# Algorithms for Editable Pangenome Graphs

Modern biology increasingly relies on understanding the full genetic diversity within a species. Traditionally, scientists have used a single *reference genome* — a representative DNA sequence — to study genetics. However, this approach can miss important variation present in different individuals. For example, some DNA segments may be entirely absent from the reference, leading to biased or incomplete insights.

To better capture this diversity, researchers have developed the concept of a *pangenome*, which includes all genetic sequences found across multiple individuals of the same species. Instead of a single linear sequence, a pangenome represents all variants, unique genes, and structural differences, providing a richer and more accurate resource for genetic studies.

A powerful way to represent pangenomes is through variation graphs. In this model, each node contains a DNA sequence segment, and paths through the graph correspond to individual genomes. This graph-based approach aligns shared sequences once, reducing redundancy, and naturally captures complex variations such as insertions, deletions, and rearrangements. It also provides a unified coordinate system for precise genome comparison and supports applications such as DNA read mapping and variant detection.

However, defining which DNA segments should be shared across genomes is a non-trivial task. Existing state-of-the-art methods typically rely on pairwise alignments and heuristic homology detection, which introduces ambiguity into the resulting graph structure. Moreover, most tools construct the graph all at once from a fixed set of genomes, making it difficult to incorporate new data or correct errors. Adding or removing genomes requires rebuilding the entire graph from scratch, which is computationally expensive and inefficient for large or evolving datasets.

We have previously developed a novel method that constructs a mathematically well-defined and unique graph structure based on the $k$-mer composition of the input genomes. This approach avoids the ambiguity inherent in alignment-based methods and provides a principled foundation for pangenome construction.

In this project, we aim to extend this method to support dynamic updating by developing algorithms for editable variation graphs—graphs that can be modified incrementally without losing their mathematical guarantees. The goal is to enable the insertion and deletion of genome sequences, as well as the merging of separately constructed graphs, while maintaining a unique and consistent graph structure, regardless of the order in which these operations occur.

Our approach will be based on rigorous theoretical foundations aimed at ensuring that the graph structure remains unique and consistent, meeting the construction's requirements despite successive modifications. By supporting incremental updates, it will become possible to build graphs in parallel and gradually incorporate new data over time without costly reconstruction.

This innovation will make pangenome graphs more practical and scalable for real-world applications, especially in large initiatives such as the Human Pangenome Reference Consortium, which aims to capture the genetic diversity of hundreds of human genomes. It will also benefit bacterial genomics, where precise representation of genomic diversity can aid in studying mechanisms of antibiotic resistance and tracking the spread of resistant strains. In livestock breeding, it will enable the use of advanced genetic selection methods, allowing for more efficient and precise improvement of desirable traits. Furthermore, any field requiring accurate and comprehensive genomic comparisons will gain from the increased flexibility and efficiency of these tools.