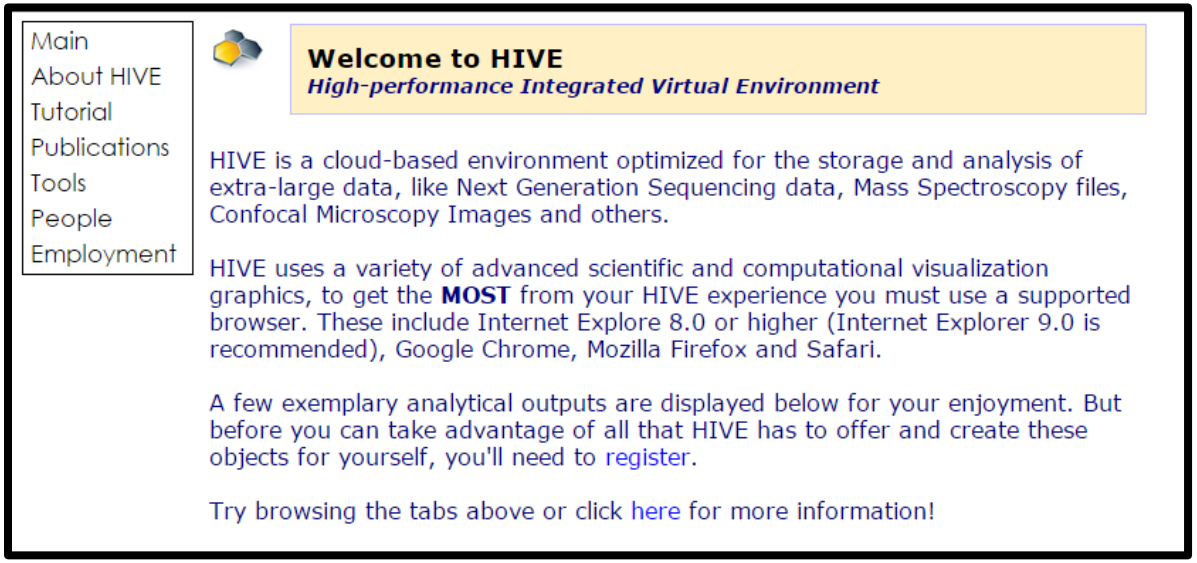
Species and Quasispecies analysis using the public/FDA-HIVE platform

Raja Mazumder mazumder@gwu.edu

Assoc. Prof. Biochemistry and Molecular Medicine; Project Lead, public-HIVE;
Co-director, The McCormick Genomic & Proteomic Center

GW

High-performance Integrated Virtual Environment (HIVE)



2 portals: Public portal (GW) & a FDA only portal

mazumder@gwu.edu | vahan.simonyan@fda.hhs.gov

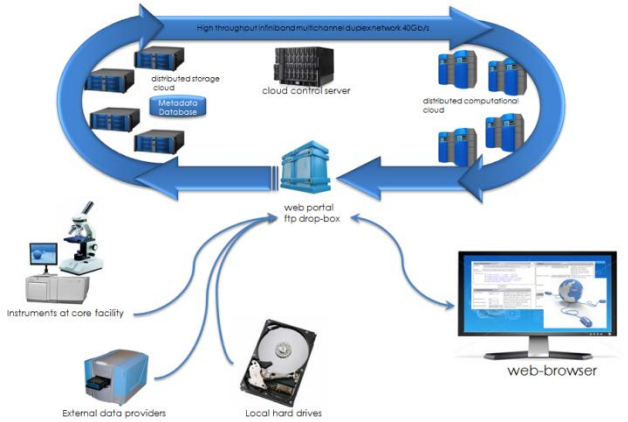
HIVE tools + Community tools

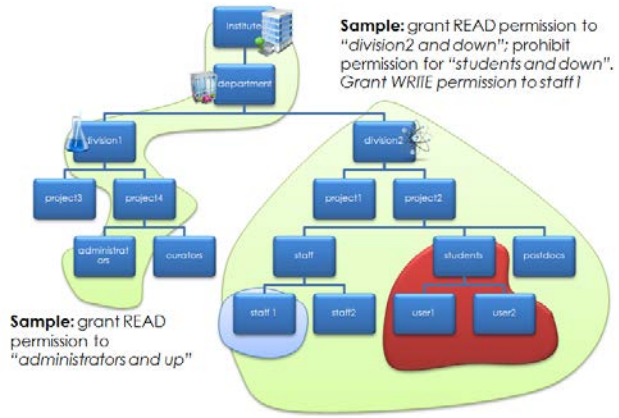
INTEGRATED
BIG DATA
STORAGE, SHARING
AND ANALYSIS
PLATFORM
ATO APPROVED

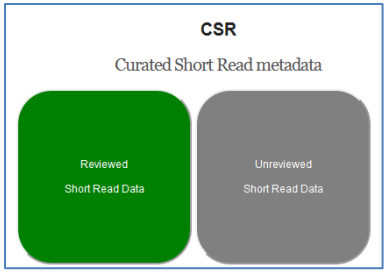
References + Standards in collaboration with community

HIVE key features and focus

- Distributed storage and computing
- Security features (FDA/industry/patient data in mind). Authorized to Operate (ATO) in Regulatory Environment
- Granular sharing capabilities
- Biocuration of content
- Compute speed



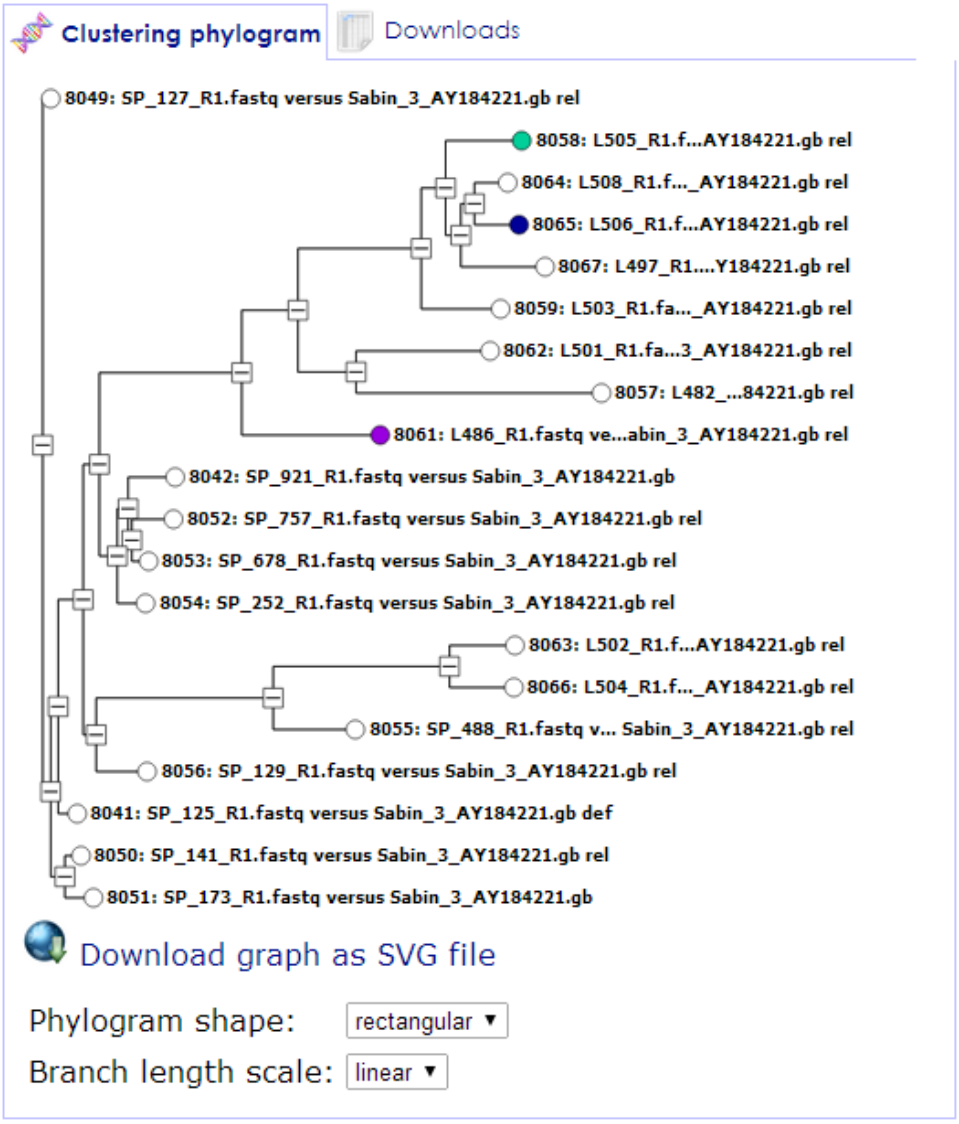


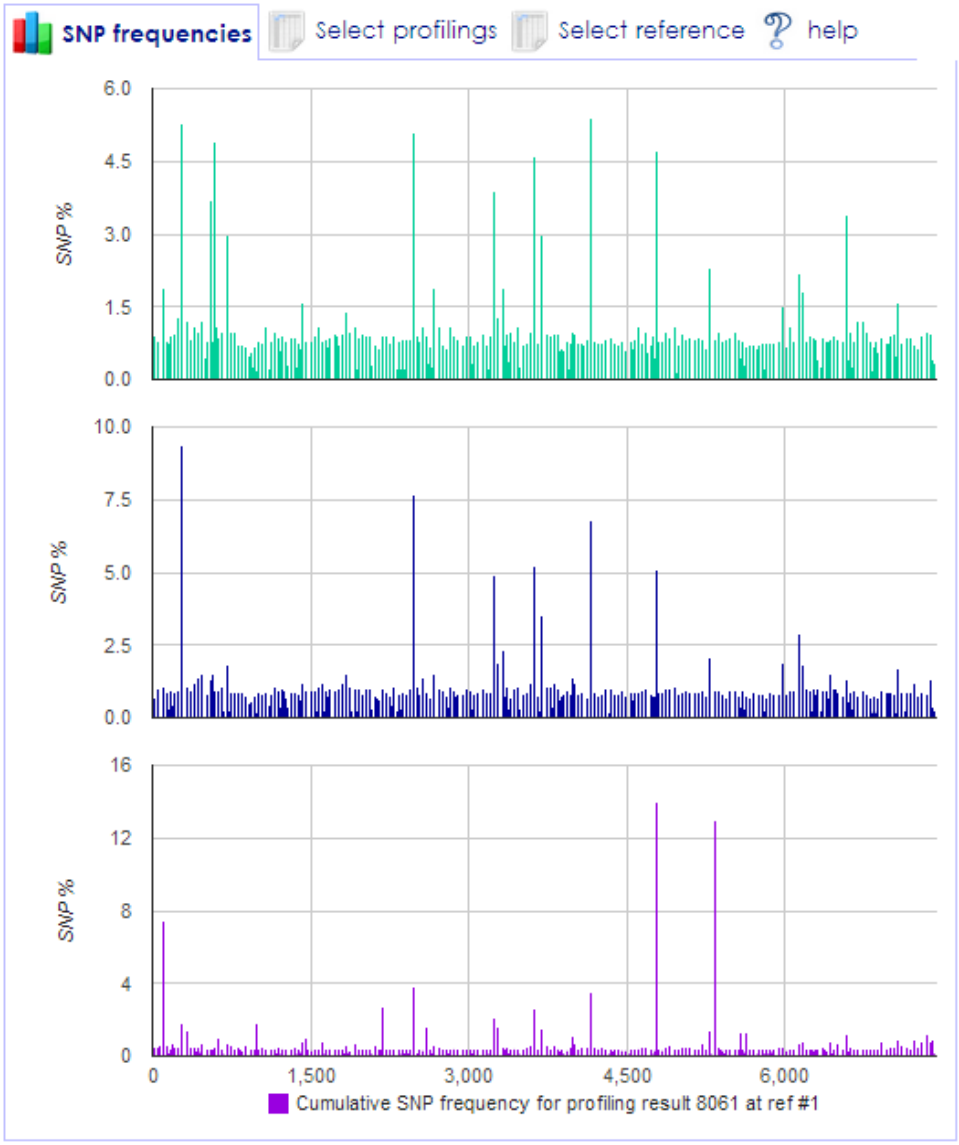


BioMuta
BioXpress

HIVE: visualizations

NGS → Trees





HIVE SNP based hierarchal clustering and phylogenetic analysis

HIVE: Pipeline

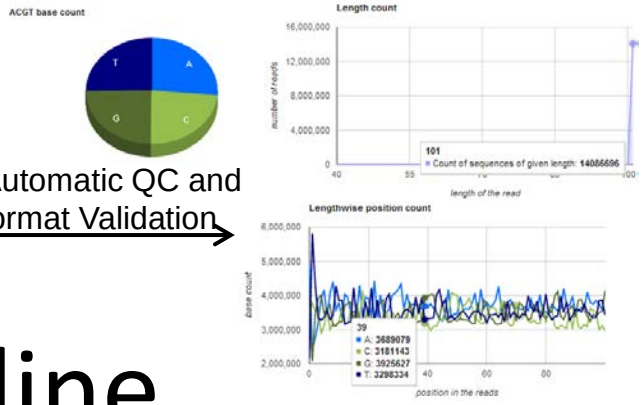
Sample Reads

12 Reference Strains

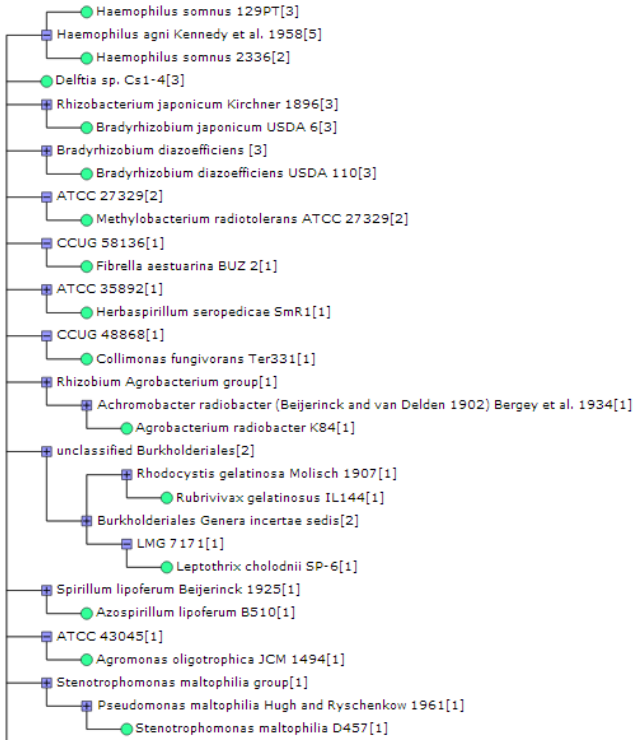
Los Alamos NATIONAL LABORATORY EST. 1943

[1]Upload to HIVE

[2]Automatic QC and Format Validation



[3]Analyze Contaminants using CensuScope



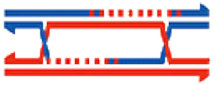
[4]Remove host/contaminant

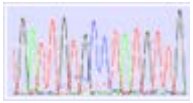
Hexagon

ITERATIVE LOOP: (Repeated within different parameters)

[5]Align samples to mutual frame (HIVE-hexagon)

[6a]Recombinant mosaic discovery (Recombination Tool)





[6b]Generate SNV profile

[6c]Use SNV based phylogenetic tree to compare multiple samples (HIVE-phyloSNP)

[6d]REPORT DOMINANT RECOMBINANT

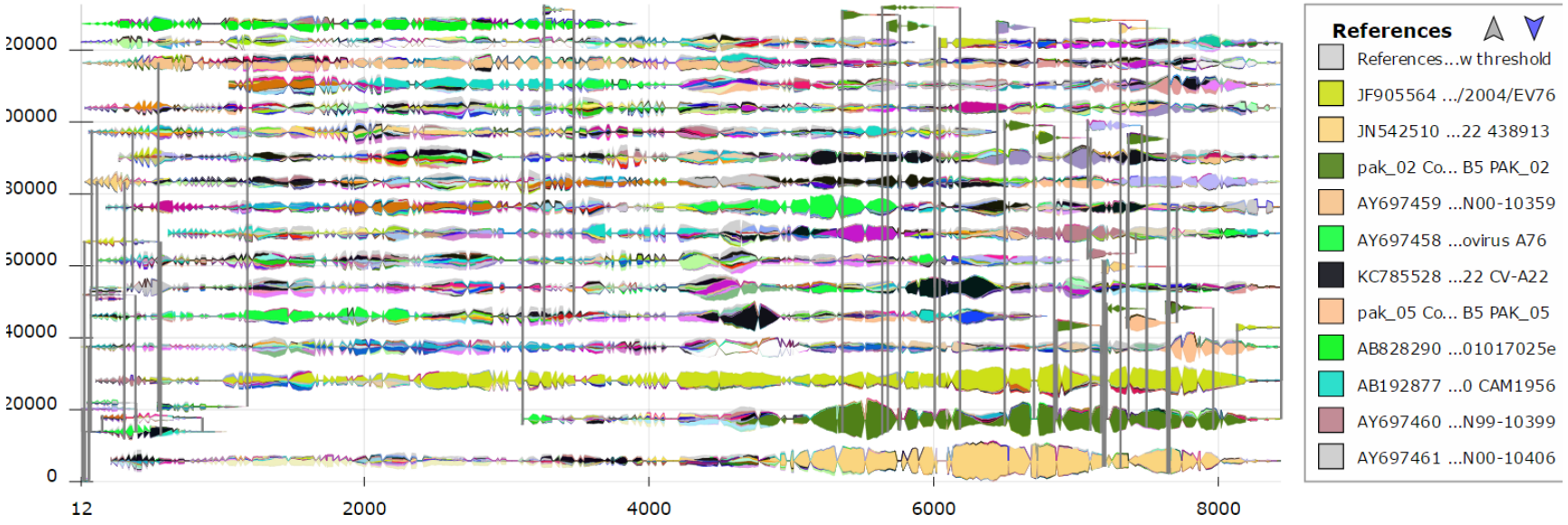
[7a]Clonal discovery

[7b]REPORT DOMINANT CLONE/S

[8]Reinitiate loop with different parameters

Environmental sample

- Stool sample aligned against database of 600 enteroviruses.
- Paired end reads of 300bp.
- Only 0.1% of reads aligned to the reference set
- Haplotype reconstruction performed with 1% mutation cutoff
- ORF validation revealed 12 out of 15 biologically meaningful global haplotypes



Metagenomic analysis for pathogen detection

Patient

Food
Pathogen?

Field

Sample
collected

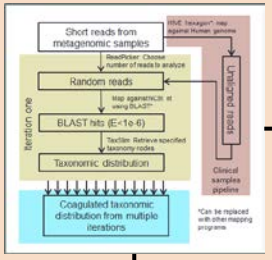
Bench-top
sequencer

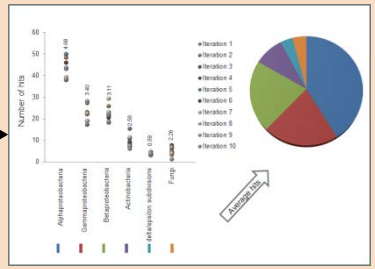
Metagenome
sequenced

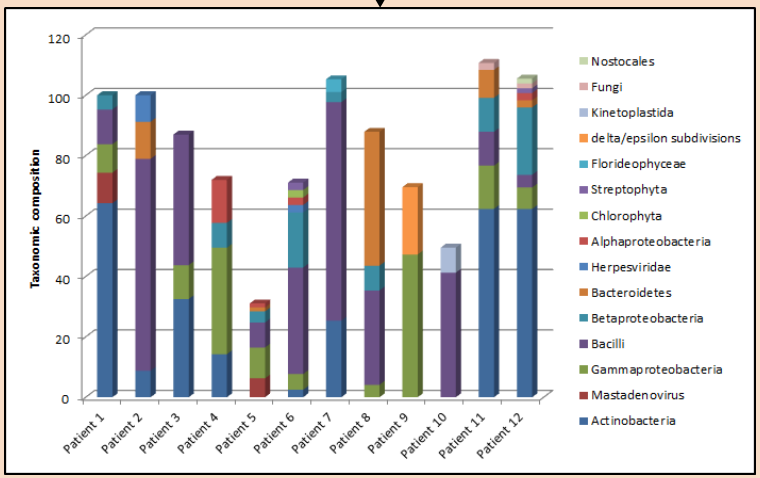
CensuScope+CUPID

Pathogen/s detected
+ comprehensive
taxonomy report

Current analysis performed by CensuScope and CUPID







Curated HIVE content

- **signatureDB**

Consists of unique organism signatures; updated every month

- **Pathogencity_islandDB and pathwaysDB**

Hand curated

- **NIAID/CDC/FDA_priority_genomesDB**

Updated regularly

- **Combination of k-mer/CensuScope**

Algorithm

- Many more community driven collaborations

Rapid pathogen detection

Core and unique functional gene markers

NIAID/CDC pathogens

UniProt ref genomes

*i*ProClass family, function, pathway from >100 databases

Step 3
Filter+Hexagon
~20 minutes

CUPID signatures

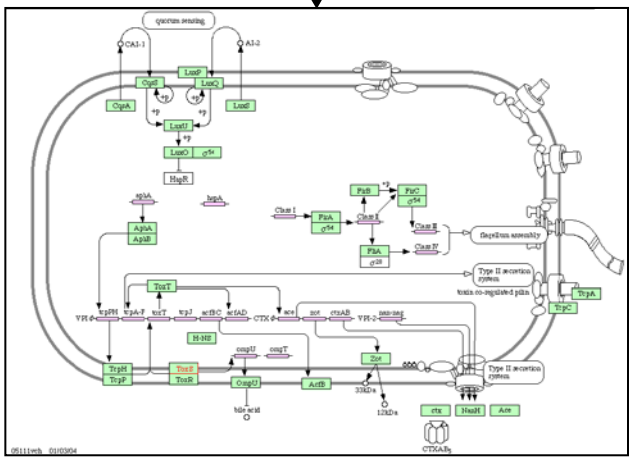
Reconstruction

Step 2
CUPID
~1min

User interface

Step 1
CensuScope
~10 min

Genbank/EMBL nucleotides

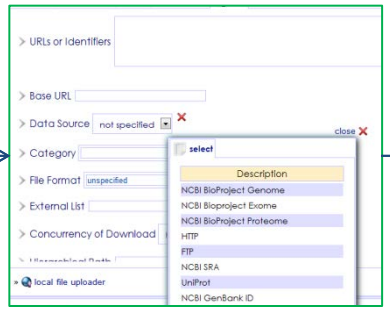


Detection pipeline

Evolving complex system
Needs to be highly flexible

Reads (NGS)

Reference/s



HIVE dmDownloader
NCBI eutils/Webservices

RG 15

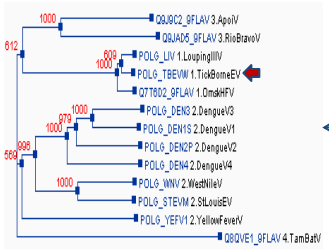
RG 75

RG ##

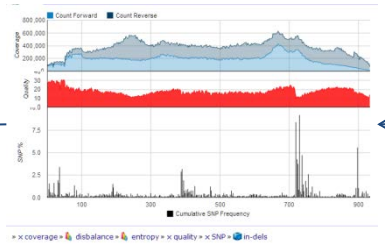
Core Unique Genes

Reference genome

Reference/Representative genome filters



Phyletic placement



SNV profiling

HIVE Hexagon mapping

RG 15

RG 75

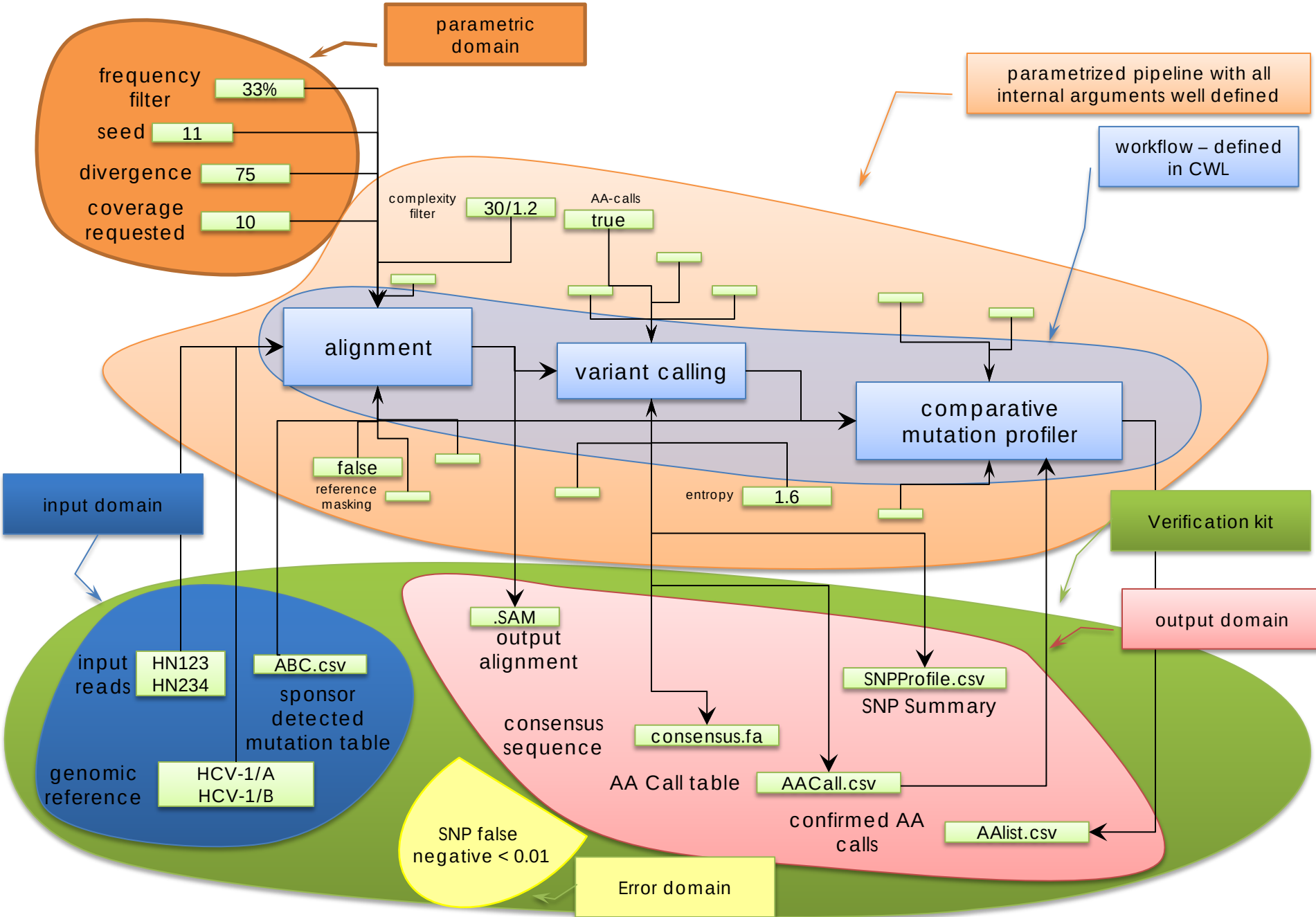
RG ##

Reference genomes

Cannot be done in isolation
Requires collaborative effort

Biocompute objects
Communicating computational analysis to FDA

Adapted from Dr. Vahan Simonyan



```
{
    "id": "obj.    ",
    "type": "antiviral_resistance_detection",
    "name": "HCV1a [taxID:31646] ledipasvir [PubChem:67505836] resistance SNP [SO:0000694] detection",
    "version": "1.1",
    "digital_signature": "905d7fce3f3ac64c8ea86f058ca71658",
    "verification_status": "unreviewed",
    "publication_status": "draft",
    "usability_domain": ["Identify baseline single nucleotide polymorphisms (SNPs[SO:0000694]),
insertions[SO:0000667], and deletions[SO:0000045] that correlate with reduced ledipasvir[PubChem:67505836]
antiviral drug efficacy in Hepatitis C virus subtype  ", "Identify treatment emergent amino acid
substitutions[SO:0000048] that correlate with antiviral drug treatment failure", "Determine whether the
treatment emergent amino acid substitutions[SO:0000048] identified correlate with treatment failure
involving other drugs against the same virus"],
    "authors": [{"name": "Eric Donaldson"},{"orcid": "0000-    -    -    "}],
    "description_domain":{
        "xref":["SO:       ","SO:       ","PubChem:        ","taxID:     " ","PMID:         ","PMID:
        "],
        "keywords": ["antiviral resistance", "SNP"],
        "pipeline_steps": {"HIVE_hexagon":{...}, "HIVE_heptagon":{...}}},
    "execution_domain": {
        "platform": "hive",
        "pipeline_version": "0.1",
        "driver": "shell",
        "url": "https://hive.biochemistry.gwu.edu/workflows/",
        "env_parameters": ["HIVEv1.3"],
        "script": "hive://workflows/antiviral_resistance.py",
        "prerequisites": [
            {"name":"HIVE_hexagon","version":"1.3"}, {"name":"HIVE_heptagon","version":"1.3"}]},
    "parametric_domain": {
        "hexagon_minimum_coverage": "0.15",
        "hexagon_seed": "14",
        "hexagon_minimum_match_len": "66",
        "heptagon_freq_cutoff": "0.10",
        "heptagon_divergence_threshold_percent": "30"},
    "io_domain": {
        "reference_uri": ["http://www.ncbi.nlm.nih.gov/nuccore/CP000139.1"],
        "input_uri_list": ["hive://nuc-read/      ", "hive://nuc-read/      "],
        "output_uri_list": ["hive://data/514769/dnaAccessionBased.csv","hive://data/514801/SNP.csv"]},
    "error_domain": ["false negative discovery <  .0 ", "false positive discovery <  . 01"]}
```


Metadata
(not needed for computation)

parametrized pipeline with all internal arguments well defined

workflow – defined in HIVE

parametric domain

Verification Kit

input domain

output domain

Biocompute workshop speakers and panelists

Acknowledgements

HIVE TEAM MEMBERS (GW+FDA), COLLABORATORS, USERS

Funding sources for our projects

NIH, FDA, GW, NSF, Pharma

Special thanks

Konstantinos Karagiannis, Vahan Simonyan, Alin Voskanian, Konstantin Chumakov, Viswanath Ragupathy, Naila Gulzar, Krista Smith

Contact

mazumder@gwu.edu

HIVE jamboree 2016