All The World's A Sequencer: Global Hepatitis Outbreak Surveillance Technology (GHOST) Cloud Platform

FY15-AMD-81 Project

S Sims[1], DS Campo[1], I Rytsareva[1], Y Zheng[2, 3], A Longmire[1], P Skums[1], Z Dimitrova[1], M Mirabito[4, 5], S Wang[4, 5], R Tracy[4, 5], T Sukalac[4, 5], C Lynberg[5], Y Khudyakov[1].

[1]Molecular Epidemiology and Bioinformatics Laboratory, Division of Viral hepatitis. [2]Eagle Medical Services, LLC, [3]AMD bioinformatics, NCEIZD, [4]Northrop Grumman. [5]NCHHSTP Informatics Office. [6]Research & Development, OD.

Introduction

- The depth of sequencing afforded by recent Next-generation sequencing (NGS) technologies is capable of detecting large numbers of viral variants carried by infected individuals. This data offer novel prospects for understanding both intra- and inter-host pathogen evolution.
- As well as offering a wealth of epidemiologically relevant information, NGS provides an opportunity to implement a molecular surveillance of infectious diseases for accurate and comprehensive description of disease dynamics, detection of transmissions, monitoring of epidemic progress, and providing informed guidance for planning public health interventions.
- However, the greatest scientific strength of NGS systems is their greatest weakness in practice. Collection and processing of the volume of NGS data presents significant challenges and stress to legacy information technology (IT) systems. To be able to transfer, analyze, and store the torrent of data requires a fundamental rethink of the public health IT infrastructure status quo. We present GHOST, an IT system designed to collect, and analyze large scale data.

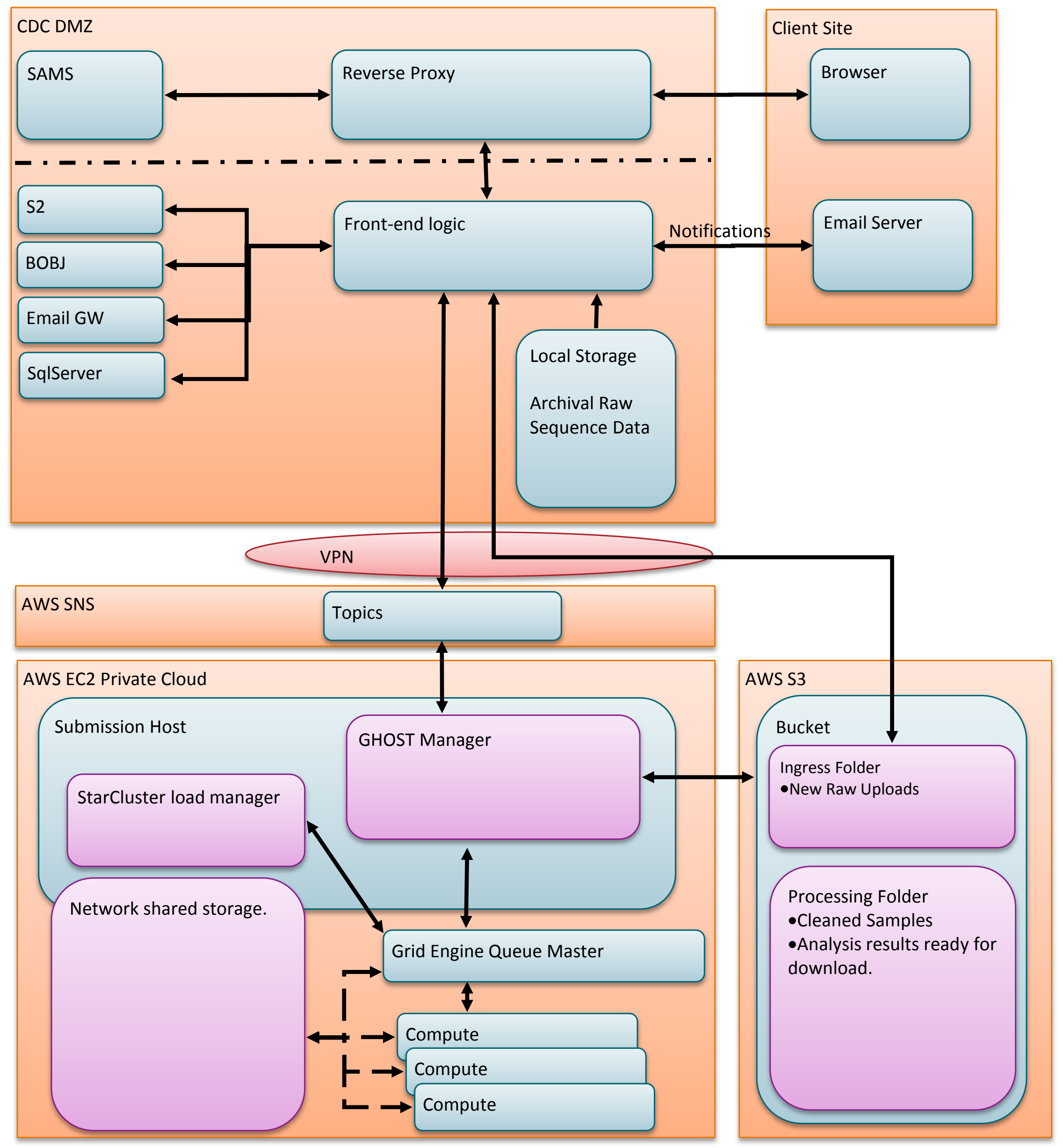
Challenges

- Build a system that is simple to use and requires minimal training.
- Large volumes of data to be stored.
- Work-load is in large bursts rather than spread out over time.
- Analysis step workload scales approximately with the square of the number of samples analyzed. i.e. analyzing 2x more samples requires 4x more work.
- Information Security considerations significantly limit the allowable system structure.
- High availability, fault tolerance, and disaster recovery.
- Pilot the new FedRAMP Amazon cloud with the help of ITSO.

Solutions

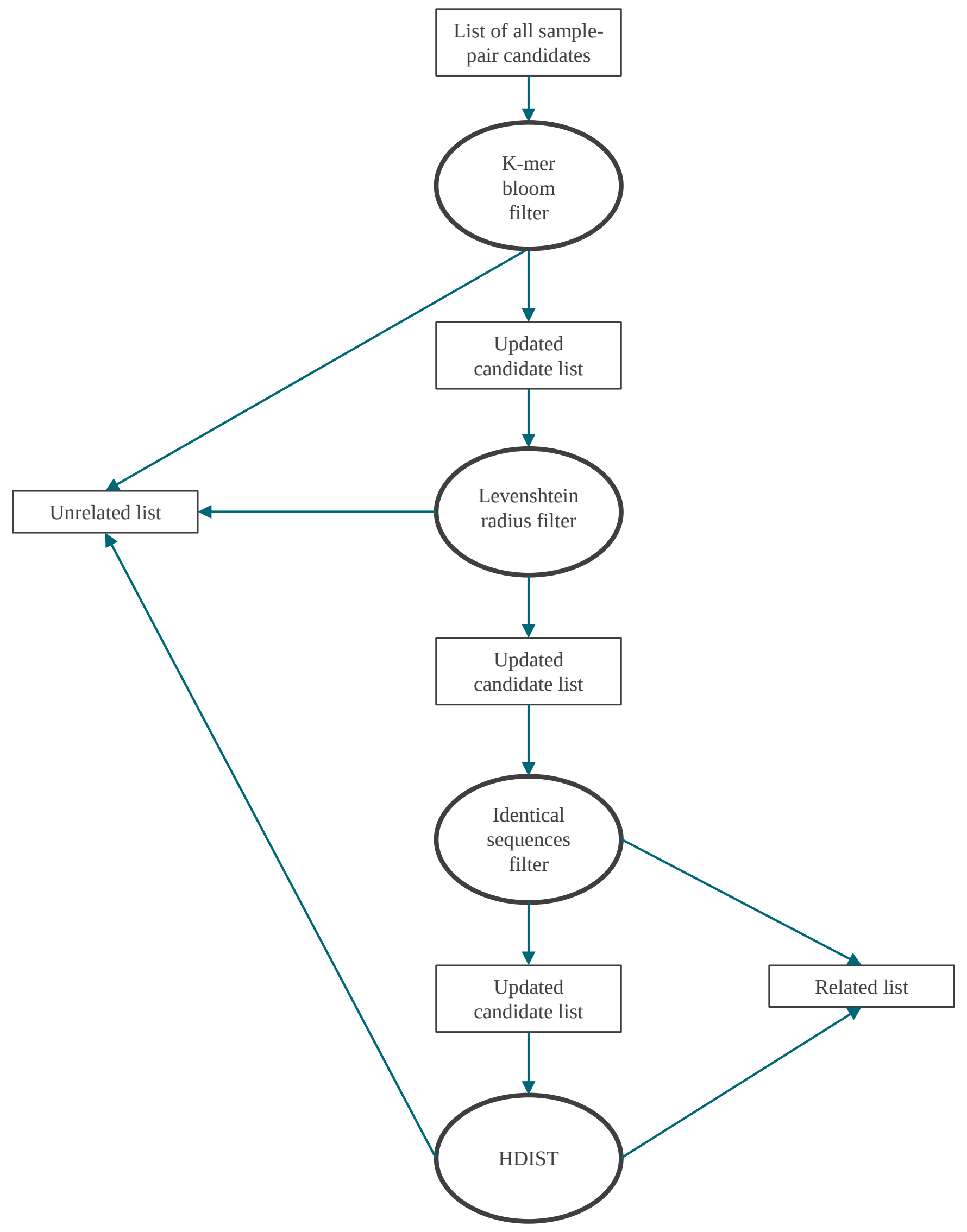
- Build a system that is simple to use and requires minimal training
- Hybrid design with user management and presentation at CDC, together with high computational loads in the Amazon cloud.
- Simple web based interface with graph based visualization.
- Storage of raw data at CDC; cloud storage for computational data.
- Use the Amazon cloud's scalability and judiciously designed heuristic algorithms which reduce the workload.
- Standardized laboratory protocol with regional GHOST centers accepting samples.

GHOST Cloud Architecture Diagram



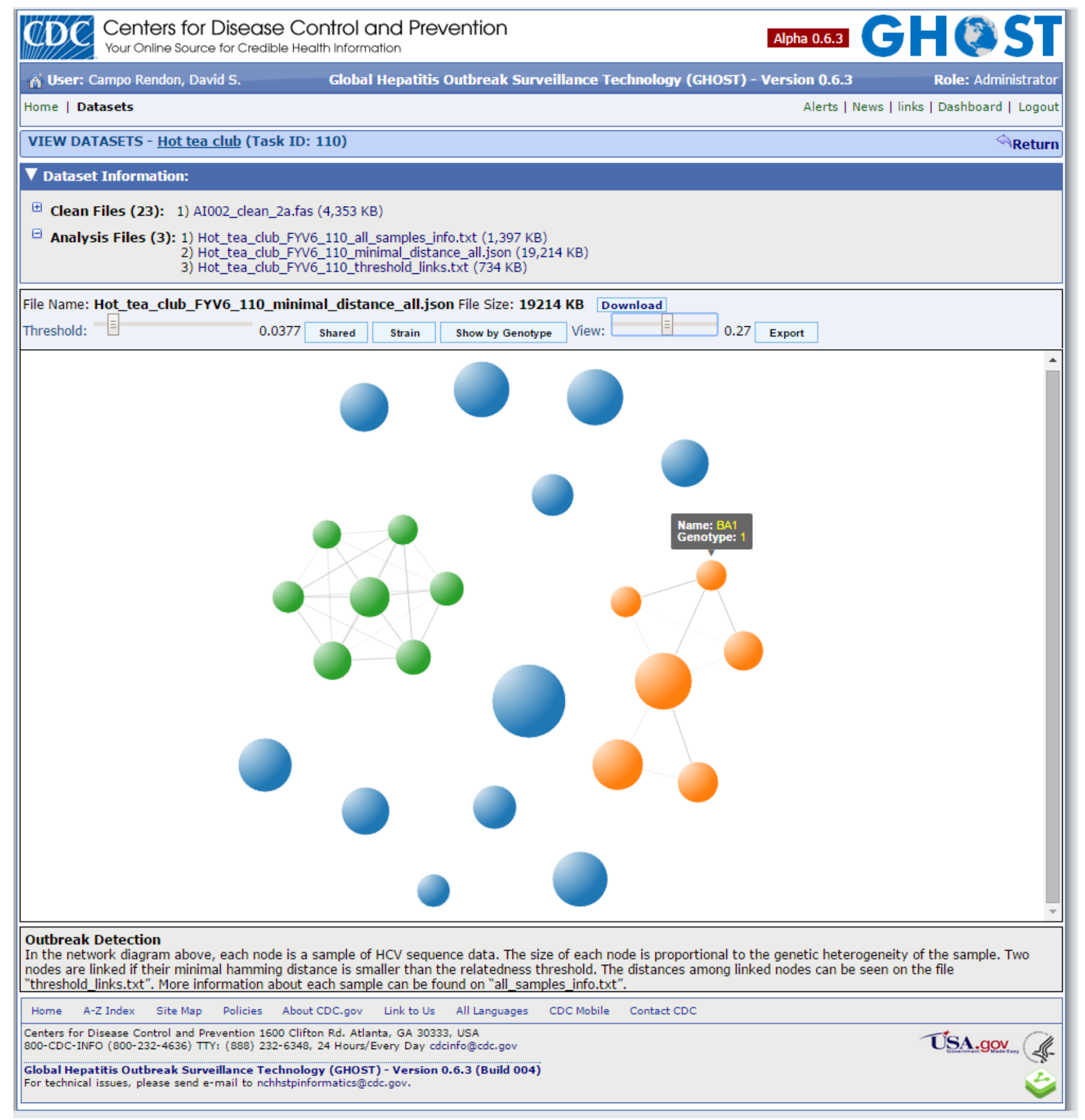
Filtering of candidate pairs

- The naïve, simplest form of our transmission link detection algorithm requires that the set of sequences for each endpoint of a putative transmission link be pooled and aligned. As the most time-consuming step of the pipeline, this represents a significant opportunity for algorithmic optimization.
- As such, we have focused on classifying possible links between samples without resulting to the full distance measurement. We also take advantage of significant pre-computation in a common time/space tradeoff.
- The first filter uses a string metric, the edit distance, to define a sphere in sequence space that encompasses each set. By measuring the distance between the center of a pair of sets, and subtracting the radius of each hyper-sphere, sets that do not overlap are quickly pruned from the list.
- Secondly, for sequence sets to fall below the relatedness threshold they must also share a minimum length of k-mer. Any links that do not share any k-mers are pruned.
- We also check for sets that share an exact sequence, allowing us to add the pair to the set of links with a distance of zero without performing the full distance calculation.



Our fast and efficient three-step filtering strategy removes 91.0% of all pairwise sequence comparisons, accurately establishing which pairs of HCV samples are below the relatedness threshold.

Screenshot of GHOST's visualization page


CDC Centers for Disease Control and Prevention
Your Online Source for Credible Health Information
Alpha 0.6.3
GHOST
Home | Datasets
Alerts | News | links | Dashboard | Logout
Return
Shared
Strain
Show by Genotype
Download
Export
Name: BA1
Genotype: 1
Home A-Z Index Site Map Policies About CDC.gov Link to Us All Languages CDC Mobile Contact CDC
USA.gov

User: Campo Rendon, David S. Global Hepatitis Outbreak Surveillance Technology (GHOST) - Version 0.6.3 Role: Administrator

VIEW DATASETS - Hot tea club (Task ID: 110)

Dataset Information:

Clean Files (23): 1) AI002_clean_2a.fas (4,353 KB)

Analysis Files (3): 1) Hot_tea_club_FYV6_110_all_samples_info.txt (1,397 KB)
2) Hot_tea_club_FYV6_110_minimal_distance_all.json (19,214 KB)
3) Hot_tea_club_FYV6_110_threshold_links.txt (734 KB)

File Name: Hot_tea_club_FYV6_110_minimal_distance_all.json File Size: 19214 KB

Threshold: 0.0377 View: 0.27

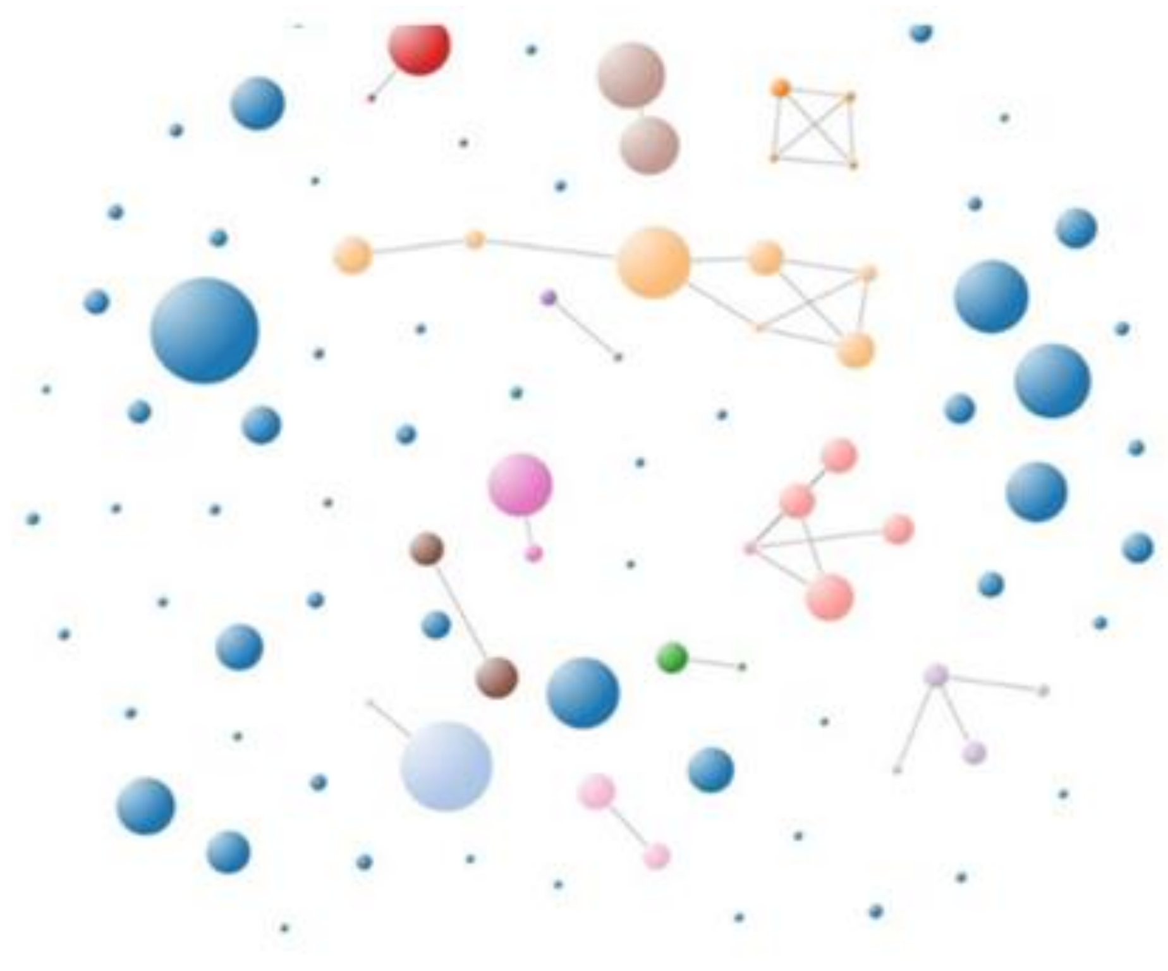
Outbreak Detection
In the network diagram above, each node is a sample of HCV sequence data. The size of each node is proportional to the genetic heterogeneity of the sample. Two nodes are linked if their minimal hamming distance is smaller than the relatedness threshold. The distances among linked nodes can be seen on the file "threshold_links.txt". More information about each sample can be found on "all_samples_info.txt".

Centers for Disease Control and Prevention 1600 Clifton Rd. Atlanta, GA 30333, USA
800-CDC-INFO (800-232-4636) TTY: (888) 232-6348, 24 Hours/Every Day cdcinfo@cdc.gov

Global Hepatitis Outbreak Surveillance Technology (GHOST) - Version 0.6.3 (Build 004)
For technical issues, please send e-mail to nchhstpinformatics@cdc.gov.

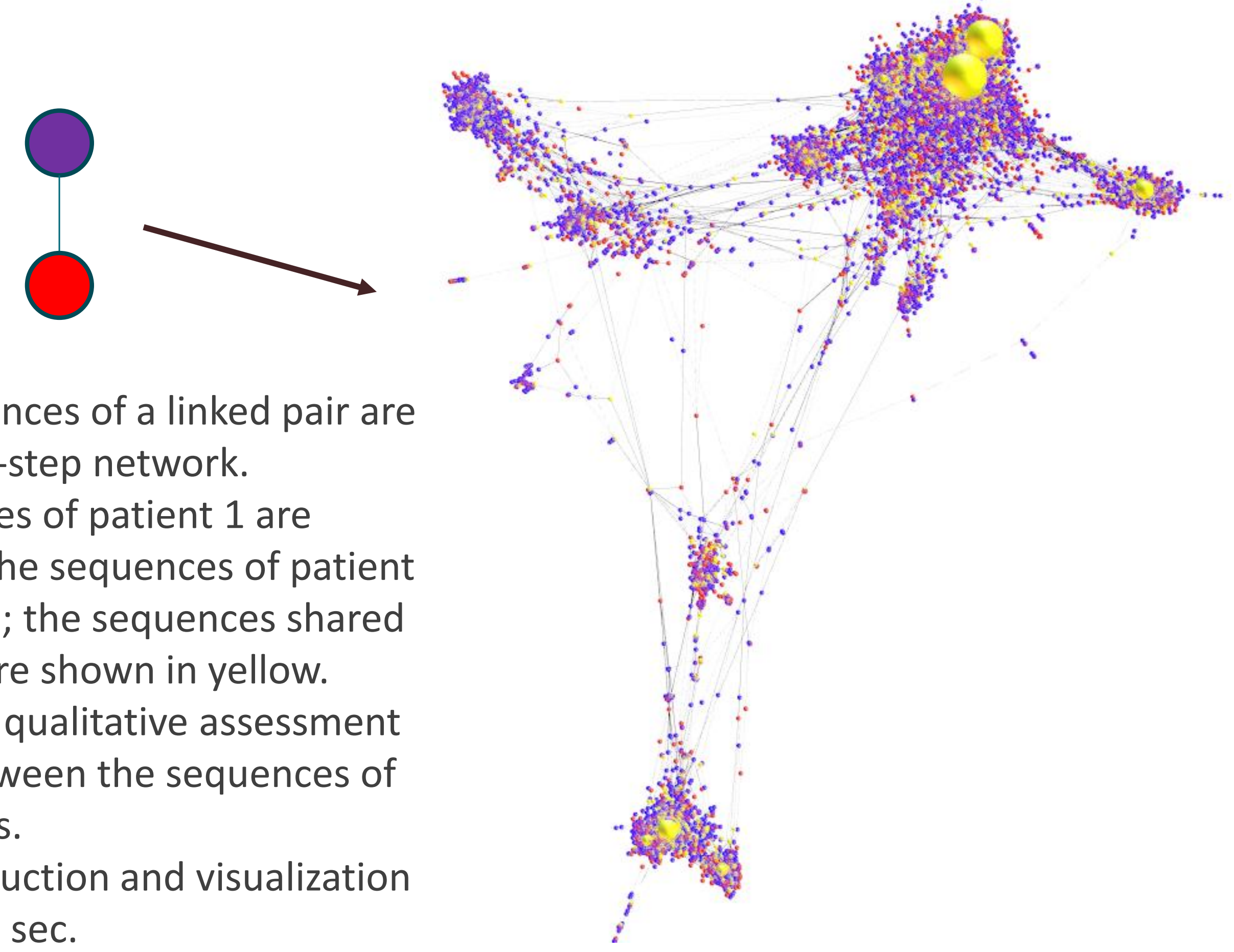
- GHOST is available for beta testers at: https://webappx.cdc.gov/GHOST/

Detail of the Indiana outbreak



- GHOST allowed timely and cost-effective analysis of HCV strains among PWID (n=240) in the recent HIV IN Outbreak.

New visualization tools



- All the sequences of a linked pair are visualized with a k-step network.
- The sequences of patient 1 are shown in purple; the sequences of patient 2 are shown in red; the sequences shared by both patients are shown in yellow.
- This allows a qualitative assessment of the overlap between the sequences of two linked patients.
- k-step construction and visualization are done under 10 sec.

National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention
Division of Viral Hepatitis

www.cdc.gov | Contact CDC at: 1-800-CDC-INFO or www.cdc.gov/info
The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.