HIVE: High-performance Integrated Virtual Environment Provides Solutions for Next-Generation Sequencing Data Storage and Analysis

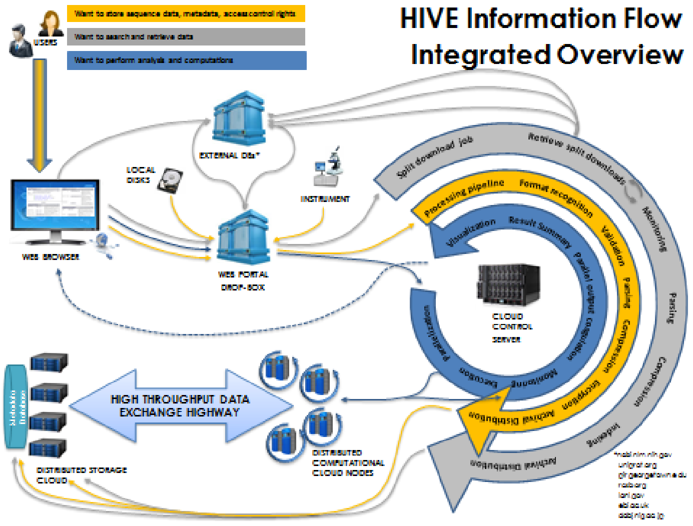
Highlights:

- Integrated storage and computational nodes minimizes data transfer and removes I/O bottlenecks
- HIVE-hexagon alignment algorithm delivers outputs comparable to industry standards while improving speed and sensitivity
- HIVE-honeycomb data model allows determination of data access privileges in a finely granular manner
- System easily extended and scaled-up to support future data requirements
- Operating Models allow HIVE to be installed in a variety of environments

Issues associated with Next-Generation Sequencing Data

The decreasing cost, increasing speed and improving accuracy of next-generation sequencing (NGS) over the last decade has led to significant advances in genomic technologies. Due to the petabyte (and growing) scale of resultant reads and attendant data, the challenges associated with storing, accessing and computing on NGS and other biomedical data create bottlenecks for downstream analysis.

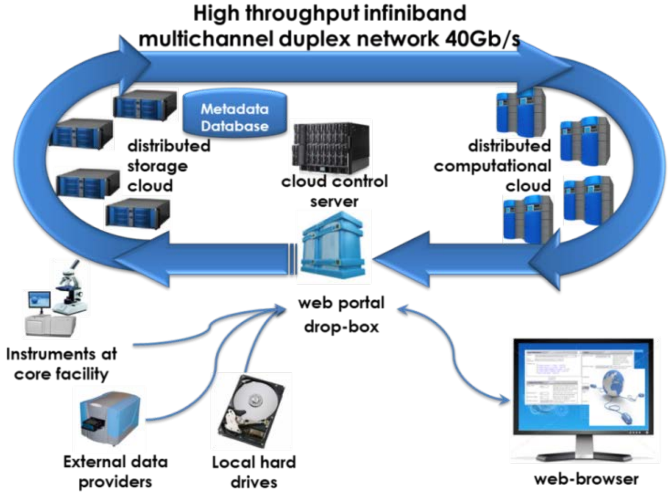
The High-performance Integrated Virtual Environment (HIVE) was designed to address the efficiency of such big-data transfer and computational complexity. HIVE was architected to be a multicomponent cloud infrastructure with seamlessly linked distributed storage library and computational powerhouse. This environment provides web access for authorized users to deposit, retrieve, store, annotate and compute on NGS data.



HIVE Information Flow
Integrated Overview

Infrastructure

HIVE is a massively parallel computing environment that is both robust and flexible due to the colocation of storage and the metadata database on the same network. The distributed storage layer of software and drivers is the key component for file and archive management and the backbone for the deposition pipeline. The data deposition backend allows automatic uploads and downloads of external datasets to HIVE data repositories. The metadata database can be used to maintain specific information about short reads obtained from NGS experiments.

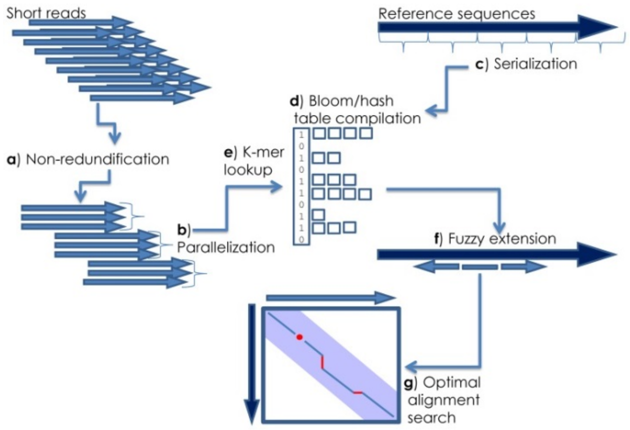

High throughput infiniband multichannel duplex network 40Gb/s
Metadata Database
distributed storage cloud
cloud control server
distributed computational cloud
web portal drop-box
Instruments at core facility
External data providers
Local hard drives
web-browser

The honeycomb engine developed for HIVE can represent objects as either flat files or networked files, maintaining the interrelatedness through a novel flat schema. Unlike other object oriented databases, HIVE implements unified API interfaces to search, view, and manipulate data of all types. The honeycomb engine also facilitates a highly secure hierarchical access control and permission system, allowing determination of data access privileges in a finely granular manner without flooding the security subsystem with a multiplicity of rules. This model, designed for sensitive data, provides comprehensive control and auditing functionality in accordance with HIVE's designation as a FISMA moderate system.

Algorithms

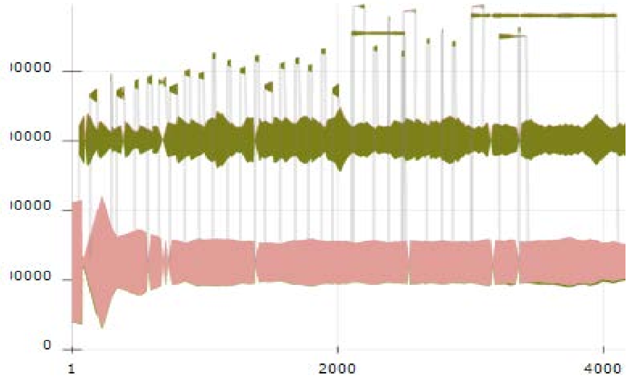
Specific algorithms developed which employ these properties are:

- **HIVE-pentagon**- Modern NGS platforms typically generate numerous short reads that map to the same genomic position, resulting in a high degree of redundancy. HIVE-pentagon is a highly parallelized, innovative algorithm designed to remove redundancy and self-similarity among reads to decrease memory footprints and accelerate processing.
- **HIVE-hexagon**- HIVE-hexagon is a massively parallel alignment algorithm specifically designed to work with NGS alignments to reference genomes. This HIVE-native algorithm outperforms alternative aligners in speed and sensitivity due to a linearized diagonal adaptation of Smith-Waterman optimization and other modifications to conventional heuristic seeding algorithms.

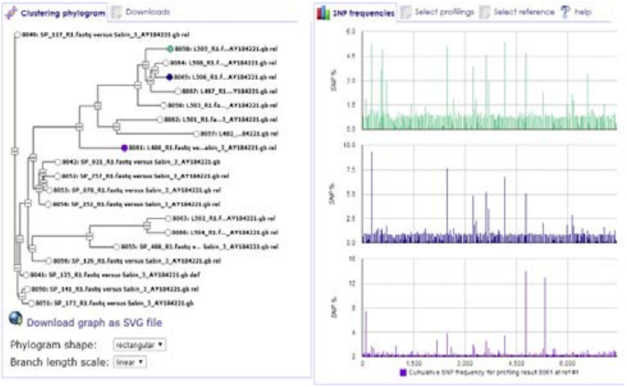


- **HIVE-heptagon**- HIVE-heptagon is a sequence profiling tool that performs base-calling and SNP-calling and provides quality and noise assessment profiles. Post-alignment quality control (QC) procedures are reported based on a positional base-frequency entropic information content paradigm to validate alignment results and distinguish computational artifacts from real variability.
- **HIVE-octagon**- Profile outputs from HIVE-heptagon can be used as inputs to conduct comparative analysis. HIVE-octagon generates and compares sequences of SNP frequencies with respect to reference positions and produces a phylogram displaying the computed hierarchical clustering.
- **Reference Recombination Tool**- Study of viral, microbial and other complex populations of environmental samples is enabled by this tool, which facilitates discovery of genetic recombination and allows resolution of populations into subgroupings.

- **Clonal Population Discovery Tool (Hexahedron)**- Bifurcations of read mappings along the specified reference are displayed following reference-assisted de novo assembly. A Sankey diagram provides a graphical visualization of all possible clones.



- **CensuScope**- CensuScope is a robust subsampling-based algorithm that detects the taxonomic composition of metagenomic datasets.
- **PhyloSNP**- Inputs of variation data are analyzed and output in phylogenetic trees.


Clustering phylogram
Downloads
8040: SP_127_R1.fastq versus Sabin_3_AY184221.gb rel
8058: L505_R1.f...AY184221.gb rel
8064: L508_R1.f..._AY184221.gb rel
8065: L506_R1.f...AY184221.gb rel
8067: L467_R1...Y184221.gb rel
8059: L503_R1.fa..._AY184221.gb rel
8062: L501_R1.fa...3_AY184221.gb rel
8057: L482_...84221.gb rel
8061: L488_R1.fastq ve...abin_3_AY184221.gb rel
8042: SP_931_R1.fastq versus Sabin_3_AY184221.gb
8052: SP_757_R1.fastq versus Sabin_3_AY184221.gb rel
8053: SP_070_R1.fastq versus Sabin_3_AY184221.gb rel
8054: SP_252_R1.fastq versus Sabin_3_AY184221.gb rel
8063: L502_R1.f...AY184221.gb rel
8066: L504_R1.f..._AY184221.gb rel
8055: SP_488_R1.fastq v... Sabin_3_AY184221.gb rel
8056: SP_129_R1.fastq versus Sabin_3_AY184221.gb rel
8041: SP_135_R1.fastq versus Sabin_3_AY184221.gb def
8050: SP_141_R1.fastq versus Sabin_3_AY184221.gb rel
8051: SP_171_R1.fastq versus Sabin_3_AY184221.gb
Download graph as SVG file
SNP frequencies
Select profilings
Select reference
help
SNP %
6.0
4.5
3.0
1.5
0.0
SNP %
10.0
7.5
5.0
2.5
0.0
SNP %
16
12
8
4
0
0
1,500
3,000
4,500
6,000
Cumulative SNP frequency for profiling result 8061 at ref #1

Phylogram shape: rectangular ▾
Branch length scale: linear ▾

- **Table Query Analyzer**- A sophisticated table parser embedded with an arsenal of selection and statistical tools allows dynamic analysis of tabular datasets, for example, post-market data.
- **FDA**- Functional discriminant analysis can separate data with respect to a number of characteristics.

Database Projects

- **Curated Short Read Archive (CSR)**- Curated metadata associated with publicly available short read sequences, including data hosted by TCGA, CGHub, and NCBI SRA, is housed in this database.
- **DisVar**- This database stores single-nucleotide variations (SNVs) and biomarkers associated with disease phenotypes from sources including GWAS and clinical data, variation databases, and literature mining.
- **BioMuta**- This database stores curated, cancer-related, non-synonymous single nucleotide variation (nsSNV) information. Sequence feature information is collected from a variety of sources. Only nsSNVs with associated PubMed ID (PMID) are maintained in the cancer-centric database.
- **BioXpress**- This is a gene expression and cancer association database where expression levels are mapped to genes using high-quality RNA-Seq data obtained from TCGA, ICGC, Expression Atlas, and publications.

Interfaces

New interfaces have been developed to optimize the usability of the underlying infrastructure and algorithmic layers. HIVE is equipped with a web-portal for user and administration access. The portal's client side applications are built using technologies, such as JavaScript and HTML5, driving CGI requests to the backend servers which perform computations using highly efficient C/C++ native programs.

Visualizations

The HTML5 vector scalable graphical engine is at the core of the HIVE scientific visualization package. HIVE provides its own graphical abstraction layer to ensure independence from particular platform implementations in developments of object oriented web applications. An interactive event handling layer facilitates interaction with graphical primitives on the screen using both mouse and keyboard. This functionality enables a user to communicate with a rendered scene and send commands or retrieve additional information about the visual representations. Generic sequence, annotation, and visualization controls also allow the user to create and view sequence annotations supplied by the user prior to analysis or generated by HIVE computations.

Performance

Speed

Speed of HIVE-hexagon aligner was measured for alignments of query reads to the corresponding reference genomes of three species: hepatitis C virus, mycoplasma and human. Full viral genome mapping was completed in 12 seconds, and full human genome mapping in 23 minutes. (NOTE: The maximum number of CPUs used in these trials is 48.)

Accuracy

One million synthetic reads were generated directly from genomes with a randomly distributed noise of 0, 1 or 5%. Genomes used include influenza, a bacterial mixture and human. HIVE-hexagon has shown the ability to fully align all error-free reads for influenza and other small organisms with high similarity to the reference. All three compared tools leave some error-free human reads unaligned but HIVE-hexagon reports significantly fewer (16) than Bowtie (150) and BWA (147).

Purpose	**% Reads Unaligned**		
	HIVE-hexagon	**Bowtie**	**BWA**
Influenza – 0% noise	0.0000	0.0000	0.0000
Influenza – 1% noise	0.0013	0.5171	0.4930
Influenza – 5% noise	0.4645	16.7000	21.3228
Human – 0% noise	0.0016	0.0150	0.0147
Human – 1% noise	5.8383	18.8320	18.4592
Human – 5% noise	15.3918	62.8203	62.4555
Bacterial mix – 0% noise	0.0000	0.0000	0.0000
Bacterial mix – 0% noise	0.0002*	0.2963	0.4078
Bacterial mix – 5% noise	0.3539*	16.2000	20.3869

* with repeat and transposition subsearch turned on

Scalability

The design of HIVE's hardware, software and network resources accounts for the dynamic nature of NGS technology and can accommodate scalable expansion. The following elements allow extension of HIVE to support expected future data needs:

- **Multipathing**- Multiple connections between compute and storage nodes increase network thickness, facilitating more efficient communication.
- **Locality Scaling**- Expansion is accomplished by organization of data into local clusters such that additional clusters, not individual nodes, are used to increase capacity, decreasing the total compute + store + network scaling costs.
- **Preferential launch of jobs at location of data**- A combination of data sorting and the ability to move computations to data at known locations optimizes parallelization efficiency by minimizing time required for data transfer.
- **Map-wrap**- Unlike the Hadoop "map-reduce" paradigm, HIVE uses wrapper objects as indices to access information. Files and computational results are only reduced when requested for download.

Operating Models

HIVE can operate in multi-thousand core supercomputer centers or in field offices with small compute stations.

For More Information

HIVE public website:
https://hive.biochemistry.gwu.edu

High-performance Integrated Virtual Environment (HIVE)

Contact:

Raja Mazumder	mazumder@gwu.edu
Vahan Simonyan	vahan.simonyan@fda.hhs.gov