Bridging Linear to Graph Alignment for Whole Genome Population Reference

Mohamed Gunady, Stephen Mount, Hector Corrada Bravo, Sangtae Kim

WORK IN A SHELL

Do we need a Whole-Genome Population Reference Graph?

Our approach takes advantage of representing population haplotypes as a graph, and efficiently linearizes the graph using Yanagi's segmentation.

Segments empower linear aligners with a graph representation, while avoiding the expensive computational overhead of aligning over graphs.

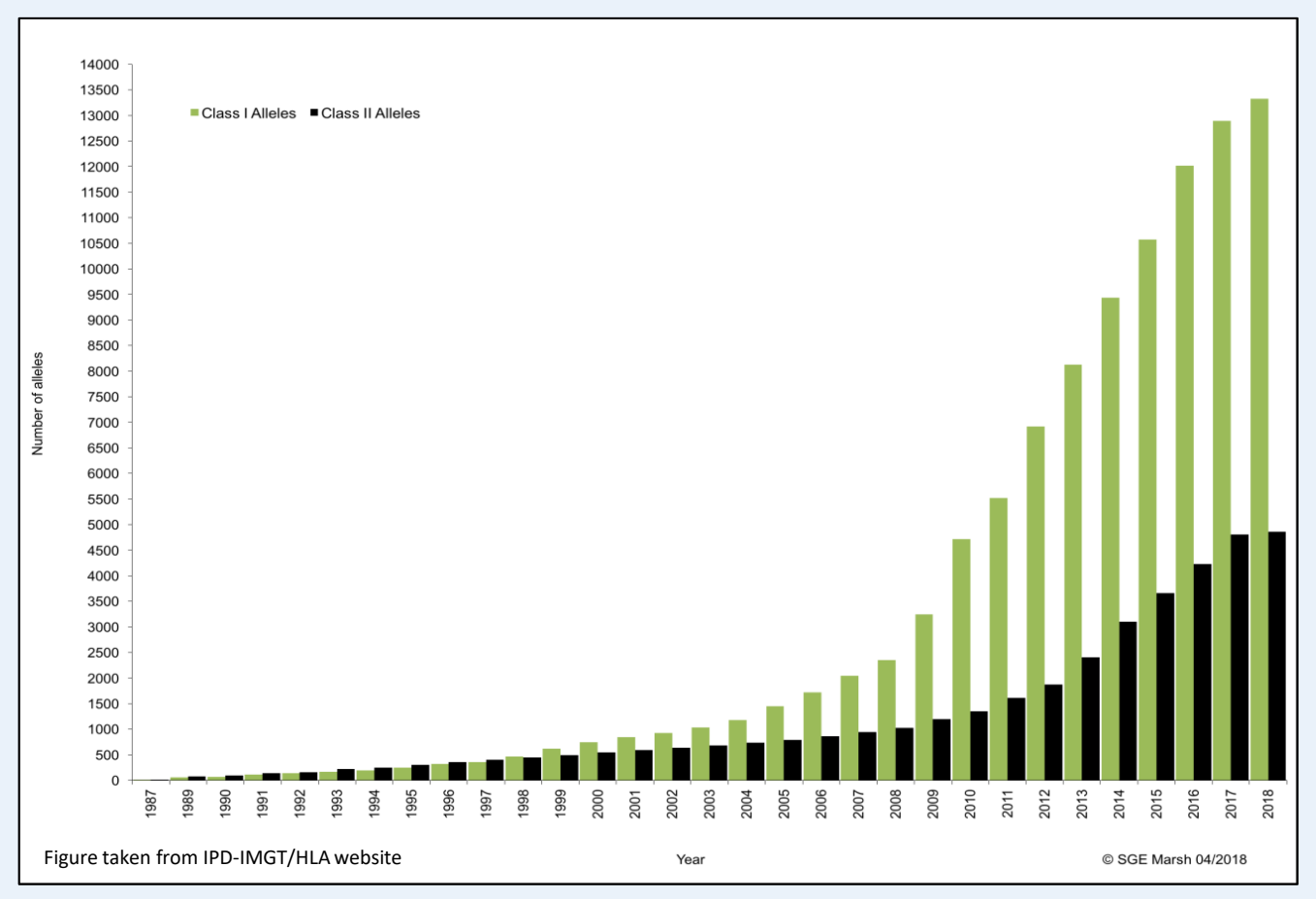
Preliminary results of using segments with linear aligners and kmer-based lightweight aligners show comparable performance to graph aligners

INTRODUCTION

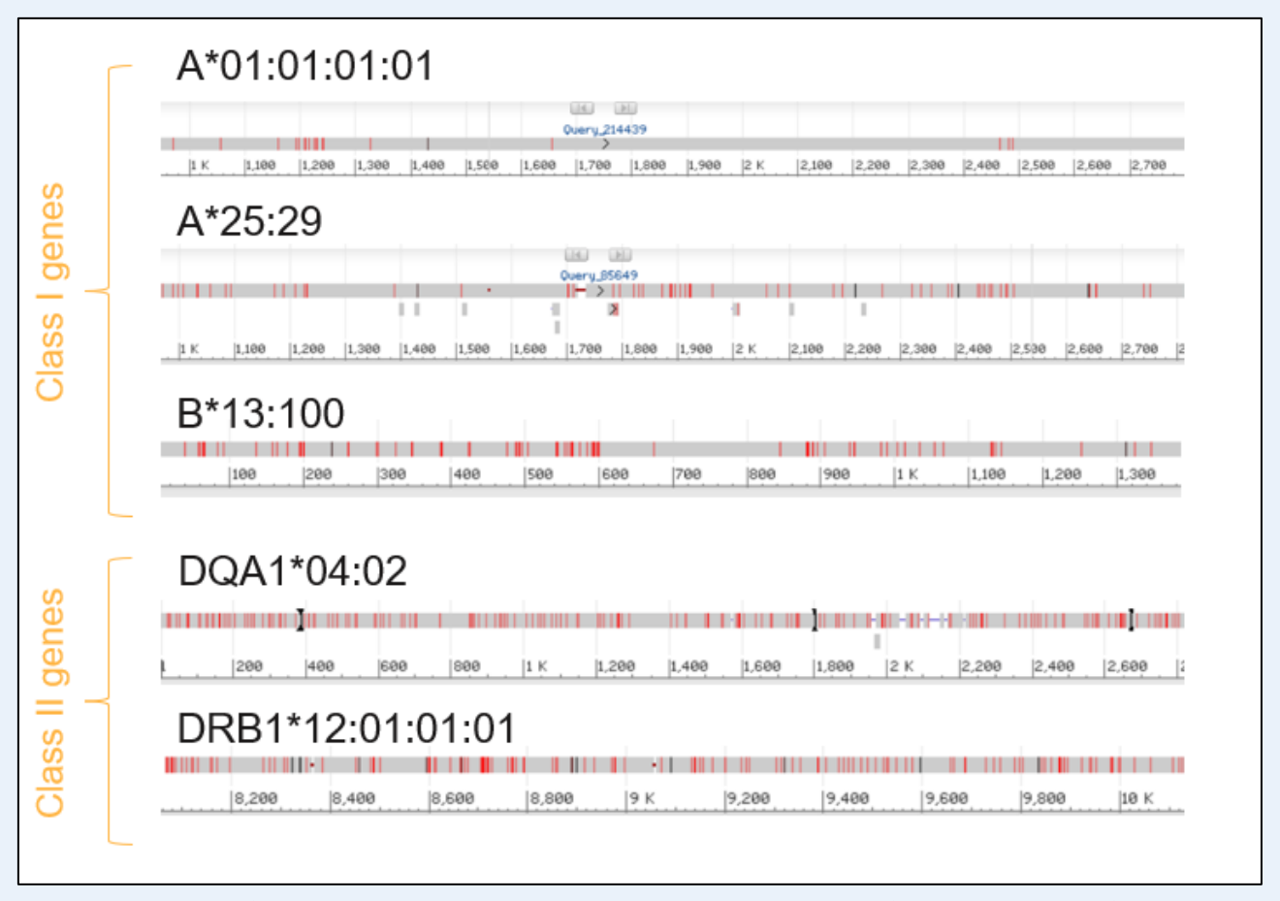
❑ Rapidly-growing databases of Genomic Variants

E.g. IPD-IMGT/HLA database currently has 18,363 allele sequences

❑ HLA genes are highly polymorphic



Yearly growth of IPD-IMGT/HLA database



Examples of HLA alleles. Red marks variants from the reference

❑ Two directions to incorporate alleles into alignment

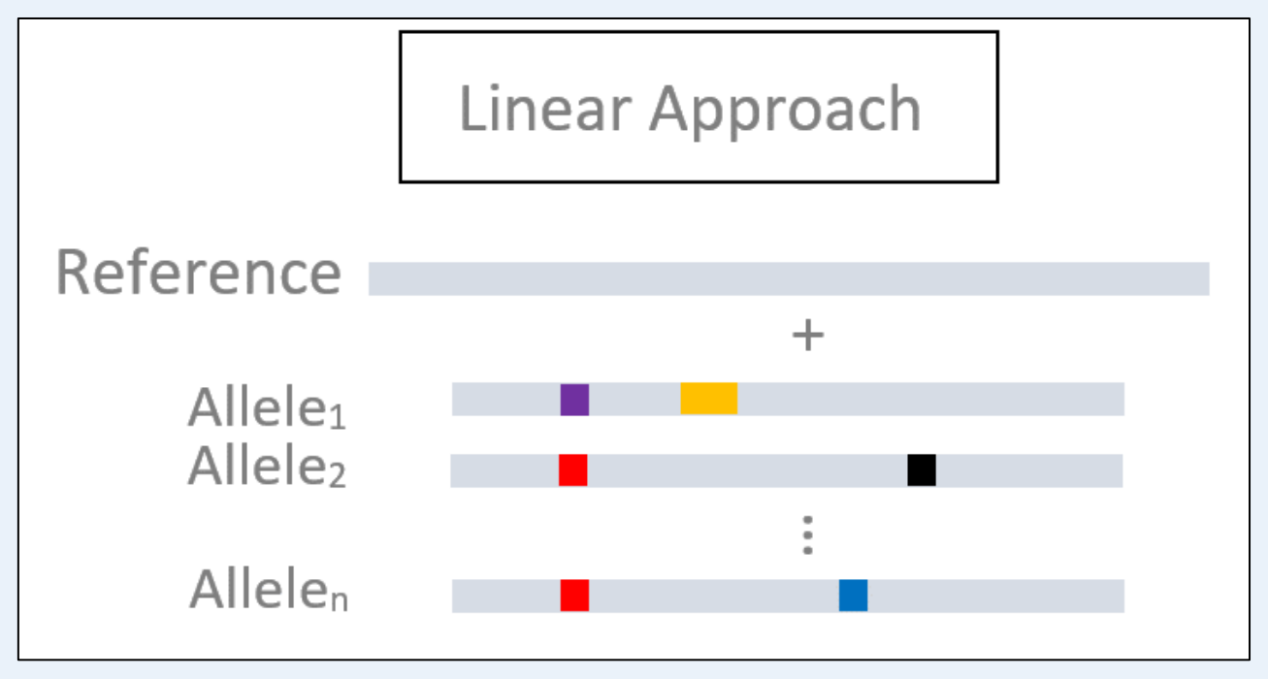
1. Linear Population Reference

Pros:

- Literature and tools well established
- Relatively fast and less expensive

Cons:

- Duplicates major portion of sequences
- Causes ambiguity assigning multi-mapped reads
- No homology relationship between sequences



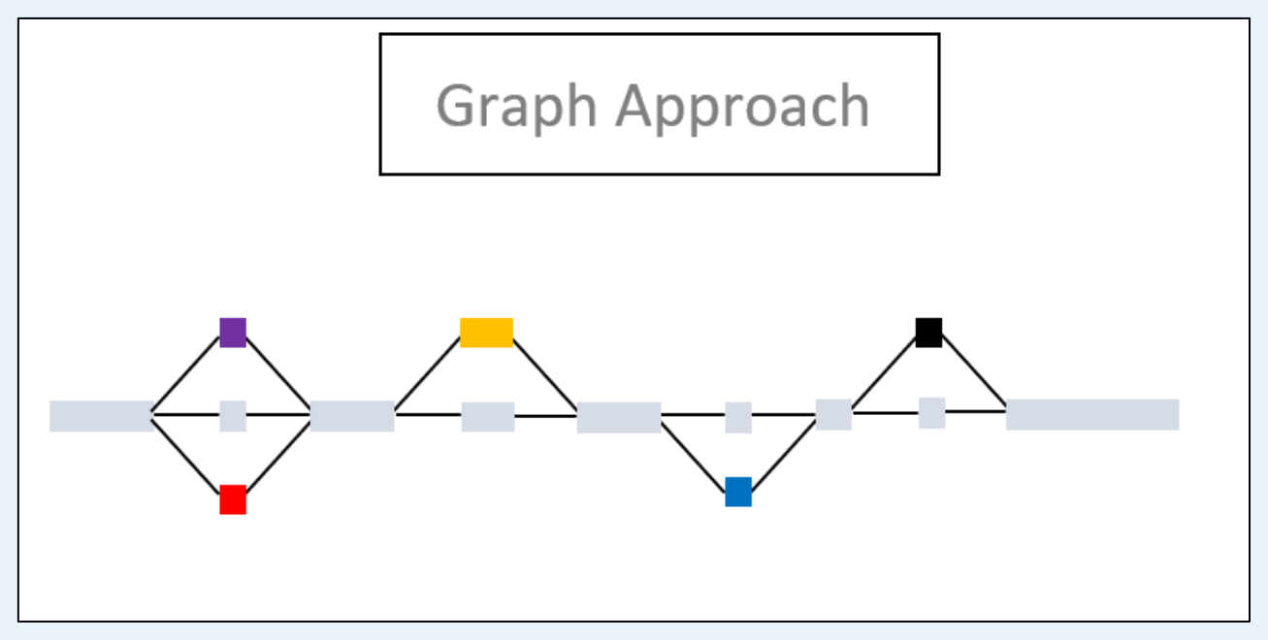
2. Graph Population Reference

Pros:

- Shared sequences represented once
- Preserves structure of the alternative alleles

Cons:

- Graph-based aligners are not mature yet
- Current implementations are computationally expensive



OBJECTIVES

❑ Build population reference genome that takes advantage of the graph structure and properties.

❑ Yet avoid the overhead of aligning over a graph.

❑ Achieve graph approaches accuracy, while maintaining linear approaches speed and flexibility

METHODS

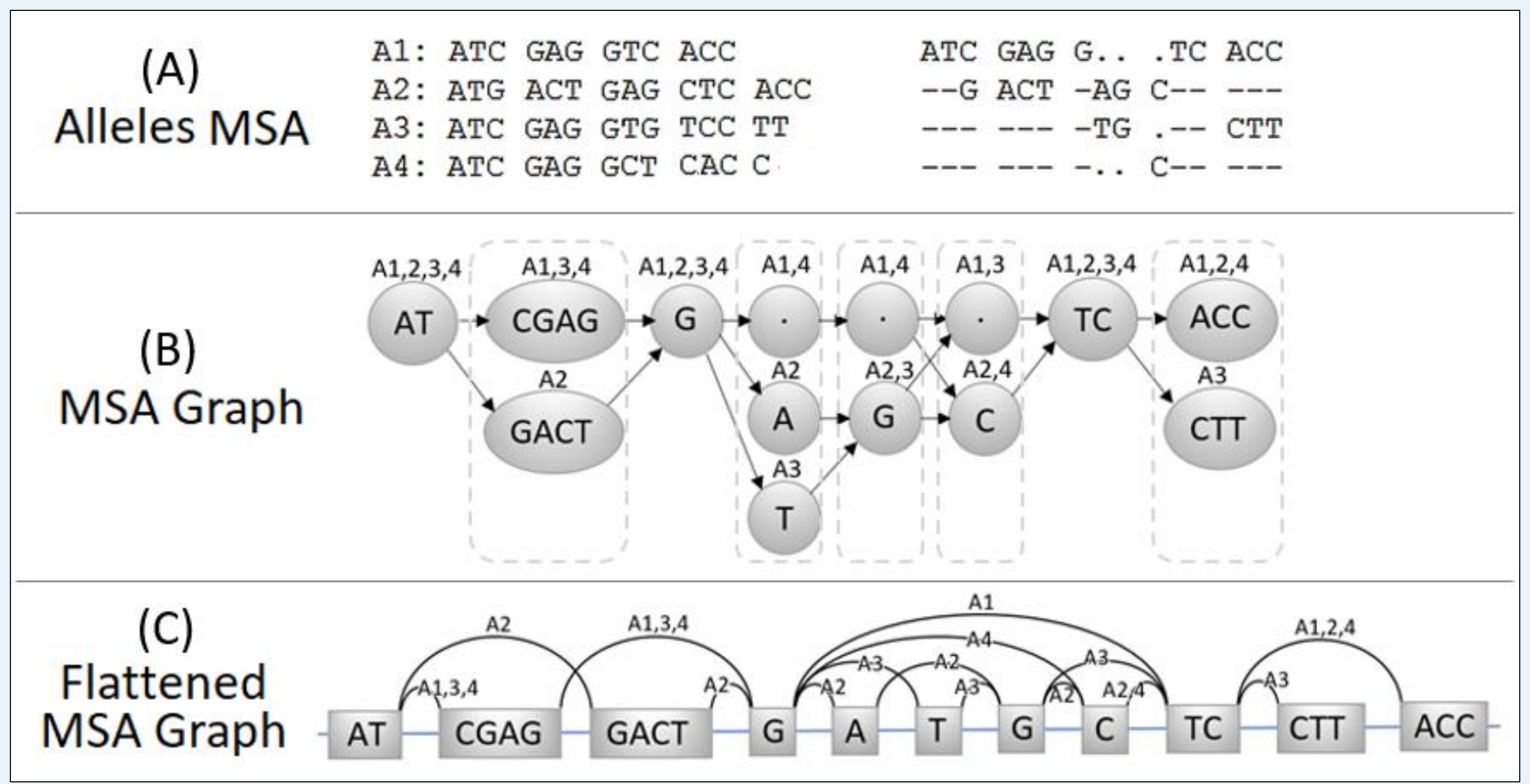
Method Outlines

- Build population genome graph
- Linearize the graph into set of segments
- Use segments as reference for alignment

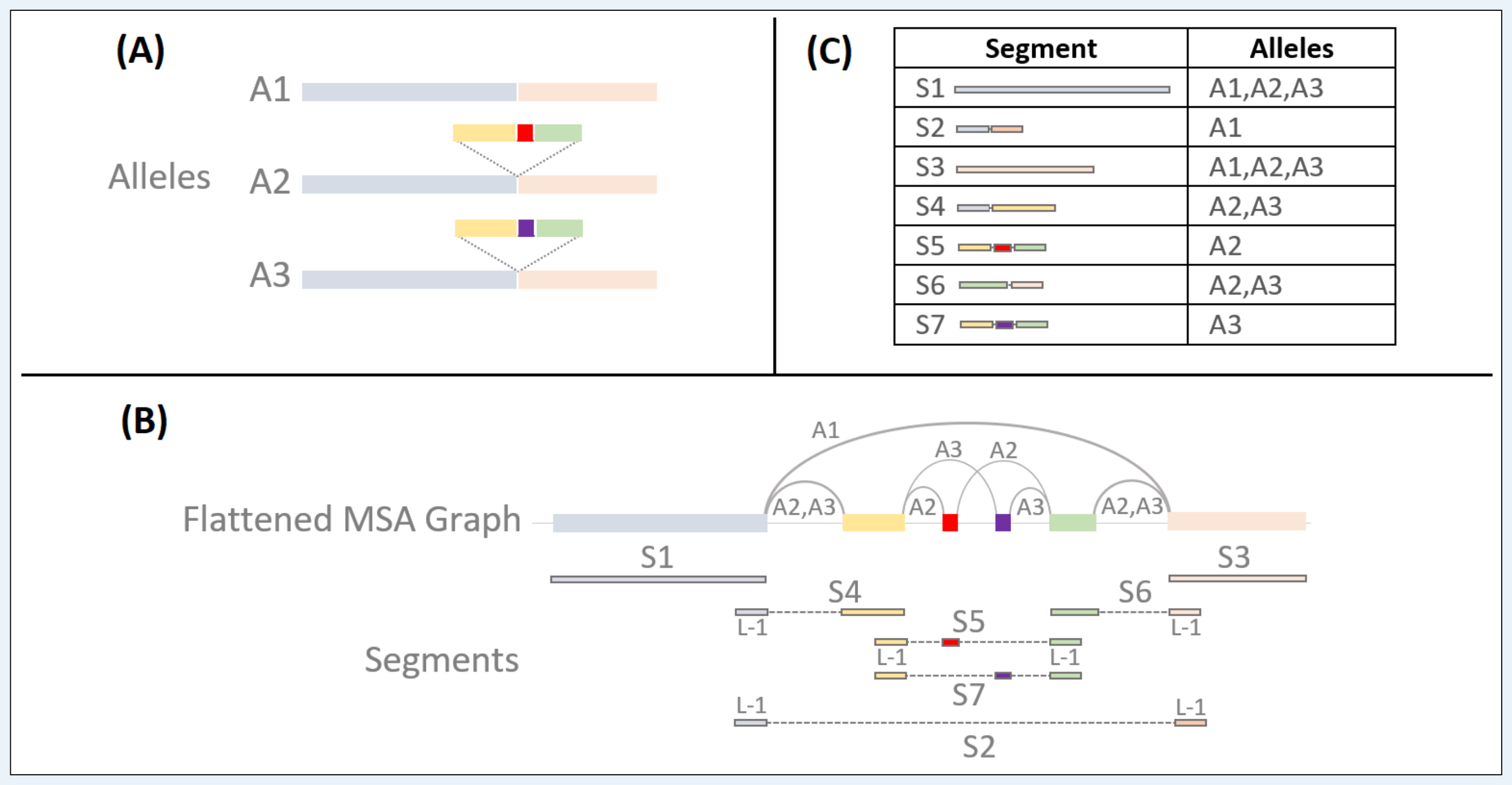
Borrowed Ideas from RNA-seq

Similar challenges seen with transcriptome alignment and similarities between Population Graphs and Splice Graphs motivated our approach.

Linearizing the graph is based on our segmentation approach and tool, *Yanagi**, for RNA-seq



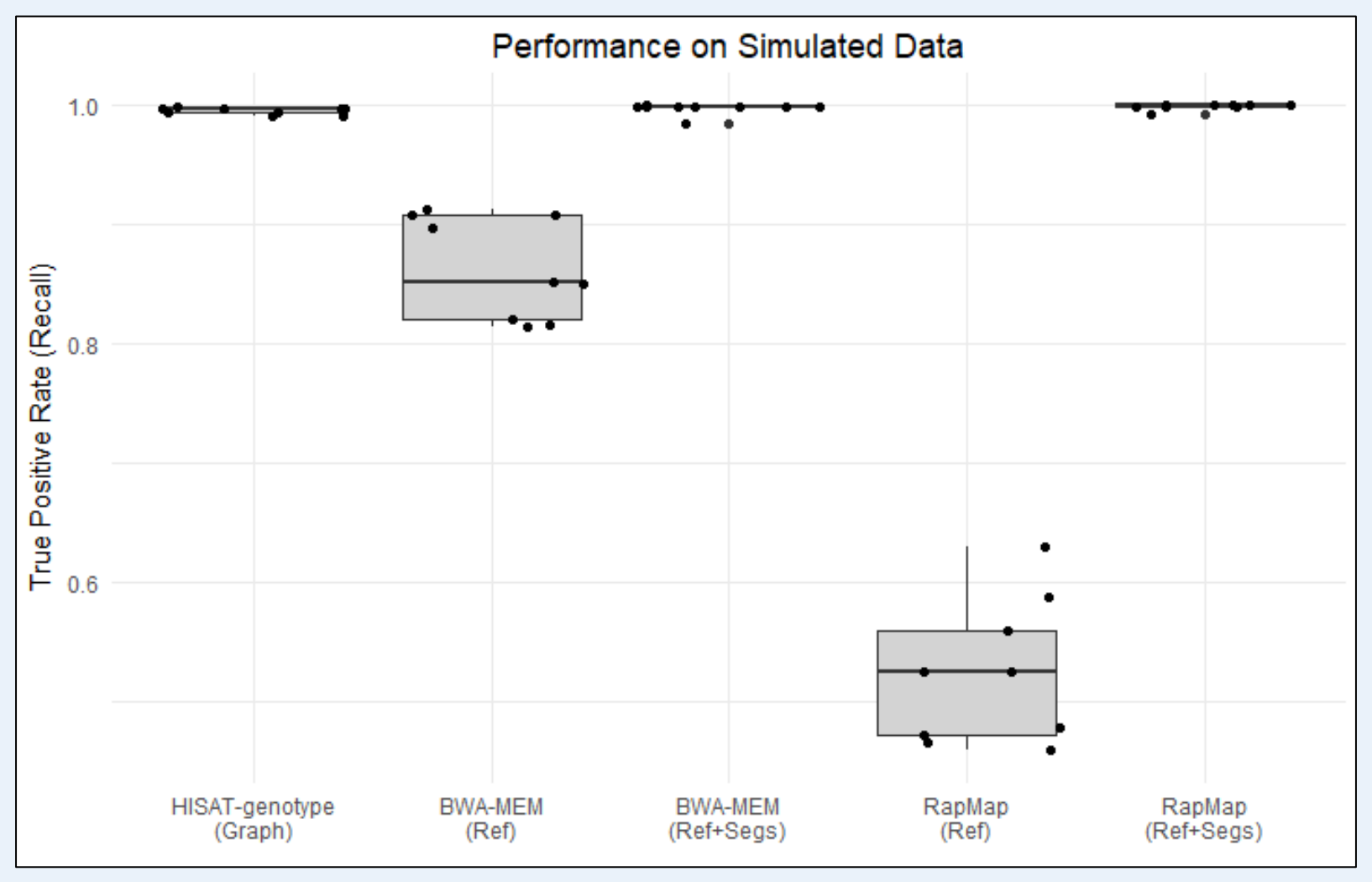
The process of preparing the population graph into Yanagi's graph format. (A) Starts by preparing the alleles MSA. (B) Builds MSA Graph of each genes' alleles as a partial-order graph. (C) Flattening MSA Graphs and replacing INDEL vertexes by new edges.



Segment	Alleles
S1	A1,A2,A3
S2	A1
S3	A1,A2,A3
S4	A2,A3
S5	A2
S6	A2,A3
S7	A3

An example of alleles segments. (A) The sequence structure of the three alleles. (B) Flattened MSA Graph of the alleles along with the seven generated segments (S1..S7) aligned to their corresponding sequences. (C) A summary of the seven segments and their allele membership.

RESULTS



Simulation Dataset of 10 samples. Each sample of 56k HLA reads simulated from two randomly selected alleles for each of the six HLA genes. Plot shows recall rates of extracted HLA reads using 5 approaches: HISAT-genotype (graph aligner), BWA-MEM (linear aligner) with HG38 reference only, BWA-MEM with reference + HLA segments, RapMap (kmer-based lightweight aligner) with reference only, RapMap with reference + HLA segments. Both linear aligners performance are elevated compared to graph aligner when HLA segments are used.

Table 1 Genome library size for the six HLA genes using reference+alleles concatenated, and reference+segments (L=150) in three metrics (reference-only and graph-based library sizes are provided as a reference when applicable). In case of graph, number of bases is estimated as the summation of bases of the graph nodes.

	Reference	Ref+Alleles	Ref+Segments	Graph
Number of bases (Gb)	0.045	9.25	2.39	0.048
Number of sequences	6	2,094	45,609	2,094
FASTA file size (MB)	0.03	10	2.4	NA

Table 2 Running time for alignment of sample NA12878 (24 threads on Dual E5-2690 2.90GHz)

	HISAT-genotype (Graph)	BWA-MEM (Ref+Segs)	RapMap (Ref+Segs)
Running Time	20 hours	8 hours	2 hours