Bridging Linear to Graph Alignment for Whole Genome Population Reference

Mohamed Gunady, Steve Mount, Hector Bravo, Sangtae Kim

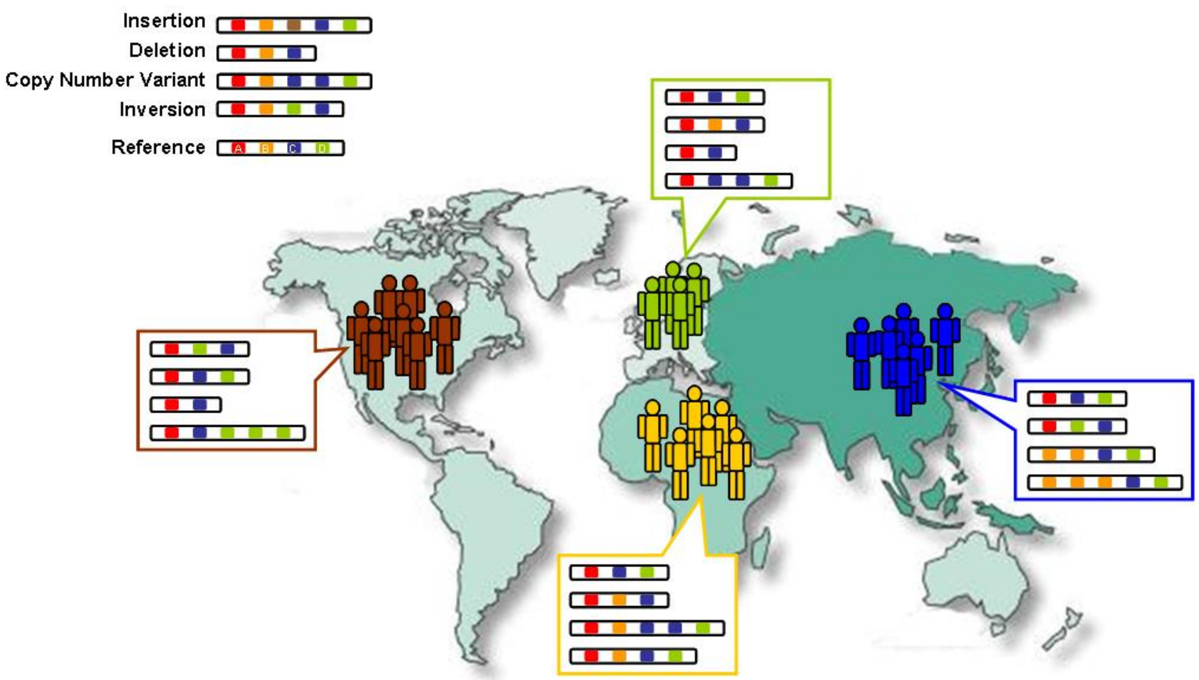
Department of Computer Science

University of Maryland – College Park

CENTER FOR BIOINFORMATICS & COMPUTATIONAL BIOLOGY

Background

- Whole Genome Population Reference
  - A challenge handling population diversity

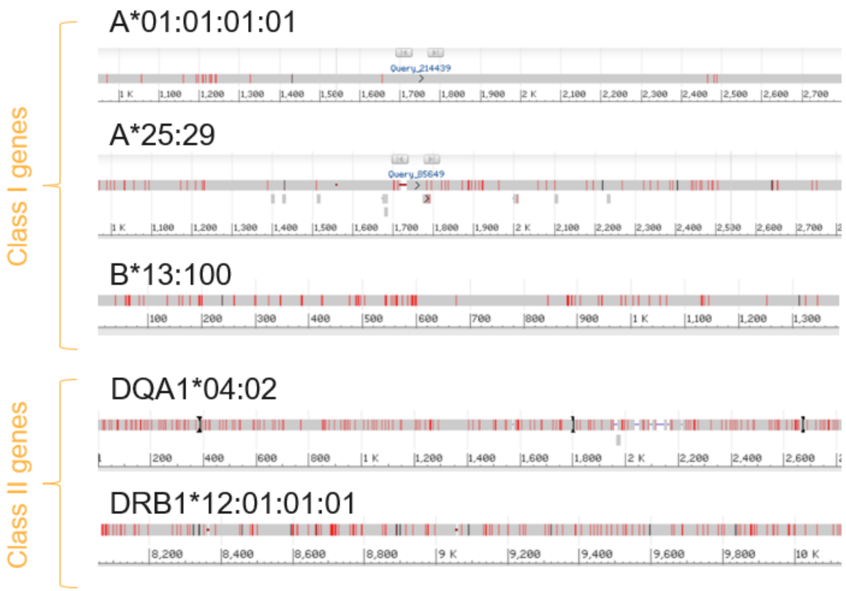


1000 Genomes Project

Background

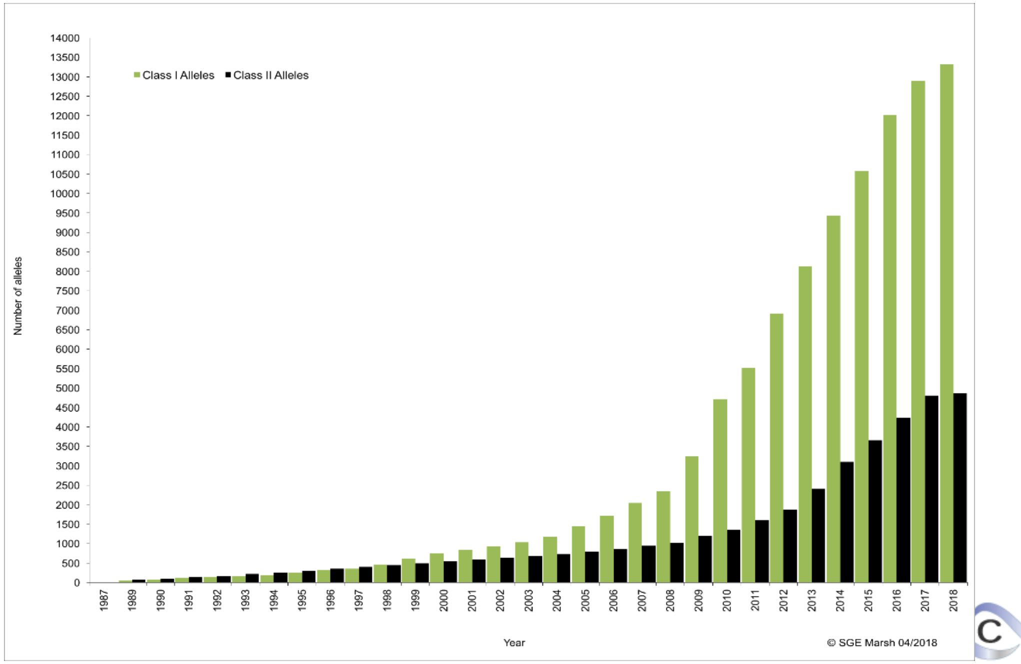
- Some genes are highly polymorphic
  - E.g. Human Leukocyte Antigen (HLA) system
  - Regulates the human immune system, so of significant medical importance

- Alignment with reference only, can miss significant amount of reads originating from HLA genes



Background

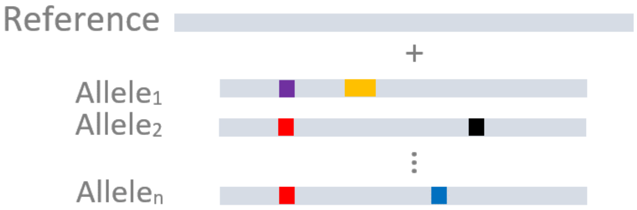
- Projects providing catalogs of known genomic variants, e.g.
  - IPD-IMGT/HLA Database
  - 1000 Genomes Project
- IPD-IMGT/HLA Database
  - Rapidly growing, provides 18,363 allele sequences for public access



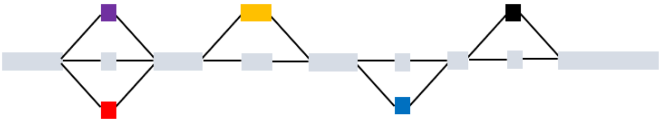
Background

- Two directions to incorporate alleles into alignment

Linear Approach



Graph Approach



Alt-aware Aligners

e.g. BWA-MEM

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

Graph Aligners

e.g. HISAT-genotype

Pros:

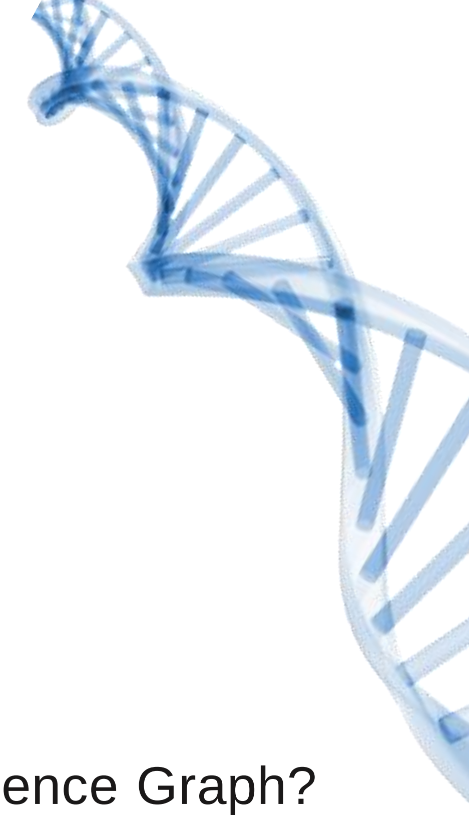
- Shared sequences represented once
- Preserves structure of the alternative alleles

Cons:

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

Our Approach

Population Graph Segmentation



Our Approach

Population Graph Segmentation

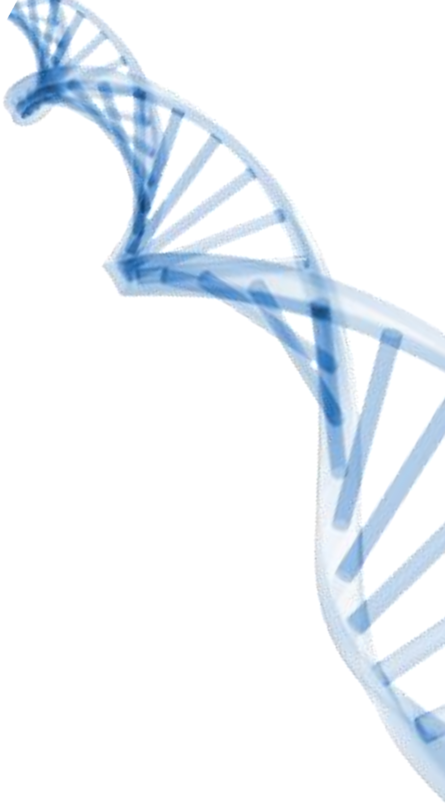
Question:

Do we need a Whole-Genome (WG) Population Reference Graph?

Can we preserve graph’s advantages while maintaining linear approaches speed and flexibility?

Our Approach

Population Graph Segmentation

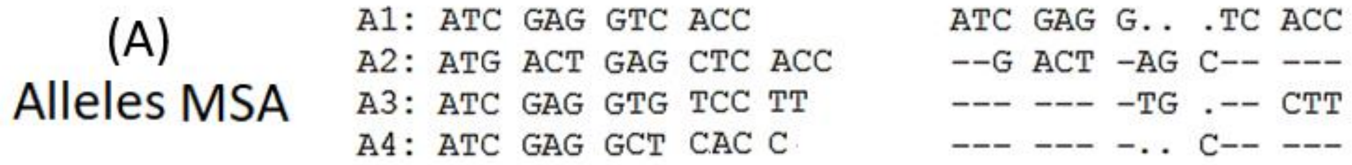


- Method Outlines:
  1. Build population genome graph
  2. Linearize the graph into set of segments
  3. Use segments as reference for alignment

Our Approach

Population Graph Segmentation

1. Build population genome graph



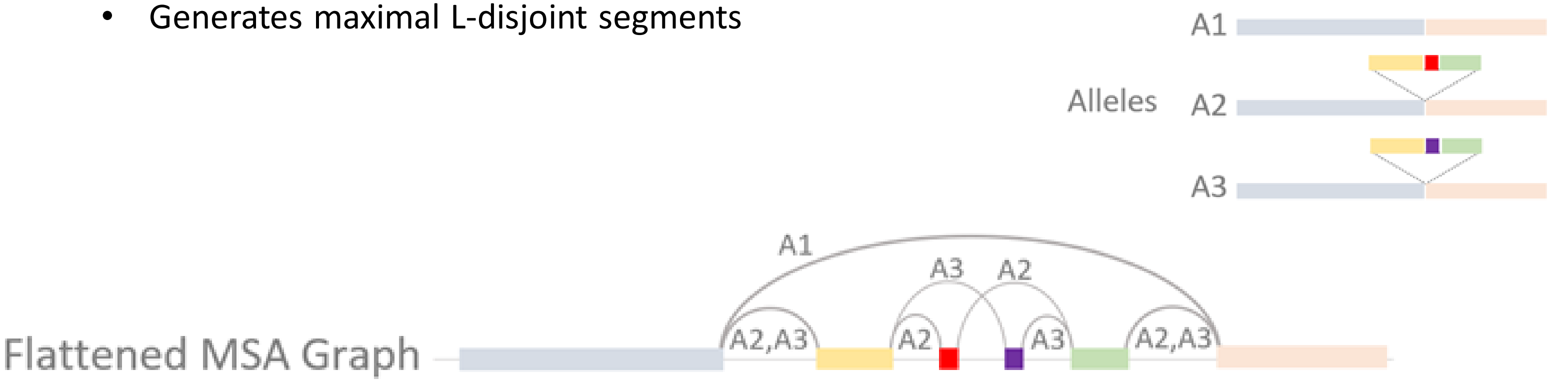
(A)
Alleles MSA

```
A1: ATC GAG GTC ACC           ATC GAG G.. .TC ACC
A2: ATG ACT GAG CTC ACC       --G ACT -AG C-- ---
A3: ATC GAG GTG TCC TT        --- --- -TG .-- CTT
A4: ATC GAG GCT CAC C         --- --- -.. C-- ---
```

Our Approach

Population Graph Segmentation

2. Linearize the graph into set of segments
   - Adapt our transcriptome segmentation approach (Yanagi*)
     - Generates maximal L-disjoint segments

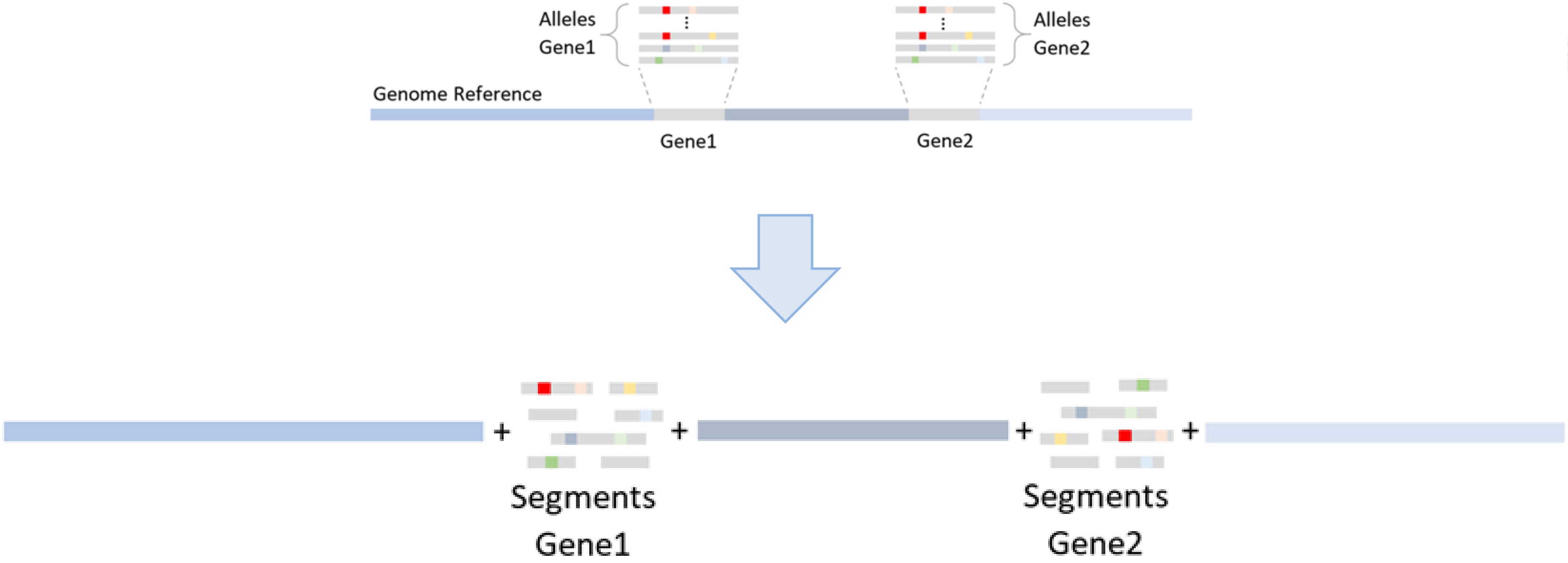


* Gunady, M.K., Cornwell, S., Mount, S.M., Bravo, H.C.: Yanagi: Transcript Segment Library Construction for RNA-Seq Quantication. (WABI 2017)

Our Approach

Population Graph Segmentation

3. Use gene segments as its reference for alignment

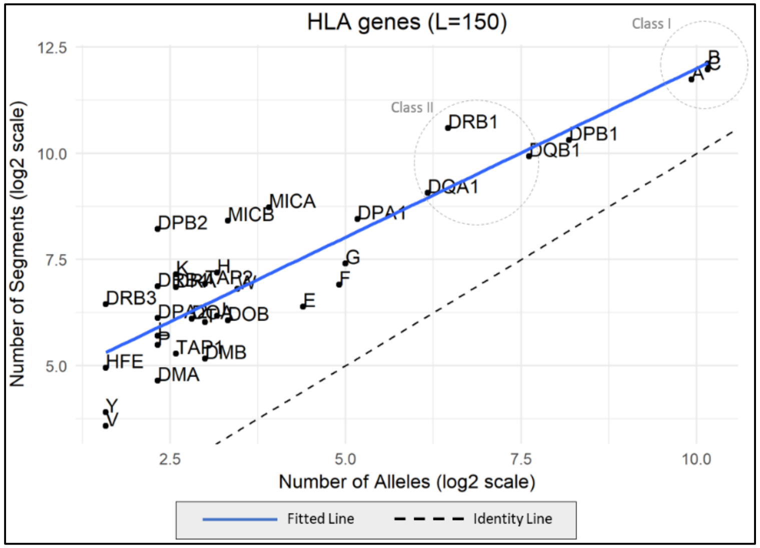


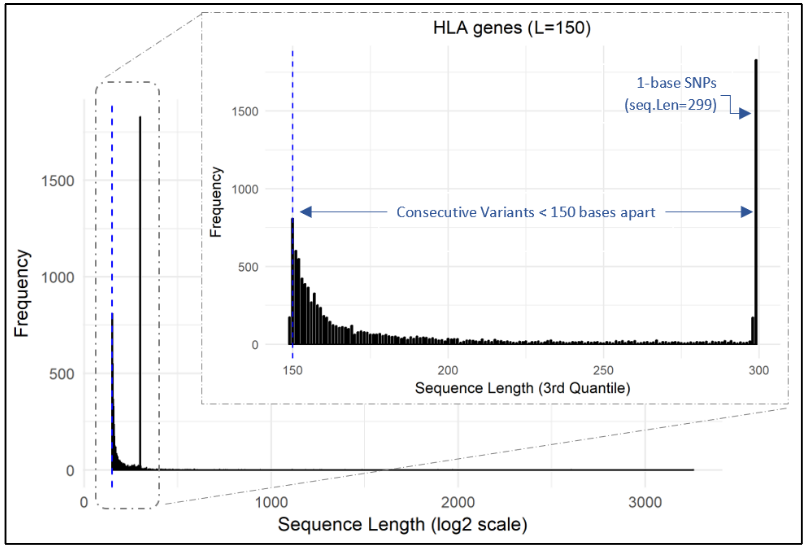
Experiments

HLA Class I and Class II genes

HLA Segments Analysis

- HLA Segments (L=150)





Class I genes

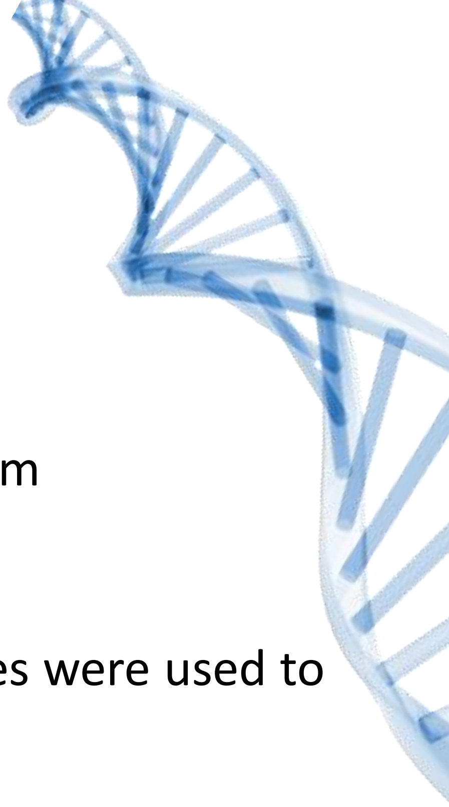
Class II genes

	A	B	C	DQA1	DQB1	DRB1	Total
Num. of 16-mers	19,115	18,865	20,691	19,274	26,830	53,076	148,764
New 16-mers (%)	45.7%	49.8%	49.6%	19.2%	40.7%	27.9%	36.6%

University of Maryland

CENTER FOR BIOINFORMATICS & COMPUTATIONAL BIOLOGY

HLA Reads Extraction



- Simulated Data
  - 10 Simulated samples of combining reads simulated from
    - 6 HLA genes (-A, -B, -C) and (-DQA1, -DQB1, -DRB1)
    - Non HLA genes
  - Per sample, per HLA gene: Two randomly selected alleles were used to simulated reads
    - Paired-End
    - Length 150bp
    - Average coverage of x40
  - A sample contains ~56k HLA reads and 2M non-HLA reads

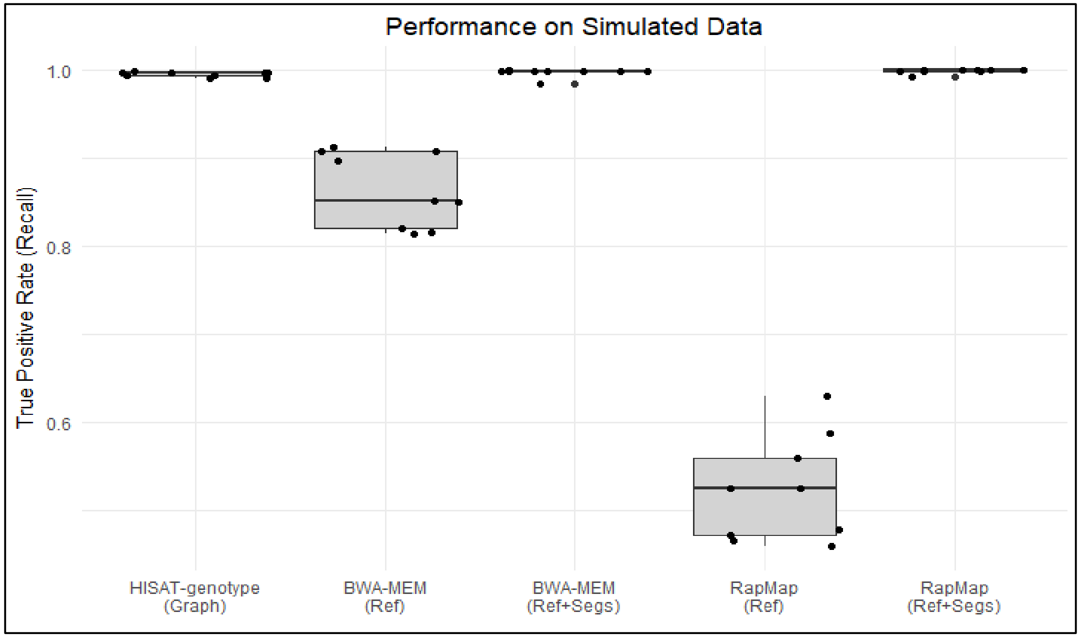
HLA Reads Extraction

- Simulation Results
  - $Recall = \frac{HLA\ reads\ mapped\ to\ HLA\ genes}{All\ HLA\ reads}$

Graph aligner

Alt-aware linear aligner

Kmer-based pseudo aligner



HLA Reads Extraction

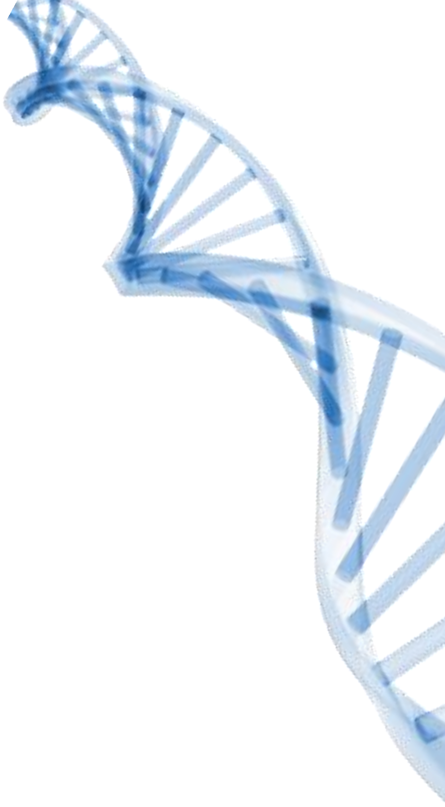
- Real Data Running Time
  - Sample NA12878
  - (24 threads on Dual E5-2690 2.90GHz)

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

Summary

- We introduced an approach of linearizing population haplotypes graph using Yanagi's segmentation.
- Linear aligners with allele segments can achieve graph aligners' performance, while avoiding the expensive computational overhead of aligning over graphs.
- Yanagi's approach opens the door for bridging the gap between linear and graph representations of catalogs of sequences in different domains.

Future Extensions



- Experiments on other polymorphic genes
  - E.g. selected genes from 1000 Genomes Project
- Handle complex repeats
- Handle complex Structural Variants

Acknowledgments

- UMD Team
  - Mohamed Gunady
  - Hector Bravo
  - Stephen Mount

- Illumina Team
  - Sangtae Kim
  - Chris Sanders

Thank you!

Questions?

See Poster A-148

Contact: mgunady@cs.umd.edu

Yanagi Github: https://github.com/mgunady/yanagi