

提升基於 Spaced k-mer 的基因分型工具 MaskedPanGenie 之效率與準確率

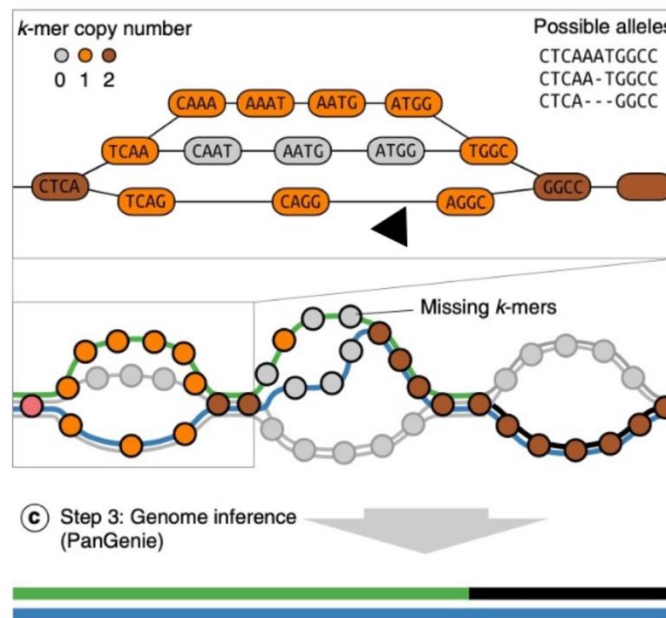
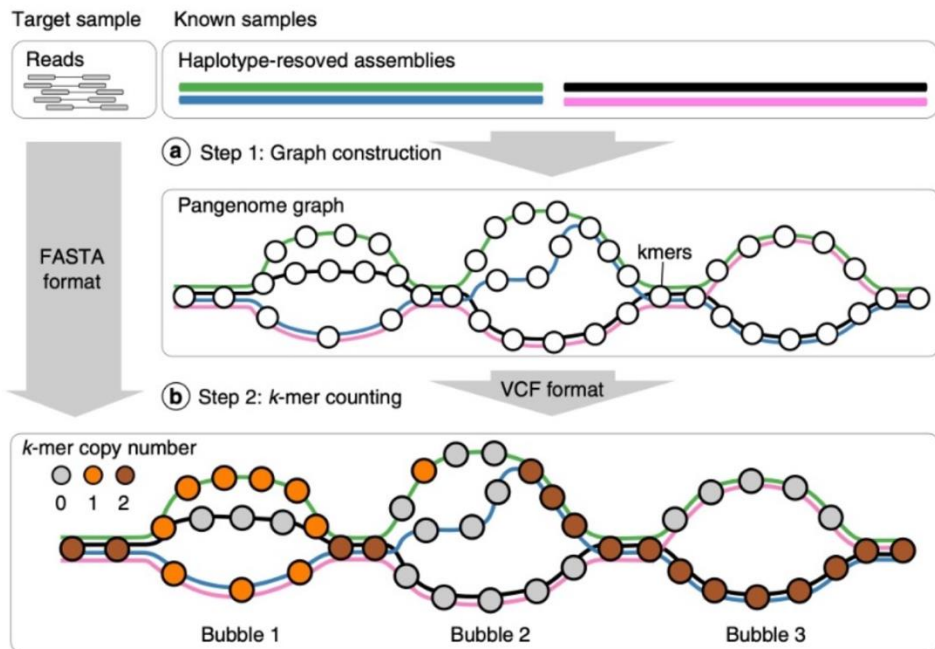
專題生 ：王宇桐

指導教授 ：賀保羅 老師

Background

基因分型：判斷基因序列中各個基因的變異情況

PanGenie：基於 k-mer 做基因分型 MaskedPanGenie：改成非連續 k-mer



Spaced Seed : 10101

Sequence : ATCGATATC

k-mers : ACA, TGT, CAA, GTT, AAC

▲ PanGenie Flow Chart

Ebler, J., Ebert, P., Clarke, W.E. et al. Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes. Nat Genet 54, 518–525 (2022). <https://doi.org/10.1038/s41588-022-01043-w>

Motivation

1. 執行時間較長

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

2. 怎麼定義非連續 k-mer pattern (Seed)

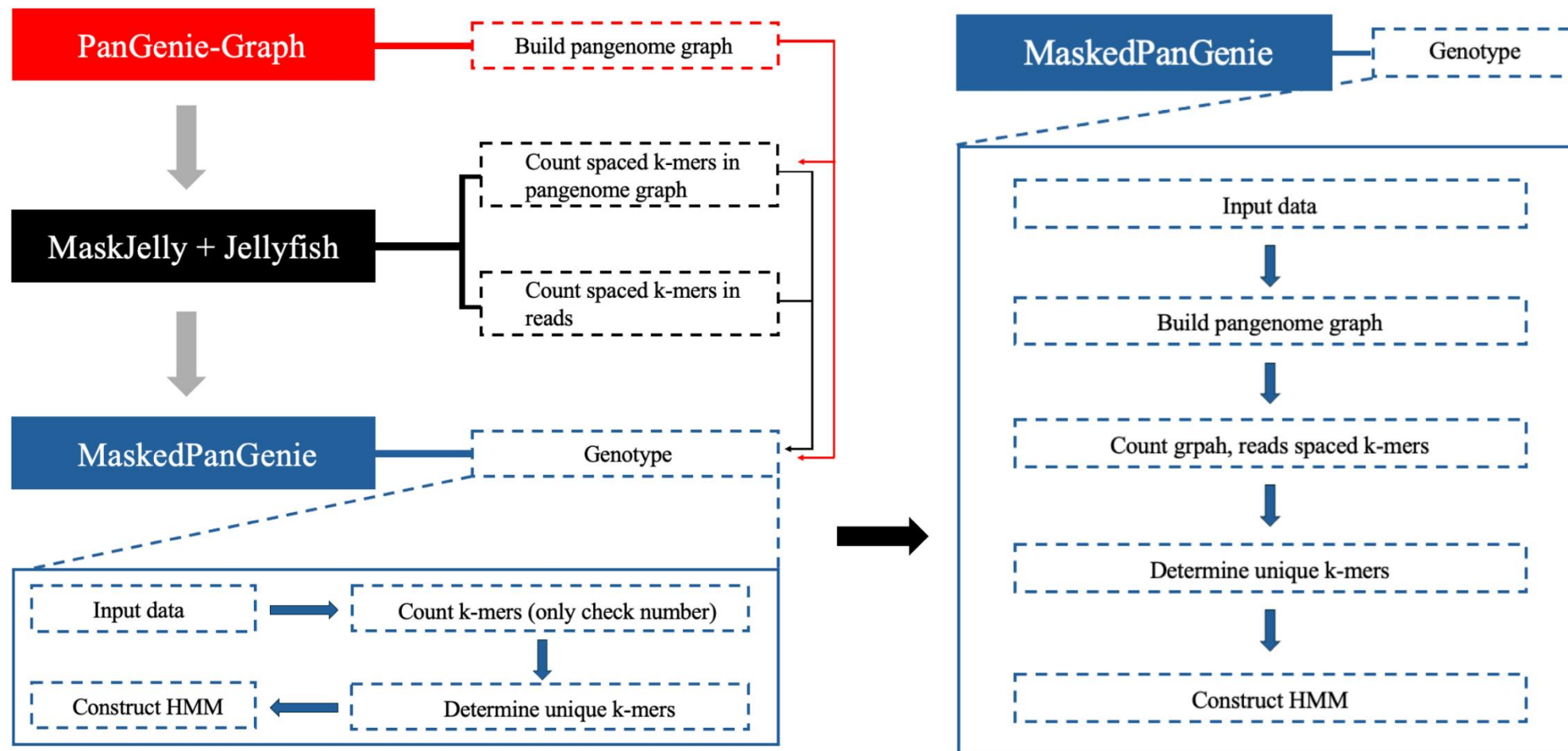
Sensitivity : 評價 Seed 對 match 區域的覆蓋度

S₁: 111011100101010011010011111110010110010101001110111

S₂: 111101010011100100110101111101011001001110010101111

Method

1. 優化系統架構



Method

2. Fast Spaced Seed Hashing (FSH) algorithm

Seed: 1 1 0 1 0 1 1 0 1

Sequence A C G T A G A T C G A

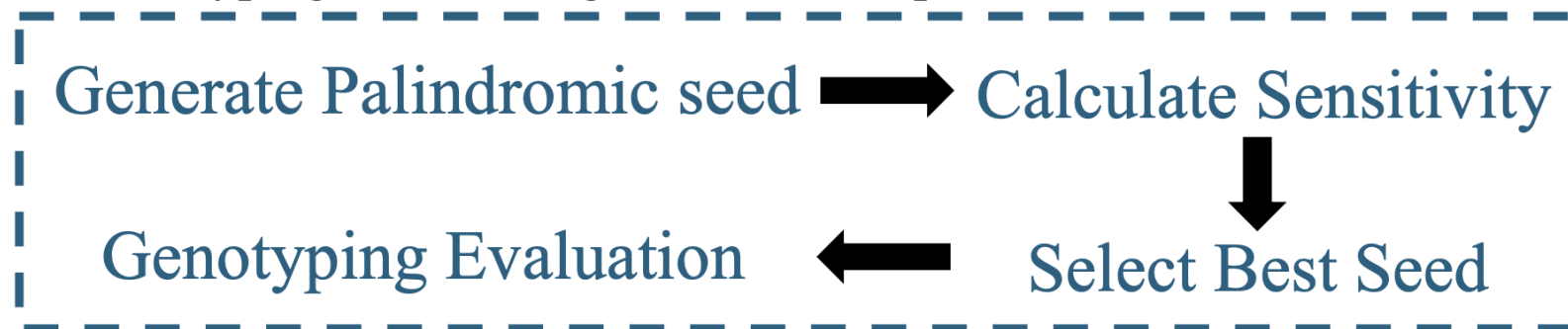
A C T G A C

C G A A T G

Shift to next k-mer

Two positions are double-counted

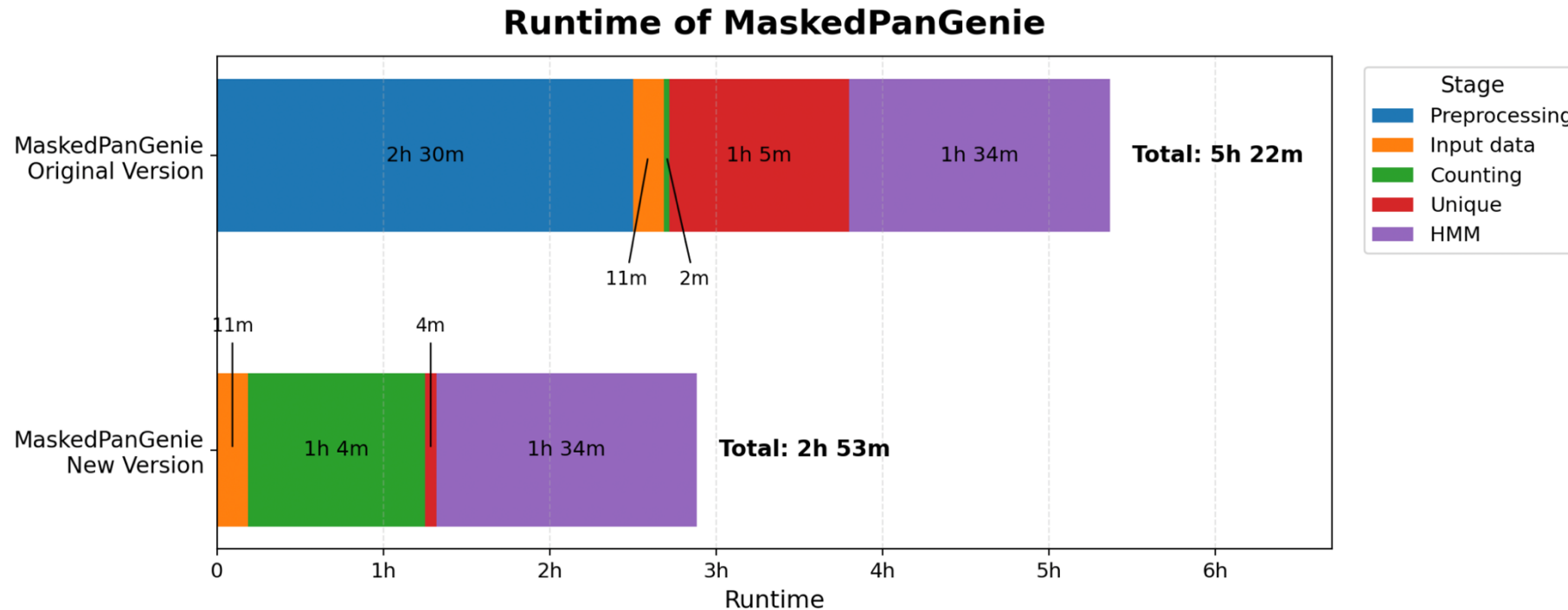
3. 嘗試不同 Seed 做基因分型



Result

Runtime Compare

	Preprocessing	Input data	Counting	Unique	HMM	Total
MaskedPanGenie (Original version)	2 h:30 min	11 min	2 min	65 min	94 min	5 h:22 min
MaskedPanGenie (New version)	-	11 min	1 h:4 min	4 min	94 min	2 h:53 min



Result

Accuracy Compare

