Computational Pipeline Engine in FDA HIVE: Adventitious Agent Detection from NGS Data

Ilya Mazo, Alexander Lukyanov, Anton Golikov, Luis Santana- Quintero
Center for Biologics Evaluation and Research (CBER)

INTRODUCTION

Testing for the absence of adventitious agents is an important part of QA in vaccine development and manufacturing. Next Generation Sequencing (NGS) has become the method of choice for detection of viral and microbial contamination in the product matrices. In this project we followed the guidelines of the Advanced Virus Detection Technologies Interest group (AVDTIG)[1] to implement a computational genomics pipeline based on the FDA HIVE[3] platform for the detection of adventitious agents in NGS data.

METHODS

We have implemented a mechanism in the FDA HIVE that supports the assembly of individual algorithms and tools into computational pipelines. Importantly, the pipeline execution can be run on multiple compute and data nodes. We have configured a genomics pipeline to: (i) perform the mapping of the reads from the very high coverage datasets (>10,000) to the reference genomes (e.g. African Green Monkey (AGM)), (ii) assemble unaligned reads into contigs, (iii) map the unaligned reads and contigs to the Reference Viral Database (RVDB)[2]. The hits to RVDB have been further analyzed individually by BLAST and several classes of false positives have been found.

DISCUSSION / CONCLUSION

The pipeline support mechanism in HIVE enables rapid development and configuration of genomics pipelines for both scientific and regulatory applications. The Adventitious Agent Detection as a part of HIVE is designed to facilitate and standardize the vaccine data review process.

Reads from AGM WGS and transcriptome systematically produce false positive hits against RVDB due to presence of repeats (e.g. alpha satellites), retroelements and poly A stretches. Additional steps to be added to the pipeline to post-filter the hits to remove such false positives.

AGM reference genome has gaps and when used as a reference should be complemented with simian rRNA dataset.

REFERENCES

1) Advanced Virus Detection Technologies Interest Group (AVDTIG): Efforts on High Throughput Sequencing (HTS) for Virus Detection. Arifa S Khan et al. PDA J Pharm Sci Technol (2016)
2) A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. Goodacre N, et al. mSphere. (2018)
3) High-performance integrated virtual environment (HIVE): a robust infrastructure for next-generation sequence data analysis. Simonyan V et al. Database Oxford. (2016)

RESULTS

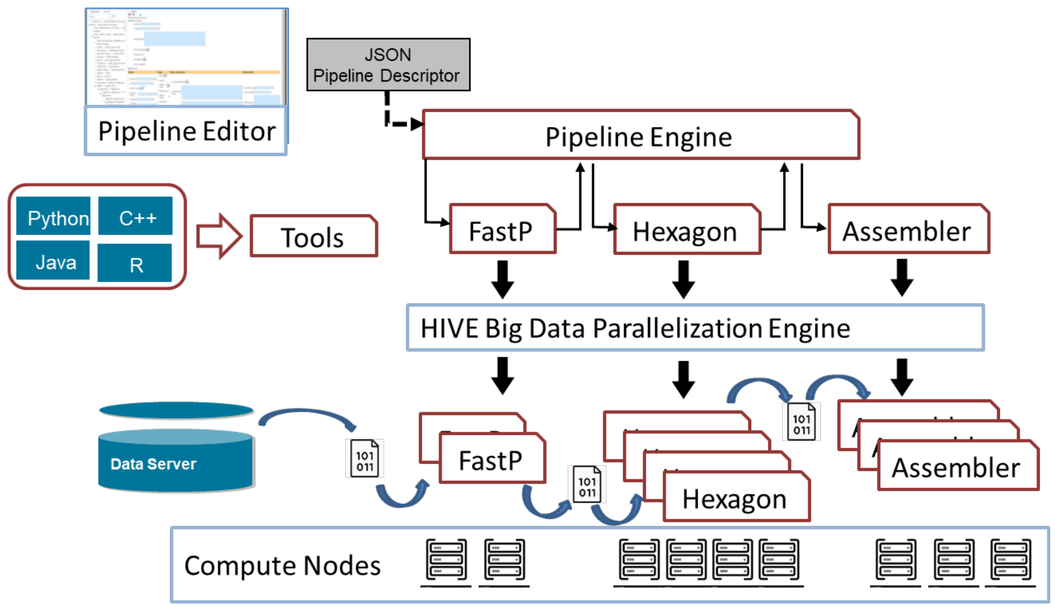
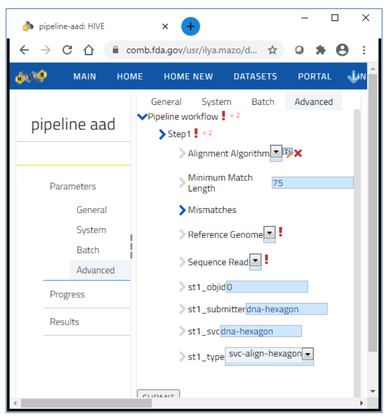


FIGURE 1. The pipeline execution engine in HIVE Third party or internal algorithms are implemented as tools and serve as building blocks for the new pipelines. The pipeline steps and parameters are described through GUI and provided to the engine in JSON format. The data output from each step of the pipeline can be saved to the database.

VERO cells from African Green Monkey kidney are used for production of viral vaccines.
NGS data from vaccine samples is predominantly the mixture of vaccine and AGM reads.



Pipeline steps and parameters

Pipeline Step	Alignment Step 1	Step 2	Step 3
Hexagon Mode	Virus specific	Human genome	Customized
Minim. Match	75	75	75
Mismatches	15	3	15
Matches to Keep	Random vote between equally best alternative matches	First Match	Random vote between equally best alternative matches
Seed K-mer	11 letters	14 letters	14 letters

We use two-step filtering approach: (i) stringent alignment to the reference AGM genome to remove the majority of mammalian reads, (ii) competitive alignment with relaxed parameters to discriminate between RVDB and endogenous virus-like sequences.

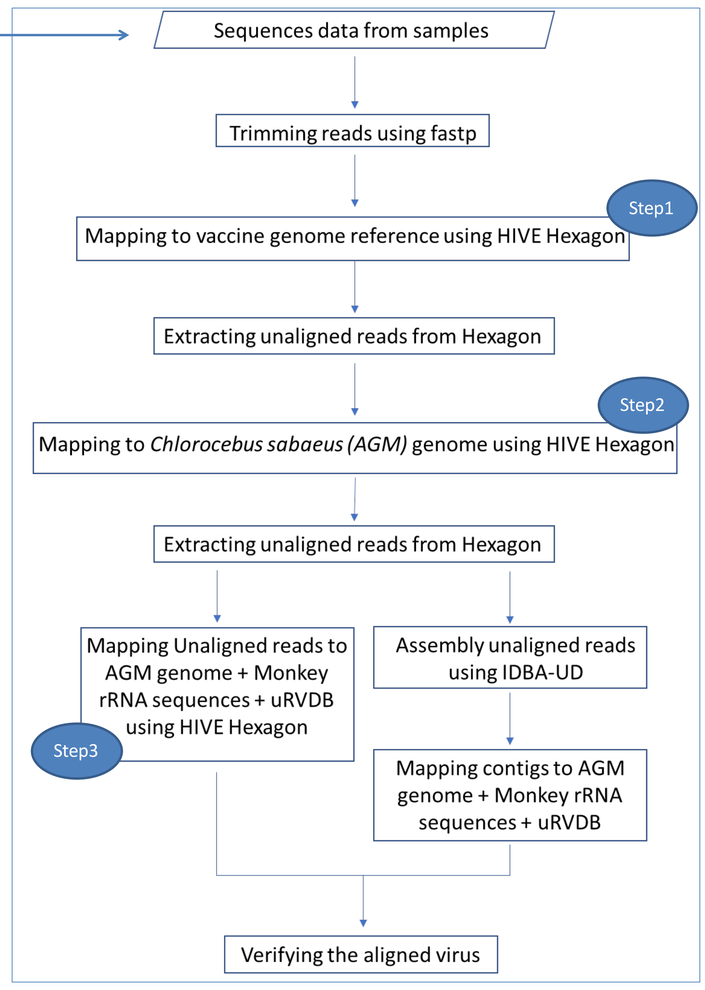


FIGURE 2. The scheme of Adventitious Agent Detection pipeline. The known challenge in adventitious virus detection is to discriminate between endogenous virus-like sequences and actual pathogenic viruses.

Organism in RVDB	Hits	Verification	Actual Origin
Pteropus lylei-associated alphaherpesvirus	13953	FALSE	TA repeat + simian alpha satellite
Macaca mulatta polyomavirus 1	7304	FALSE	Simian alpha satellite fragment
Macaca mulatta polyomavirus 1	2307	FALSE	Simian alpha satellite fragment
Guanarito mammarenavirus	1438	FALSE	Human lncDNA or simian rRNA
Macaca mulatta polyomavirus 1	662	FALSE	Simian DNA for autonomously replicating sequence ors15
Guanarito mammarenavirus	374	FALSE	Simian rRNA
Tragelaphus spekii	204	FALSE	ERV from endogenous retrovirus
Lampetra aepyptera	162	FALSE	Simian retrotransposon
Bovine alphaherpesvirus 5	150	FALSE	Bos DNA
Orf virus	82	FALSE	Bos DNA
Bunyavirus sp.	65	FALSE	poly T
Jingmen tick virus	49	FALSE	Bos DNA
Human alphaherpesvirus 1	29	FALSE	poly C
Macaca mulatta endogenous virus SERV-K1	28	FALSE	ERV from endogenous retrovirus
Semliki Forest virus	19	FALSE	Clone vector
Macaca mulatta polyomavirus 1	16	FALSE	Simian alpha-satellite fragment
Human alphaherpesvirus 1	15	FALSE	poly G
Human alphaherpesvirus 1	13	FALSE	poly C

FIGURE 3. False positive hits from African Green Monkey NGS data with high homology to RVDB entries. The AGM NGS dataset used (SRA DRA002256). The RVDB hits from the pipeline were individually tested by BLAST and found to be of mammalian origin.

Presented virtually at the 8th Annual FDA Scientific Computing Days (Sep 29-30, 2020)