



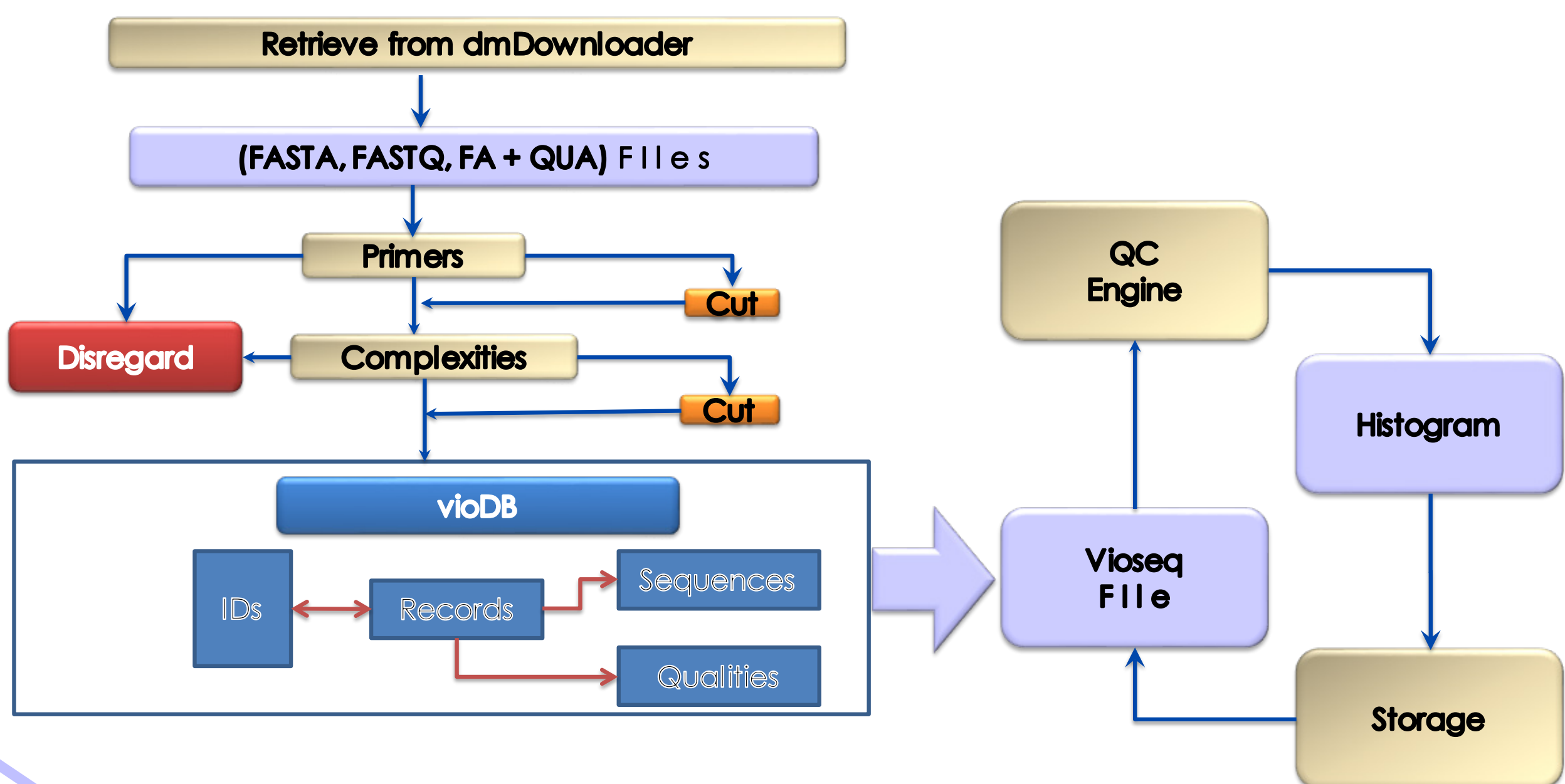
High-performance Integrated Virtual Environment (HIVE): A Robust Infrastructure for Next-generation Sequence Data Analysis

Elaine E. Thompson¹, Vahan Simonyan¹, Raja Mazumder² and the HIVE Team^{1,2}

¹Food and Drug Administration, Center for Biologics Evaluation and Research, Silver Spring, MD, 20993

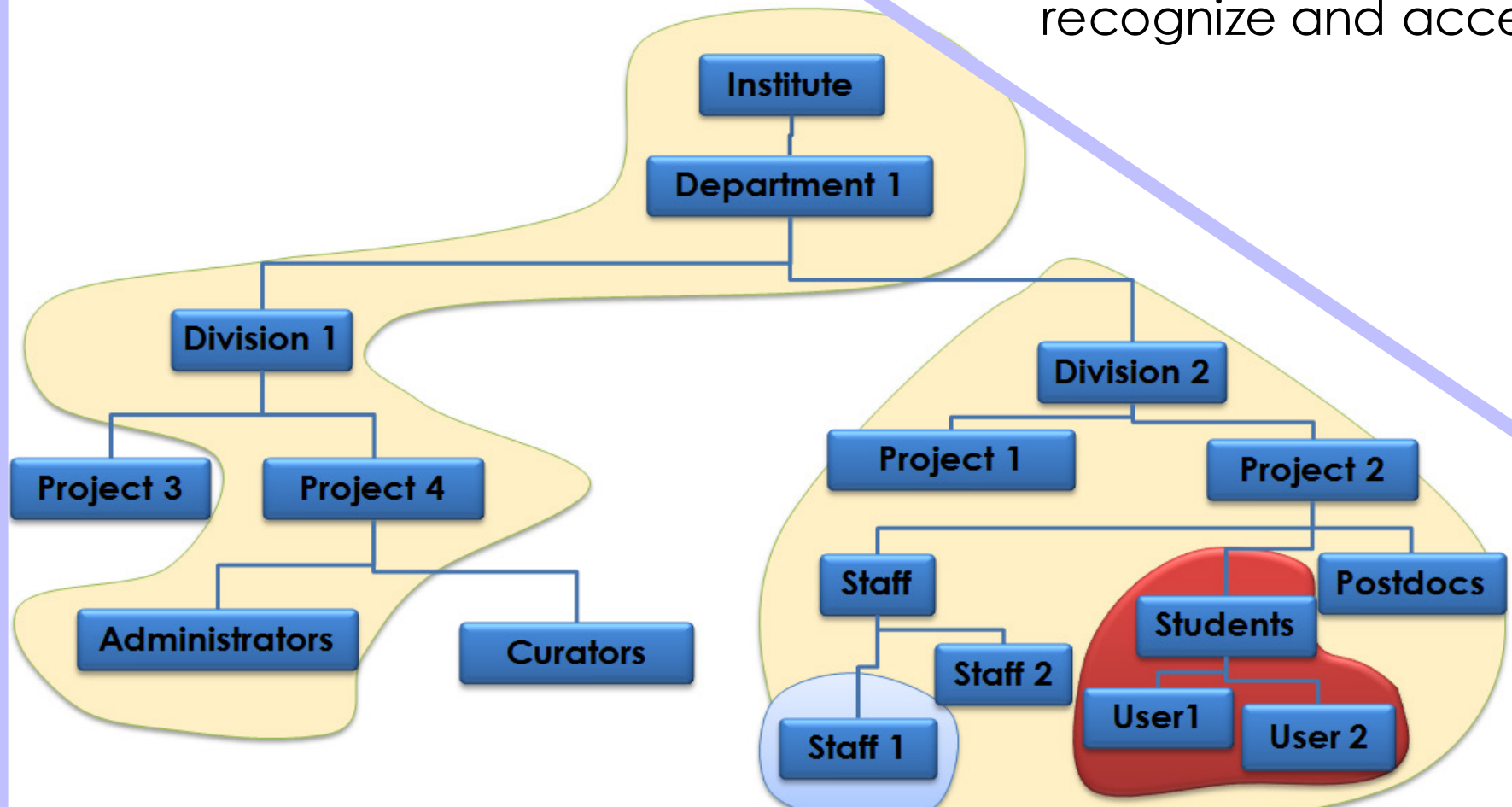
²Department of Biochemistry and Molecular Biology, The George Washington University, Washington, DC, 20037

THE GEORGE
WASHINGTON
UNIVERSITY
WASHINGTON, DC



Data Deposition Pathway

Data deposition into HIVE is accomplished by two different tools: the downloader retrieves the file from user specified location; the archiver identifies file type and executes the parser to identify and analyze the content of each file and extract the information to be stored in the system. Each field is assigned a unique ID that is used internally to recognize and access the path to a specific file.



Hierarchical Security Model

Users, groups, files, processes, metadata, visual elements and algorithms are all treated as security-enabled objects. A hierarchical security model allows granting of permissions down the hierarchy, up the hierarchy or to specific subjects with a minimal number of access control rules.

Who	When	What	Result	Comment
		READ		...
		DELETE		...
		SHARE		...

Collaborative Environment

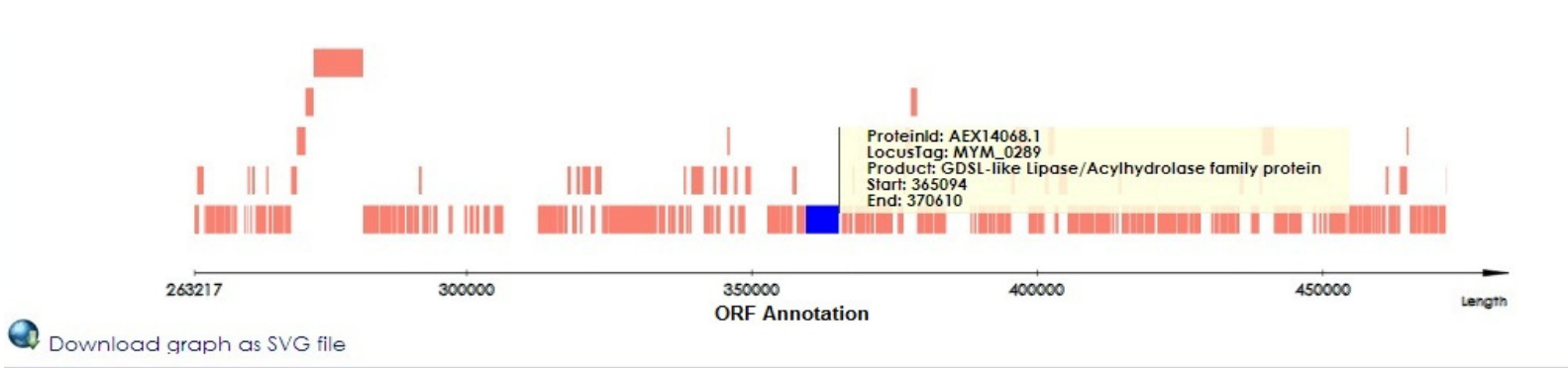
All available resources allow for the implementation of a collaborative environment. This creates an optional environment for users who want to enjoy free exchange of ideas and projects using the powerful HIVE engine.

Visual interfaces are implemented in a Data Driven Document model using JavaScript and HTML-5 which are natively implemented in all modern browsers



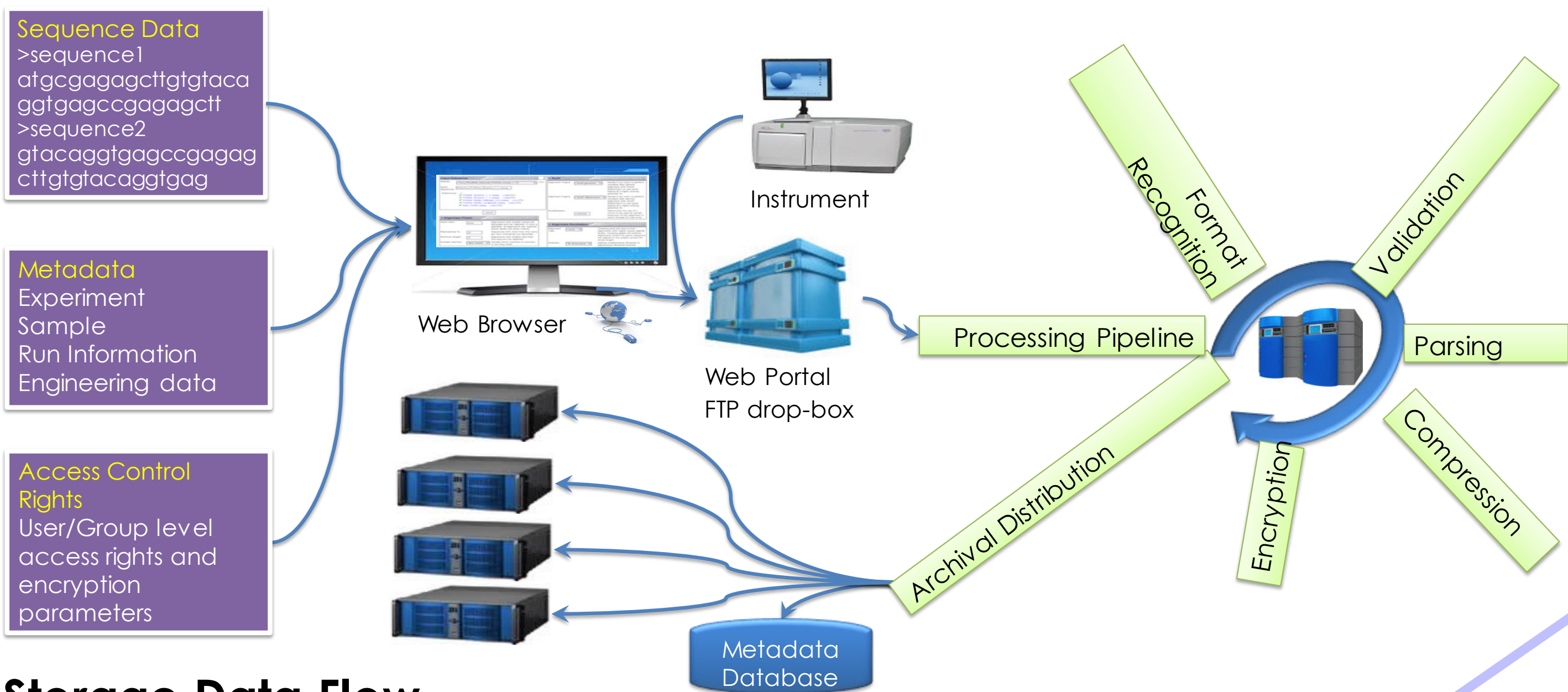
Example: Genomic-Proteomic Mapping

Once the SNP positions in a genomic sequence are discovered, a follow up question which may be asked is how the mutation affects the protein sequence. As a result, HIVE provides the graphic representation that shows the mapping between the genomic and proteomic space.



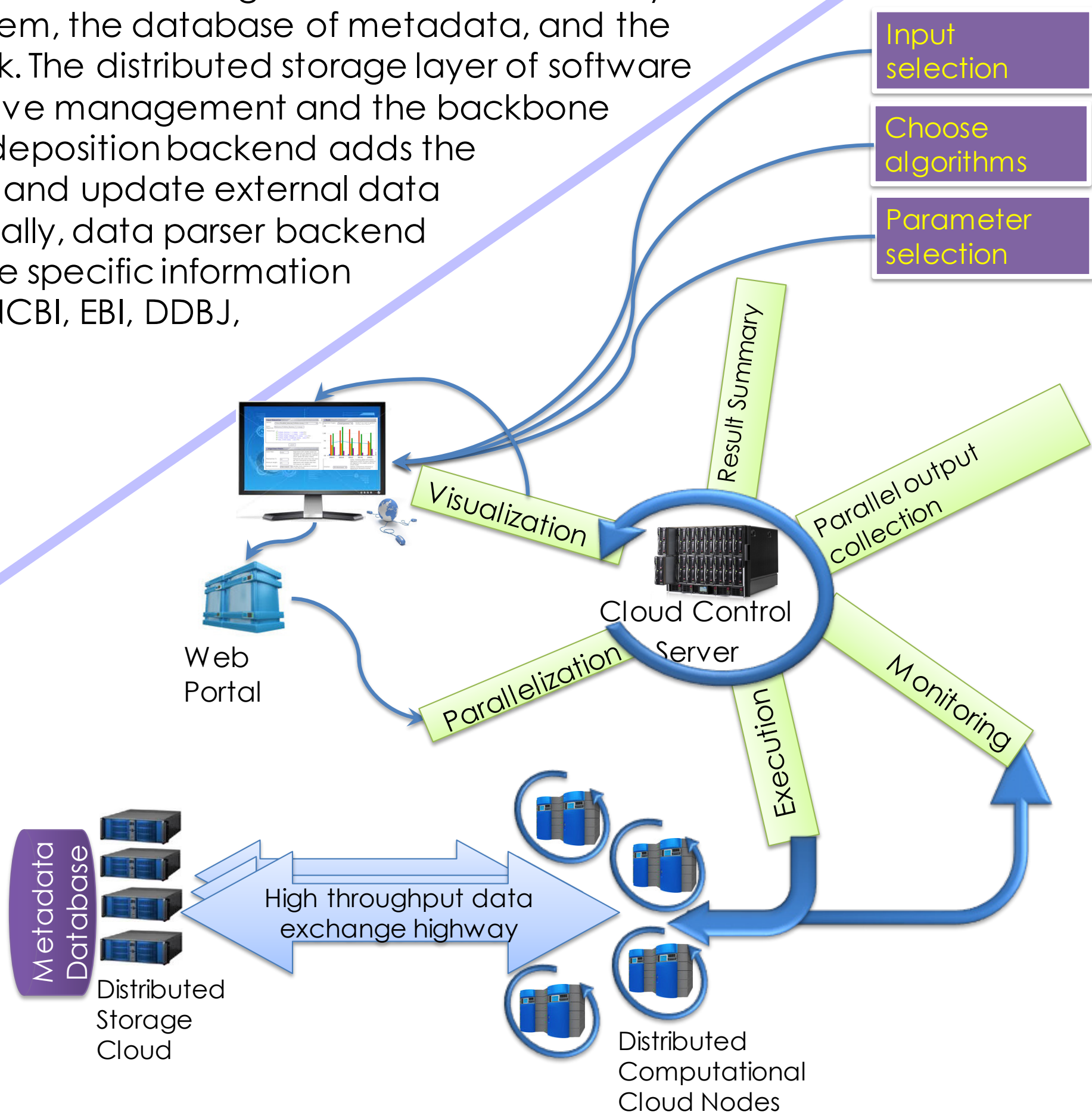
Example: SNP Profile Mapping

The SNP profile calculates the frequency of individual bases plotted against either the number of bases in the reference genome (the index) or the number of bases in the consensus sequence.



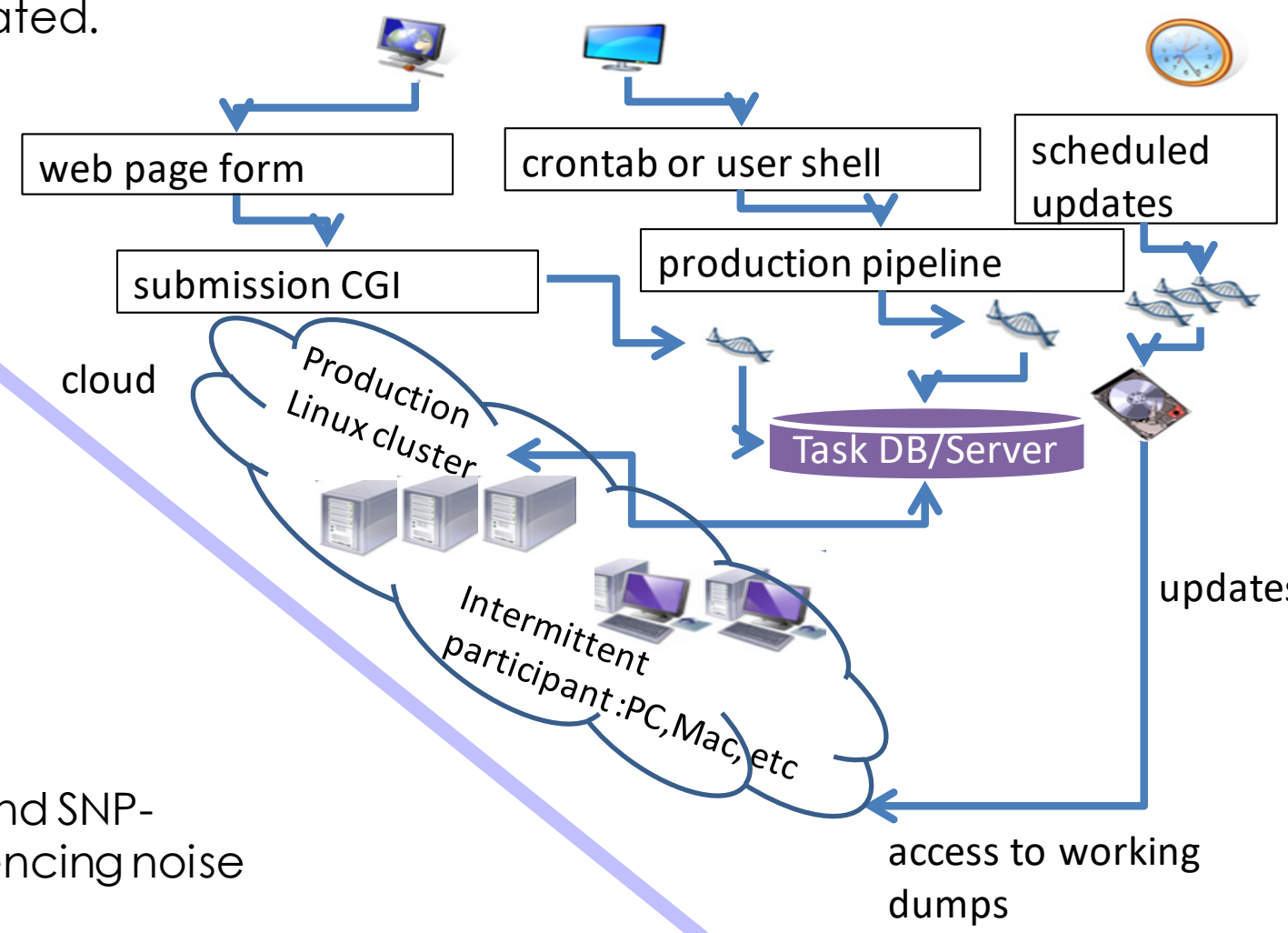
Storage Data Flow

An elegant solution to the data transfer bottlenecking issue has been found by integrating the distributed storage system, the database of metadata, and the compute nodes into the same network. The distributed storage layer of software 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 procedures enable the users to retrieve specific information from external data sources including NCBI, EBI, DDBJ, UniProt, PIR and others.



Computational Pipeline

The submission process, usually a web-page CGI process or a client program, submits the initial information from HTML form into the HIVE cloud server queue and retrieves a unique request identification number. Using this value, the request can be tracked in the system and associated information can be retrieved or updated.



dna-hexagon – massively parallel, efficient algorithm designed to work with NGS alignments to reference genomes. It is comparable to well-recognized algorithms in its quantitative and qualitative outcomes with significant improvements in speed and sensitivity.

Base-calling and SNP-profiler – utility that allows base-calling and SNP-calling and reports statistical significance, quality profile, sequencing noise profile.

Post-alignment Quality Control Procedures – set of tools 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 Tool – utility to study microbial populations of environmental samples, to discover genetic recombination events and resolve subspecies.

Clonal Population Discovery Tool – a powerful de novo utility capable of tracking bifurcation diagrams of read mappings along the reference-assisted profile and extracting tens or hundreds of reference genomes from a soup of meta-genomic material where multiple co-related species exist.

Comparative SNP Phylogenetic Analysis Tool – algorithm designed to show hierarchical information or graphs/networks of interconnected data on the screen. The visuals produced by this tool are capable of showing interactive controls for user to explore the data.

XL- table data analysis tool – Excel like backend tool capable of executing information retrieval inquiries from multi-million row tables: sorting, filtering, categorization, non-redundification and more...

Support, adoption and adaptation of existing industry standard tools and utilities for next gen analysis.

ACKNOWLEDGEMENTS

Mark Walderhaug, Carolyn Wilson, **HIVE team:** Grace Alterovitz, Amanda Bell, Prince Birring, Ting-Chia Chang, Stefan Dabic, Hiral Desai, Hayley Dingerdissen, Lama Elzohary, Marianna Faradzheva, Brian Fochtman, Tigran Ghazanchyan, Anton Golikov, Naila Gulzar, Robel Kahsay, Konstantinos Karagiannis, Charles Hadley King, Ilya Mazo, Reza Mousavi, Rahi Navelkar, Ekaterina Osipova, Aleksey Pshenichnov, Yao Ren, Alexandre Rostovtsev, Xutian Ruan, Luis Santana-Quintero, Michelle Shen, Amanraj Singh, Krista Smith, Valery Tkachenko, Phuc VinhNguyen Lam, Jeet Kiran Vora, Alin Voskanian