Bridging Linear to Graph-based Alignment with Whole Genome Population Reference Graphs

Recently, several attempts are devoted into building comprehensive catalogues of known genomic variants. However, read alignment approaches that efficiently utilize them are scarce. Since the catalogues contain hundreds of alleles which in general share most of their sequences except where the instant variations appear, that makes a graph of these alleles a reasonable and efficient representation of the data. Unfortunately, the lack of efficient implementations and algorithms for graph-based alignment makes graph-based approaches computationally expensive for practical application.

Our approach takes advantage of graph representation in obtaining prominent levels of data compressions, and efficiently linearizes the variants graph by sacrificing a portion of the compression ratio. Our model for linearizing the variants graph depends on our previous work in transcriptome segmentation for RNAseq. For each gene of interest, we start from the multiple sequence alignment (MSA) of the individual alleles (which can be already provided in the catalogue or derived from VCF files). Then we use Yanagi[1] to generate a set of *maximal L-disjoint* segments representing the linearized MSA graph (Figure 1). The segments library is then used by any alt-aware linear alignment tool.

The advantage of using our approach over the standard alt-aware aligners that uses a reference of the genome sequence appended by the population haplotypes is that segments sequences are highly compressed which is space efficient and speeds up the alignment process (Table 1). On the other hand, our approach is potentially flexible such that the generated segments can be used with most linear aligners rather than being limited to a specific graph model. Moreover, it avoids the expensive computational demands of aligning over graphs.

As a proof of concept, we test our approach on IPD-IMGT/HLA database to study six class I and class II HLA genes[2] with significant medical importance. In addition to testing using graph aligners (HISAT-genotype[3]), linear aligners (BWA-MEM), and linear aligners with segments (BWA-MEM), we also included a test for using fast and lightweight RNAseq aligners (RapMap) to examine the possibility of using fast RNAseq aligners for the task of read extraction. We simulated three datasets simulating three scenarios: 1) ClassI-Easy: reads are simulated from HLA class I alleles that are not very different from the reference genome, 2) ClassI-Hard: uses HLA class I alleles that are different from the reference, and 3) ClassII-Hard: uses HLA class II alleles that are much different from the reference. Preliminary results (Table 2) showed that the more divergent the samples are, the harder for linear aligners to correctly align reads. However, assisting linear aligners with the population segments can achieve comparable results to graph aligners without compromising the space and computational requirements (Table 3). Although RNAseq aligners with the reference alone performed the worst, adding segments elevated its performance back.

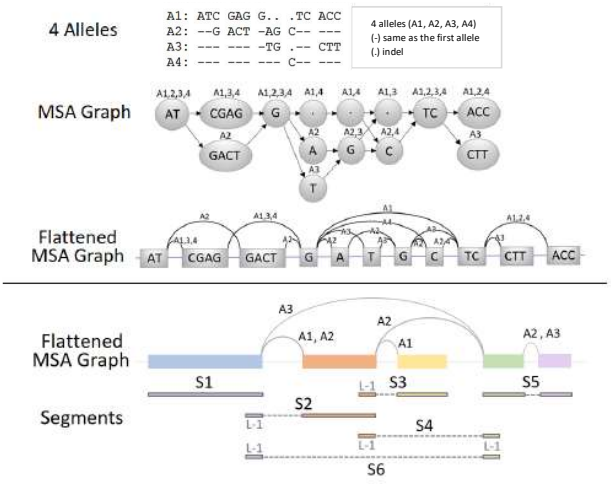

Figure 1: Illustrative examples for our segmentation model in two steps. (Top) Construct a flattened MSA graph of the gene's alleles. (Bottom) Create *maximal L-disjoint* segments of the population graph using Yanagi.

Table 1: Genome library size for the six HLA genes. In case of graph, number of bases is counted as the bases sum 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: Number of correctly aligned reads from simulated reads using: HISAT-genotype (graph aligner), BWA-MEM (linear alt-aware aligner), and RapMap (RNAseq lightweight aligner). In case of both BWA-MEM and RapMap, results are shown either when using only the reference genome or the reference combined with yanagi's segments for the six HLA genes.

	Num.	HISAT-	BWA-MEM		RapMap	
	Reads	genotype	Ref	Ref+Segs	Ref	Ref+Segs
ClassI-Easy	6,000	5,900	6,000	6,000	4,163	5,990
ClassI-Hard	6,000	5,966	5,797	6,000	3,553	5,990
ClassII-Hard	14,000	13,844	12,232	13,997	7,628	13,975

Table 3: Running time for alignment of real sample NA12878.

	HISAT-genotype (Graph)	BWA-MEM (Ref+Segs)	RapMap (Ref+Segs)
Running Time	10 hours	3 hours	2 hours

[1] Gunady, Mohamed K., et al. "Yanagi: Transcript Segment Library Construction for RNA-Seq Quantification." LIPIcs-Leibniz International Proceedings in Informatics. Vol. 88. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017. (WABI-2017)

[2] The six genes: HLA-A, HLA-B, HLA-C, HLA-DQA1, HLA-DQB1, HLA-DRB1. We use L=150 for Yanagi in all experiments.

[3] Kim, Daehwan, Joseph M. Paggi, and Steven Salzberg. "HISAT-genotype: Next Generation Genomic Analysis Platform on a Personal Computer." bioRxiv (2018): 266197.