Computational Genomics Lab, Kim Lab (kim-lab.org)

Daehwan Kim (PI), Chanhee Park (software engineer), Jongjun Lee (postdoc) and Chris Bennett (postdoc)

Lyda Hill Department of Bioinformatics

UT Southwestern Medical Center

Next-Generation Analysis Platform for Individual Human Genomes

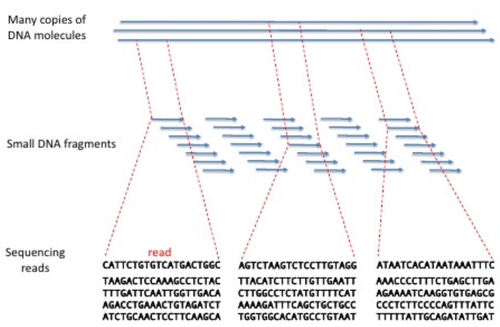
The Kim Lab pursues research at the intersection of genomics and computer science. Recent rapid advances in next-generation sequencing (NGS) technologies have proven to be very effective in performing genome-scale analyses such as variant calling, gene expression analysis, and the study of transcript structure. Computational approaches have played a key role in conducting these analyses. The ever-increasing size in sequencing data sets make it especially critical to build fast and scalable computational analysis systems.

We recently developed a novel indexing scheme using a graph approach that captures a wide representation of genomic variants and has low memory requirements. We have built a new alignment system, HISAT2, that enables fast search through the index. HISAT2 is the first and only practical method available for aligning sequencing reads to a graph at the human genome scale that can be executed on a desktop. The graph-based alignment approach enables much higher alignment sensitivity and accuracy than linear reference-based alignment approaches, especially for highly polymorphic genomic regions such as HLA genes and STRs.

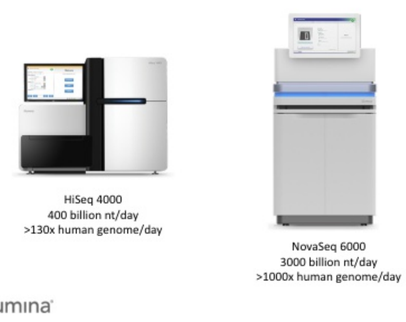
Human Reference Genome

- Assembled using the genomes of a few individuals
- Haploid reference, consisting of a single set of chromosomes
- About 3 billion base pairs long (bps)

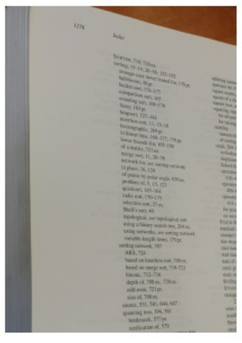
DNA Sequencing



Sequencing

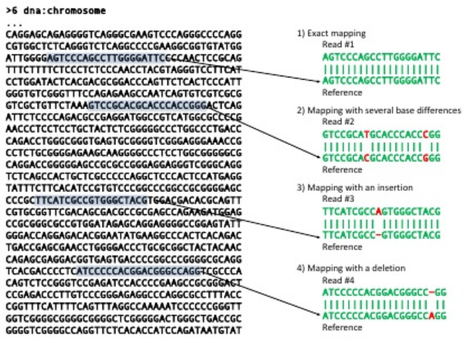


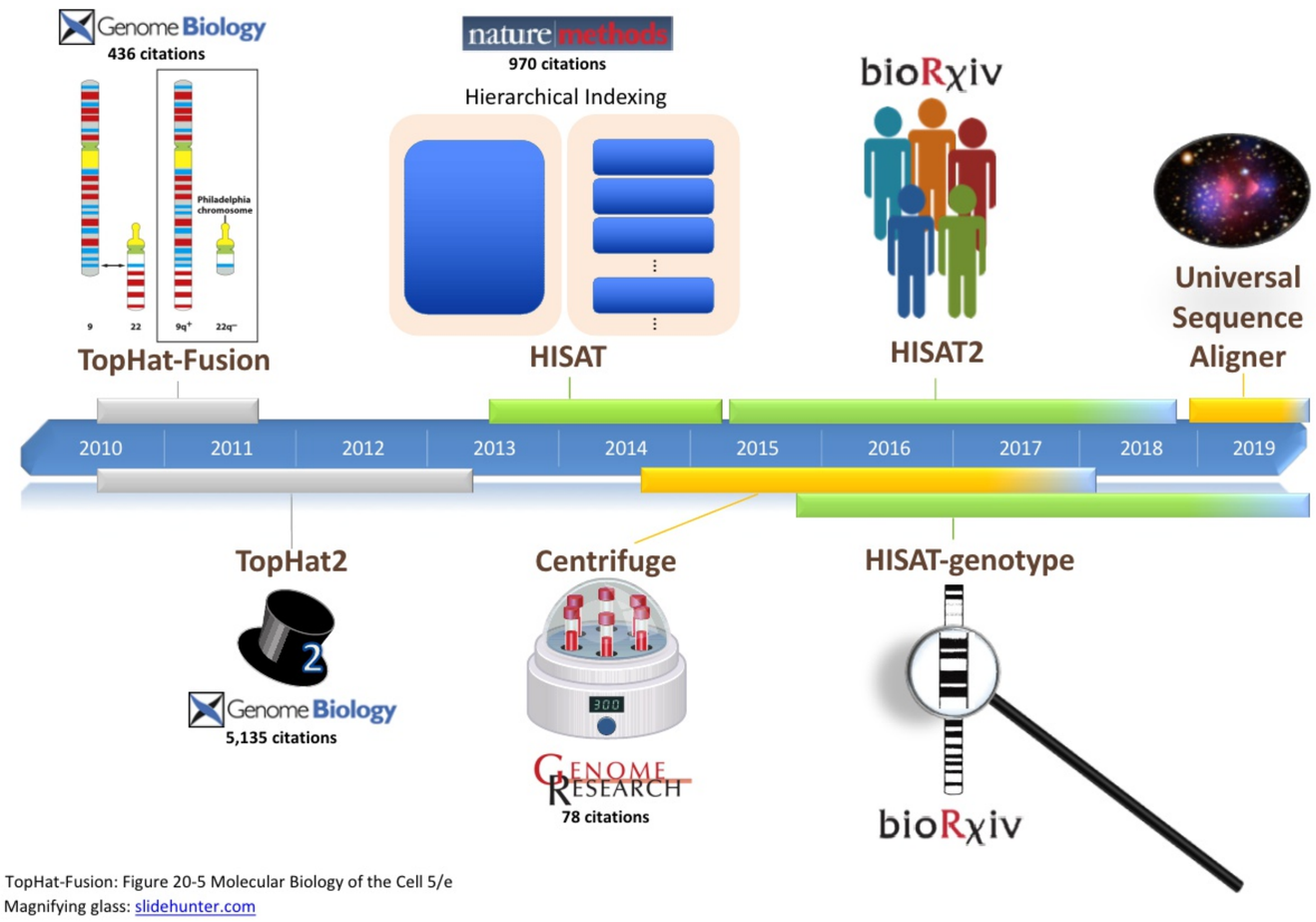
What is an index?



- K-mer table
- Suffix Array
- BWT/FM index
  - Low memory footprint
  - Relatively fast search

Mapping/Alignment Types





TopHat-Fusion: Figure 20-5 Molecular Biology of the Cell 5/e

Magnifying glass: slidehunter.com

Kim Lab

Daehwan Kim, Ph.D.

Principal Investigator

Postdoc in Genetic Medicine at Johns Hopkins University, School of Medicine

Ph.D. in Computer Science at University of Maryland, College Park

Chanhee Park

Scientific Software Engineer (January 2018 to Present)

B.S. in Computer Science and Engineering at POSTECH

Jongjun Lee, Ph.D.

Post-doctoral Researcher (November 2017 to Present)

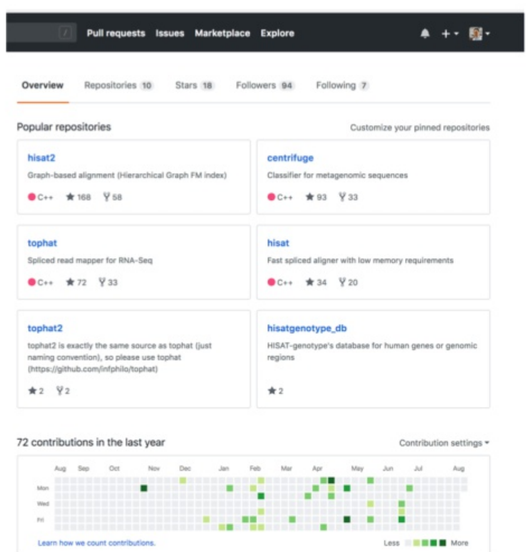
Ph.D. in Mathematical Statistics at University of Maryland, College Park

B.S. in Mathematics at KAIST

Christopher Bennett, Ph.D.

Post-doctoral Researcher (April 2018 to Present)

Ph.D. in Molecular Biology at University of Colorado, Boulder



Building off of HISAT2, we plan to develop a practical software solution that can accurately analyze an individual's genome and its >20,000 genes within a few hours on a desktop computer. The genetic information about individuals made available through this proposed work is essential to promoting personalized medicine. The software will enable researchers to more efficiently perform unbiased analyses for next-generation sequencing experiments, further improving our understanding of tumorigenesis and finding personalized treatments for cancer patients.

Acknowledgements

This work is supported by the Cancer Prevention Research Institute of Texas (CPRIT - $2,000,000) under grant RR170068 and the UT Southwestern Endowed Scholars Program ($1,200,000) to Daehwan Kim.