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

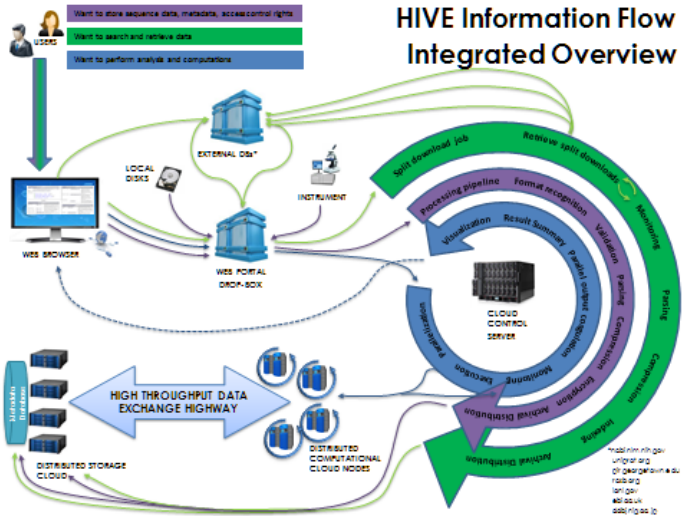
Highlights:

- blueHIVE Technology Group provides NGS data solutions
- 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
- The honeycomb data model allows determination of data access privileges in a finely granular manner, providing capabilities of securing every object in the system.

Issues associated with Next-Generation Sequencing Data

The decreasing cost, increasing speed and improving accuracy of next-generation sequencing (NGS) production over the last decade has led to significant advances in genomic science and technology. This progress facilitated a data explosion, generating huge amounts of sequence information over a very short period of time. Due to the now petabyte-scale size of the reads (and expectations of doubling that scale within a few years), the challenges associated with storing, accessing and computing on NGS data are increasingly difficult and have created a bottleneck in the downstream analysis.

To address these issues, HIVE, the High-performance Integrated Virtual Environment, is an implementation of a multicomponent cloud infrastructure where the distributed storage library (sequence archive) and the distributed computational powerhouse are linked seamlessly. This environment provides web access for authorized users to deposit, retrieve, annotate and compute on NGS data.



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

blueHIVE Technology Group

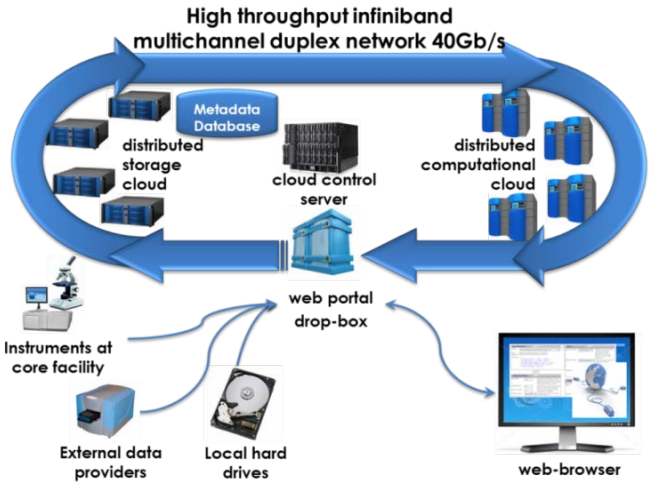
High-performance Integrated Virtual Environment (HIVE)

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

Infrastructure

Unlike many massively parallel computing environments, HIVE uses a cloud control server which virtualizes services, not processes. It is both very robust and flexible due to the introduced abstraction layer between computational requests and OS processes. The novel paradigm of moving computations to the data instead of moving data to computation nodes implemented in HIVE has proven to be significantly less taxing for hardware and network infrastructure.

The location of the distributed storage system and the database of sequence read archive metadata on the same network is an elegant solution to the issue of data transfer bottlenecks. 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 adds the capability to automatically download and update external data sets to HIVE data repositories. Additionally, data parser backend processes enable the users to retrieve specific information from external data sources including NCBI, UniProt, PIR and others. This utility facilitates filtration and receipt of data based on the user's query. The database of sequence read archive metadata is used to extract significant and specific data from short reads obtained from NGS experiments. This database collects and classifies all experiment- and study-specific information into categories designed as part of the metadata.



The honeycomb data model developed for HIVE differs from traditional relational databases by coalescing the metadata into an object oriented model, but unlike other object oriented databases implements unified API interfaces to search, view and manipulate all data regardless of type. This model simplifies the addition of new data types and minimizes the necessity for restructuring of the database, streamlining the developments of new integrated information systems. The honeycomb model implements 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 multiplicity of rules. This model, designed for sensitive data, provides capabilities of controlling and auditing all operations to every object in the system.

Algorithmics

Modern next-generation sequencing platforms are capable of generating a large number of short reads mapped into the same genomic position with corresponding low error rates. For shorter genomes, like those of viruses and bacteria, this results in a high degree of redundancy. Innovative consideration of this redundancy and self-similarity between reads allows the algorithm to minimize the memory footprint by removing repeats and 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 and accelerate alignment of each individual read by:

- considering less high-scoring segment pair (HSP) candidate positions for optimal alignment
- memory and CPU cache usage is increasingly optimal
- the algorithm itself using previous sequence alignments to quickly squeeze the scope of highly rated HSPs

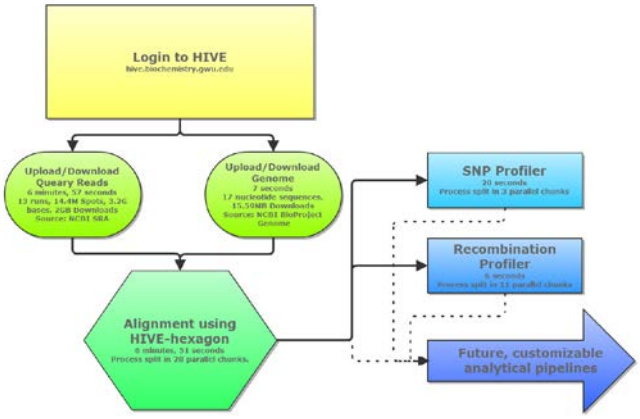
Unlike existing analogues, the diagonal of the Smith-Watermann algorithm compute matrix is allowed to float together with the running high scoring path. This allows generation of longer, multiple insertion/deletion-containing alignments to take advantage of the dynamic matrices diagonalization method.

Modifications to conventional heuristic seeding have been made and the resultant refined approach has

achieved high levels of parallelization using HIVE infrastructure. Data transfer is minimized, thus removing most of the IO bottlenecks.

Some specific algorithms developed which employ these properties are:

- **HIVE-hexagon** - massively parallel, efficient algorithm specifically designed to work with NGS alignments to reference genomes. This algorithm is comparable to well-recognized algorithms in its quantitative and qualitative outcomes but is greatly superior in speed and sensitivity.
- **Base-calling and SNP-profiler** - allows performing base-calling, SNP-calling, reports statistical significance, quality profile, sequencing noise profile.
- **Post-alignment quality control procedures** - developed based on positional base-frequency entropic information content paradigm from information theory methodologies to validate the results of alignment algorithms and to distinguish artifacts from real biological variability.
- **Meta-genomic recombination analysis** - study viral and bacterial population dynamics.



Interfaces

New interfaces have been developed to optimize usability of the underlying infrastructure and algorithmic layers. HIVE is equipped with a web-portal for user and administration access. The portals client side applications are built using technologies such as Javascript, HTML5, Java driving the requests to backend servers using C/C++ and PHP CGI internal built applications using Ajax.

Account Registration and Management

User self-registration web pages allow creation of a new account, capable of accessing the system resources. A registration authentication and supervisor confirmation notification system adds the capability to manage account registration confirmation emails and provides functionality for group administrators to verify joining members. Once logged in, the HIVE secure web-portal home page includes access to sequence datasets, user files, available algorithmic utilities, analysis pipelines, and results of computations.

Tools and Algorithms

The NGS alignment web-interface by default uses HIVE-hexagon, allowing quick and efficient identification of millions of short reads. Usage of HIVE-hexagon is recommended as the native aligner to the system because it has been developed and optimized specifically for use in High Performance Cloud computing environments. However, a number of other industry-standard tools (BLAST, Bowtie and others) have been adapted and embedded within HIVE to facilitate optimal compatibility and performance.

The interface for NGS data-profiling on reference genomes allows statistically accurate base-calling and SNP discovery. Viral and bacterial DNA recombination site interface aids in visualization of recombination events in the pool of viral/bacterial isolates. The interface allows visualization of subsets of data based on user preference and inputs.

The quality control interface displays preview summary information of uploaded sequence files and outcomes of different post-alignment QC procedures for validation and verification of NGS data.

Visualizations

HTML-5 graphical engine is at the core of the graphical visualization package. HIVE has its own graphical abstraction layer making the development of object oriented web applications independent of particular implementations of a platform. 2D and 3D graphical primitives provide basic graphical fundamentals of drawing applets. A matrix transformation stack library is implemented for geometric operations together with support for graphical attribute abstraction layers. Series/plot/layer/presentation model creates the next layer of graphical abstraction library providing a bridge between scientific data and graphical objects rendered on the screen.

Interactive event handling layer provides a capability to interact with graphical primitives on the screen using mouse and keyboard. Using this functionality a user can communicate with rendered scene, send commands or retrieve more detail information about visual representations of scientific data while sliding through the scientific graphs and images. Using the underlying core graphical library HIVE builds the next level of abstraction providing to its developers the following native chart visuals and sequence controls.

Customizable, native line/area/pie/scatter/bar/err-chart visuals are the fundamental scientific presentation elements. Using this library, HIVE developer only funnels the appropriately formed data from the server into the visualization element on the user's screen. Together with a set of predefined configuration elements defined on web–pages, this data gets translated into the final scientific plots.

Generic sequence, annotation and visualization controls are a list of web-tools which allow the user to create or view sequence annotations obtained from remote resources or as a result of HIVE computational results. The basic interface is designed to interactively select ranges of sequences, assign tags and categories and enter curated information describing those domains.

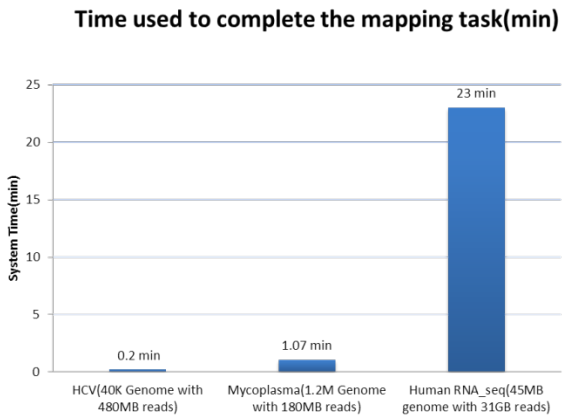
Performance

HIVE performance is competitive with other industry-standard systems of comparable capabilities. Tests were performed to measure the speed, accuracy and reproducibility of the two major tools, HIVE-hexagon mapping and SNP-calling sequence profiler.

Speed

Speed of HIVE-hexagon aligner was measured for alignments of three species' query genomes to the corresponding species' reference genomes. Species considered include hepatitis C virus, mycoplasma and human.

- Full viral genome mapping in 12 seconds
- Full human genome mapping in 23 minutes



NOTE: The maximum number of CPUs used in these trials is 48. The number of CPUs is flexible per implementation and will affect the overall speed of processing.

Accuracy

Synthetic reads of known content were randomly generated and blindly mapped to a set of reference genomes using HIVE-hexagon. Following alignment, these same sequences were profiled and analyzed for SNP calling accuracy.

- High accuracy for both mycoplasma and hepatitis C mapping
- Perfect accuracy for SNP-calling in both species