Pathogen Detection Using Next Generation Sequencing

L. Albayrak [1,2,3], M. Rojas [1,2], K. Khanipov [1,2,3], G. Golovko [1,2], M. Pimenova [1,2], G. J. Berry [4], M. Loeffelholz [4], I. Pavlidis [3], Y. Fofanov [1,2]

[1] Department of Pharmacology and Toxicology, University of Texas Medical Branch, Galveston, TX 77555
[2] Sealy Center for Structural Biology, University of Texas Medical Branch, Galveston, TX 77555
[3] Department of Computer Science, University of Houston, Houston, TX 77004
[4] Department of Pathology, University of Texas Medical Branch, Galveston, TX 77555

INTRODUCTION

Progress in Next Generation Sequencing technology in the recent past has resulted in dramatic cost reduction and the improvement of quality throughput allowing it to be used by universities and even individual labs rather than being exclusive to large sequencing centers. The latest "benchtop" sequencers have progressed into extremely cost-effective tools that can be used for pathogen detection. In order to transform NGS technology from a powerful research tool to instruments that are routinely used for pathogen detection with applications such as clinical diagnostics, several challenges, however, must be resolved, including:

1)The storage, manipulation, and transfer of the large amounts of data produced by High-Throughput Sequencing (HTS) instruments;

2)The costly computational resources and time requirements of the tools used in NGS analysis and absence of standard data analysis algorithms and reference databases;

3)The large and complicated output of available NGS data analysis pipelines which usually requires Ph.D. level scientists to interpret the analysis results.

Our goal is to create data structures and algorithms that enable analysis of NGS samples for pathogen detection using affordable computational infrastructure.

1. SEQUENCING DATA COMPRESSION

Sequencing instruments produce large volumes of data and the cost of sequencing has dropped dramatically over the years. Millions of sequencing reads of typically 36 to 150 base pairs are generated in every run.

Platform	Number of reads / run	Cost / 100 cycle run	Costs / 1M reads
Illumina HiSeq 2500	~3,000,000,000	$17,000	$5.6
Illumina HiSeq 1500	~1,360,000,000	$11,200	$8.23
Illumina MiSeq	~23,500,000	$800	$34.04

CoCo, is a sequencing data compression tool developed in-house. It can significantly reduce the space and memory footprint of sequencing files, generating non-redundant binary representations of the sequencing reads. CoCo takes advantage of the quality score distribution and the presence of identical sequences and common prefixes present in sequences. ASCB (Array Subsequences Compressed Binary) is CoCo's proprietary compressed binary format for storing sequencing data.

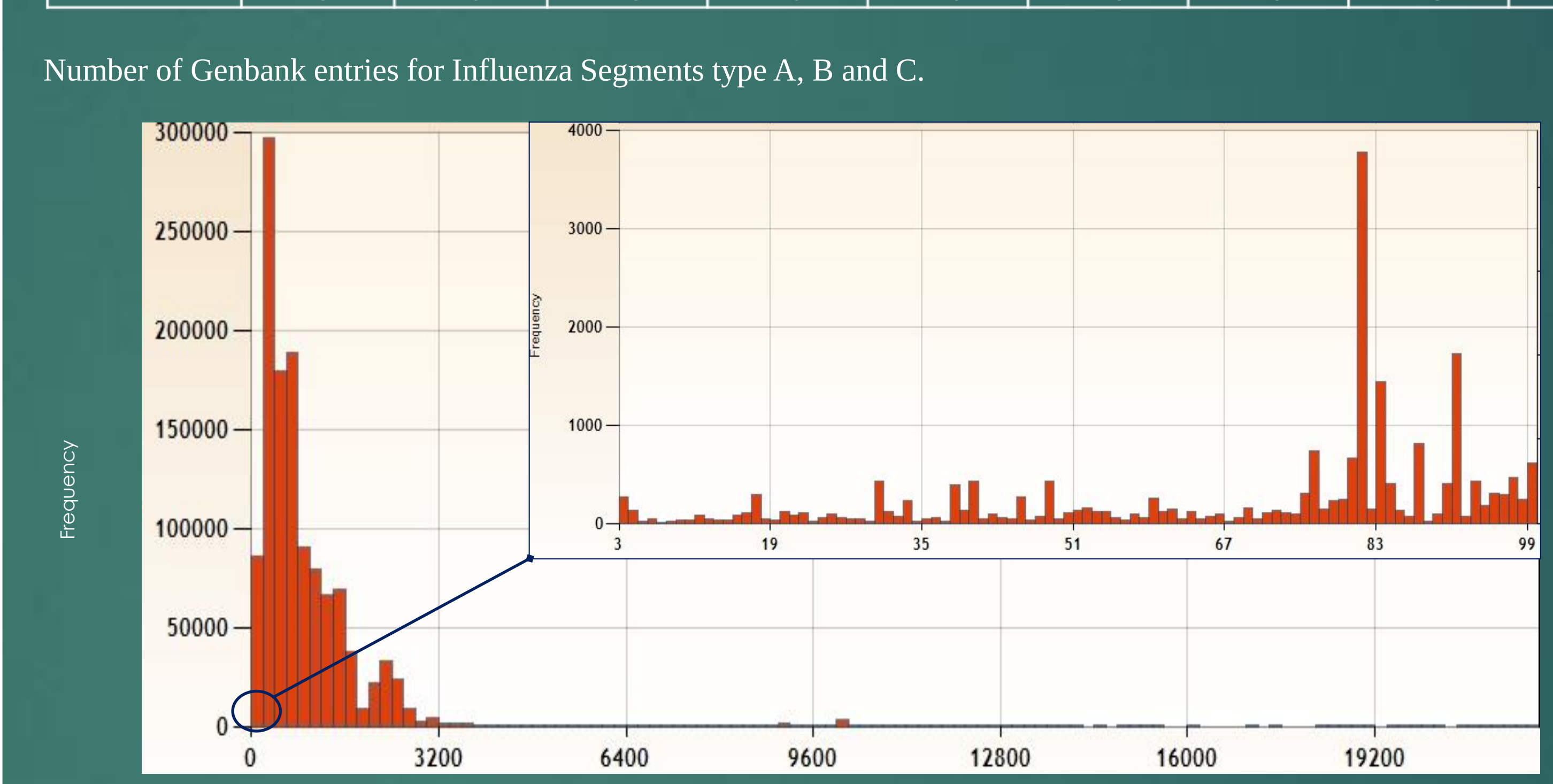
Data Set	Bacteria Genome (mb)	Human Transcriptome (mb)	Human Genome (mb)	Groundwater Metagenome (mb)
FASTQ	2,819	14,520	105,142	20,902
FASTA	1,246	5,814	51,153	9,264
All Reads				
ASCB	271	59	11,870	2,134
Low Quality Reads Excluded				
ASCB	191	30	3,635	1,130

2A. GENE CLUSTER DATABASE

Reference databases such as GenBank are large and redundant. Some entries are not suited to be used for pathogen detection (such as very short sequences).

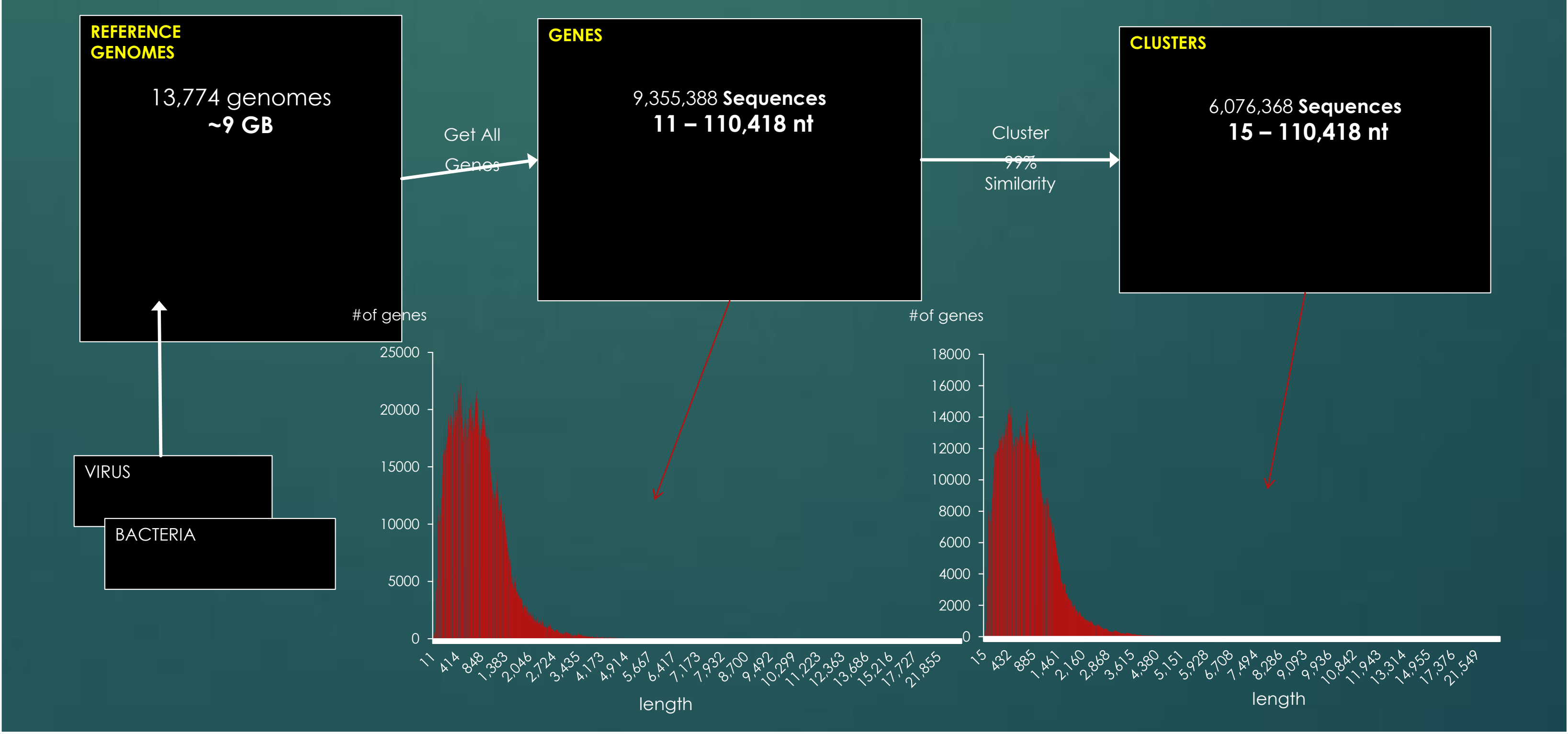
Segments	1	2	3	4	5	6	7	8	Totals
Type A	15,478	15,930	16,337	24,154	16,125	22,539	18,304	17,761	**146,628**
Type B	596	597	605	4,405	618	2,154	623	937	**10,535**
Type C	90	88	88	167	96	118	123	-	**770**
Totals	**16,164**	**16,615**	**17,030**	**28,726**	**16,839**	**24,811**	**19,050**	**18,698**	**157,933**

Number of Genbank entries for Influenza Segments type A, B and C.

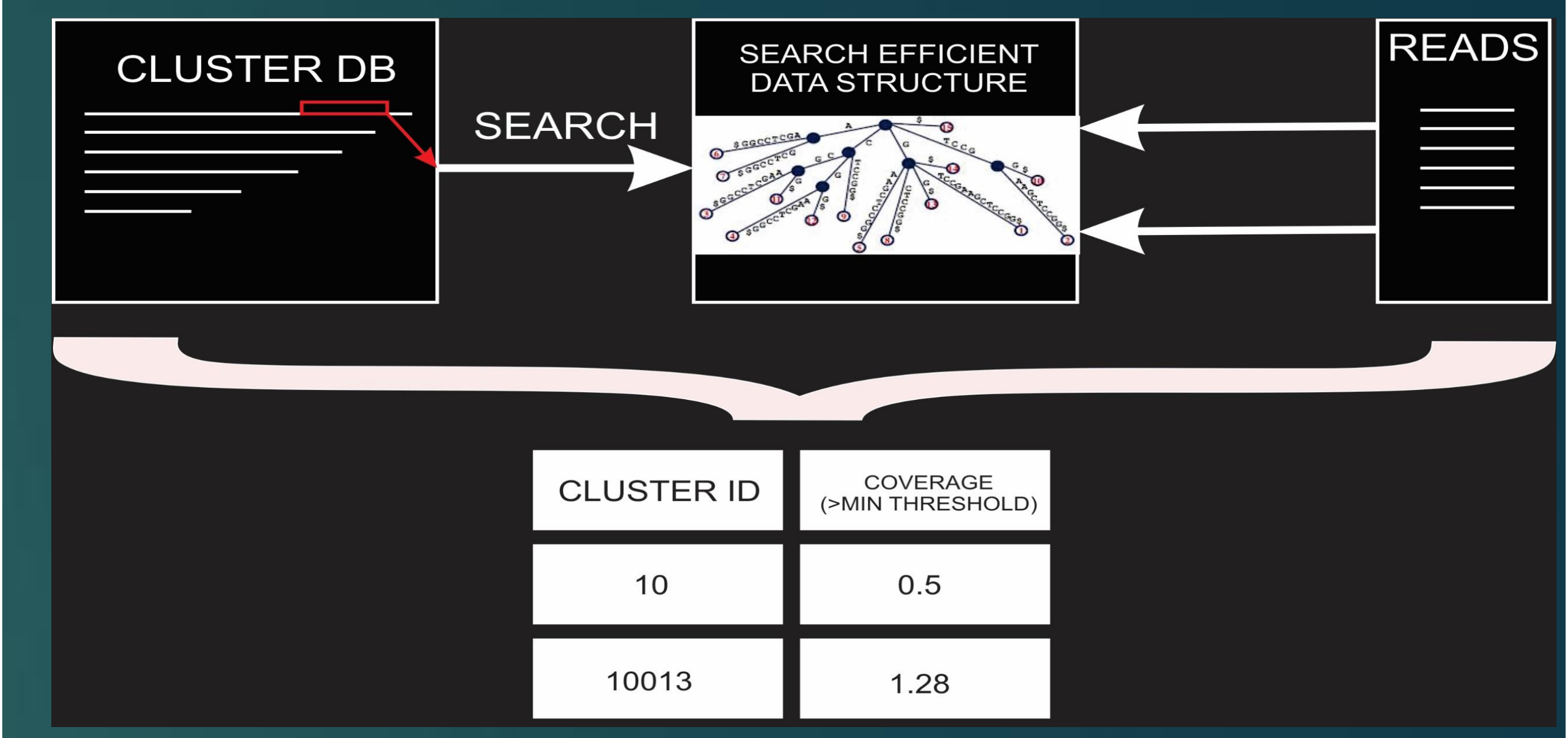


Length distribution of viral sequences in NCBI. A Large number of sequences that may not be useful for pathogen detection due to their short length can be observed. In this example we can see that there are more than 100 copies of 3 base long (3-mer) sequences catalogued in NCBI. The maximum number of unique 3mers is 4^3=64

Database search of reference genomes is necessary in order to associate sequencing reads to organisms. On average, a bacterial pathogen genome is 4-5 million nucleotides long. If a pathogen is present in a sample, a subset of its genes are expected to be sequenced by the instrument. Using only gene sequences instead of entire genomes greatly reduces the size of the required reference nucleotide sequence database. Similar genes across organisms exist and the redundancy among gene sequences can be greatly reduced by clustering them based on nucleotide similarity and representing groups of genes with a single representative sequence. Similarity between gene sequences can be determined using a global alignment algorithm such as Longest Common Subsequence (LCS). However, global alignment is a very computationally demanding process and efficient pairwise alignment of ~9.3m gene sequences is a challenging task. Our approach utilizes a sorting strategy along with a cascading Markov Chain filter to avoid unnecessary pairwise alignments and a modified LCS algorithm that is space efficient and allows early termination of the alignment if it can be determined that a desired similarity threshold cannot be met. Using a 99% sequence similarity threshold, we were able to create a non-redundant database of representative genes consisting of ~6m sequences out of ~9.3m using a centroid based progressive clustering algorithm (greedy). Using clustered genes as a reference database allows us to create gene profiles of samples and focus analysis on these profiles.



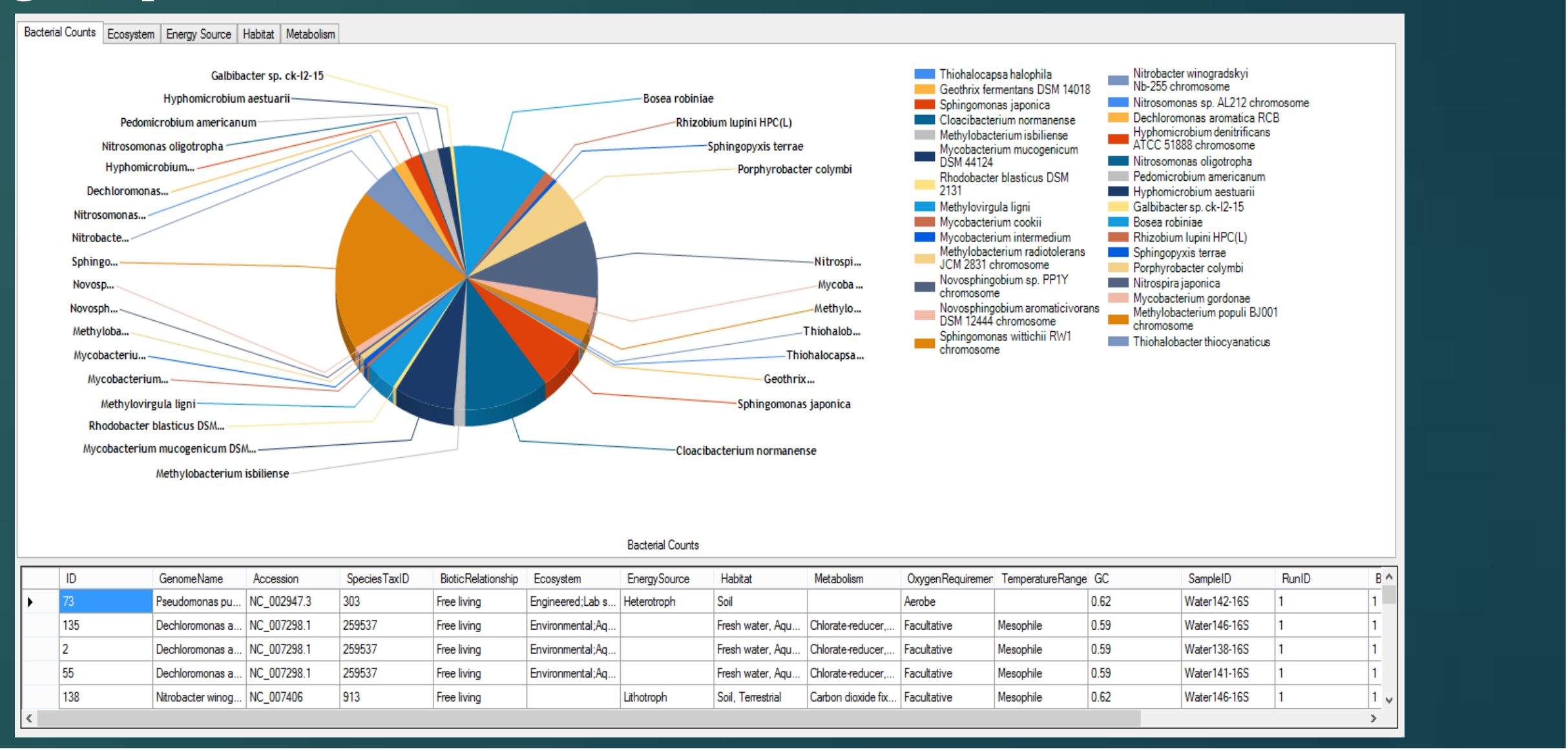
2B. MAPPING AND SEARCH STRATEGY



The majority of algorithms such as BWA, Bowtie and SHRiMP are based upon a **read-by-read search** against reference databases and produces results in a reads-centric format such as BAM/SAM. The time complexity of read-by-read search is proportional to the number of reads in the data set and the algorithms must rely on raw computing power and heuristics in order to get results in a reasonable time. Our approach performs an exhaustive search of reference database sequences against a database of sequencing reads (**reference-by-reference search**) using a Suffix-Tree like data structure. The time complexity is proportional to the number of reference sequences (or genes/clusters) in the database as opposed to being proportional to the number of sequencing reads. The output of such an algorithm is much smaller than a read-centric approach and it is tied to the reference sequences instead of sequencing reads.

3. REPORTING

It is important that user-oriented NGS data analysis strategies and reporting protocols be designed and implemented. Practical knowledge (essential to understanding the needs and operations of diagnostic labs in real clinical settings) that physicians possess can be used in combination with the sample analysis output of reference centric gene/cluster profiles and genome coverage maps.



CONCLUSION

The described combination of clustering, curation, compression and exhaustive reference-by-reference mapping strategy enables the creation of fast and accurate diagnostics tools for pathogen detection that can be used to evaluate collections of HTS samples with relatively low computational infrastructure requirements. Reducing time and space complexity of NGS analysis paves the way to bringing NGS based pathogen identification methods to clinical and field diagnostic settings suitable to run on benchtop computers. Reference centric output makes analysis and results easy to interpret.

ACKNOWLEDGEMENTS

This work was partially supported internally by the Institute for Human Infections and Immunity (IHII); 9/1/2014-8/31/2016