Scalable variant detection in pangenome models

Francesco Porto [a] , Flavia Villani [b], Andrea Guarracino [c], Christian Fischer [d], Hao Chen [e], Robert W. Williams [d] , Gianluca Della Vedova [a] , Erik Garrison [f] , and Pjotr Prins [d]

[a] Department of Informatics, Systems, and Communication, University of Milano - Bicocca, Milan, Italy
[b] National Research Council, Institute of Genetics and Biophysics 'A.Buzzati-Traverso', Naples, Italy
[c] Centre for Molecular Bioinformatics, Department of Biology, University Of Rome Tor Vergata, Rome, Italy
[d] Department of Genetics, Genomics and Informatics, College of Medicine, UTHSC
[e] Department of Pharmacology, Addiction Science, and Toxicology, The University of Tennessee Health Science Center, Memphis, TN, USA
[f] Genomics Institute, University of California Santa Cruz, Santa Cruz, CA, United States

Genomic studies typically assume a single linear reference genome. This assumption can make it difficult to observe sequences in genomes that are divergent from reference, limiting the accuracy and completeness of analyses. Pangenome models address this limitation, representing the mutual relations between many genomes without using any of them as the point of reference. These models can encode genomes and their mutual alignments in a graph-based structure. Using these models requires the development of new graph-aware methods for many basic bioinformatics tasks. Here, we focus on the problem of identifying variable sites in the pangenome.

We implement a two-step approach: first we construct a graph pangenome from a graphical fragment assembly (GFA) file that stores the fragments, where each fragment corresponds to a vertex of the graph, then we analyze the graph to detect all bubbles. In pangenome variation graphs, genetic variants appear as bubbles. These sites have a common starting context (a single inbound node), a common exit point (a single outbound node), and a diversity of possible paths that connect the two, each of which represents an allele. This approach allows us to produce variant call format (VCF) files directly from graphs. To do so, we need to choose a specific reference genome, and project the variants (i.e. bubbles of the graph) on it.

Our variant finding procedure is based on a breadth-first visit of the graph and on an analysis of the resulting tree. While this idea allows us to detect all simple bubbles, we are unable to detect some nested bubbles. To overcome this problem we are introducing a more sophisticated approach based on cactus trees and related data structures (Paten *et al.* 2018; PMID: 29461862).

The main dataset for testing our approach is currently a SARS-CoV2 dataset, that is composed of sequences in GFA format of approximately 1.2 GBytes and with 78571 fragments, obtained from 15127 genomes. Our tool has also been tested on a dataset that contains all alternative sequences of the human leukocyte antigen (HLA) complex. Since this region is highly polymorphic, and with common approaches for discovering sequence

variants some reads cannot be aligned to a reference genome, our tool can prove to be a compelling alternative.

We have implemented our approach in Rust, a programming language that combines the efficiency of languages such as C and C++ with an increased attention to memory safety. Our code is capable of obtaining a VCF file from the SARS-CoV2 dataset in ~16 minutes on a machine with 256GB RAM, and we plan on further increasing its speed by leveraging parallel computing, for which Rust is well suited. This, combined with the improved bubble detection algorithm, will allow us to increase both speed and accuracy of our tool.