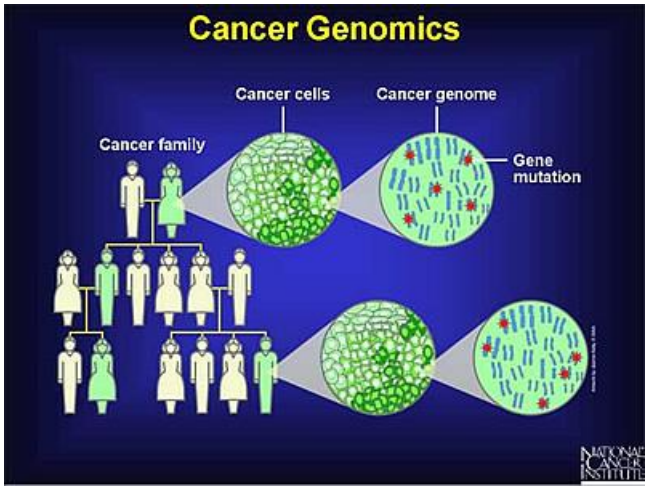
Applications of Graph Segmentation Algorithms for Quantitative Genomic Analysis

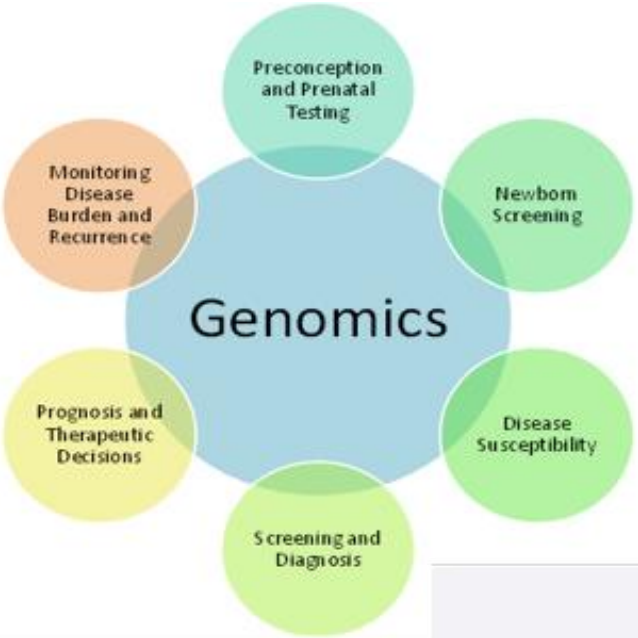
Mohamed Gunady

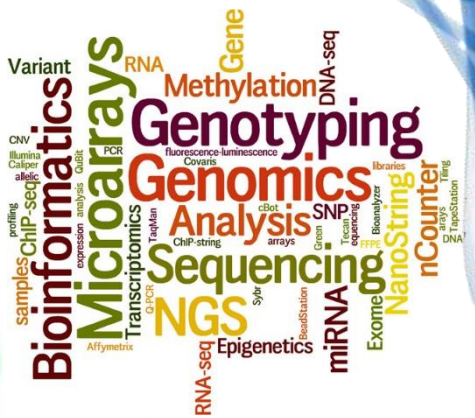
PhD Preliminary Exam Talk

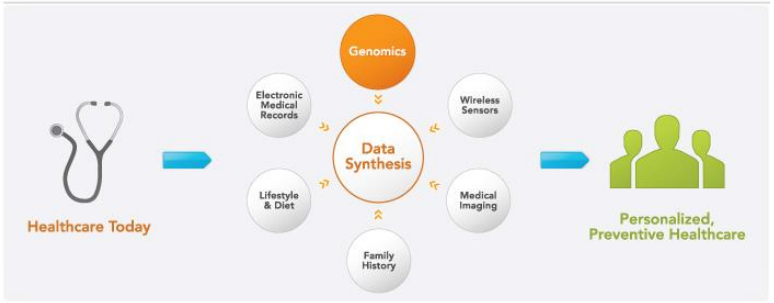
CENTER FOR BIOINFORMATICS & COMPUTATIONAL BIOLOGY

The Age of Genomics





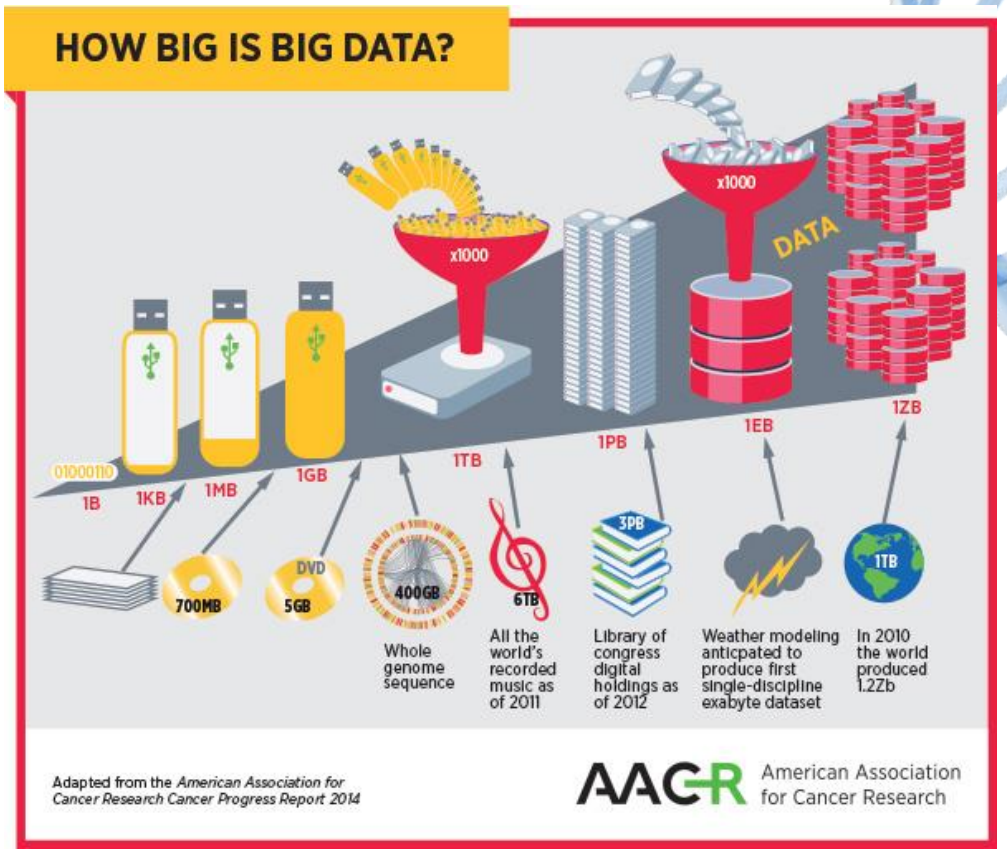




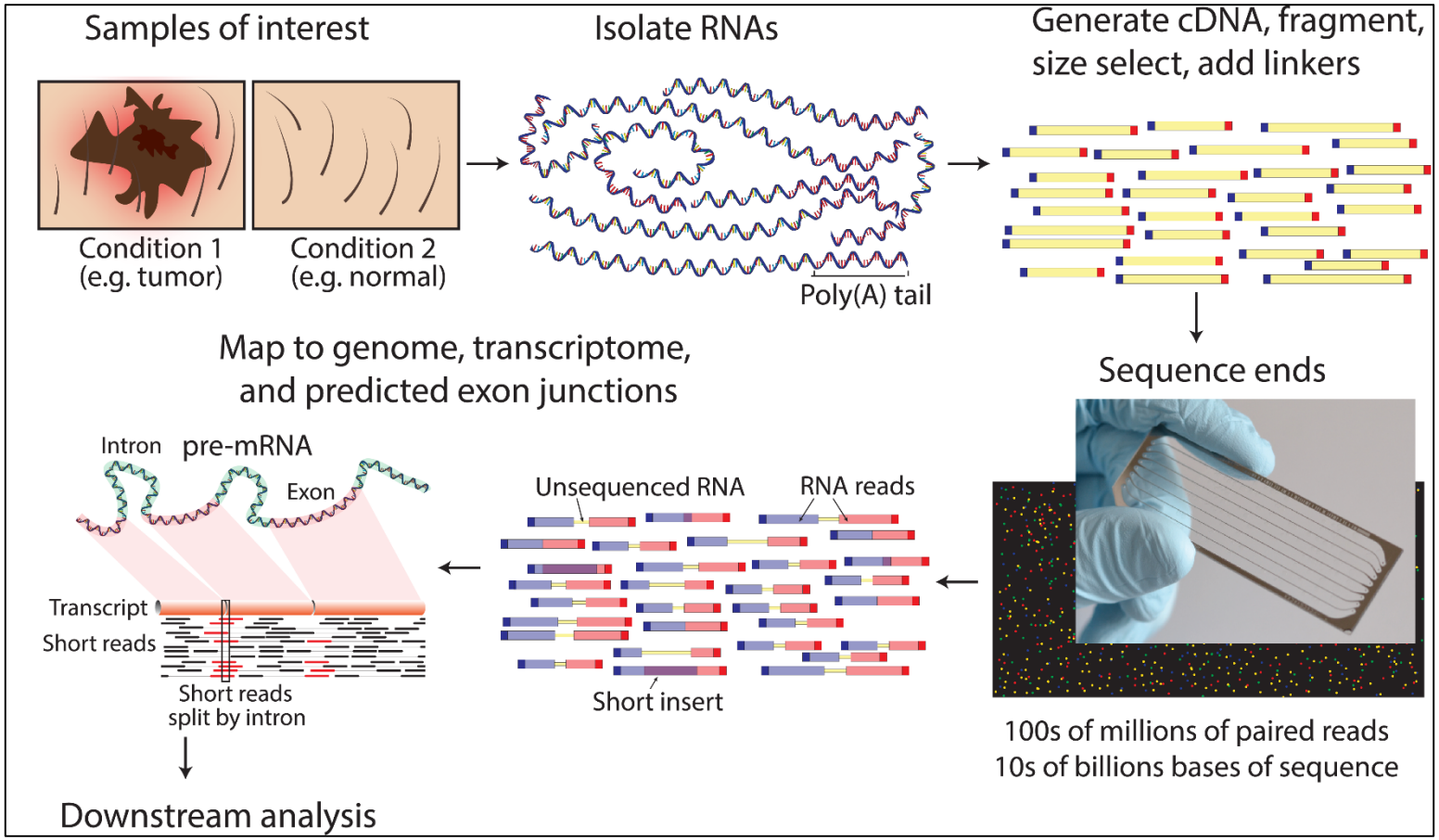
Big Data in Genomics

- Genomic Analysis is a typical Big Data

- Bioinformatics Challenges:
  - Develop faster pipelines
  - Lightweight and efficient
  - Interpretability and Accuracy



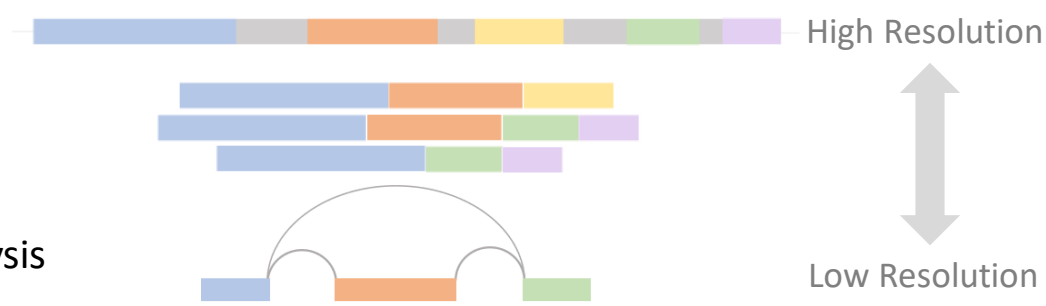
RNA-seq & Transcriptomics



https://en.wikipedia.org/wiki/RNA-Seq#/media/File:Journal.pcbi.1004393.g002.png

Overview

- Introduce a graph segmentation approach
  - Implemented in *Yanagi*, an efficient tool for transcriptome segmentation
- Show case of using Yanagi in:
  - RNA-seq down stream analysis in the three resolutions
    - Gene-level Analysis
    - Transcript-level Analysis
    - Alternative Splicing Analysis



  - Building population reference genomes for WGS

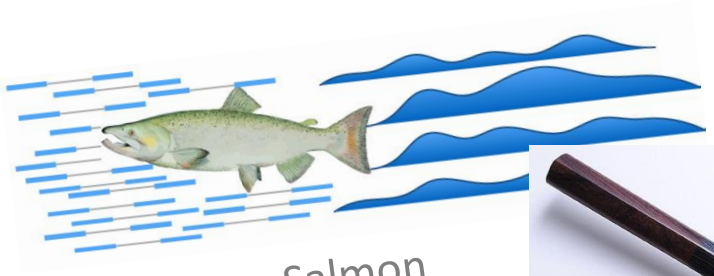
Transcriptome Segmentation

Yanagi: Transcript Segment Library Construction for RNA-Seq Quantification

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

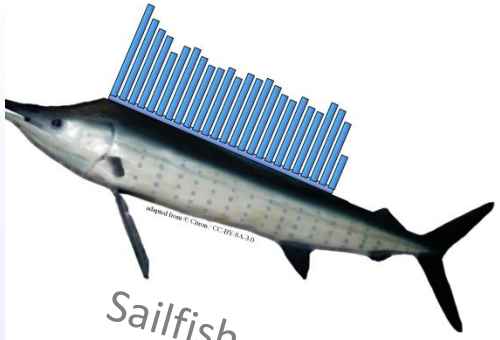
CBCB
Center for Bioinformatics & Computational Biology

Yanagi?



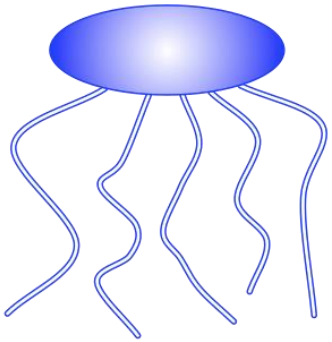
Salmon

Yanagi



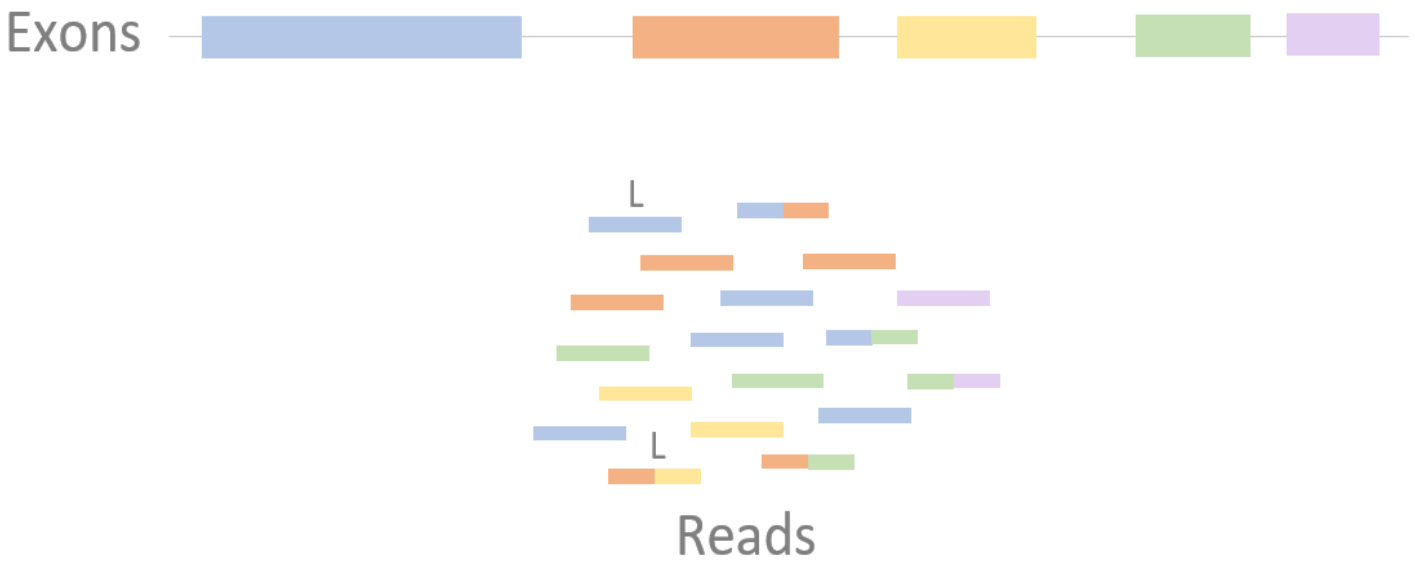
Sailfish

Jellyfish



RNA-seq example

- For some gene with 5 exons



RNA-seq example

- For some gene with 5 exons
  - Has 3 possible isoforms
  - 95% of human genes with multiple exons undergo Alternative Splicing (*Jorge Vaquero-Garcia et al. 2016*)

RNA-seq example

- For some gene with 5 exons
  - Has 3 possible isoforms
  - 95% of human genes with multiple exons undergo Alternative Splicing (*Jorge Vaquero-Garcia et al. 2016*)

- Challenges aligning over the transcriptome:
  - Ambiguity due to multi-mapped reads
    - Usually resolved using probabilistic models like EM
  - Local splicing variations may lead to a combinatorial number of transcripts (*Haas B.J. et al. 2003*)
    - Standard Annotations list only a minimal subset
  - Short-read sequencing does not provide information for correlation between distant splicing events (*Hagen Tilgner et al. 2015*)

Yanagi's Approach

- Our Vision:
  - Eliminate multi-mapping trivially caused by the significant share of genomic regions.
  - Building sufficient statistics describing individual splicing events.
    - Independently from the estimation of transcript abundances.
  - Utilize the graph representation of the transcriptome.
    - Without building a special graph-based aligners.

Yanagi's Approach

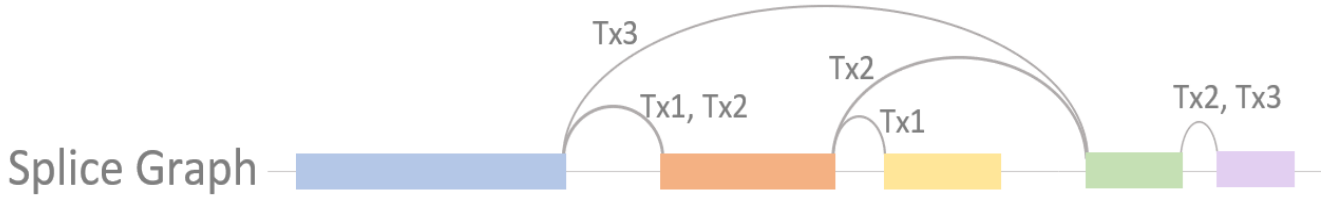
- Idea Overview:
  - Segment the transcriptome into a set of disjoint regions.
  - Without losing any possible transcriptome sub-sequences.
    - I.e. Linearizing the splice graph
  - Then use the generated segments as a reference instead of the transcriptome.

Yanagi's Approach

- Starting from the Splice Graph
- Segments are *L-Disjoint*

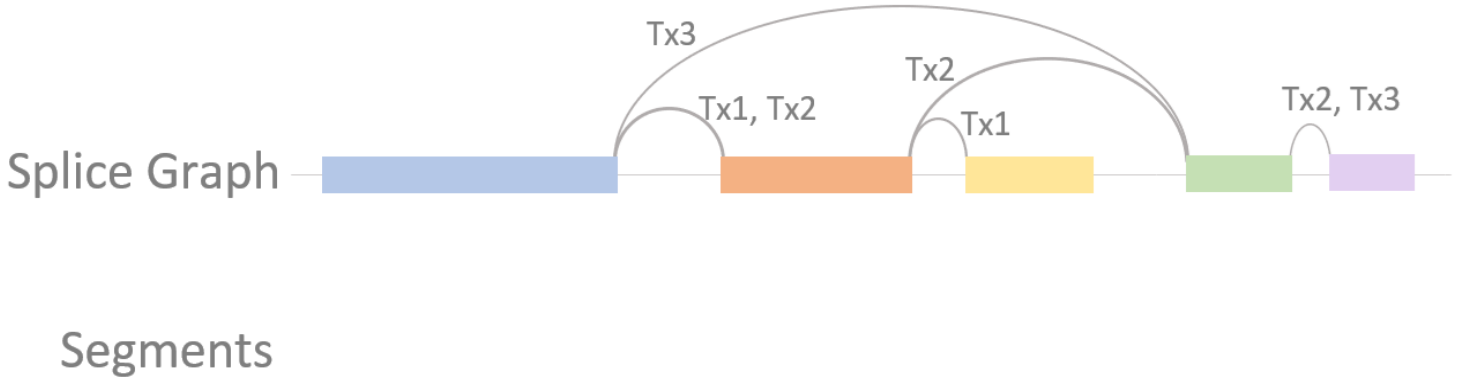
$$width\left[overlap\left(seg_i, seg_j\right)\right] < L; i \neq j$$

- *L* corresponds to the read length
- No read of length at least *L* can map to two segments
  - Ignoring sequencing errors and paralogs for now!



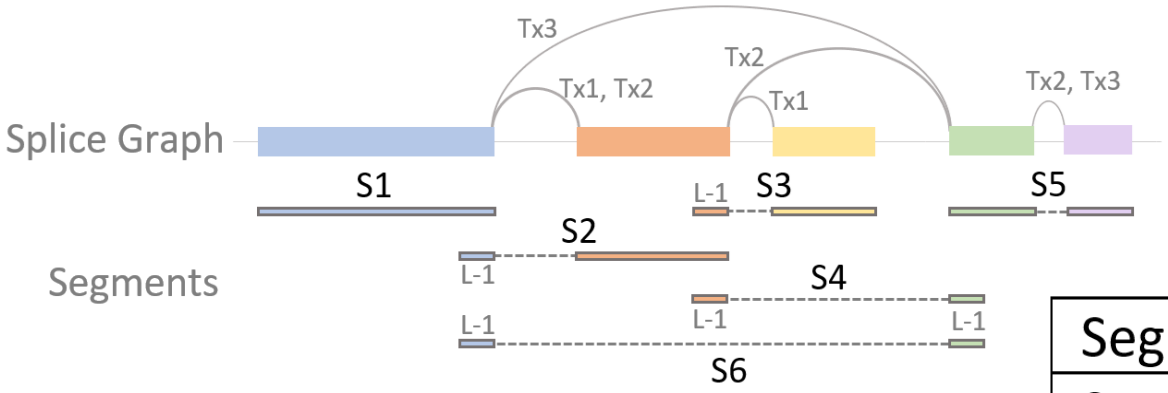
Yanagi's Approach

- Segmentation Algorithm



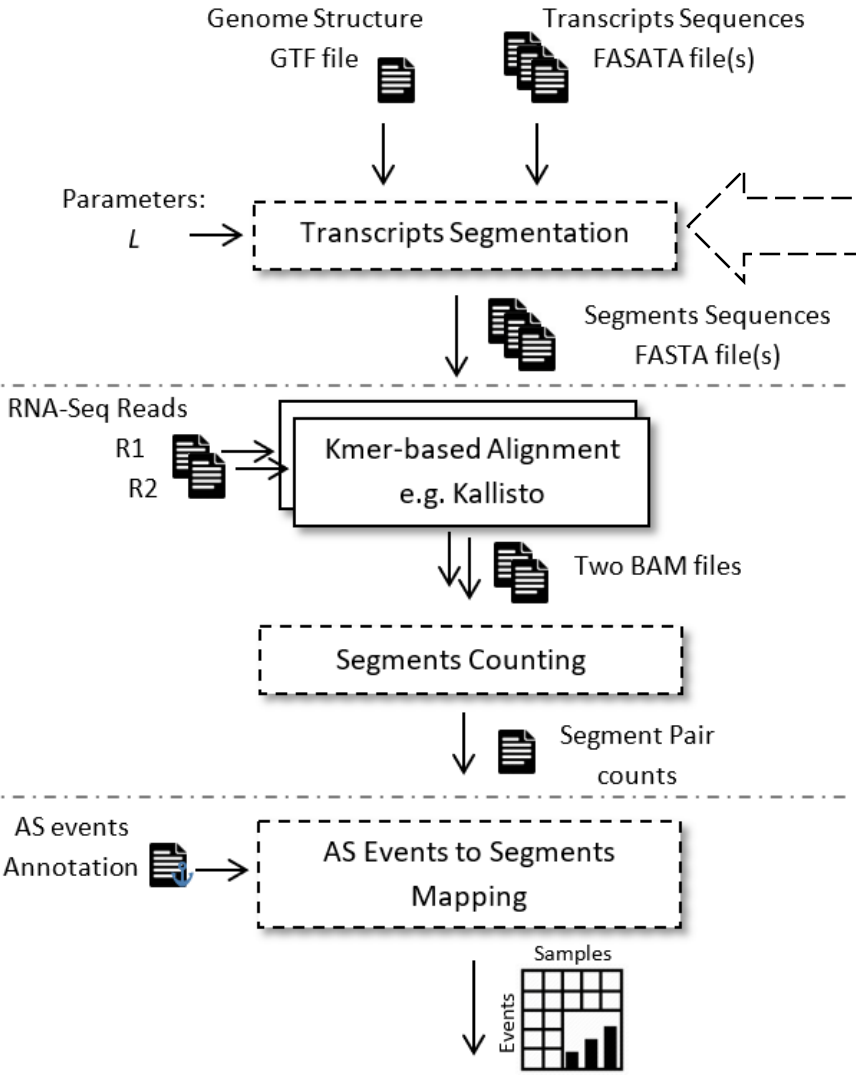
Yanagi's Approach

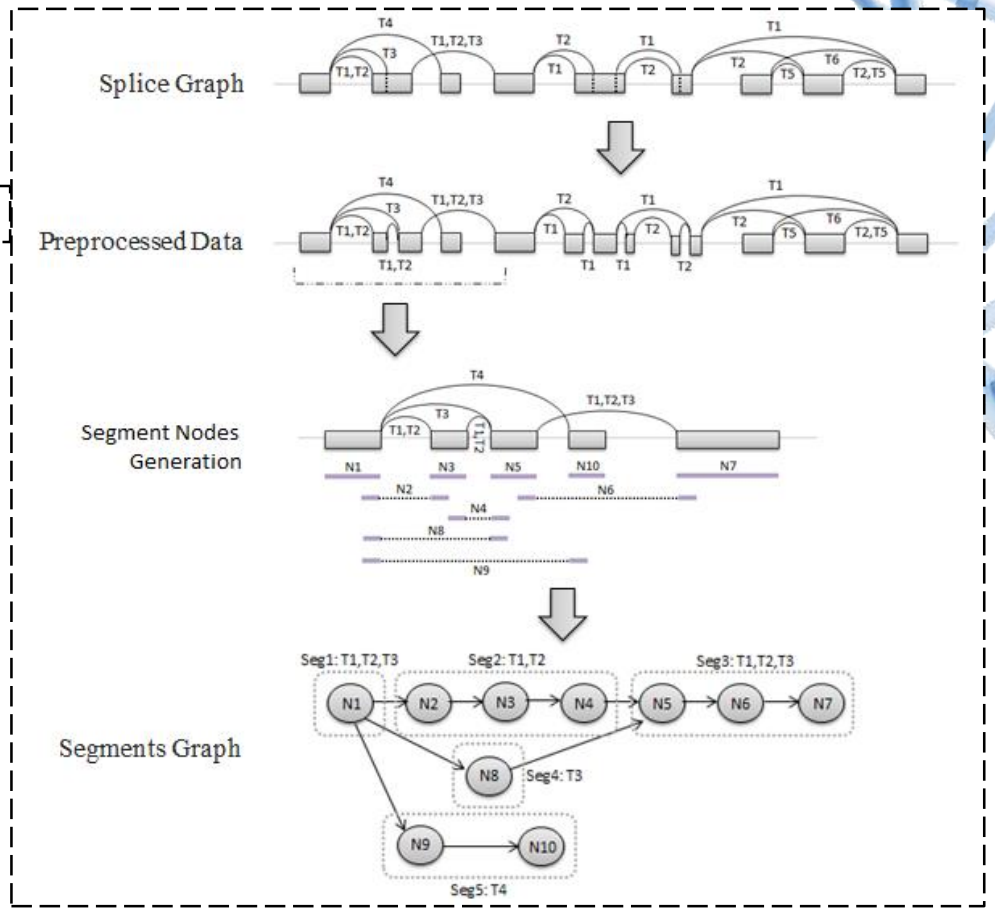
- Segmentation Algorithm



Segment	Reads	SC
S1		5
S2		4
S3		3
S4		1
S5		4
S6		1

Yanagi-based Workflow





CBCB
Center for Bioinformatics & Computational Biology

Experiments

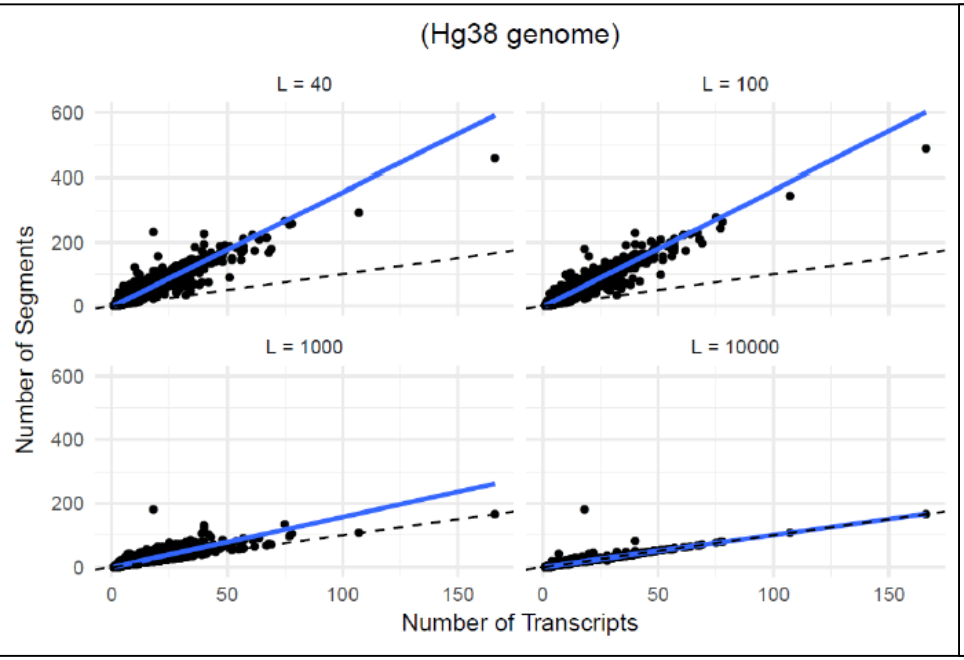
Segments Analysis

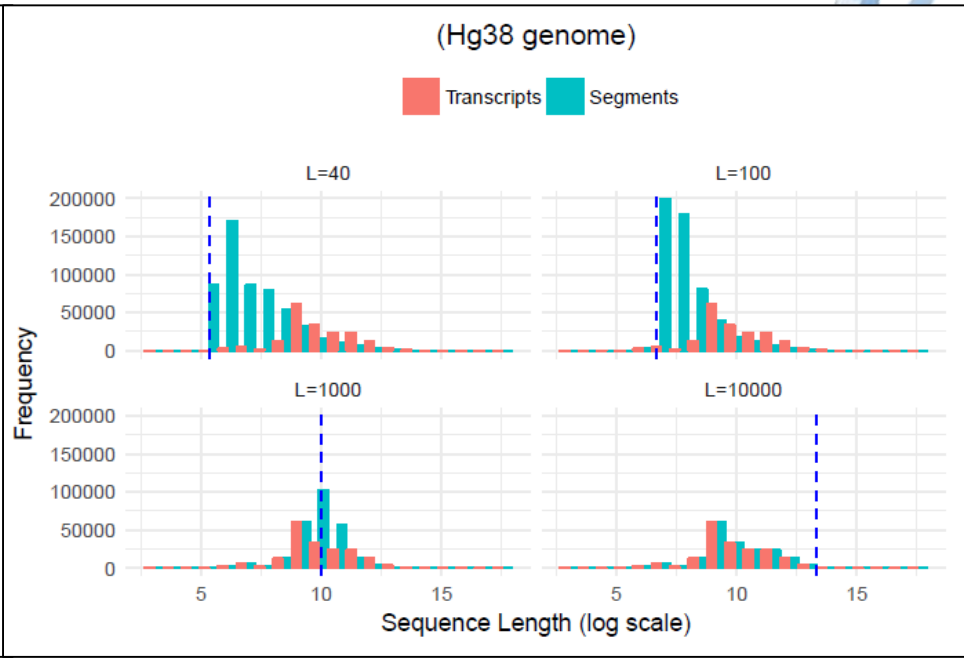
Yanagi on Github:
https://github.com/mgunady/yanagi

CBCB
Center for Bioinformatics & Computational Biology

Segments Analysis

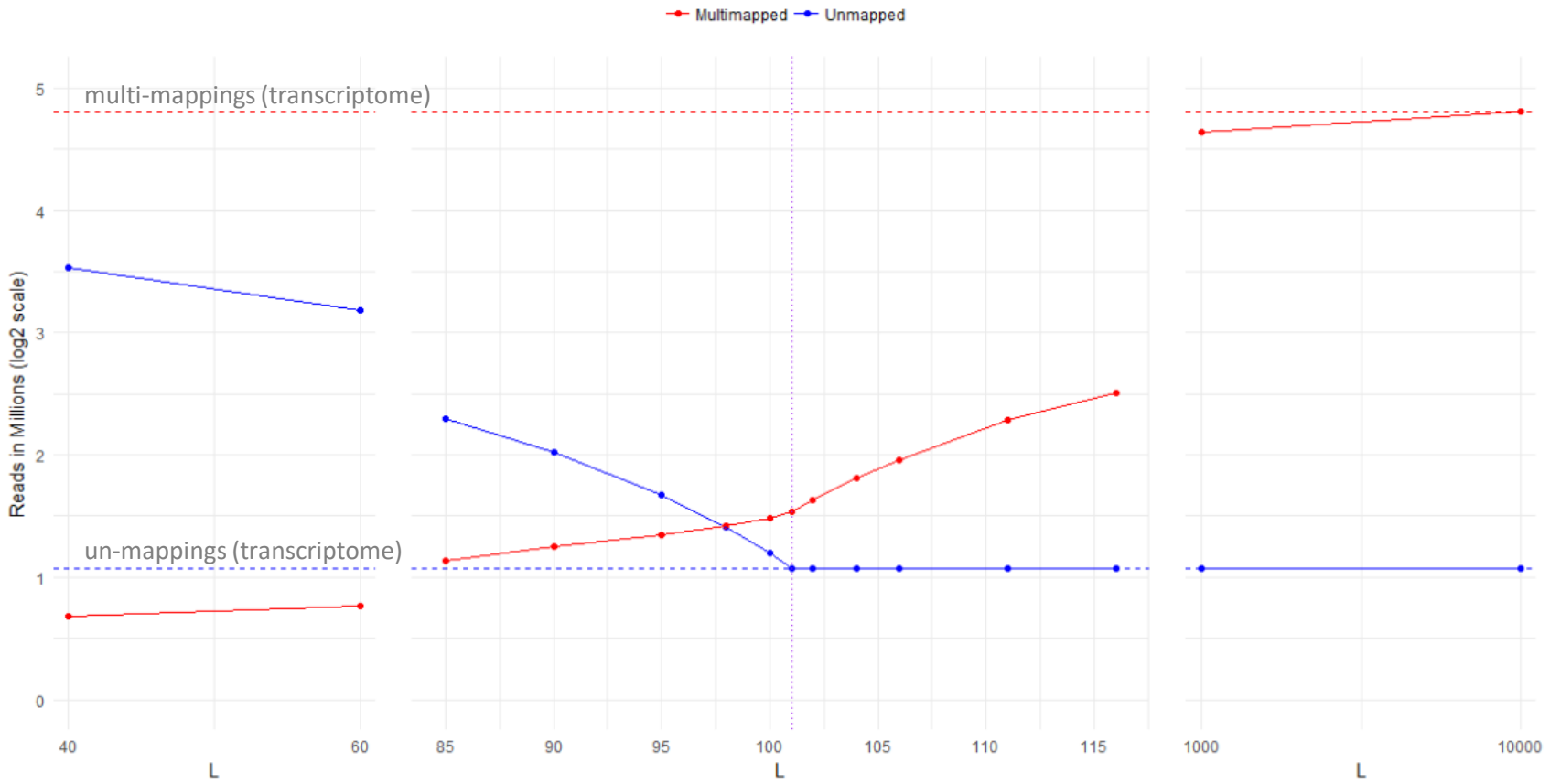
- Segments vs. Transcripts





Segments Analysis

- Impact on number of multi-mapped reads (40M reads of length 101)



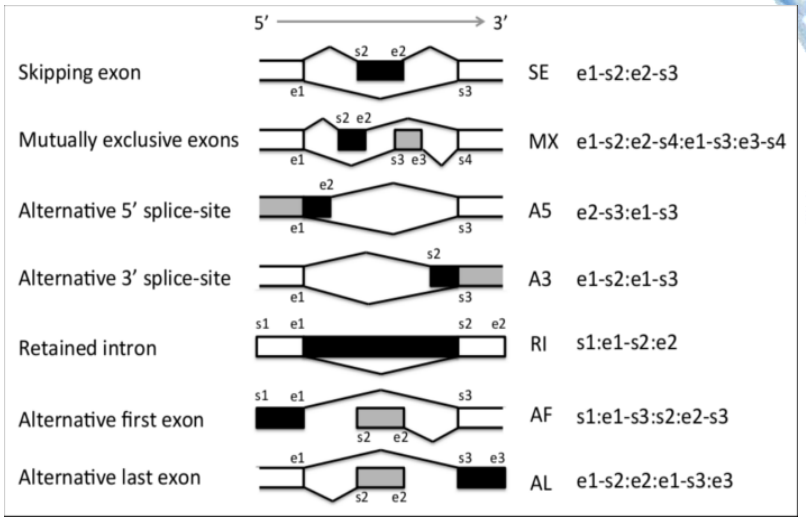
Use Case: Alt. Splicing Quantification

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

CBCB
Center for Bioinformatics & Computational Biology

Alternative Splicing Analysis

- Two directions:
  1. Counting-based Approaches:
     - E.g. rMATS, MAJIQ, DEXSeq
     - Calculates PSI values from local read counts
     - Requires mapping over the genome
     - Generally Slow



  2. Transcript-based Approaches:
     - E.g. SUPPA, DiffSplice, CuffDiff
     - Calculates PSI values based on transcripts estimated abundances
     - Can utilize fast and lightweight kmer aligners (e.g. SUPPA)
     - Can be several folds faster

     - Depends on the accuracy of estimated transcripts abundances
     - Issues handling coverage biases

SUPPA: https://github.com/comprna/SUPPA

Segment-based Alternative Splicing Analysis

- Segment-based PSI values:

For event e in sample x:

$$PSI(e,x) = \frac{\sum_{s \in S_i(e)} SC(s,x)}{\sum_{s \in S_i(e) \cup S_e(e)} SC(s,x)}$$

where $S_i(e)$ and $S_e(e)$ are inclusion and exclusion segments, respectively, and $SC(s,x)$ is the segment count

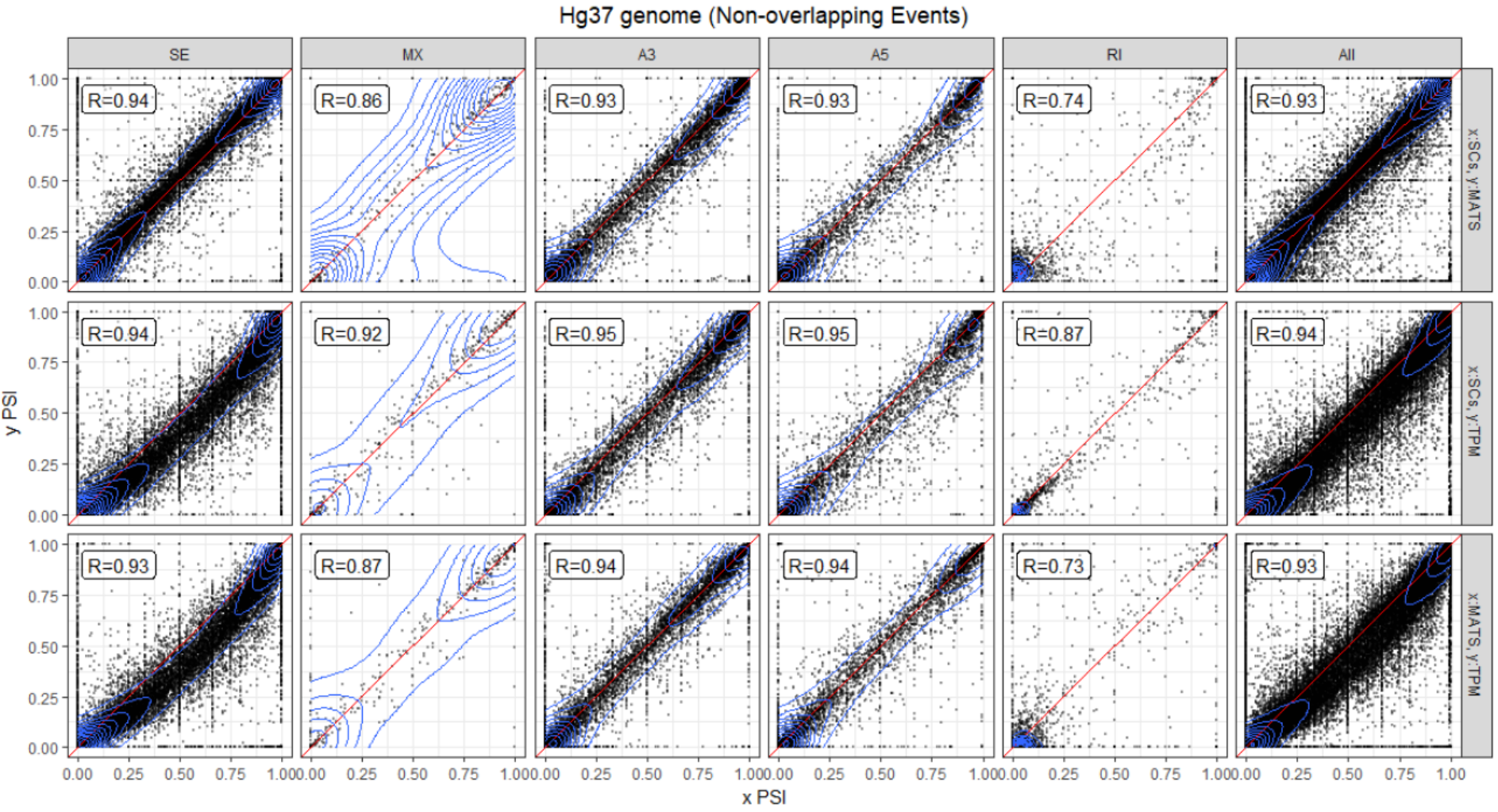
Segment-based Alternative Splicing Analysis

- Synthetic Data: (*Charlotte Soneson et al. 2016*)
  - 2 conditions, 3 replicas each.
  - Simulated reads are based on real RNA-Seq data.
  - For 1000 genes with at least two transcripts.
  - Transcription levels of the most abundant two transcripts are switched across conditions.

- Differential Analysis:
  - 5 Events Types (SE, MX, A3, A5, RI)
  - Simple Linear Model (Using Limma-Voom)
  - However, more complex model can be used

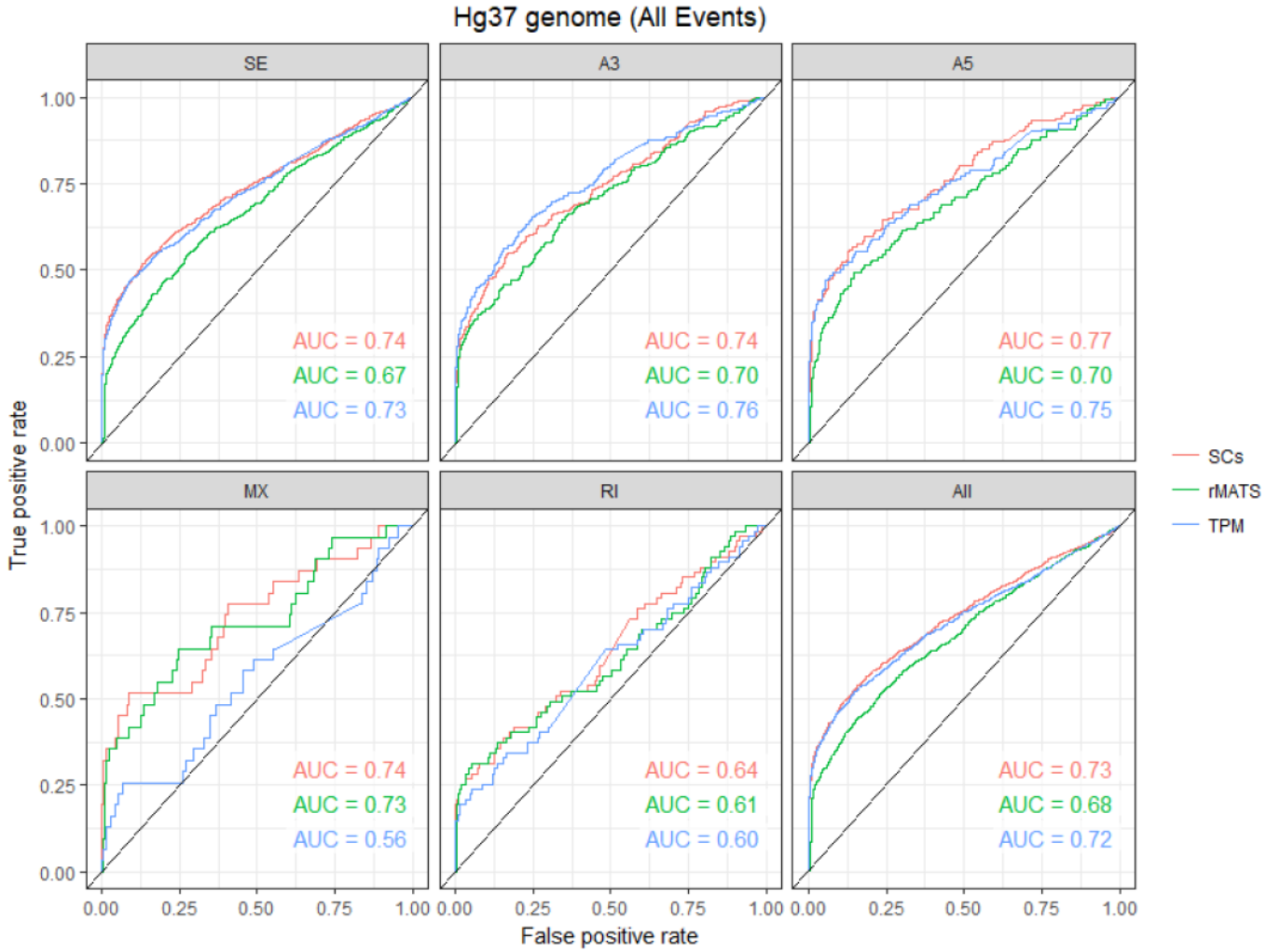
Segment-based Alternative Splicing Analysis

- Segment-based PSI values:

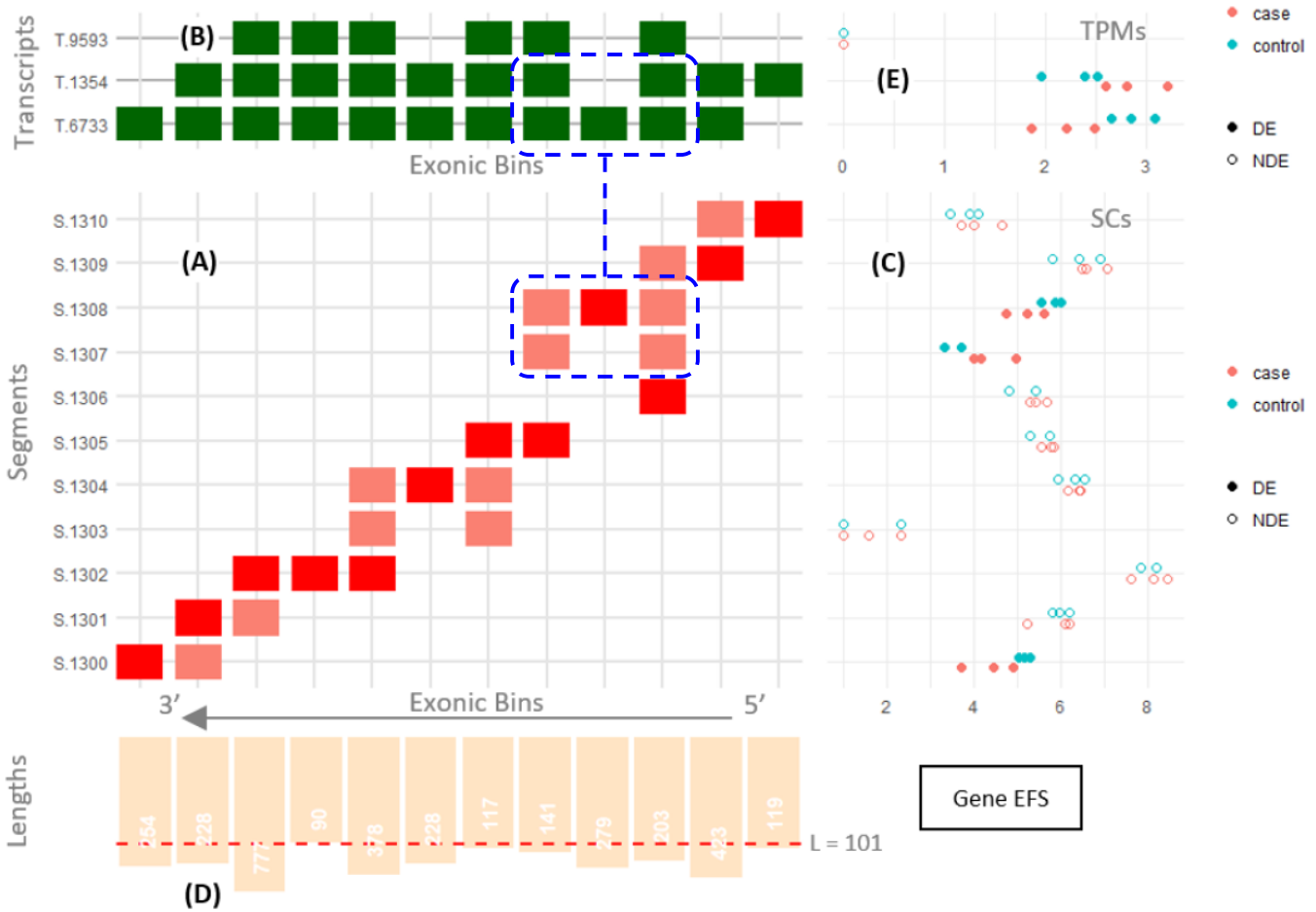


Segment-based Alternative Splicing Analysis

- Differential Analysis



Segment-based Gene Visualization



RNA-seq Summary

- Yanagi perform a transcriptome segmentation into *L*-disjoint segments.
- Enable fast and lightweight pseudo-alignment tools to provide fine-grained statistics in the resolution of local splicing.
- Segment-based AS analysis can achieve count-based approaches accuracy with the speed of transcript-based approaches.

Building population reference genome

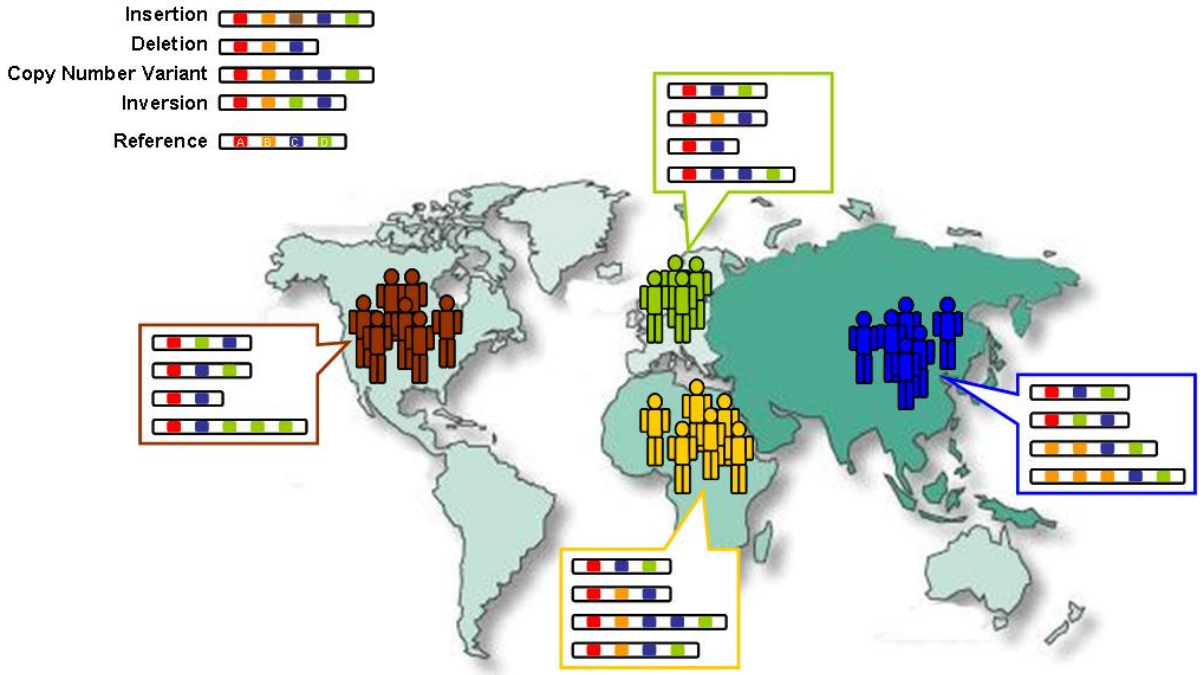
Aligning Over Genomic Variants for WGS

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

illumina

Background

- Whole Genome Population Reference
  - A challenge handling population diversity

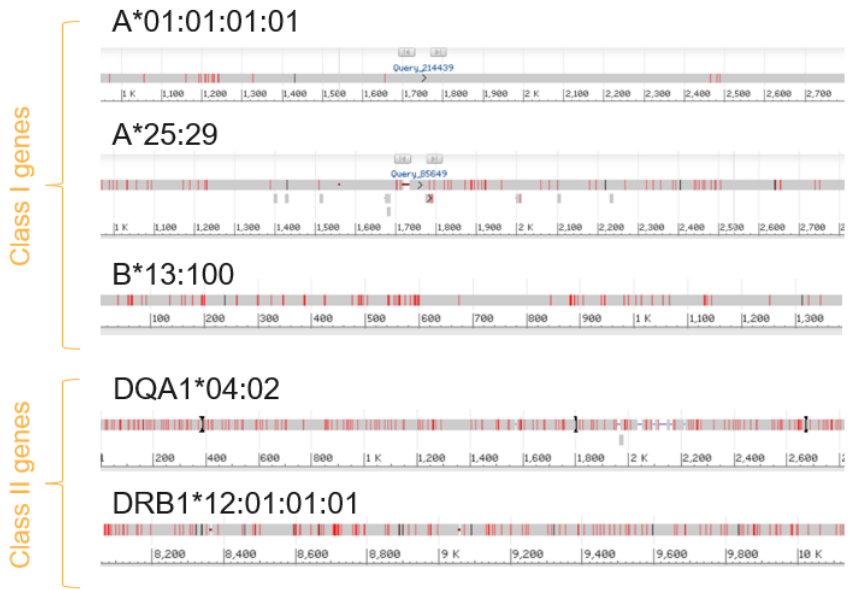


1000 Genomes Project

Background

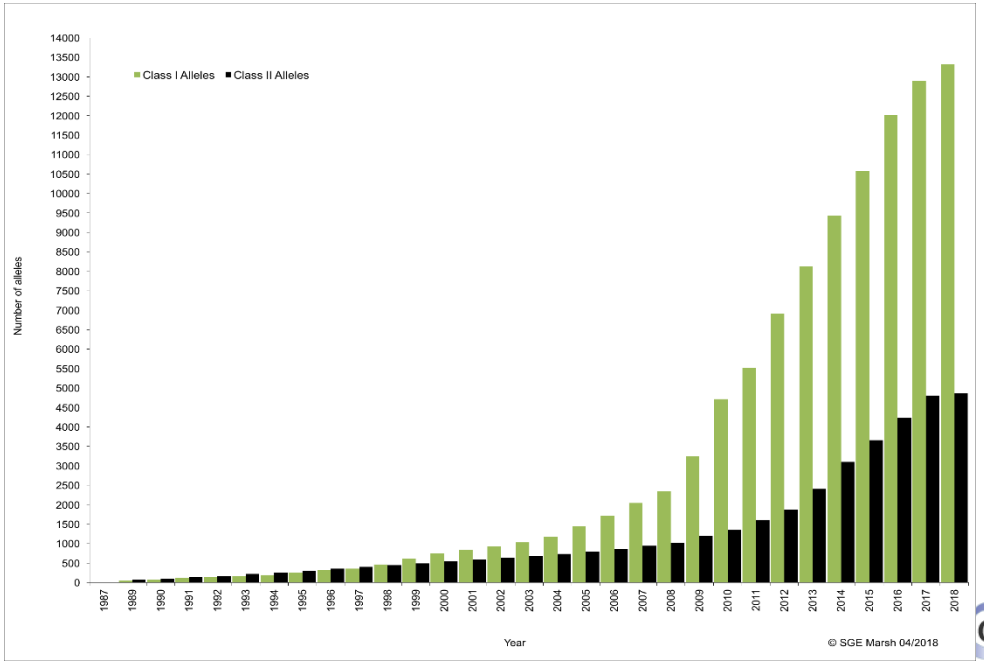
- Some genes are highly polymorphic
  - E.g. Human Leukocyte Antigen (HLA) system
  - Regulates the human immune system, so of significant medical importance

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



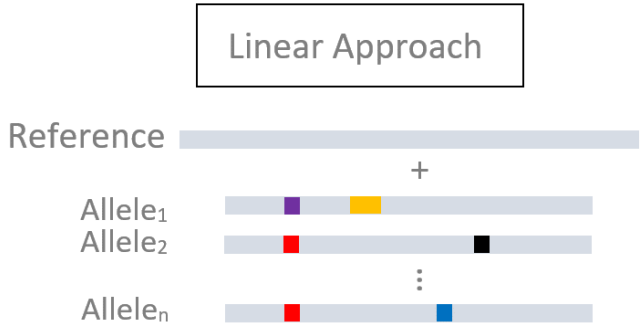
Background

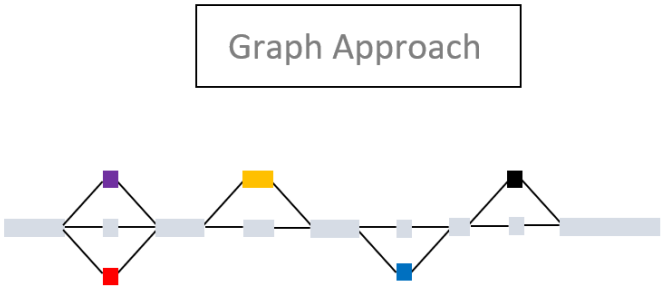
- Projects providing catalogs of known genomic variants, e.g.
  - IPD-IMGT/HLA Database
  - 1000 Genomes Project
- IPD-IMGT/HLA Database
  - Rapidly growing, provides 18,363 allele sequences for public access



Background

- Two directions to incorporate alleles into alignment





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

Segment-based Population Genome Reference

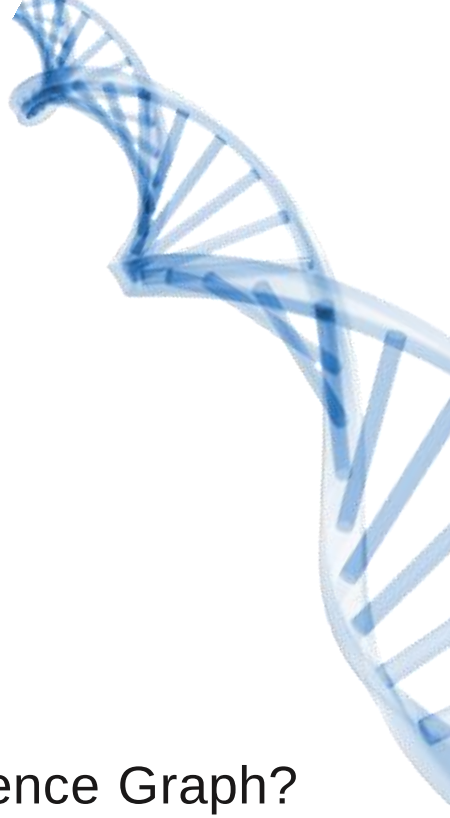
Population Graph Segmentation

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

illumina

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

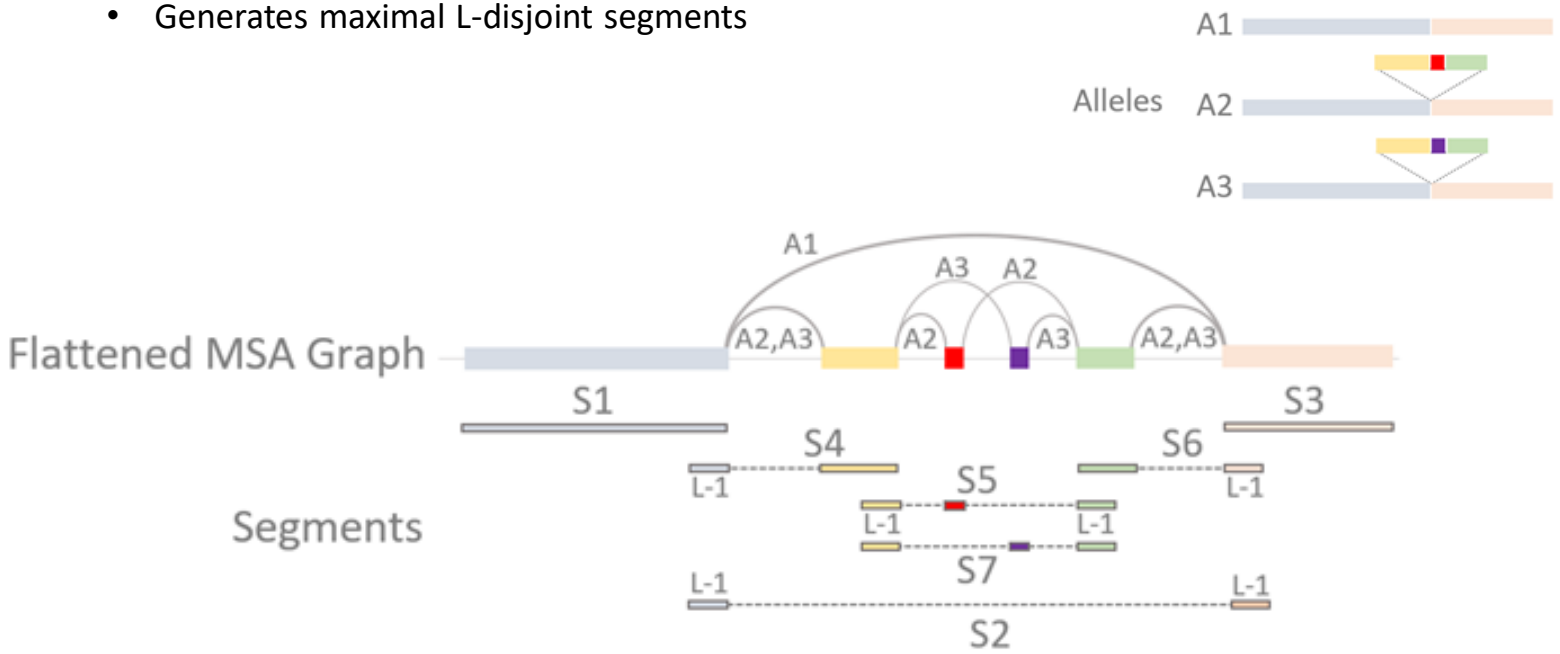
(A)
Alleles MSA

```
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

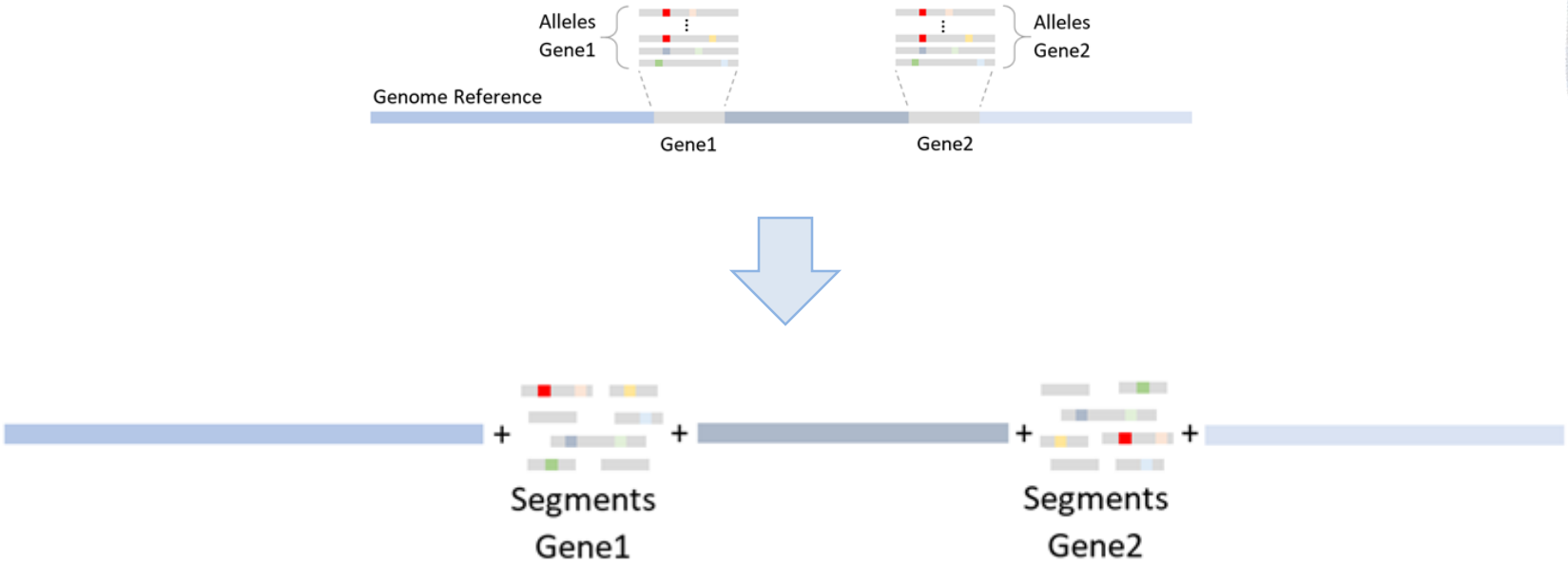
2. Linearize the graph into set of segments
   - Adapt our transcriptome segmentation approach (Yanagi)
     - Generates maximal L-disjoint segments



Our Approach

Population Graph Segmentation

3. Use gene segments as its reference for alignment



Experiments

HLA Class I and Class II genes

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

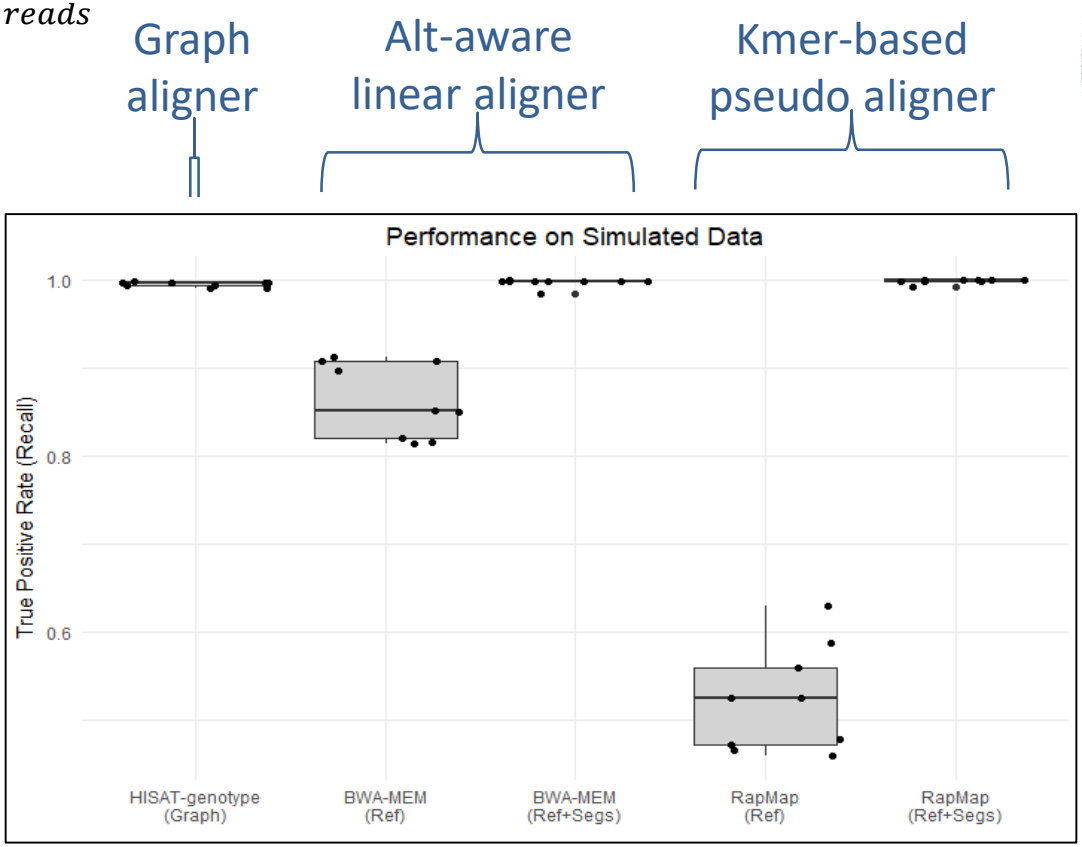
illumina

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



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

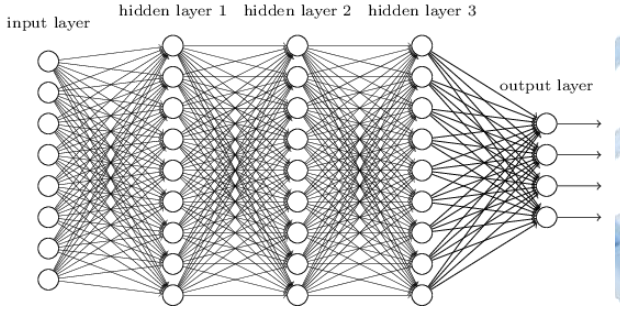
- We introduced an approach of linearizing population haplotypes graph using Yanagi's segmentation.

- 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.

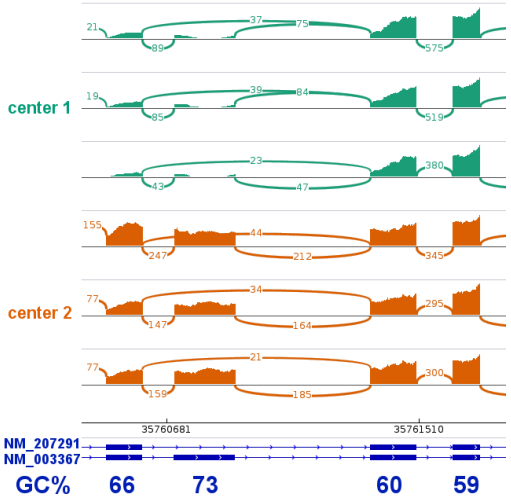
Proposed Work

1. Machine Learning models that use and predict segments expression
   - Predict tissue-specific expression using segment counts as targets in a deep network model based on sequence and chromatin measurements.
   - Use segment counts obtained from single-cell data to perform trajectory inference



2. Segment-based Transcripts Abundance Estimation
   - Estimate transcript abundances from segment counts
   - Challenges handling sources of bias
   - Use segment counts to discover unannotated junctions

3. Interactive Segment-based Gene Visualization
4. Segments Representation of Catalogs of Genomes



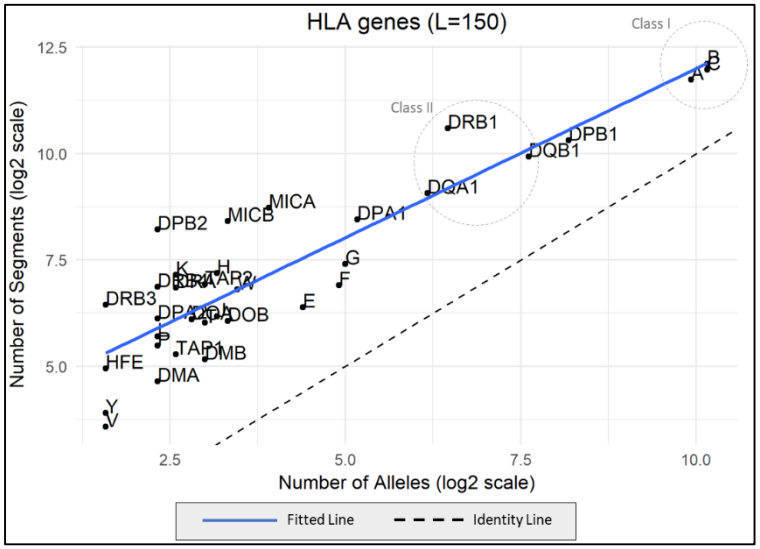
(*Michael Love et al. 2016*)

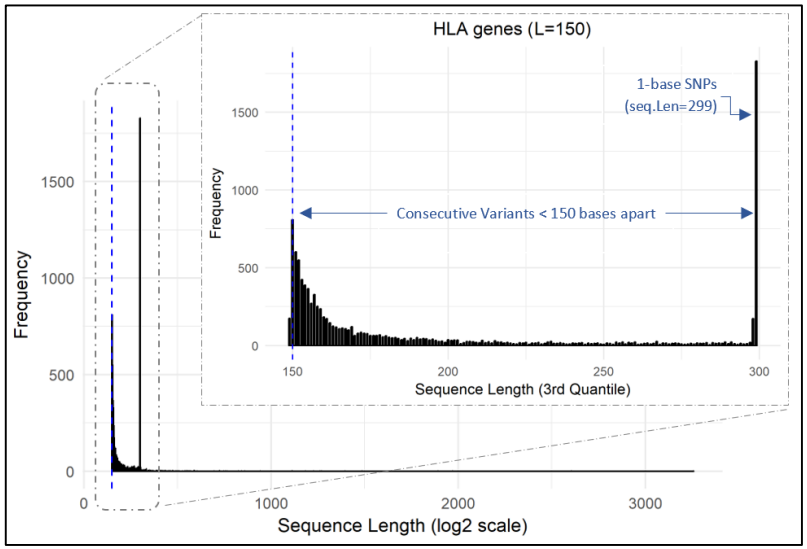
Thank you!

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

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%