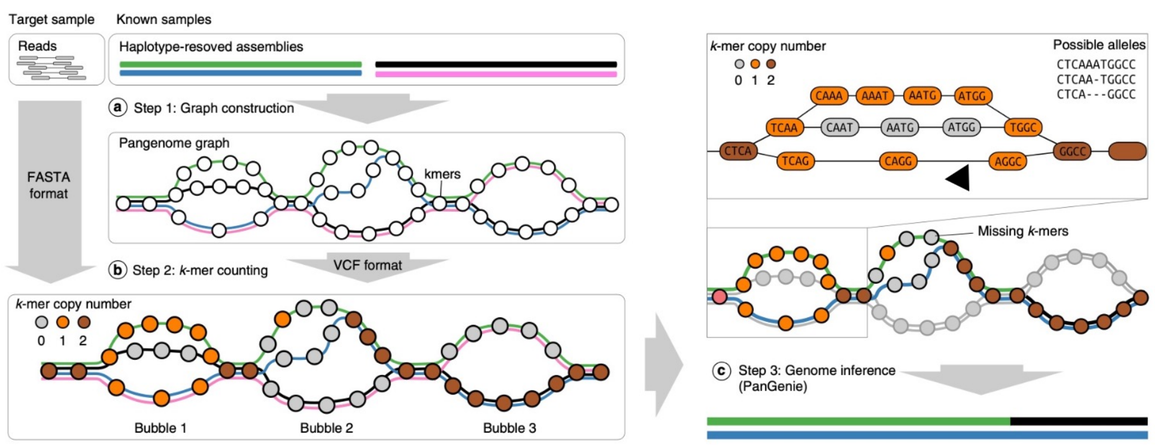
Improving the Efficiency and Accuracy of MaskedPanGenie for Spaced k-mer based Genotyping

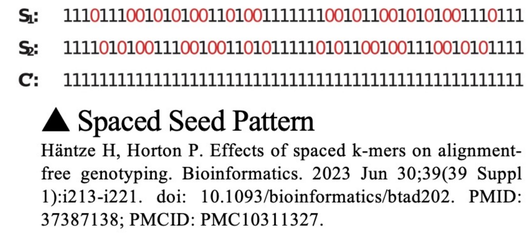
Yu-Tong Wang[1], Paul Horton*
Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan

Introduction

Genotyping : Identifying genetic variants in an individual genome from sequencing data, including SNPs, indels, and structural variants.
PanGenie : pangenome-based genotyping tool that represents shared variants from multiple individuals as a pangenome graph. Instead of performing read alignment, PanGenie compares k-mers from sequencing reads with allele-specific k-mers in the graph to infer variants. It further integrates Hidden Markov Model (HMM) and linkage disequilibrium (LD) information to reconstruct the complete genotype.
Based on PanGenie, our laboratory developed MaskedPanGenie, which introduces spaced seeds (non-contiguous k-mers). By defining "match" and "ignore" positions within k-mers, spaced seeds improve tolerance to sequencing errors and increase variant detection sensitivity.



▲ 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



▲ Spaced Seed Pattern
Häntze H, Horton P. Effects of spaced k-mers on alignment-free genotyping. Bioinformatics. 2023 Jun 30;39(39 Suppl 1):i213-i221. doi: 10.1093/bioinformatics/btad202. PMID: 37387138; PMCID: PMC10311327.

Spaced Seed : 10101
Sequence : ATCGATATC
k-mers : ACA, TGT, CAA, GTT, AAC
▲ Spaced Seeds k-mer

Motivation

Although MaskedPanGenie improves error tolerance using spaced seeds, it still faces two major limitations.

1. Longer execution time

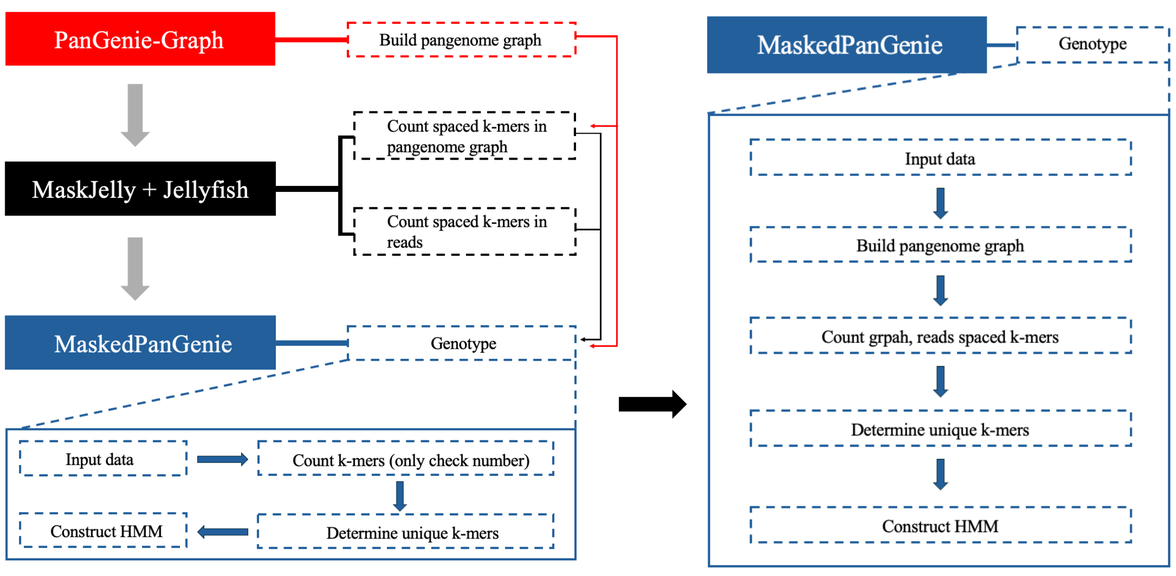
Workflow	Preprocessing	Execution	Total
PanGenie ($k=31$)	–	3 h:07 min	3 h:07 min
PanGenie ($k=51$)	–	3 h:12 min	3 h:12 min
MaskedPanGenie	4 h:24 min	3 h:27 min	7 h:51 min

2. Maybe not the best Spaced k-mer pattern
Sensitivity : Higher sensitivity increases the likelihood of detecting true variants.

SpEED ⟶ Half Seed $\xrightarrow{\text{Mirror}}$ Palindromic seed

Methodology

Project Background
1. Optimize system structure



2. Fast Spaced Seed Hashing (FSH) algorithm
Using the overlap of adjacent spaced k-mers to avoid redundant computation

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

My Contributions
1. Combine the two optimized versions of the program. Change the Spaced k-mer counting algorithm to FSH, then integrate the program into the main program.
2. Genotyping results using different seed patterns

Generate Palindromic seed ⟶ Calculate Sensitivity ⟶ Select Best Seed ⟶ Genotyping Evaluation

Data and Statistics

9 human genome datasets, including sequencing reads and corresponding variant call sets and do the leave-one-out validation. Runtime was used to measure computational performance and wGC, precision, recall, and F-score were used to assess genotyping accuracy.

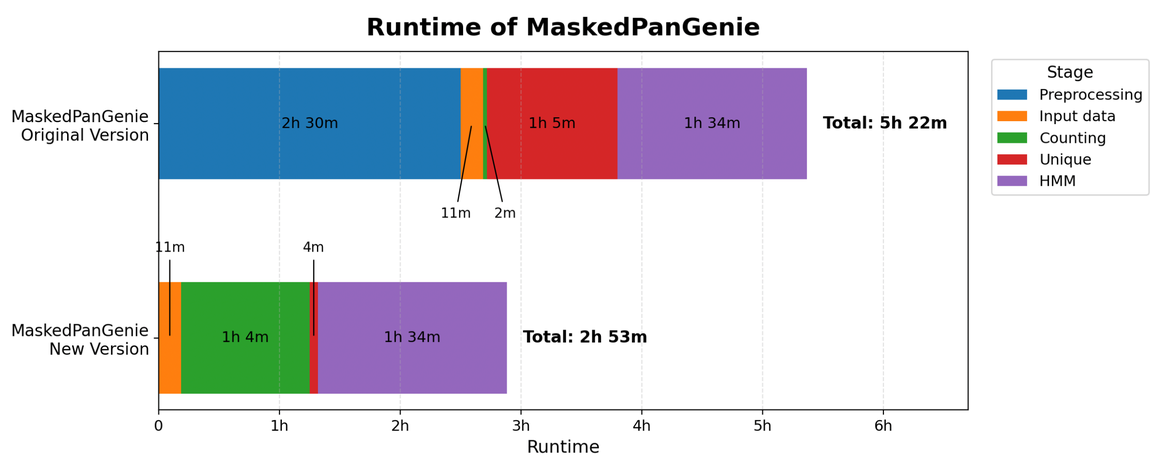
$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad \text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \qquad \text{conc}(x|y) = \frac{T_{x|y}}{T_{x|y}+F_{x|y}}$$

$$F-\text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision}+\text{recall}} \qquad \text{wGC} = \frac{\text{conc}(0|0)+\text{conc}(0|1)+\text{conc}(1|1)}{3}$$

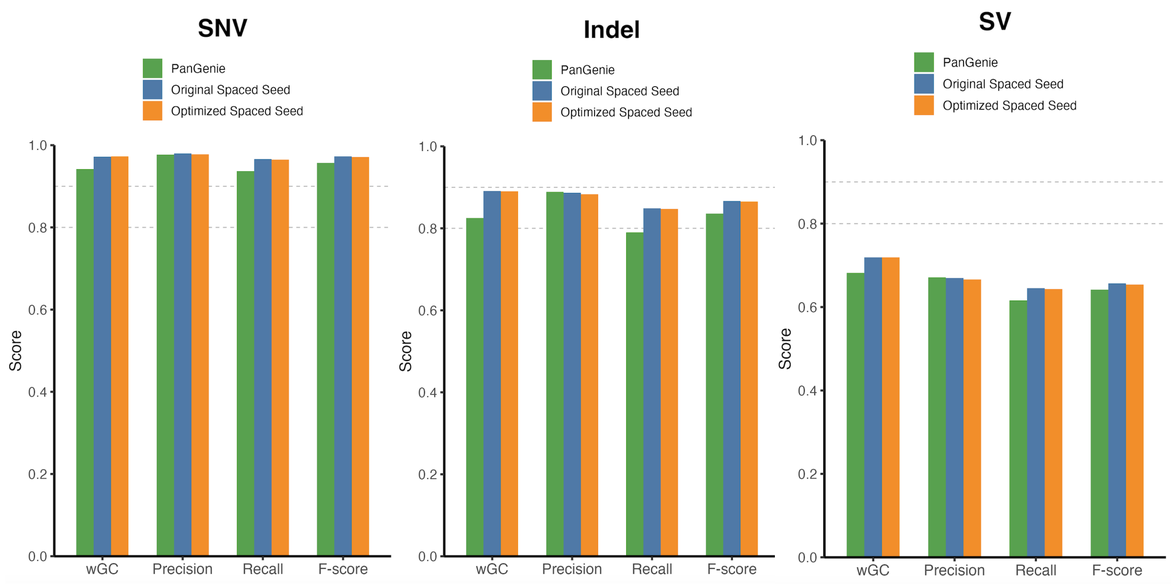
Result

1. Execution time comparison

	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



2. Accuracy



Conclusion and Future Work

System-level optimization and FSH-based spaced k-mer counting substantially improve the efficiency of MaskedPanGenie. In contrast, increasing seed sensitivity yields only marginal gains in genotyping accuracy. While MaskedPanGenie performs well on SNVs and indels, structural variant genotyping remains challenging. Future work will focus on improving the detection and genotyping accuracy of structural variants.

1. Hartmut Häntze, Paul Horton, Effects of spaced k-mers on alignment-free genotyping, *Bioinformatics*, Volume 39, Issue Supplement_1, June 2023, Pages i213i221, https://doi.org/10.1093/bioinformatics/btad202
2. 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

國立成功大學 National Cheng Kung University