驗證「spaced seed」雜湊函數對基因分型的效果

Validation of "spaced seed" hashing technique for genotyping

指導教授：賀保羅
專題成員：林鈺婕
開發工具：Bash shell script
測試環境：Ubuntu 24.02.2 LTS

一、簡介：

本專題聚焦於探討 alignment-free 基因型判定方法中的 spaced k-mers 技術，並驗證 Hartmut Häntze 與 Paul Horton 於 2023 年發表之論文《Effects of spaced k-mers on alignment-free genotyping》中所提出的方法與工具。該研究將 spaced k-mers 應用於 alignment-free genotyping 工具 PanGenie，並改良為 MaskedPanGenie，以提升基因型判讀的效率與準確度。本研究依據該論文所描述之實驗流程，實作並重現部分分析流程，並與 Genome in a Bottle（GIAB）提供的標準資料集進行比對，以驗證其方法在實際資料下的準確性與一致性，進一步評估其可行性與實際應用效果。

驗證架構方面，選用與原論文相同的受試者 NA24385 的 whole genome sequencing（WGS）資料，依照論文所述流程，使用 MaskedPanGenie 搭配 spaced k-mers 進行 variant calling。分析過程中的參數設定與資料處理盡可能與原論文保持一致，並將 variant calling 的結果與 GIAB 所提供的 variant benchmark 進行比對，從 wGC、Precision、Recall 與 F-score 四個面向，評估 spaced k-mers 在 alignment-free genotyping 中的實際表現。

在實驗中共採用了四種不同設定的 k-mer 種子序列（seed），分別為：

- Contiguous seed C（連續種子，長度為 31）
- Contiguous seed C'（連續種子，長度為 51）
- Spaced seed S1 （權重為 31，總長度為 51）
- Spaced seed S2（權重為 31，總長度為 51）

透過這些不同種子設計，進一步探討其在 variant calling 效能上的差異。

透過本次驗證性專題研究，期望能確認 MaskedPanGenie（即在 PanGenie 架構下應用 spaced k-mers）在標準資料集上的效果，建立一套可行的評估流程，進一步補充原論文在此部分的實驗結果。

以下為實驗流程：

1. 從 pangenome 構建 variant graph
2. 計算 variant graph 中 spaced k-mers 數量
3. 計算 reference pangenome reads 中 spaced k-mers 數量
4. 計算受試者 WGS 資料中 spaced k-mers 數量
5. 執行基因分型

二、測試結果：

- Result

受試者 NA24385之 WGS 資料使用 MaskedPanGenie 與 spaced seed S1進行 variant calling 後，與 GIAB 之 variant benchmark 比對之結果

Seed	precision	recall	F1-score
S1	0.271	0.885	0.415

Seed	Total alleles	Correct	Incorrect	Proportion
S1	13320475	3583945	9736530	0.269055345

- Runtime

論文《Effects of spaced k-mers on alignment-free genotyping》中使用的處理器為 AMD Ryzen 9 5950X 16-Core Processor，其執行時間如表所示。

Workflow	Preprocessing	Execution	Total
PanGenie (k = 31)	–	3 h:07 min	3 h:07 min
PanGenie (k = 51)	–	3 h:12 min	3 h:12 min
MaskedPanGenie	4 h:24 min	3 h:27 min	7 h:51 min

此專題於搭載 Intel Xeon Gold 6242（2.80GHz）之伺服器環境下執行所有測試，其執行時間如表所示。

Workflow	Preprocessing	Execution	Total
PanGenie (k = 31)	–	5 h:03 min	5 h:03 min
PanGenie (k = 51)	–	5 h:14 min	5 h:14 min
MaskedPanGenie	8 h:15 min	5 h:03 min	13 h:18 min