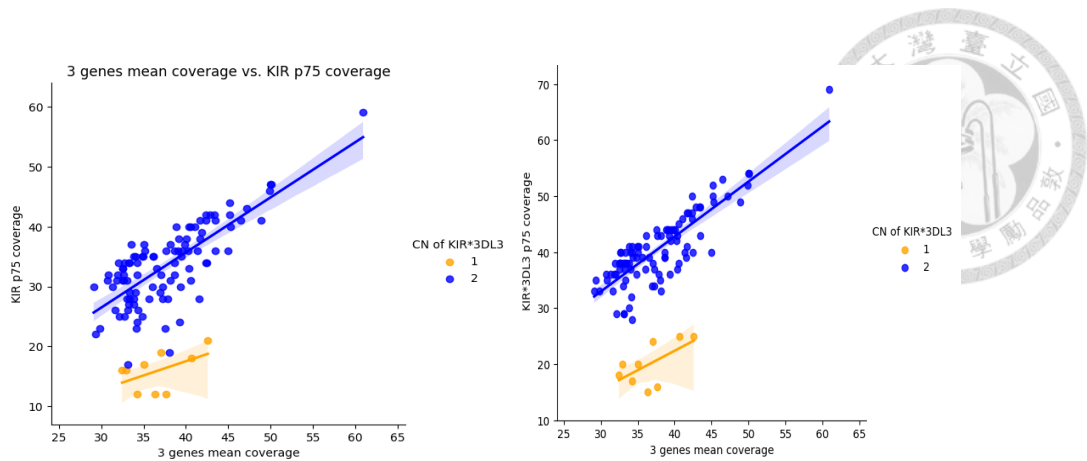


圖 9 參考二倍體基因平均深度與 KIR 區域基因深度之關聯性

4.4.2 台灣人拷貝數分布統計

圖 10 為 1492 台灣人中，每個人 KIR 基因中拷貝數分布狀況的統計。前十常見的拷貝數分佈中，共有 1230 人的拷貝數分布情形符合 Graph-KIR 拷貝數預測模型對拷貝數分布機率的假設，二拷貝的數目大於一拷貝，而三拷貝發生的機率小於一拷貝，占總樣本的 82.4%。然而，在第四、第七和第八常見的等拷貝數分布狀況中出現了拷貝數為一的基因數大於拷貝數為二的基因數的情形，共 131 人，約占總樣本的 8.8%。不過總體而言，拷貝數為三在樣本中的出現的機率和數目都非常少，模型只要能將拷貝數為一和二基因的深度聚落正確適配到標準差較小的拷貝數一二常態分佈，且模型更新後存在對二倍體深度的參考，在拷貝數為一的基因數大於拷貝數為二的基因數的情形要正確分辨出拷貝數一二兩聚落不會是困難的事情。

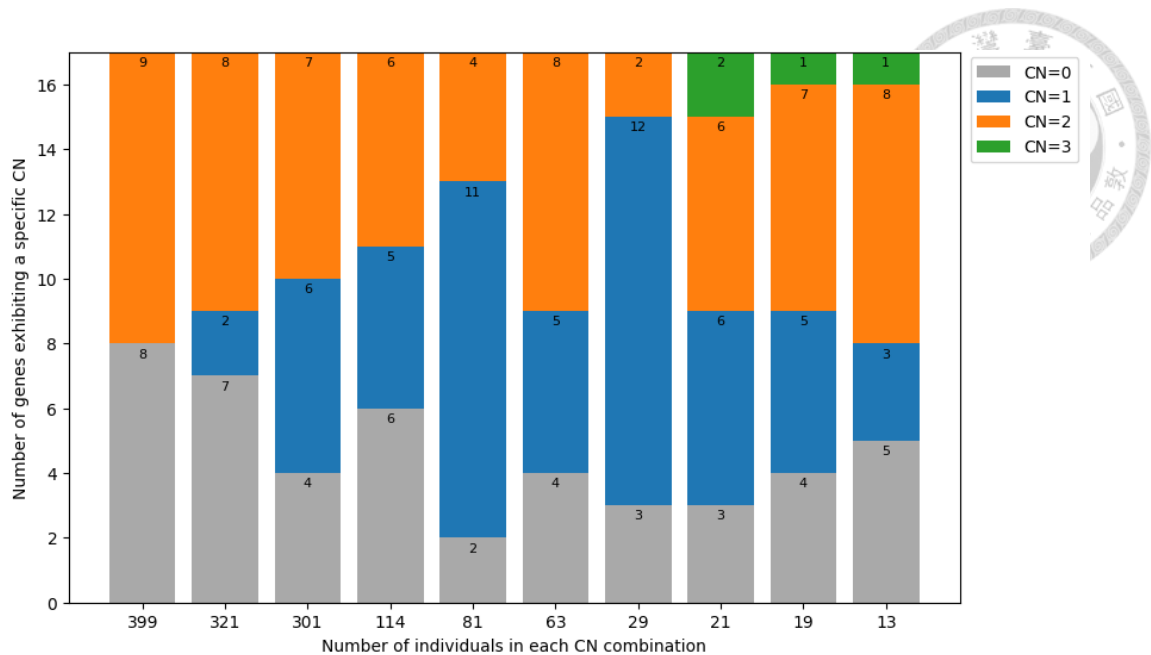


圖 10 台灣 1492 人前 10 常見個體 KIR 基因拷貝數分布

4.4.3 台灣人等位基因型統計

運用最新版本的 Graph-KIR，本研究統計了 1492 台灣人中的 7 碼等為基因分型結果。圖 11 顯示在台灣人各 KIR 基因最常出現等位基因頻率大於 5% 的等位基因型。本結果呈現了 Graph-KIR 對族群進行高精度等位基因分型統計的應用，以及本工具促進未來族群特異精準醫療的可能性。比較台灣族群利用 Graph-KIR 統計的常見等位基因型和先前研究使用 PING 對北美族群 KIR 統計的結果，發現在各基因中，部分等位基因同樣是常見的等位基因型，但常見程度並不相同，整體來看，兩個族群中最常見的等位基因型也不盡相同，代表在不同族群間常見 KIR 等位基因型是有差異存在的。由於當前工表現仍有進步的空間，加上真實資料中勢必存在未紀錄於 IPD-KIR 資料庫的新穎等位基因型，單單以最相似的 IPD-KIR 等位基因型標注樣本等位基因，無法最完善的呈現族群 KIR 基因的樣貌。面對上述情況，若未來工具在輸出最相似已知等位基因型時，亦能夠報導出存在的新穎變異，或是提供輸出等位基因型和樣本等位基因型的相似度分數，將有助於使用者更加充分的了解樣本基因型的全貌，並讓使用者對結果可靠度的有所參考，作為後續對資料應用的重要評估依據。

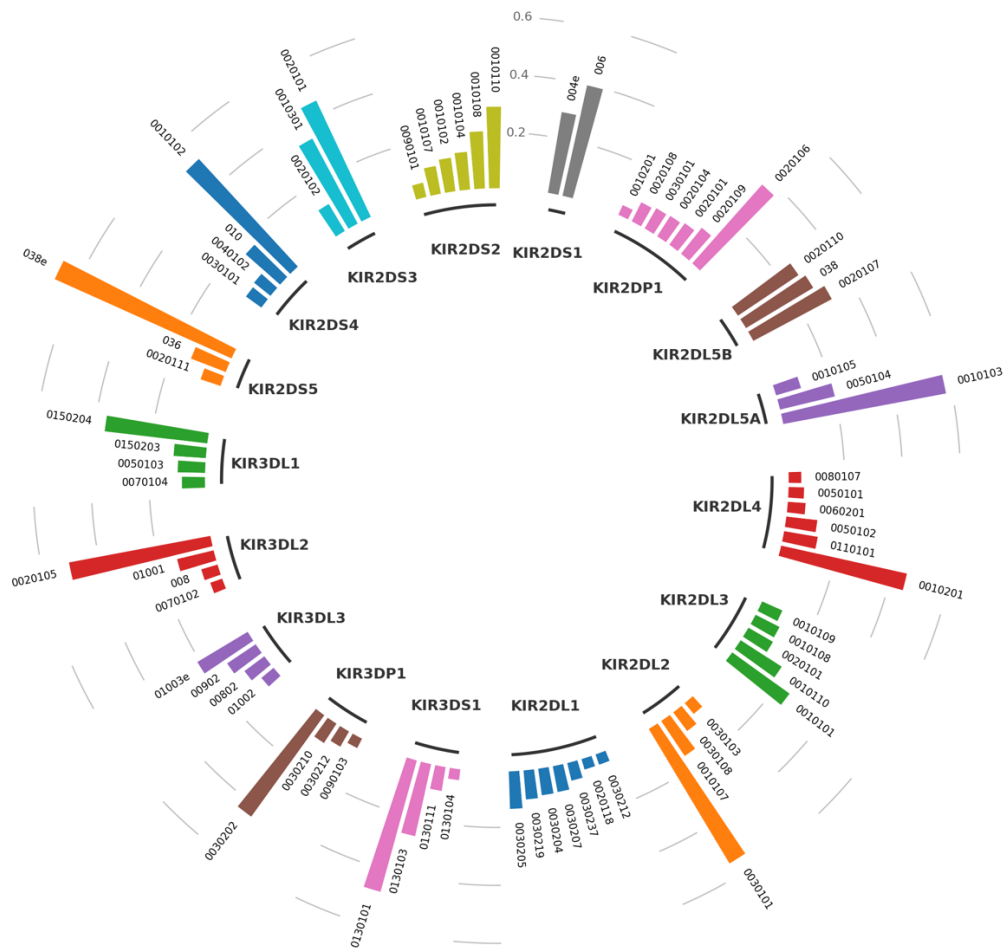


圖 11 台灣 1492 人各 KIR 基因最常見的等位基因型，等位基因頻率大於 5% 的等位基因顯示於圖上

4.5 未來工作

本研究透過對算法的更新來應對 Graph-KIR 缺失變異錯誤回貼造成的影響，初步提升工具的表現。然而當前算法的更新並未從根本解決錯誤回貼的問題，如何讓算法能真正避免缺失變異區域的影響或是做到對缺失變異的正確識別才是真正的挑戰。*KIR3DP1* 同分預測情形對分數的影響是最大的，本研究僅評估了該問題的影響，嘗試識別出該狀況下 *KIR3DP1* 的變異序列，來使同分等位基因組得以鑑別，有機會使算法準確度躍進。本研究專注在真實樣本已知等位基因分型的效能，尚未進一步對工具在新穎變異存在情況下分型的表現做同樣深入的檢驗。未來，提升算法對新穎等位基因標註的表現，以及開發工具對新穎變異報導的功能，是在真實樣本中存在新穎變異且資料庫不完備情況下，工具發展的關鍵方向。此外，目前工具僅接受高通量短序列作為輸入資料，將來可以調整工具的流程，使其能更廣泛地應用於各種資料類型，如目標區預測序列（Targeted sequencing）。



第五章 結論



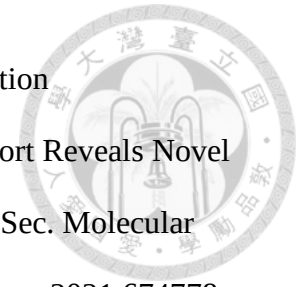
本研究釐清了目前 Graph-KIR 等位基因分型流程所面臨的幾個挑戰。首先，利用等位基因分型算法區分合併基因的個別拷貝數時容易出錯；其次，Graph-KIR 在缺失變異區域的回貼效果不佳，導致答案等位基因組分數無法突出；最後，拷貝數預測模型缺乏能將拷貝數分布模型限制在合理深度區間的機制，使不尋常拷貝數預測發生。算法問題的深度識別是 Graph-KIR 算法優化與校正的重要基礎，本研究發現 *KIR3DP1* 因為缺失變異區域回貼不良，易輸出多個同相似度分數可作為答案的等位基因組，若允許所有同分輸出皆被納入預測結果，Graph-KIR 在 HPRC 非新穎等位基因階層 7 碼準確度將由 0.711 上升至 0.811，雖然同分預測問題未被真正解決，但造成真實資料準確度下降一大原因得以被釐清。同時，本研究修改等位基因分型分數算法，降低錯誤缺失變異的影響，並增強對同型合子基因的分型能力，準確度由 0.817 上升至 0.866。在拷貝數預測模型加入二倍體深度資訊校正後，Graph-KIR 能更自動化地產生合理的拷貝數預測，此改動在大規模樣本群的預測尤為重要，在台灣 1492 人中，36 個不正常預測被自動校正，9 個 KIR 區域基因回貼深度較淺樣本被發現並重新校正。最終，本研究使用優化後的算法對台灣族群的 KIR 基因型進行統計分析，首次報導了該族群的首個 7 碼等位基因型統計數據，為未來台灣族群 KIR 研究提供重要的基礎。未來工具準確度的進步、對新穎變異分型算法的開發和資料庫新穎等位基因型的擴充，都有望進一步增加該統計結果的參考價值，促進免疫基因體領域個人化精準醫療的發展。

參考文獻



- Robinson, J. (2012). IPD--the Immuno Polymorphism Database. In *Nucleic Acids Research* (Vol. 33, Issue Database issue, pp. D523–D526).
<https://doi.org/10.1093/nar/gki032>
- Hong-YeLin, Hui-WenChuang, Tsung-KaiHung, Ting-JianWang, Ching-JimLin, Jacob ShujuiHsu, Chia-LangHsu, Ya-ChienYang, Pei-LungChen, ChienYuChen. Graph-KIR: Graph-based KIR Copy Number Estimation and Allele Calling Using Short-read Sequencing Data. *bioRxiv.*, 2023 Dec.
- Marin, W. M., Dandekar, R., Augusto, D. G., Yusufali, T., Heyn, B., Hofmann, J., Lange, V., Sauter, J., Norman, P. J., & Hollenbach, J. A. High-throughput Interpretation of Killer-cell Immunoglobulin-like Receptor Short-read Sequencing Data with PING. *PLoS Computational Biology*, 17(8), e1008904. 2021.
- Li Song, Gali Bai, X Shirley Liu, Bo Li, and Heng Li. T1k: efficient and accurate kir and hla genotyping with next-generation sequencing data. *Genome Res.* 2023 Jun.
- Yen-Chen Anne Feng and Chia-Yen Chen. Taiwan Biobank: A rich biomedical research database of the Taiwanese population. *Cell Genomics*, 12;2(11):10019. 2022.
- Wesley M Marin and Jill A Hollenbach. Software update: Interpreting killer-cell immunoglobulin-like receptors from whole genome sequence data with ping. *HLA*, 2022.
- 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 Biotechnology*, 37(8), 907–915.

Leonardo M. Amorim and Danillo G. Augusto. (2021). High-Resolution Characterization of KIR Genes in a Large North American Cohort Reveals Novel Details of Structural and Sequence Diversity. *Front. Immunol., Sec. Molecular Innate Immunity* Volume 12 - 2021 | <https://doi.org/10.3389/fimmu.2021.674778>



Ting Wang, Lucinda Antonacci-Fulton, Kerstin Howe, Heather A Lawson, Julian K Lucas, Adam M Phillippy, Alice B Popejoy, Mobin Asri, Caryn Carson, Mark JP Chaisson, et al. The human pangenome project: a global resource to map genomic diversity. *Nature*, 604(7906):437–446, 2022.

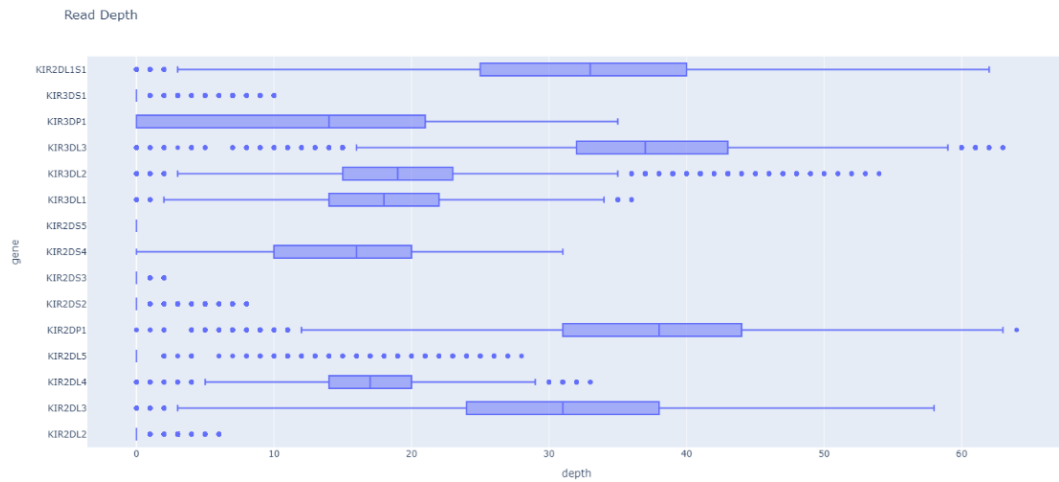
Tsung-Kai Hung, Wan-Chi Liu, Sheng-Kai Lai, Hui-Wen Chuang, Yi-Che Lee, Hong-Ye Lin, Chia-Lang Hsu, Chien-Yu Chen, Ya-Chien Yang, Jacob Shujui Hsu, and Pei-Lung Chen. Genetic diversity and structural complexity of the killer-cell immunoglobulin-like receptor gene complex: A comprehensive analysis using human pangenome assemblies. *bioRxiv*, 2023. doi:10.1101/2023.11.12.566753.

Helga Thorvaldsdóttir, James T. Robinson, and Jill P. Mesirov. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 2013 Mar

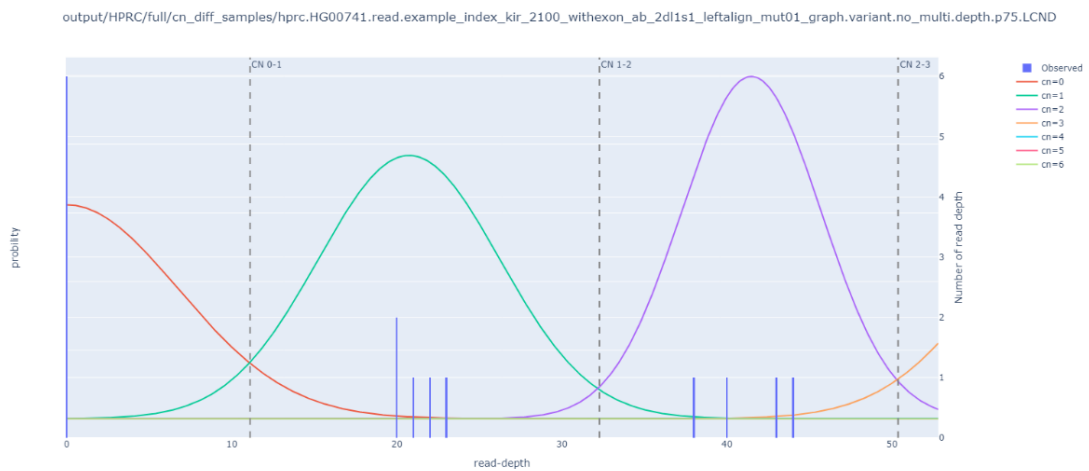
Seung-Been Lee, Marsha M Wheeler, Karynne Patterson, Sean McGee. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genet Med.*, 2019 Feb.

MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* 2014;42(Database issue):D986–992.

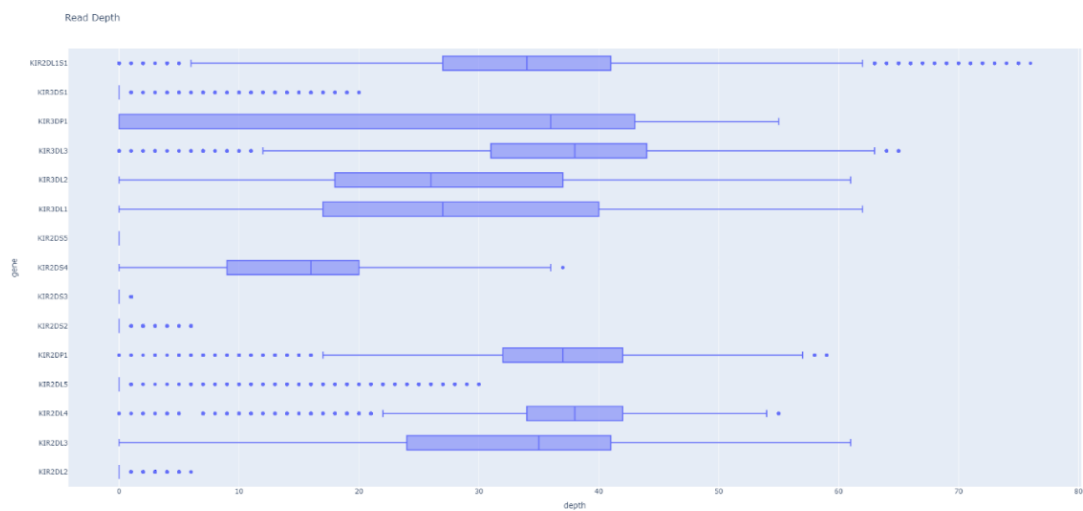
附錄 A



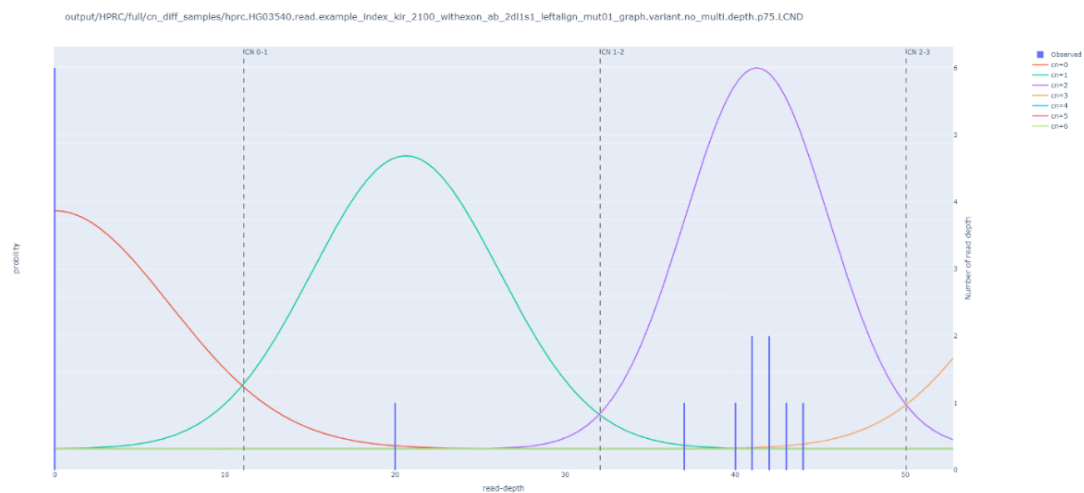
附圖 A-1-1 HPRC 樣本各 KIR 基因深度，樣本編號：HG0741



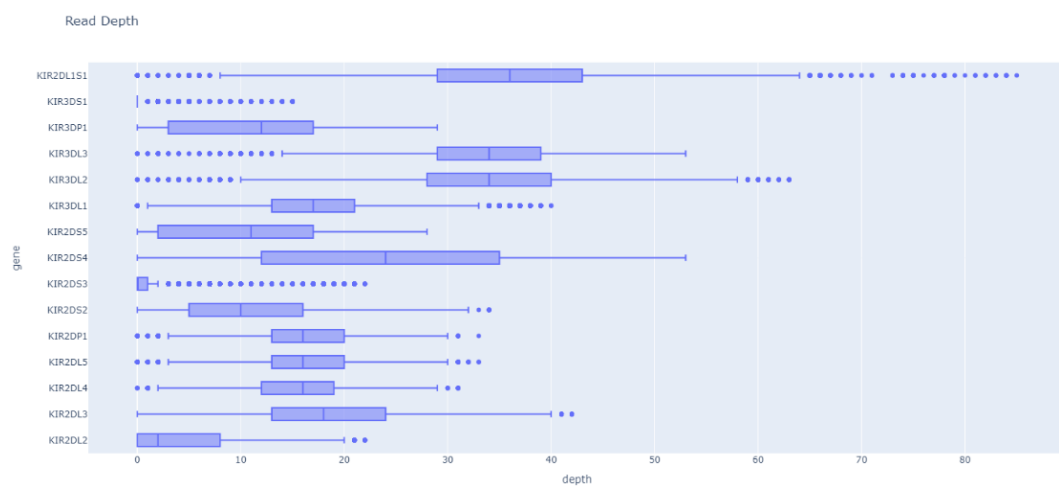
附圖 A-1-2 HPRC 拷貝數錯誤預測結果，樣本 ID: HG0741



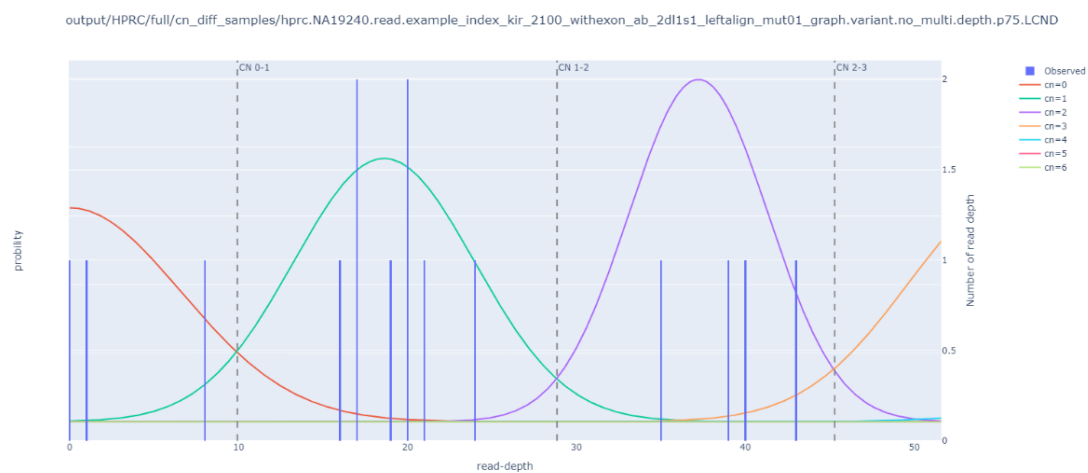
附圖 A-2-1 HPRC 樣本各 KIR 基因深度，樣本編號：HG03540



附圖 A-2-2 HPRC 拷貝數錯誤預測結果，樣本 ID: HG03540

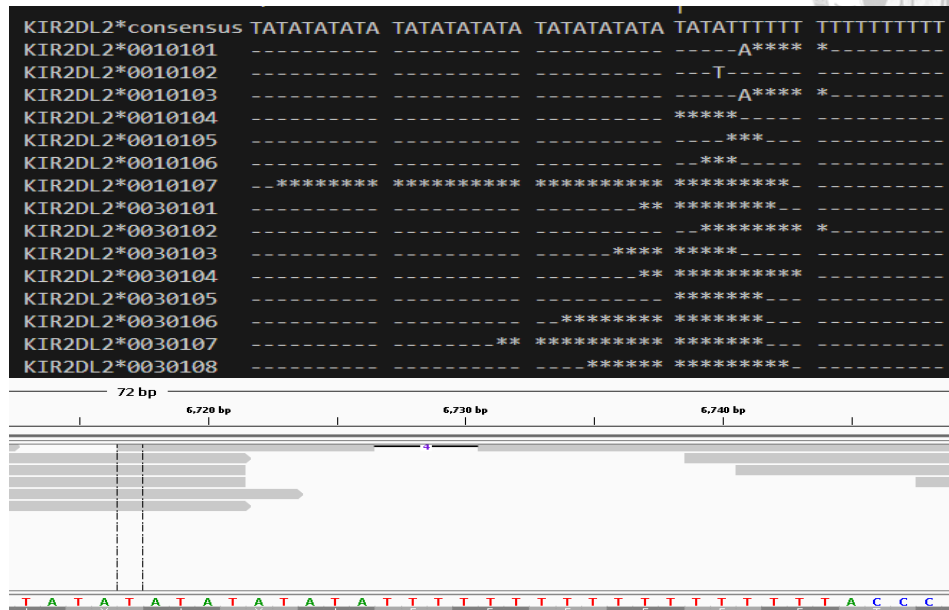


附圖 A-3-1 HPRC 樣本各 KIR 基因深度，樣本編號：NA19240



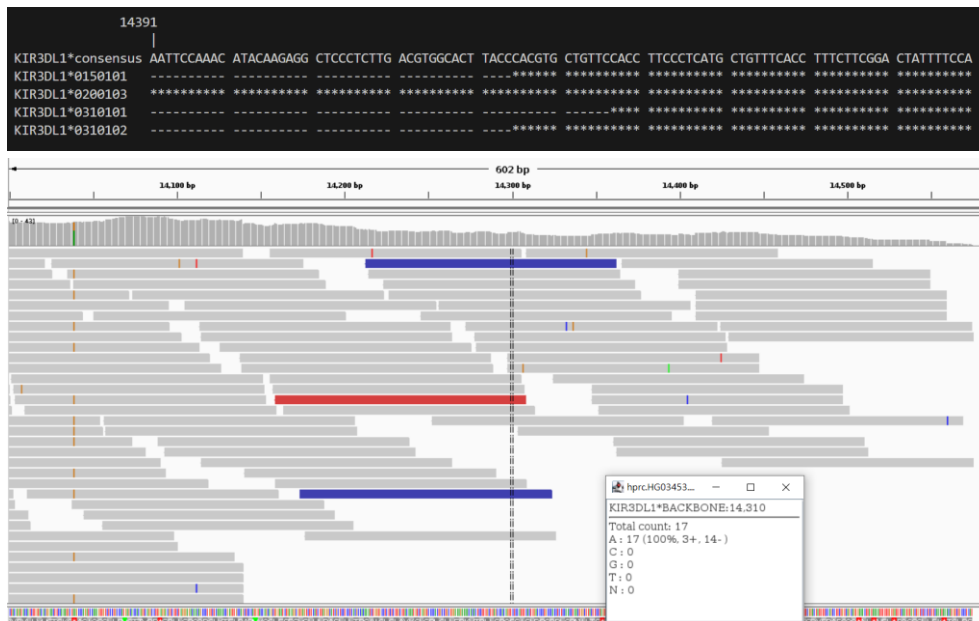
附圖 A-3-2 HPRC 拷貝數錯誤預測結果，樣本編號：NA19240

附錄 B



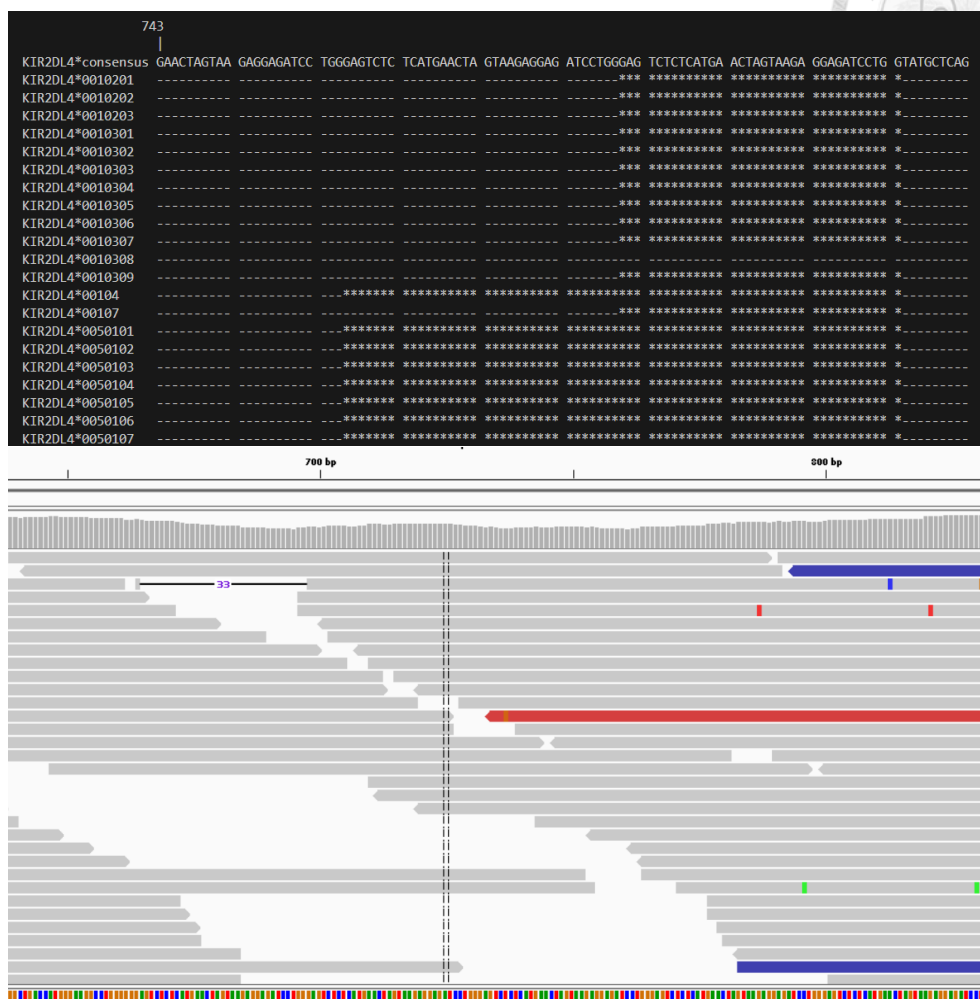
附圖 B-1 範例：常見缺失變異區域短序列無回貼，樣本：HG01175

，SKIRT 標註答案：KIR2DL2*0030101



附圖 B-2 基因末段長片段缺失存在序列回貼，樣本：HG03453

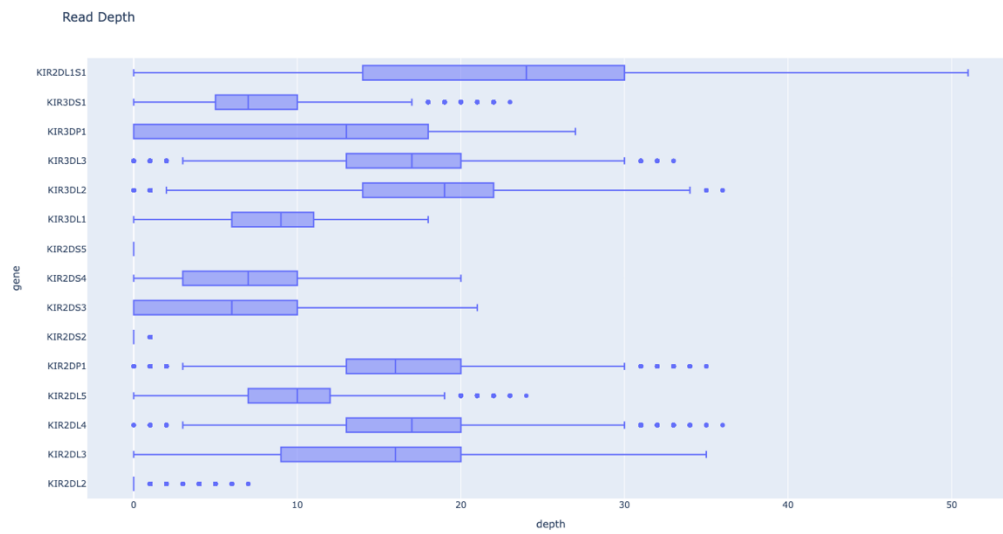
，SKIRT 標註答案：KIR3DL1*0150101、KIR3DL1*0310101



附圖 B-3 長片段重複的缺失無法被偵測，樣本：HG00438

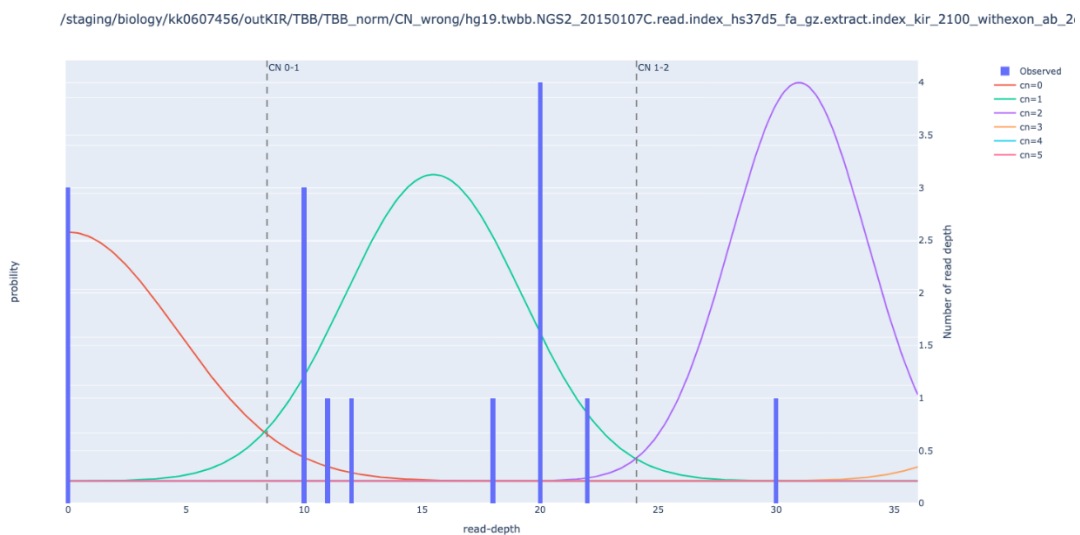
，SKIRT 標註答案：KIR2DL4*0010201、KIR2DL4*0010201

附錄 C



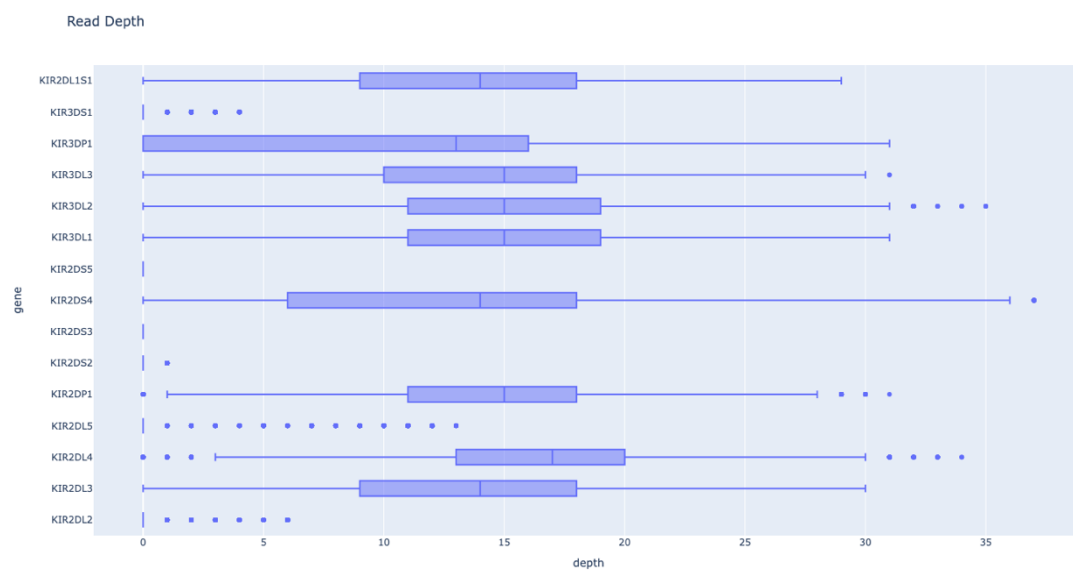
附圖 C-1-1 KIR 區域深度較淺樣本各基因圖形回貼深度，樣本 ID:

NGS2_20150107C



附圖 C-1-2 KIR 區域深度較淺樣本校正前拷貝數預測結果範例，樣本 ID:

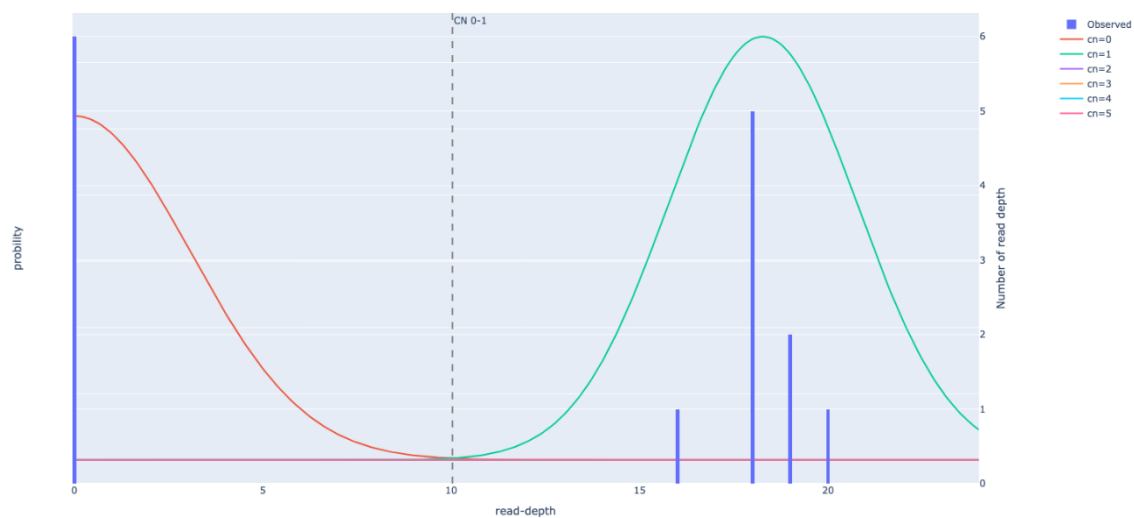
NGS2_20150107C



附圖 C-2-1 KIR 區域深度較淺樣本各基因圖形回貼深度，樣本 ID:

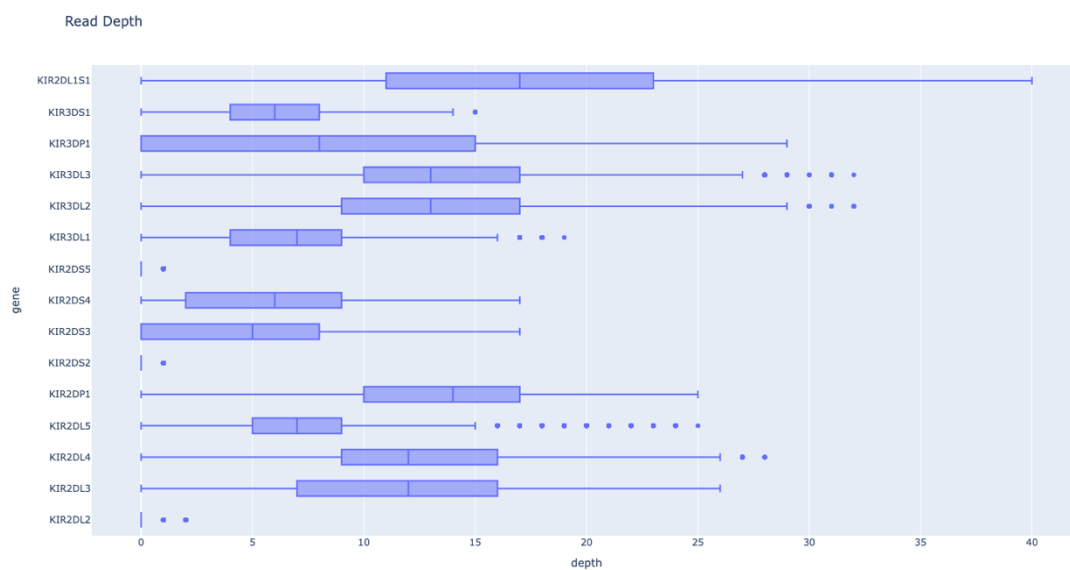
NGS2_20150107D

/staging/biology/kk0607456/outKIR/TBB/TBB_norm/CN_wrong/hg19.twbb.NGS2_20150107D.read.index_hs37d5_fa_gz.extract.index_kir_2100_withexon_ab_2



附圖 C-2-2 KIR 區域深度較淺樣本校正前拷貝數預測結果範例，樣本 ID:

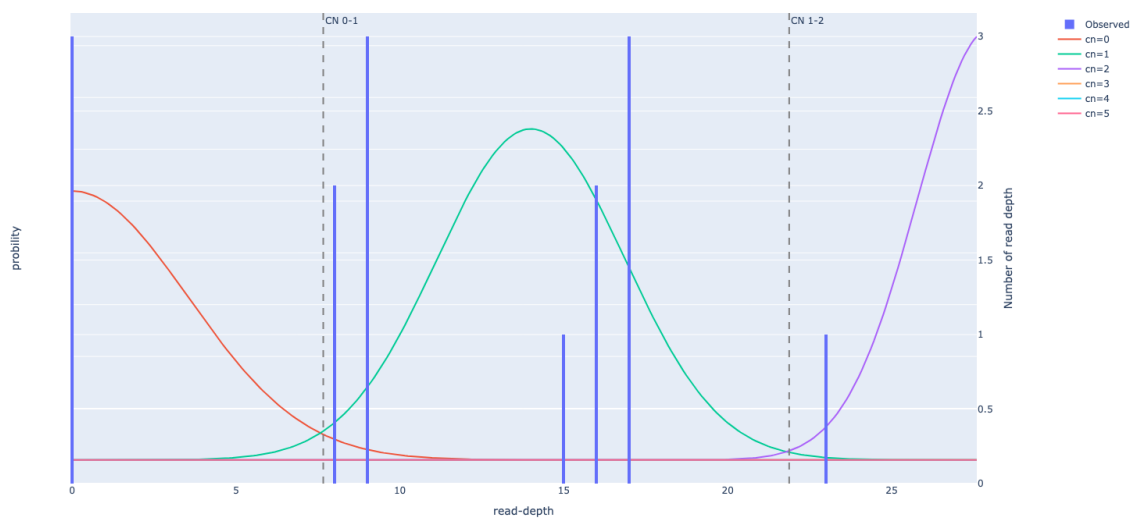
NGS2_20150107D



附圖 C-3-1 KIR 區域深度較淺樣本各基因圖形回貼深度，樣本 ID:

NGS2_20150208E

/staging/biology/kk0607456/outKIR/TBB/TBB_norm/CN_wrong/hg19.twbb.NGS2_20150208E.read.index_hs37d5_fa_gz.extract.index_kir_2100_withexon_ab_2i



附圖 C-3-2 KIR 區域深度較淺樣本校正前拷貝數預測結果範例，樣本 ID:

NGS2_20150208E