# HIVE: Highly Optimized Efficient Approaches of Next-gen Data Analysis

Hayley Dingerdissen[1], Alin Voskanian[2], Luis Santana-Quintero[2], Raja Mazumder[1], Vahan Simonyan[2]

[1]Department of Biochemistry and Molecular Biology, The George Washington University, Washington, DC, 20037
[2]Food and Drug Administration, Center for Biologics Evaluation and Research, Rockville, MD, 20852

Best poster award
Bio-IT Conference
2013

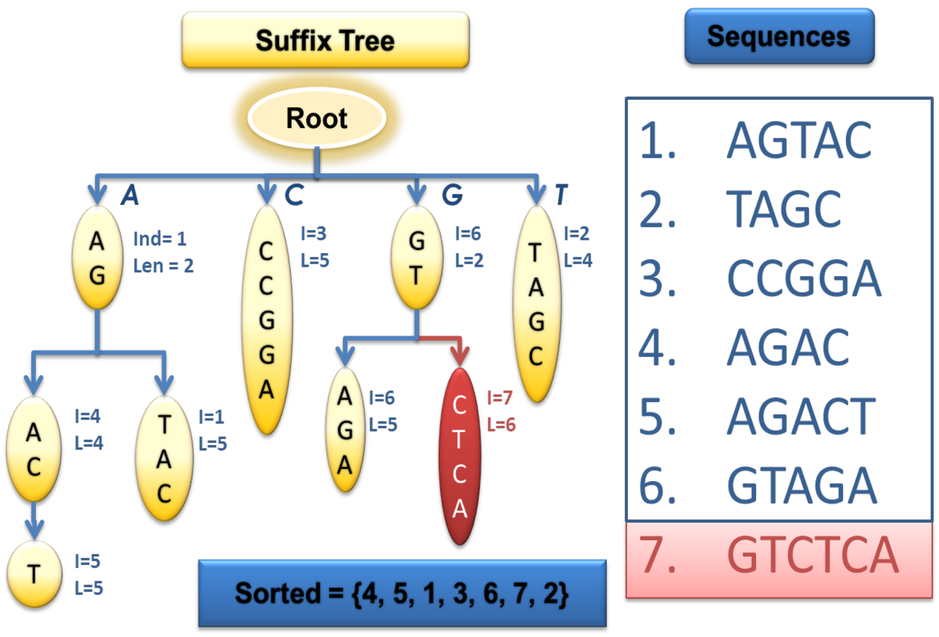
## ABSTRACT

Modern next-generation sequencing platforms generate a large number of short reads mapped to the same genomic position with low error rate, resulting in a high degree of redundancy. Innovative usage of redundancy identification and self-similarity between reads allows minimization of the memory footprint by removing repeats and by allowing a higher rate of vertical compression.

Novel prefix- tree algorithms used to discover self-similarity decrease the number of individual reads in the alignment thereby accelerating the alignment process. Modifications to the optimal alignment searching Smith-Waterman allow the diagonal to float together with the running high scoring path, thus taking full advantage of the dynamic matrices diagonalization method. A refined approach to conventional heuristic seeding involving high levels of parallelization minimizes data transfer and removes most of the I/O bottlenecks.

## NONREDUNDIFICATION ALGORITHM

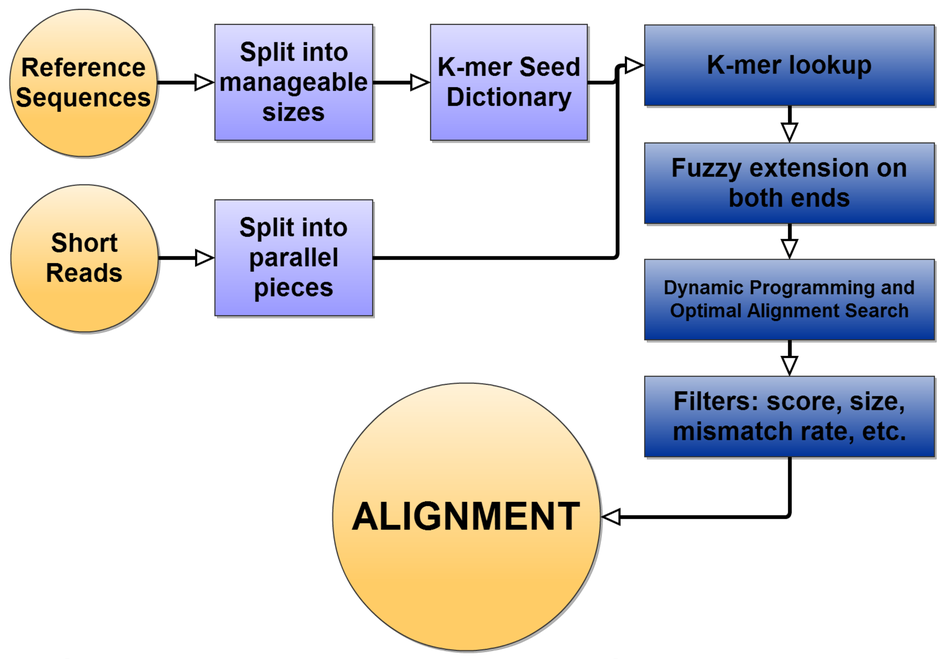
Nonredundification: smaller footprint, faster computations




### FIGURE 1. Identifying and removing redundant sequences

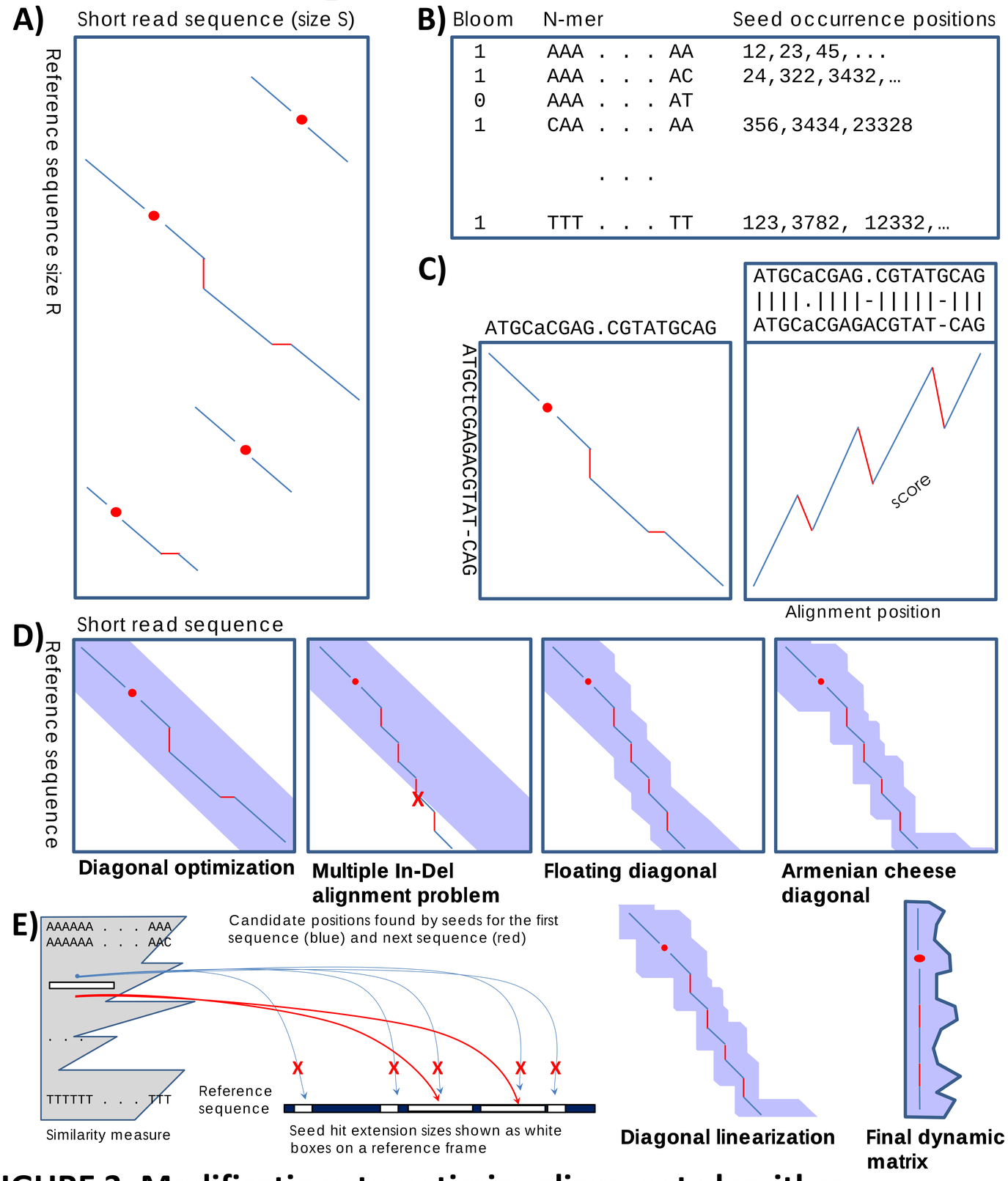
We want to build a tree representation of the following 6 sequences: AGTAC, TAGC, CCGGA, AGAC,AGACT and GTAGA. The tree contains all the information of the 6 sequences, including index and length for each node. The ROOT branches into four nodes because there are four possible first letters in the list, "A", "C", "G" and "T". The node "A" branches into two because, starting with the two-character prefix "AG", there are two possible choices for the third character, "A" and "T" for sequences 4 and 1, respectively. If the new sequence GTCTCA is added, it is first compared with the existing sequences. A difference in the third base from sequence 6 means a new node has to be created. This new node will only keep the information of the sequence: CTCA.

## ALIGNMENT ALGORITHM




### FIGURE 2. Workflow for HIVE-hexagon alignment utility

Generation of seed dictionary and maintenance of a bloom binary table of sorted read universe K-mer seeds allows optimal efficiency in memory and cache usage because of the ability to retrieve K-mers sequentially. This, combined with dynamic programming of alignment matrices and filtering procedures, drastically reduces the time required for extension and high-quality, optimal alignments.




### FIGURE 3. Modifications to optimize alignment algorithm

**A)** For any set of short reads and references, we assume the optimal alignment will be along the diagonal with respect to sequence size. **B)** HIVE-hexagon maintains a bloom lookup table where each K-mer is represented only by a single bit signifying the presence or absence of that K-mer. **C)** The fuzzy extension algorithm allows accurate definition of the alignment frame and also filters significant number of accidental K-mer hits. **D)** HIVE-hexagon implements a floating diagonal approach where the diagonal of the computation is maintained along the two sides of current highest scoring value of the matrix. **E)** Using self-similarity of short reads, we filter out seed hits with low rate of successful alignment in advance.

## ACKNOWLEDGEMENTS

Jean Thierry-Mieg, PhD, Office of the Director, NIH NLM
Carolyn A. Wilson, PhD, Associate Director for Research, Office of the Center Director, FDA CBER
Konstantin Chumakov, PhD, Associate Director for Research, Office of Vaccines Research and Review, FDA CBER
Implementation: blueHIVE technology group bluehivescience@gmail.com (Mazumder and Simonyan groups)