National HIV Surveillance System (NHSS)

Attachment 4e.

Detecting HIV Transmission Clusters

Technical Guidance for HIV Surveillance Programs

Detecting HIV

Transmission Clusters

HIV Incidence and Case Surveillance Branch
Atlanta, Georgia

Table of Contents

Acknowledgements

This document was adapted and updated based on a larger document that was first developed in 2016/2017 and revised in 2018. Contributors include:

The Centers for Disease Control and Prevention (CDC) contributors: Anne Marie France, Alexa Oster, Amy Board, Cheryl Ocfemia, Nivedha Panneer, Pat Sweeney, Erica Dunbar, Stacey Muckleroy, Phil Peters, Bill Switzer, Laurie Linley, Meg Watson, Tianchi Zhang, Chenhua Zhang, Neeraja Saduvala, Kyle Bernstein, Donato Clarke, Cindy Getty, Kischa Hampton, Kevin O’Connor, Tobey Sapiano, Sheryl Lyss, Richard Kline, Michal LaFlam

Health Resources and Services Administration (HRSA) contributors: Susan Robilotto, Marlene Matosky

Health Department contributors: Bridget Anderson, New York State Department of Health; Kathleen Brady, Philadelphia Department of Public Health; Mary-Grace Brandt, Michigan Department of Health and Human Services; Heidi Jenkins, Connecticut Department of Public Health; Melissa Miller, Philadelphia Department of Public Health; Analise Monterosso, Texas Department of State Health Services; Jen Reuer, Washington Department of Health; Heather Bronson, Virginia Department of Health; Lauren Ostrenga, Louisiana Department of Health; Randall Collura, New York State Department of Health; Lucia Torian, New York City Department of Health and Mental Hygiene

Academic contributors: Joel Wertheim, University of California, San Diego

Introduction

A critical step toward bringing the nation closer to the goal of no new infections is identifying and responding to clusters of active, ongoing HIV transmission. HIV transmission clusters are groups of persons with HIV who have an epidemiological connection related to HIV transmission; clusters include persons with diagnosed or undiagnosed HIV. Such clusters can be identified through multiple approaches, including partner services, astute providers, and surveillance, including HIV nucleotide sequence data reported as part of HIV surveillance. Evidence shows that HIV surveillance can identify transmission clusters that would otherwise go unrecognized. Information about these transmission clusters and the associated risk networks can help us to focus proven HIV prevention tools where they are needed most. In this way, expanded use of HIV surveillance has the potential to significantly improve HIV prevention efforts.

This document describes the use of HIV surveillance data to detect transmission clusters through the identification of HIV diagnoses clustered in time and space (i.e., time-space clusters) and clusters of HIV infections with closely related strains (i.e., molecular clusters). It also describes the mechanisms behind detecting molecular clusters using HIV nucleotide sequence data and the relationship of a molecular cluster to the underlying transmission cluster and risk network. A brief introduction into the methodology behind time-space cluster detection is also presented.

The focus of this document is cluster detection; cluster response is covered elsewhere. See the CDC's HIV cluster and outbreak detection and response webpage for accompanying tools and resources regarding cluster and outbreak response.

Definitions and context

What is a transmission cluster?

- A **transmission cluster** is a group of persons with HIV (diagnosed or undiagnosed HIV) who are connected by HIV transmission. Transmission clusters can represent recent and ongoing HIV transmission in a population, and prevention efforts could prevent new infections. Section 2, How can identifying transmission clusters help focus prevention efforts?, describes the importance of identifying transmission clusters for prevention efforts in more detail.
- A transmission cluster represents a subset of a **risk network**. A risk network includes the group of persons among whom HIV transmission has occurred and could be ongoing. This network includes persons who are not HIV infected but may be vulnerable to infection, as well as persons with HIV who are in the transmission cluster. Transmission clusters present opportunities for prevention in the larger risk network.
- Transmission clusters can be identified through multiple mechanisms:
  - **HIV case surveillance data.** An increase in diagnoses in a particular geographic area or population (i.e., a time-space cluster). In areas with low incidence of HIV (like many rural communities in the United States), transmission clusters might be more easily detected through HIV case surveillance. Improved timeliness of reporting may improve a jurisdiction's ability to detect a transmission cluster. It is important to note, however, that an increase in the number of diagnoses may not reflect an increase in transmission. Rather, an increase in diagnoses may reflect an increase in HIV testing that has diagnosed infections that may be longstanding. The use of HIV

case surveillance data to identify transmission clusters is discussed in more detail in Section 3, Identifying growing transmission clusters using surveillance data.

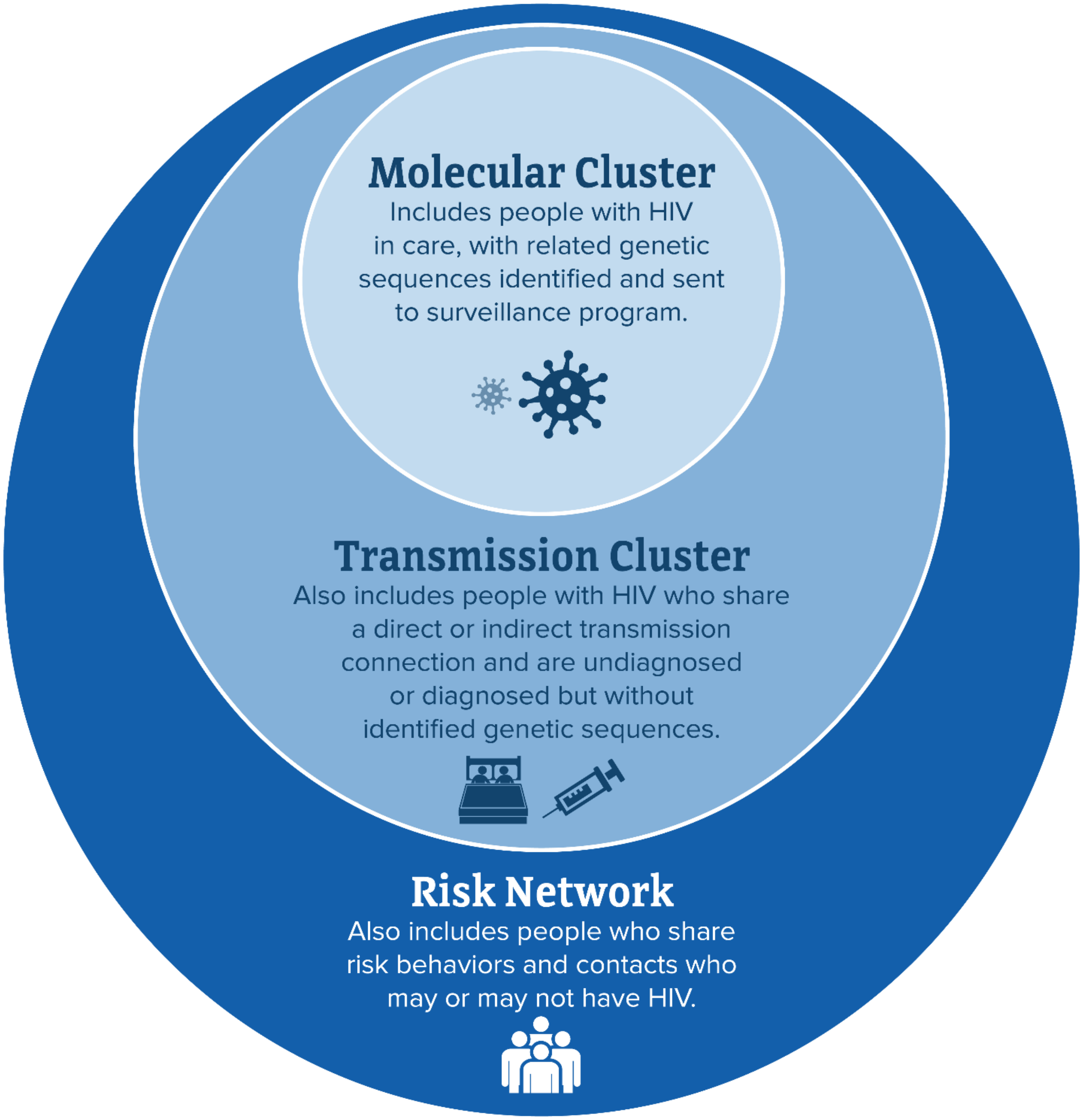
  - **HIV partner services and contact investigations.** Partner services staff (referred to as disease intervention specialists [DIS] in many jurisdictions) routinely perform investigations for persons with newly diagnosed HIV infection, interviewing them to elicit information about their partners, who can then be confidentially notified of possible exposure and potential risk. Partner services activities can also include prevention counseling, testing for HIV and STDs, and linkage or referral to medical care. Because DIS work intensively in local communities, they are positioned to notice unexpected patterns or increases in HIV diagnoses.
  - **Molecular HIV sequence data.** Analysis of molecular HIV sequence data reported to surveillance can identify clusters of cases with closely related HIV strains (i.e., molecular clusters). This method may be particularly useful in identifying transmission clusters that are not detected through other mechanisms. Examples include transmission clusters occurring in an area with a high incidence of HIV infection, that involve multiple jurisdictions, or that are in populations in which persons do not provide contact tracing information to DIS.
  - **Astute health department staff, care providers, or community members.** HIV transmission clusters may be initially detected through astute observations from frontline staff at the health department or clinical providers. Observations of increases in HIV diagnoses call for further investigation to determine if and how these persons are connected and the extent of other connections they may have in a community.

What is a molecular cluster, and how does it relate to a transmission cluster?

- Identification of **molecular clusters** provides a tool to identify transmission clusters. A **molecular cluster** is a group of persons with diagnosed and genetically similar HIV infection. HIV is constantly evolving; therefore, persons whose HIV infections are genetically similar may be closely related by transmission. For more information on HIV evolution, see Appendix B.
- A **molecular cluster** contains only those people for whom molecular data are available and can be analyzed; it is typically a subset of a larger transmission cluster**.**
- Molecular clusters are identified through analysis of HIV molecular sequence data, information that is generated from HIV drug-resistance testing. Drug-resistance testing is conducted to identify mutations in HIV associated with resistance to HIV antiretroviral medications and to help the HIV care provider select an appropriate treatment regimen. This testing is recommended for all persons with diagnosed HIV infection and should be conducted at entry to HIV care.
- As a result, molecular clusters include persons with diagnosed HIV infection who have entered care, have had genetic resistance testing, and have had sequences transmitted to the health department for analysis.
- A molecular cluster is typically a subset of a larger transmission cluster, which can also include:
  - Persons with diagnosed HIV infection who do not have a sequence available for analysis because:
    - They did not enter care
    - They entered care, but have not had a genetic resistance test
    - They entered care and have had a genetic resistance test, but the sequence was not transmitted to the health department for analysis, or was of poor quality and could not be analyzed
  - Persons with undiagnosed infection

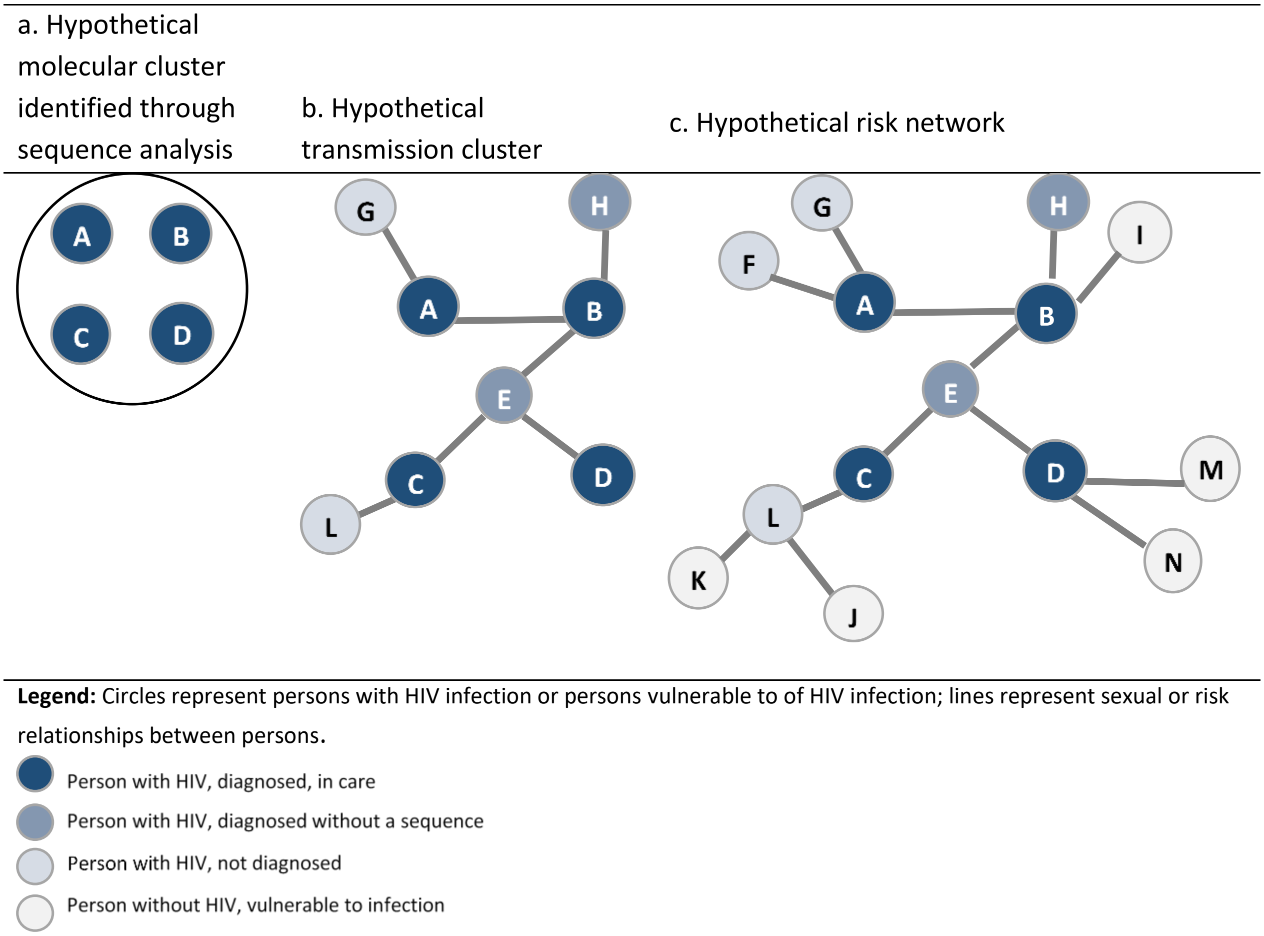
- In addition to the persons in the transmission cluster, the risk network will include:
  - Persons who are not HIV infected but are vulnerable to acquiring HIV

Figure 1-A. Molecular cluster and its transmission cluster and risk network.


Molecular Cluster
Includes people with HIV in care, with related genetic sequences identified and sent to surveillance program.
Transmission Cluster
Also includes people with HIV who share a direct or indirect transmission connection and are undiagnosed or diagnosed but without identified genetic sequences.
Risk Network
Also includes people who share risk behaviors and contacts who may or may not have HIV.

- Molecular data cannot reveal which cases are directly related by transmission or determine the direction of transmission. This limitation is because two persons with genetically similar HIV strains are not necessarily directly linked by transmission: the relationship could be indirect, and there could be unidentified persons involved in transmission relationships.
- Use of molecular sequence data to identify molecular clusters is described in detail in Section 3, Identifying growing transmission clusters by using surveillance data.
- Once a molecular cluster is identified, the corresponding transmission cluster and risk network can only be identified through investigation.

Figure 1-B. Hypothetical molecular cluster (a) and corresponding transmission cluster (b) and risk network (c).



a. Hypothetical molecular cluster identified through sequence analysis

b. Hypothetical transmission cluster

c. Hypothetical risk network

Legend: Circles represent persons with HIV infection or persons vulnerable to of HIV infection; lines represent sexual or risk relationships between persons.

● Person with HIV, diagnosed, in care

◉ Person with HIV, diagnosed without a sequence

◯ Person with HIV, not diagnosed

○ Person without HIV, vulnerable to infection

What is a time-space cluster, and how does it relate to a transmission cluster?

Analysis of case surveillance data to find **time-space clusters** provides another tool to identify transmission clusters. A **time-space cluster** occurs when the number of diagnoses of HIV infection in a particular geographic area exceeds levels expected given previous patterns. In some cases, time-space clusters may reflect one or more transmission clusters that have not yet been identified through molecular data or other approaches.

Why are time-space clusters important?

Surveillance systems should systematically use all data and methods (time-space and molecular sequence-based approaches) available to detect clusters and outbreaks. Reported diagnoses, which are typically more timely and complete, can complement sequence-based techniques by detecting increases in diagnoses clustered in time and space. Routine use of this systematic method in near real time can automate detection of increases in HIV diagnoses that potentially merit further investigation and help state and local health departments prioritize and target HIV prevention efforts for maximal public health impact.

Time-space and molecular cluster detection are complementary

	Time-space cluster detection	Molecular cluster detection
Strengths	• Can be conducted in jurisdictions where sequence data are not yet available • May detect increases in diagnoses concentrated in time and space earlier than molecular clusters can • May detect increases in HIV diagnoses that are not due to a single transmission cluster but are nonetheless concerning	• Can identify transmission clusters that are not geographically concentrated • Can detect transmission clusters in geographic areas where HIV diagnoses overall are decreasing or stable • Correspond to transmission clusters
Limitations	• Clusters that are not concentrated geographically can be missed • Incremental increases in diagnoses across time may not generate an alert • In some jurisdictions, provisional surveillance data could result in observed increases that are later determined to be data artifacts	• May be less timely than time-space cluster detection • Detection limited based on incomplete data (i.e., sequences are only available for persons for whom a resistance test has been ordered, and for whom the test result has been reported to the health department)

Time-space clusters may represent recent and ongoing HIV transmission. In some cases, time-space clusters may reflect transmission clusters that have not yet been identified through molecular data or other approaches. Time-space increases may indicate a single transmission cluster or multiple, smaller transmission clusters, both of which are important to investigate for prevention interventions. Increases in the number of diagnoses may also reflect an increase in HIV testing that has identified longstanding infections, which can also indicate a need for focused prevention efforts. Following the identification of time-space clusters, the review of additional data is important to determine whether investigations and interventions are needed.

For those time-space clusters that appear likely to represent recent and ongoing HIV transmission, steps should be taken to investigate and intervene by using many of the same principles that are outlined in the remainder of this document for molecular clusters.

Analysis of HIV surveillance data to identify time-space clusters can complement analysis of molecular data because time-space clusters can be detected in areas where collection of HIV nucleotide sequences is incomplete or delayed. Time-space cluster detection methods may be particularly useful for subgroups of HIV transmission that might warrant different investigative and intervention approaches because specific analyses can look at time-space clusters specifically among these groups. Notably, infections attributable to injection drug use (IDU) constitute a small proportion of total diagnoses, so the ability to identify potential IDU transmission clusters by analyzing IDU-attributable infections separately is a strength of this method.

Purpose

How can identifying transmission clusters help focus prevention efforts?

- Transmission clusters can identify risk networks that are concerning because of ongoing transmission, poor outcomes, or other reasons, such as transmission in a particularly vulnerable or underserved population, or transmission of drug resistance. Networks of concern include:
  - *Networks in which HIV transmission occurred rapidly* (with multiple new infections occurring within months of one another) and within a recent time window (within ~1–2 years). Recent rapid transmission suggests extremely high-risk transmission networks and could represent an ongoing outbreak; public health intervention could interrupt transmission and prevent future infections.
  - *Networks with characteristics suggesting high potential for ongoing transmission,* such as identification of risk behaviors, including IDU or coinfection with STDs or hepatitis.
  - *Networks characterized by poor outcomes*, such as late diagnosis, lack of viral suppression, or coinfection with STDs, hepatitis, or other comorbities; this could suggest poor access to care and could indicate a network in which persons with HIV infection not yet diagnosed are contributing to ongoing transmission.
  - *Networks representing vulnerable or underserved populations*, such as pregnant women, adolescents, rural populations, persons who inject drugs (PWID), foreign-born persons, or other groups defined by local epidemiology and context.
  - *Networks in which drug-resistant strains of HIV are being transmitted*, particularly networks with resistance to preexposure prophylaxis (PrEP) regimens.
  - *Networks not reached by testing efforts*, as evidenced by large proportions of infections that were diagnosed through incidental testing, such as screening in plasma centers, emergency departments, or correctional institutions; this could indicate other infections in the network that have not yet been diagnosed and are contributing to ongoing transmission.
- Investigation of transmission clusters can identify key characteristics of the risk network to guide intervention efforts to improve outcomes and prevent additional infections.
  - Investigation includes the examination of existing data, including partner services data, or collection of new data to identify factors associated with transmission
- Intervening in risk networks can improve outcomes; activities that interrupt transmission include:
  - Identifying persons with diagnosed HIV infection in the transmission cluster who are out of care, and ensuring that these persons are linked to or re-engaged in care
  - Identifying persons with undiagnosed infection who are part of the transmission cluster, and linking these persons to care
  - Identifying persons without HIV infection in the risk network who are vulnerable to acquiring HIV and offering effective prevention interventions, such as PrEP
  - Interventions at the transmission cluster or population-level to address social-structural or programmatic factors that contributed to transmission (for example, investigation of a transmission cluster could lead to the recognition of gaps in existing prevention programs)
- By expanding our knowledge of transmission dynamics, transmission cluster data can be a powerful tool to help target the interventions we know are effective (engagement in care, HIV testing, PrEP).

HIV sequence reporting

How are molecular sequence data generated and collected?

Generation of molecular sequences

- Molecular sequences are generated through drug-resistance testing.
- Drug-resistance testing is conducted to identify mutations associated with viral resistance to antiretroviral medications and to help the HIV care provider select an appropriate treatment regimen. This testing is recommended for all persons with diagnosed HIV infection and should be conducted at entry to HIV care.
- Drug-resistance testing is typically ordered by providers at entry to HIV care, but can also be ordered at a later time (for example, if a patient is on treatment but does not have a suppressed viral load).
- The final output of drug-resistance testing is a report identifying known mutations that confer drug resistance, which is sent to the care provider. The HIV molecular sequence is generated as a part of the testing process, and laboratories can retrieve this information for surveillance reporting purposes. Current testing methods generate sequences by using a sequencing method called Sanger sequencing.

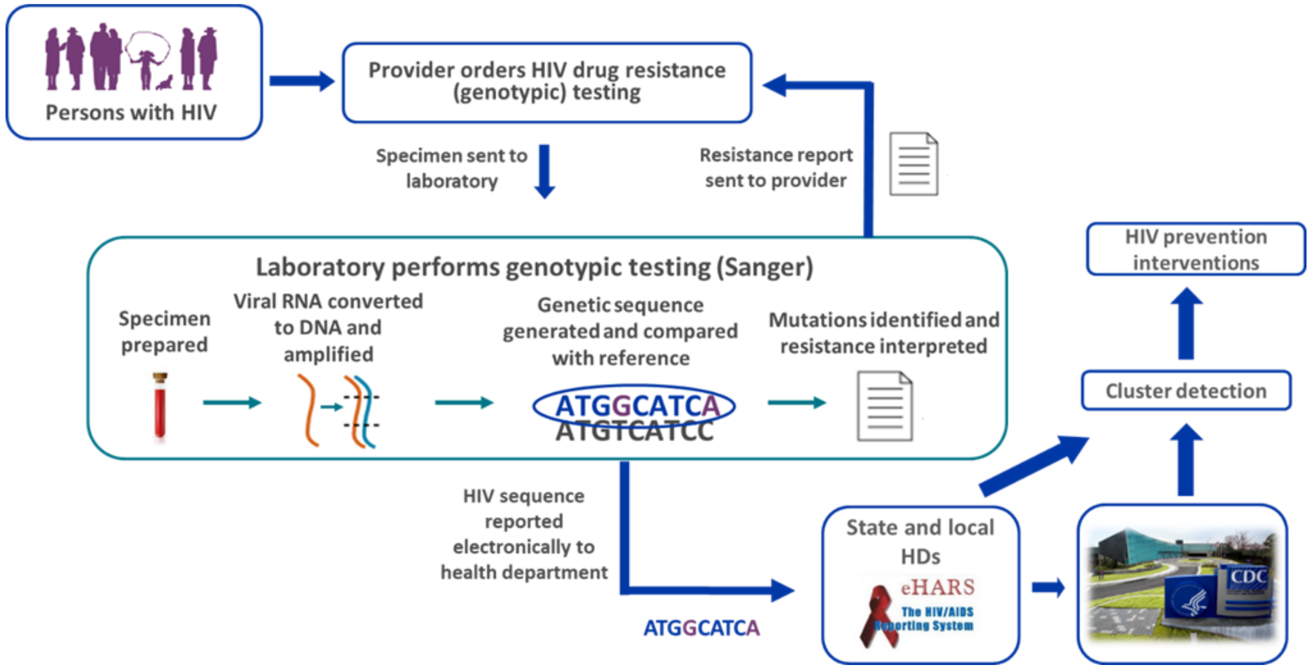
Collection of molecular sequence data

- Laboratories report HIV molecular sequence data to HIV surveillance jurisdictions; these data are an integrated component of the National HIV Surveillance System in all jurisdictions.
- Health departments report all HIV case information collected by HIV surveillance to CDC (demographics, transmission category, CD4 results, viral load results, HIV molecular sequence) without identifying information (name, street address). See Figure 3-A.
- Collection of HIV sequence data is monitored as part of the National HIV Surveillance System; the goal for sequence reporting is ≥60% of persons with diagnosed HIV infection. Achieving high sequence reporting completeness is essential in order to detect clusters and to capture the greatest extent of molecularly linked cases in a cluster. Jurisdictions should refer to the process and outcome standards listed in the file 'Evaluation and Data Quality'
- For more information on nucleotide sequence reporting, see the file 'Reporting'

***PS18-1802 Measure 1.2.11:** ≥60% of cases for a diagnosis year have an analyzable molecular sequence, assessed at 12 months after the diagnosis year.*

Jurisdictions are should achieve molecular HIV sequence reporting of ≥60% of persons with diagnosed HIV infection each year. Achieving high sequence reporting completeness is essential in order to detect clusters and to capture the greatest extent of molecularly linked cases in a cluster.

Figure 3-A. Collection of HIV molecular sequence data


Persons with HIV
Provider orders HIV drug resistance (genotypic) testing
Specimen sent to laboratory
Resistance report sent to provider
Laboratory performs genotypic testing (Sanger)
Specimen prepared
Viral RNA converted to DNA and amplified
Genetic sequence generated and compared with reference
ATGGCATCA
ATGTCATCC
Mutations identified and resistance interpreted
HIV sequence reported electronically to health department
ATGGCATCA
State and local HDs
eHARS
The HIV/AIDS Reporting System
CDC
Cluster detection
HIV prevention interventions

Limitations in molecular sequence data

- Although drug-resistance testing is recommended for all persons with diagnosed HIV infection, not all persons receive a drug-resistance test.
- In some instances, even if a drug-resistance test is completed, reporting challenges prevent a health department from receiving a molecular sequence for a person.
  - For example, in some jurisdictions, sequences may not be reported for persons receiving medical care in federal systems (e.g., Veterans Affairs or federal prisons) or those in blinded clinical trials.
  - In some cases, the identifying and locating information provided by the laboratory could be so incomplete that the sequence cannot be linked to a person in the surveillance data.

Cluster detection methods

Identifying growing transmission clusters by using surveillance data

Routine analyses of surveillance data can identify growing transmission clusters that would otherwise not be identified. Transmission clusters can be identified both through molecular and time-space clusters, and surveillance systems should systematically use all data and methods (time-space and molecular sequence-based approaches) available to detect clusters and outbreaks.

How are molecular clusters identified?

HIV is constantly evolving

- The molecular sequence (also called nucleotide sequence) of HIV accumulates changes over time. Immediately following transmission of HIV between two people, the molecular sequence of the HIV strain in the recipient will be nearly identical to strains found in the transmitting person. As time passes, however, the strains infecting each person will change independently of one another and will look more and more different. In each new person infected, the virus will continue to change independently, so the HIV strains will look less and less similar over the course of a transmission chain. This relationship between the extent of the difference and the relatedness of strains is sometimes referred to as a "molecular clock." For more detail about the evolution of HIV, including the rate of change, please see Appendix B and Appendix C.
- Analysis of the molecular sequence of viral strains can determine how genetically similar the strains are.

Persons infected with viral strains that are genetically similar may be closely related (directly or indirectly) via transmission.

Analysis of sequence data

- Many approaches are available for analyzing HIV sequence data, but the current approach used by CDC that should also be used by state and local health departments is transmission network analysis. In this analysis, each HIV molecular sequence is compared to every other HIV molecular sequence to identify pairs of sequences that are extremely similar (i.e., sequences that have a very small genetic distance, or difference). The level of genetic similarity used to identify closely related pairs is referred to as the **genetic distance threshold**.
    - The genetic distance threshold applied can vary based on the goal of the analysis. For example, to identify cases related by recent and rapid transmission, a very close genetic distance threshold can be used—for example, 0.5% (which corresponds to 5 different nucleotides in a sequence that is 1,000 nucleotides long). A genetic threshold of 0.5% corresponds approximately to 2–3 years of viral evolution separating these strains (which may correspond to time since a common transmission event). By contrast, if the goal is to identify all possible cases that could be related to a given case, a larger genetic distance threshold can be used—for example, 1.5%. A 1.5% threshold corresponds approximately to 7–8 years of viral evolution separating these strains.
- Pairs of infections with similar sequences are then connected with one another to construct transmission networks and identify clusters of very closely related cases.
    - Lines are drawn between each pair of closely related sequences. This creates clusters that may have as few as two connected sequences, but can contain many more sequences that are connected.
    - Although data on potential transmission linkages between persons (i.e., pairs of people with infections that have genetically linked sequences) are useful in constructing molecular clusters, these data may be subject to misinterpretation by those not familiar with this type of analysis. As a result, CDC recommends minimizing use of these data and instead focusing on cluster-level data (i.e., considering all people in a cluster for intervention rather than focusing on people based on their position in the cluster). CDC does not recommend disseminating genetic network diagrams beyond the group of staff involved in the analysis of sequence data.

Limitations of and considerations for visualizing networks based on genetic data
Importantly, although some tools, such as Secure HIV-TRACE and MicrobeTrace, generate network diagrams of clusters based on genetic distance data, there are important limitations to drawing inferences from these data at an individual level. Although two persons infected with highly similar HIV strains could be directly linked through transmission, other transmission relationships could be consistent with this sequence similarity: both could have been infected from a third source, or they could be connected indirectly through a transmission chain including 1 or more intermediaries. Because of this scientific uncertainty, the potential for the misuse and misinterpretation of these data presents a concern. Moreover, presence of or patterns of linkages can be affected by timing of diagnoses and drug-resistance testing. Although analysis of molecular data to identify growing transmission clusters can identify important opportunities for individual- and cluster-level public health interventions, inferences about specific transmission linkages or indirect inferences about sexual or other risk behaviors should not be used to guide services or follow-up at the individual level. Because of the potential for misinterpretation of these diagrams, **it is not recommend to disseminate genetic network diagrams beyond the group of staff involved in the analysis of sequence data.**

- The period of data included in the analysis may vary depending upon the goals. CDC recommends that analyses focused on identifying clusters that represent recent HIV transmission include only sequences from infections diagnosed in recent years (e.g., the 3 most recent years); this entails restricting data to the most recent 3 years of diagnoses prior to analysis in Secure HIV-TRACE (HIV TRAnsmission Cluster Engine).
    - Using a shorter period, such as 3 years, can identify bursts of recently infected cases that indicate recent and potentially ongoing transmission. Although persons with diagnoses outside of the time window who are out of care could be sources of ongoing transmission, limiting the time window allows the analysis to flag clusters with substantial recent growth. A secondary analysis can then be conducted to identify additional potentially related persons with HIV who might be considered in the investigation.
    - Analysis conducted for other purposes, such as understanding a larger transmission network, might include cases diagnosed over a much longer period.
- For details about HIV sequence data analysis, including the selection of the genetic distance threshold to define a cluster, a description of the regions of the HIV genome included in the analysis, and other technical details, please see Appendix B and Appendix C.

How can jurisdictions identify molecular clusters?

- Analysis of molecular sequence data by state and local HIV surveillance programs can allow for identification of clusters in closer to real time and for monitoring of clusters that have been previously identified. Barriers to reporting and processing of HIV sequences to allow prompt identification of growing clusters should be addressed.
- Secure HIV-TRACE (supported by CDC, University of California, San Diego, and Temple University) is a bioinformatics tool that allows HIV surveillance programs to detect, analyze, and visualize clusters. This tool is available to jurisdictions to conduct analyses locally.
    - HIV surveillance programs should analyze their data by using Secure HIV-TRACE at least monthly.

        - HIV surveillance programs can obtain additional information and technical assistance related to Secure HIV-TRACE by emailing hivtrace@ucsd.edu.
    - Separately funded city/county HIV surveillance programs should collaborate with their respective state health department to develop standard analysis protocols (i.e., to determine if and when data will be analyzed separately or jointly).
    - An enhancement to facilitate the identification of multijurisdictional clusters in Secure HIV-TRACE is currently in development. CDC will routinely analyze national data to identify clusters that involve cases from multiple jurisdictions.

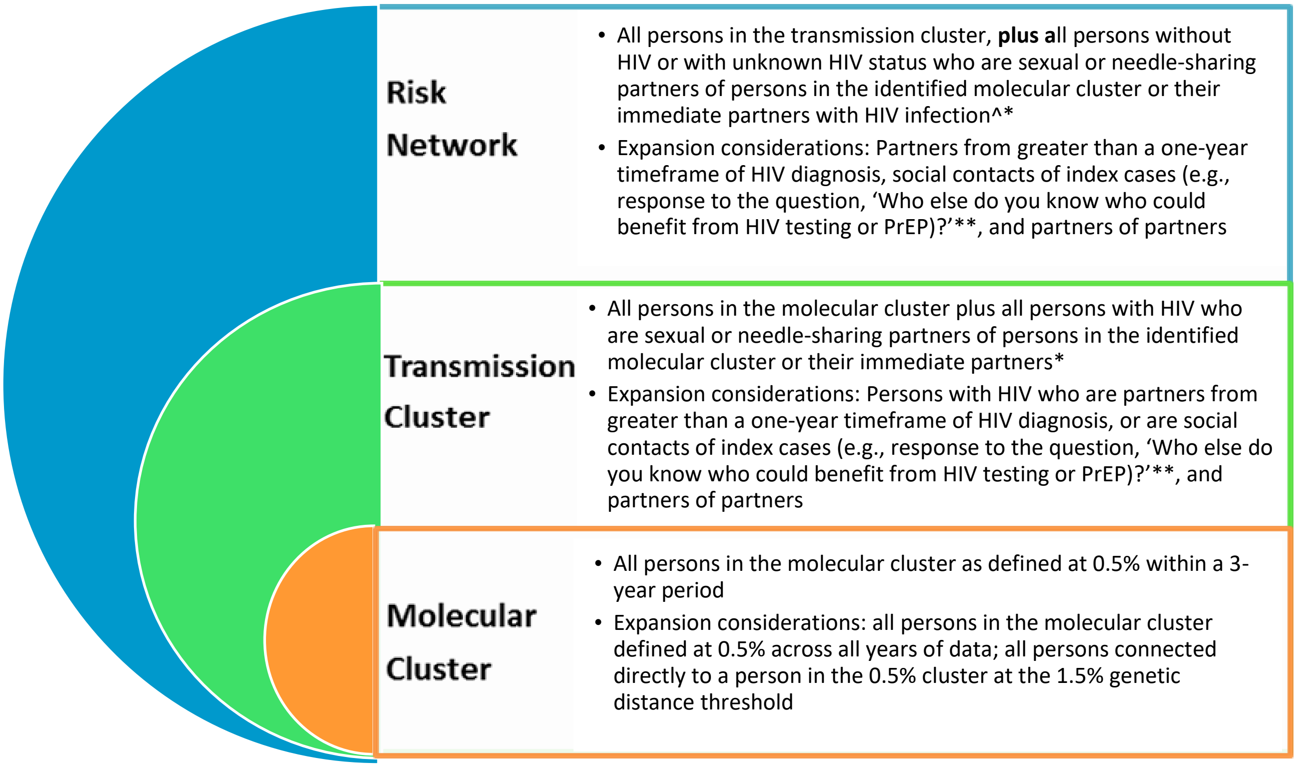
- When partnering with external collaborators (e.g., academic institutions) to analyze sequence data to identify clusters, ensure that the jurisdiction's protocols consider key factors, including data sharing and security and confidentiality of HIV-related information. Such collaborations should have a legitimate public health purpose and support the jurisdiction's HIV prevention efforts, use de-identified data if data are shared, and involve the minimum amount of information necessary. It is recommended that the parameters of such collaborations be outlined in a written project plan or agreement (e.g., a data use agreement [DUA], memoranda of agreement [MOA], memoranda of understanding [MOU], or business contract if applicable). Agreements should include a description of the project and goals, methods, data elements, access and storage requirements, roles and responsibilities, confidentiality and security provisions, disposition of the data, and a description of the dissemination plan or products. Some collaborations determined to be nonresearch or exempt from Institutional Review Board (IRB) review should still be approved by the jurisdiction's Overall Responsible Party (ORP) and may benefit from an additional review and vetting by a standing data analysis review group or public health advisory group (e.g., community planning group or advisory board) or ad hoc review group. Determining whether proposed data sharing with an academic institution supports public health and a jurisdiction's HIV prevention efforts often involves ethical and legal questions. Consulting with an autonomous body of persons experienced in public health ethics may provide insight and feedback on proposed activities.
- Any use of sequence data for research purposes must be carefully considered, be approved by the jurisdiction's ORP, and be subject to IRB approval as appropriate. Any use of identifiable surveillance data as part of any research is contingent on a clearly stated public health purpose, a demonstrated need for identifiable data, IRB and ORP approval, and the signing of a confidentiality agreement regarding rules of access and final disposition of the data and plans for dissemination or publication of results. Depending on the amount and type of data requested, the use of nonidentifiable data for research is generally permissible but because of the sensitive nature of cluster analyses, IRB approval is recommended and may be required. For more information, see Data Security and Confidentiality Guidelines for HIV, Viral Hepatitis, Sexually Transmitted Diseases, and Tuberculosis Programs.

How can decisions in the analysis to identify molecular clusters impact cluster prioritization?

Key decisions in the analyses to identify clusters can have a large impact on the number and composition of clusters identified. For example, analysis using a larger genetic distance threshold (e.g., 1.5%) can identify clusters where some or all transmissions occurred in the more distant past, and where transmission connections between cases are more likely to be indirect; these clusters would likely include more persons and be more intensive to investigate. Additionally, analysis conducted by using datasets that include cases diagnosed over many years may result in the identification of large, complex clusters comprised of many independent transmission chains, where investigation and intervention could be challenging. In general, to focus on recent and rapid transmission, we recommend using a smaller genetic distance threshold (0.5%) to identify clusters of persons with infections that are more closely temporally linked, and limiting analyses to cases diagnosed in the most recent 3-year period. CDC's criteria to identify priority clusters includes a 0.5% genetic distance threshold and the most recent 3-year period of data. Analyses of national data demonstrate the importance of these restrictions in focusing detection on clusters that represent recent and rapid transmission[1].

After a cluster that represents recent and rapid transmission has been identified, the definition of that molecular cluster can be expanded to be more inclusive and to identify other persons who might be related to this cluster. An expanded molecular cluster definition can, for example, capture persons who are likely closely connected to the transmission cluster, but whose sequence, while related, does not meet the 0.5% genetic distance threshold. Figure 3-B offers considerations for expanded definitions of the molecular cluster, transmission cluster, and risk network.

Figure 3-B. Suggested definitions for molecular clusters, transmission clusters, and risk networks, with considerations for expansion.



Risk Network

- All persons in the transmission cluster, **plus a**ll persons without HIV or with unknown HIV status who are sexual or needle-sharing partners of persons in the identified molecular cluster or their immediate partners with HIV infection^*
- Expansion considerations: Partners from greater than a one-year timeframe of HIV diagnosis, social contacts of index cases (e.g., response to the question, 'Who else do you know who could benefit from HIV testing or PrEP)?'**, and partners of partners

Transmission Cluster

- All persons in the molecular cluster plus all persons with HIV who are sexual or needle-sharing partners of persons in the identified molecular cluster or their immediate partners*
- Expansion considerations: Persons with HIV who are partners from greater than a one-year timeframe of HIV diagnosis, or are social contacts of index cases (e.g., response to the question, 'Who else do you know who could benefit from HIV testing or PrEP)?'**, and partners of partners

Molecular Cluster

- All persons in the molecular cluster as defined at 0.5% within a 3-year period
- Expansion considerations: all persons in the molecular cluster defined at 0.5% across all years of data; all persons connected directly to a person in the 0.5% cluster at the 1.5% genetic distance threshold

*Identified as partners within a 1-year timeframe of HIV diagnosis, or at any time following HIV diagnosis during which the index case was not virally suppressed.

** Note that in some cases, people with HIV who are partners identified through partner services might have discordant molecular sequences. For purposes of public health response, these persons should still be included in the transmission cluster.

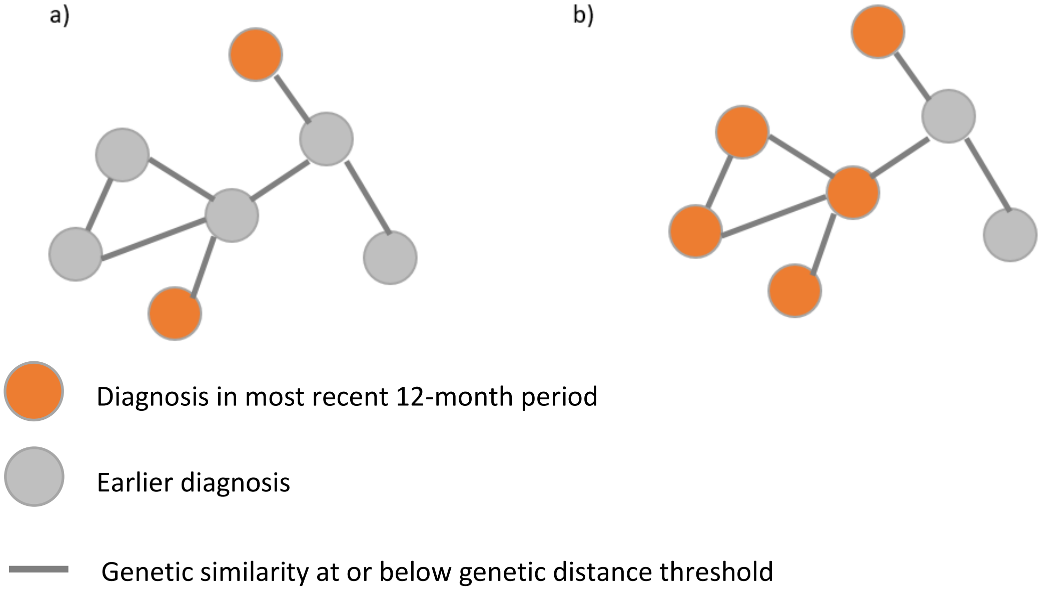
[1] Oster AM, France AM, Panneer N, Bañez Ocfemia MC, Campbell E, Dasgupta S, Switzer WM, Wertheim JO, Hernandez AL. Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data. *J Acquir Immune Defic Syndr* 2018. doi:10.1097/QAI.0000000000001856.

^*Note that the risk network will include all claimed partners, even if these persons were not named, did not have sufficient information to initiate contact, or cannot be located.

How does CDC identify molecular clusters?

- CDC conducts routine analyses to identify clusters that are concerning for recent and rapid transmission of HIV. CDC analyses will not be as timely as local analyses, because of the timeline for submission of data to CDC and data processing prior to the data becoming available for analysis. For clusters detected by the jurisdiction, CDC analyses may detect additional members from other jurisdictions, and in some cases CDC analyses will detect clusters that were not detected by any jurisdiction.
- Analyses are conducted by using national data that are available each quarter (based on data transmitted by HIV surveillance jurisdictions to CDC in March, June, September, and December).
- Prior to analysis, all HIV sequences in the national dataset are evaluated to determine the quality of the data and to remove potential contaminants. Only sequences that include protease or reverse transcriptase regions of the HIV genome and are of sufficient length are included in the analysis.
- CDC analyzes data by using a secure, local installation of HIV-TRACE, a software tool developed by University of California, San Diego and Temple University.
- With the goal of identifying clusters consistent with recent and rapid HIV transmission, these analyses include only cases diagnosed in the most recent 3-year period, and use a genetic distance threshold of 0.5%.
- National priority clusters are defined based on the burden of HIV in the jurisdiction. For lower burden jurisdictions (defined by membership in CDC's low-burden jurisdiction workgroup), priority clusters are defined as clusters with at least 3 cases diagnosed within the most recent 12-month period. For all other jurisdictions, priority clusters are defined as those with at least 5 cases diagnosed within the most recent 12-month period. Many clusters cross jurisdictional boundaries; in these cases, the priority cluster determination is made based on the number of cases diagnosed in the cluster overall, regardless of jurisdiction.
- When a cluster of concern is identified, the primary jurisdiction (the jurisdiction with the majority of cases in a cluster) is notified and a cluster snapshot or line list describing the cluster is transmitted securely via SAMS. This cluster snapshot or line list will include case count information for all cases in the cluster, regardless of jurisdiction, but will only include line-listed information for persons in the primary jurisdiction unless a specific data sharing agreement between the jurisdictions involved and CDC allows this information to be shared. A cluster snapshot companion document, showing the elements included in a cluster snapshot, can be found in Appendix D. Jurisdictions that are notified that they have persons involved in a cluster but are not the primary jurisdiction will have access to the statenos of their cases in the cluster through CDC, however the mechanism for routinely sharing this information is still being determined.
- CDC's prioritization criteria may be modified and expanded in the future, as capacity allows.

Figure 3-C. Examples of clusters that would not (a) and would (b) meet national priority criteria.

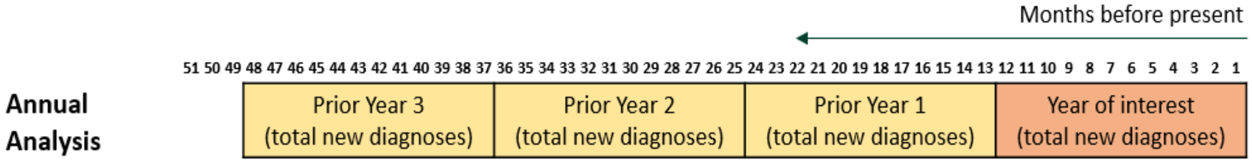

a)
b)
Diagnosis in most recent 12-month period
Earlier diagnosis
Genetic similarity at or below genetic distance threshold

How are time-space clusters identified?

- CDC, with input from some jurisdictions who conduct this type of routine analysis, has developed methods for analysis of surveillance data to detect unusual increases or changes in normal HIV diagnosis and reporting patterns.
- Jurisdictions should conduct time-space analysis locally, at least monthly.
- CDC provides a SAS program that jurisdictions can use to conduct time-space analysis at the local level. Jurisdictions may choose to use the CDC-provided SAS program, adapt it for their local purposes, or develop other approaches to identifying time-space clusters. The CDC SAS program implements the current approach:
    - Define the period of interest for analysis as the most recent 12 months of HIV diagnosis (e.g. Jan 2017–Dec 2017).
    - Define the comparison group as the previous 36 months (e.g. Jan 2014–Dec 2016).
    - Define the geographic area of interest. At a minimum, this should include the state and each county within the state. Other geographic areas of interest may include regions within the state, metropolitan statistical areas, etc.
    - Populations for time-space analysis:
        - Time-space analyses can be conducted for all diagnoses or for specific risk groups. CDC recommends that, at a minimum, jurisdictions conduct time-space analyses for:
            - Overall diagnoses
            - Diagnoses among PWID
                - Note that these analyses could be conducted for IDU only, MSM-IDU only, or IDU + MSM-IDU
            - Analyses could also be conducted for other risk groups. Note that analyses of diagnoses among MSM could be with or without MSM-IDU.

- Performing time-space analysis with the CDC-provided SAS program requires the following steps:
  - Calculate the HIV case counts for each county (or other relevant geographic area) for the most recent 12 months (or other period of interest). Jurisdictions must also calculate the average HIV case counts per year for the same areas for the previous 36 months.
  - Calculate the standard deviation for the mean number of cases during the 36-month comparison group.
  - Construct an interval of 2 standard deviations around the mean.
  - Compare the results to the most recent 12 months of data. The CDC-provided SAS code creates an “alert” for case counts that fall more than two standard deviations above the mean, as well as an increase of more than 2 diagnoses over the baseline. Jurisdictions may add additional, more stringent criteria in defining geographic and time windows to suit their needs.
- Although the primary responsibility for time-space analysis is with jurisdictions, CDC will also routinely conduct these analyses to identify clusters crossing jurisdictions.
- For very low morbidity jurisdictions, a routine manual review of data to identify increases, rather than an analytic approach, may be sufficient.

Figure 3-D. Example of analysis to identify time-space clusters, comparing number of diagnoses in the most recent 12-month period to 36-month baseline



Next steps once a cluster has been identified

Identifying transmission clusters is an important first step that should be followed by investigation and response to interrupt the spread of HIV. CDC has developed additional guidance tools and resources to assist jurisdictions with additional steps in this process; see CDC’s HIV cluster and outbreak detection and response homepage.

Outcome measures related to cluster detection

PS 18-1802 includes outcome standards related to sequence reporting, cluster detection, and cluster response. The outcome standard for nucleotide sequence reporting is described in the file ‘Evaluation and Data Quality’. The outcome standard related to cluster detection is described below. Additional outcome standards related to cluster response can be found in the PS18-1802 EPMP, under Strategy 3.

Measure 3.1.1: Analyze surveillance and other data using CDC-recommended approaches at least monthly to identify HIV transmission clusters and outbreaks.

Description: Local jurisdictions must use secure HIV-TRACE or other CDC-recommended approaches to conduct molecular analysis of HIV sequence data at least monthly, provided new data are available on a monthly basis. (For jurisdictions that have not collected HIV sequence data prior to PS18-1802, analysis of sequence data must occur once sequence data begin to be collected.) It is also required that jurisdictions analyze HIV case surveillance data monthly to identify time-space clusters in HIV diagnoses. This is an activity that can begin in all jurisdictions, including those that do not yet have molecular sequence data available. Analysis results of both molecular and time-space clusters will be used to identify new transmission clusters and to monitor the growth of existing clusters.

City-level jurisdictions may defer to the state to run this analysis monthly, as long as an agreement is in place between both parties. City-level jurisdictions must work with the state to ensure timely reporting of identified clusters and to develop plans for investigation.

Target: N/A

Data to be reported:

- Timing of analyses (when analyses were conducted)
- Type of analysis conducted (molecular and/or time-space)

Methods of Reporting Data: Jurisdictions will use the SER to report whether the analyses were conducted. Secure HIV-TRACE reports will be used to confirm the information provided in the SER.

Appendix A: List of abbreviations and key definitions

Cluster snapshot

A document developed by the CDC HIV Incidence and Case Surveillance Branch to communicate cluster and case-level data on a molecular cluster to state and local health departments.

Disease intervention specialists (DIS)

Health department personnel who are specifically trained to provide partner services. Some health departments use different titles for persons providing partner services. In addition, in certain jurisdictions, other persons (e.g., HIV counselors or clinicians) either inside or outside of the health department provide certain or all elements of partner services.

Drug-resistance testing

Conducted in order to identify mutations associated with viral resistance to antiretroviral medications and help the HIV care provider select an appropriate treatment regimen. Drug-resistance testing is recommended for all persons with diagnosed HIV infection, with the recommendation that testing be conducted at entry to HIV care. Drug-resistance testing is typically ordered by providers at entry to HIV care, but can also be ordered at later time points (for example, if a patient is on treatment but does not have a suppressed viral load). A nucleotide sequence is generated as an intermediate byproduct from a drug-resistance test.

Engagement in care

Measured by whether a person with diagnosed HIV infection has a had at least one HIV medical care visit during the analysis period

Genetic distance threshold

The level of genetic similarity used to identify closely related pairs of sequences. The genetic distance threshold used can vary based on the goal of the analysis.

HIV TRAnsmission Cluster Engine (HIV-TRACE)

A bioinformatics tool developed by researchers at the University of California, San Diego to analyze nucleotide sequences and identify clusters representing recent and rapid transmission. A secure local installation of HIV-TRACE at CDC is used to run routine analyses on national surveillance datasets.

Molecular cluster

Identified through analysis of HIV genetic sequence data that is generated through HIV drug-resistance testing. Molecular clusters contain only those people for whom molecular data is available and can be analyzed, and contains a subset of what is likely a larger transmission cluster

Molecular data

See “Nucleotide sequence”

Multijurisdictional cluster

A cluster in which coordination across jurisdictions is required for effective investigation and response. Often, multijurisdictional clusters will include cases reported from multiple states, however the jurisdictional issues involved could be relevant for clusters involving multiple counties within a single state, particularly if they include separately funded HIV surveillance or prevention programs.

National HIV Surveillance System (NHSS)

The primary source for monitoring HIV trends in the United States. The primary functions of the National HIV Surveillance System (NHSS) are (1) to provide accurate epidemiologic data to monitor the incidence and prevalence of HIV infection and HIV-related morbidity and mortality and (2) to use these data trends to assist in public health planning and policy. CDC provides federal funding to states and territories through surveillance cooperative agreements to achieve the goals of NHSS and also to assist states in developing their own surveillance programs in accordance with state and local laws and practices.

National priority cluster

A molecular cluster that has met certain criteria and which should be flagged for preliminary investigation. Currently, CDC-defined national priority clusters for high and medium morbidity jurisdictions are clusters identified at a 0.5% genetic distance threshold with ≥5 cases in the most recent 12-month period. For low morbidity jurisdictions, CDC-defined priority clusters are those identified at a 0.5% genetic distance threshold with ≥3 cases in the most recent 12-month period. Analyses of clusters meeting the abovementioned criteria indicates similar transmission rates that are approximately 11 times that of the transmission rate among HIV infected individuals in the US. In addition to using criteria for CDC-defined priority clusters, jurisdictions may also develop criteria to identify additional, locally defined priority clusters.

Nucleotide sequence

An intermediate byproduct of an HIV drug-resistance test. Analysis of nucleotide sequences can identify pairs of sequences that are extremely similar and which may be closely related in transmission

Partner services A broad array of services that should be offered to persons with HIV infection, syphilis, gonorrhea, or chlamydial infection and their partners. A critical function of partner services is partner notification, a process through which infected persons are interviewed to elicit information about their partners, who can then be confidentially notified of their possible exposure or potential risk. Other functions of partner services include prevention counseling, testing for HIV and other types of STDs (not necessarily limited to syphilis, gonorrhea, and chlamydial infection), hepatitis screening and vaccination, treatment or linkage to medical care, linkage or referral to other prevention services, and linkage or referral to other services (e.g., reproductive health services, prenatal care, substance abuse treatment, social support, housing assistance, legal services, and mental health services).

Preexposure prophylaxis (PrEP) A way for people who do not have HIV but who are at substantial risk of getting it to prevent HIV infection by taking a pill every day

Primary jurisdiction The jurisdiction with the majority of cases in a molecular cluster

PWID Persons who inject drugs

Risk network Includes the group of persons among which HIV transmission has occurred and could be ongoing. This network includes persons who are not infected with HIV but may be vulnerable to infection, as well as the persons with HIV in the transmission cluster

Secure HIV-TRACE A web-based bioinformatics tool developed by researchers at the University of California, San Diego and Temple University to analyze HIV nucleotide sequences and identify molecular clusters. Secure HIV-TRACE is available to individual public health institutions to facilitate real-time analysis by state and local health departments to better understand and respond to their specific HIV burden.

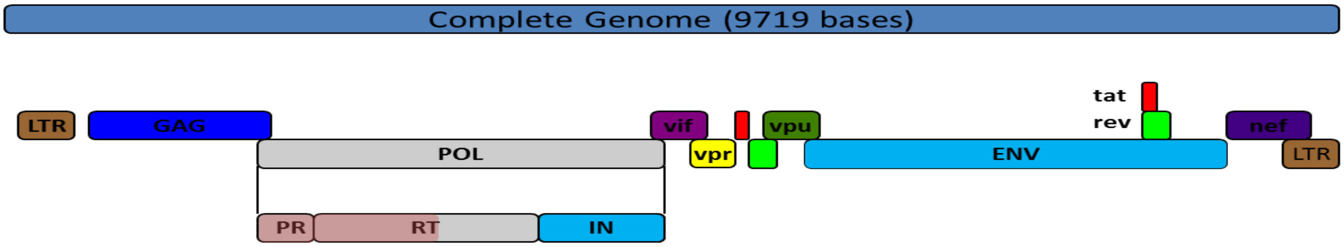
Time-space cluster A time-space cluster occurs when the number of diagnoses of HIV infection in a particular geographic area is elevated above levels expected given previous patterns.

Transmission cluster A group of persons with HIV who are connected by HIV transmission. A transmission cluster represents a subset of a risk network

Appendix B. HIV Molecular Evolution

HIV-1 Genome and Structure

The HIV-1 RNA genome is comprised of approximately 10,000 nucleotides that are the code for 9 to 10 genes that encode for 16 proteins. Tgroup specific antigen (*gag*), polymerase *(pol*) and envelope (*env)* genes encode the information needed to make the structural proteins for new viral particles. The *pol* gene also codes for enzymes for viral replication (reverse transcriptase [RT]) and integration into the host genome (integrase [IN]). The other genes encode for regulatory or accessory proteins that control replication and infectivity.



HIV-1 Genetic Evolution

HIV-1 replicates rapidly, generating about 10 billion viral particles every day in an untreated person. HIV-1 also has a high genomic evolutionary rate ranging from 1.3 x 10^{-3} to 3.5 x 10^{-3} nucleotide substitutions/site/year depending on the HIV-1 subtype and specific gene region examined. For pol, this corresponds to a rate of evolution of 1% every 10 years. The genetic distance that reflects the relative change between HIV sequences can be used as a proxy for the number of years since the HIV sequences diverged from a common ancestor or transmission event. The genetic distance applied can vary based on the goal of the analysis. For example, to identify cases related by recent and rapid transmission, a very close genetic distance threshold should be used—for example, 0.5% (which, for a sequence that is 1000 nucleotides long, corresponds approximately to 5 different nucleotides). A genetic threshold of 0.5% corresponds to approximately a maximum of 5 years of viral evolution (2–3 years for each person, because the virus is evolving in each person) separating these strains (which may correspond to time since a common transmission event). By contrast, if the goal is to identify all possible cases that could be related to a given case, a larger genetic distance threshold should be used—for example, 1.5%. A 1.5% threshold corresponds to a maximum of 15 years of viral evolution separating these strains.

The high substitution rate is believed to be caused by the low fidelity of the RT enzyme during replication and by HIV-1 genome interactions with other cellular enzymes. RT is the enzyme used by HIV to convert single-stranded HIV RNA into double-stranded cDNA allowing integration into the host genome. Because RT does not have a proofreading mechanism, transcription from viral RNA to DNA is error prone. HIV's fast replication cycle and high substitution rate of HIV-1 leads to high genetic diversity, which enables the virus to evade the immune system and to develop drug-resistant mutations.

Mutations have been found in all HIV-1 genes. When considering only the *pol*, *gag*, and *env* genes, there are small sequence regions in each that are considered genetically conserved, because mutations in those regions negatively affect the virus's ability to survive or replicate. In general, the *pol* gene is considered the most conserved gene and *env* is considered the least conserved gene likely due to *env* having a higher substitution rate. Therefore, analyses of regions other than pol may need to consider different genetic distance thresholds.

HIV-1 Subtypes and Minority Strains

HIV-1 can be classified into four groups; of which M is associated with the majority of infections worldwide. Within group M, many distinct subtypes exist (e.g., A, B, C, D, F, G, H, J and K). The sequences within any one subtype are more similar to each other than to sequences from other subtypes. These subtypes represent different lineages of HIV, and have some geographical associations. Additionally, different subtypes can combine genetic material to form a hybrid virus, known as a "circulating recombinant form" (CRF).

Although most persons with HIV are infected with a single variant of HIV-1, rapid error-prone replication over time leads to HIV-1–positive individuals being infected with an enormous pool of genetically related strains called "quasispecies." These quasispecies or variants are closely related viruses with different nucleotide sequences. Within an infected individual, HIV-1 diversity typically consists of a major, dominant strain and other less frequent genetic variants, which can change due to viral fitness, changes in immune response or drug pressure.

Minority strains are normally defined as those variants that are present in less than 20% of the total quasispecies pool. Sanger or bulk sequencing, the most common approach used for HIV drug-resistance testing in clinical settings, detects variants with a frequency of at least 20% of the total viral population. Hence, most minority strains will not be detected when testing for HIV drug resistance by using Sanger sequencing. The clinical significance of minority HIV-1 strains for development of drug resistance is not clear at the present time. About 10%–15% of newly diagnosed patients are infected with strains containing at least one drug-resistant mutation.

HIV gene regions analyzed for detecting antiretroviral drug resistance

Testing for HIV drug-resistance mutations consists of sequencing only specific positions of the *pol* gene that encode enzymes targeted by antiretrovirals, including RT, protease (PR), and integrase (IN). Currently, the PR and RT sequences are analyzed in programs, such as Secure HIV-TRACE, to identify molecular clusters. Commercial, drug-resistance detection assays were developed using HIV-1 subtype B and may not perform well with other subtypes.

Sanger Sequencing vs Next Generation Sequencing (NGS)

Sanger or "bulk" sequencing is the most common method used for HIV drug-resistance testing in clinical settings. This testing is available at commercial labs, but HIV researchers also perform this testing by using in-house assays and analysis. Sanger sequencing was developed by Dr. Frederick Sanger in the 1970s and involves the termination of DNA synthesis by selective incorporation of chain-terminating dideoxynucleotides by DNA polymerase during DNA replication, and separation and visualization of the resulting fragments by capillary electrophoresis and laser detection.

Currently, NGS, deep sequencing or massively parallel sequencing is predominantly used in HIV research studies. NGS methods are based on the "sequencing by synthesis" principle where nucleotides incorporated into a strand of DNA provide a unique signal. The unique signal in most NGS platforms is a fluorescent molecule but can also be a change in pH. NGS can sequence myriad DNA fragments simultaneously in a short period of time and uses bioinformatics programs to piece together and analyze the synthesized sequences. Several NGS platforms are available, each with their own synthesis and detection methods.

Sanger Sequencing	Next Generation Sequencing
DNA synthesis and signal detection are two separate processes and only one DNA strand (forward or reverse) can be read at a time	Synthesis and signal detection occur simultaneously by using multiple DNA templates
Cost per sample is more expensive; one sample per sequencing reaction	Cost is lower and process is faster; can sequence many samples simultaneously
No special bioinformatics infrastructure and storage capacity required	Bioinformatics infrastructure and storage capacity to store and analyze millions of sequence fragments is required
	Higher error rate but getting better
Sequence reads are longer (~700–900 bases per read per sample)	Sequence reads are shorter (< 400 bases per read depending on platform), but many more reads per sample per run are possible
Detects variants with prevalence of/greater than ~20%	Detect minority variants at a prevalence of/less than 1%
Most common method used for HIV drug-resistance testing in clinical settings	Mostly used in research settings

Appendix C. Frequently asked questions about HIV-TRACE and transmission network analysis

Adapted from “Secure HIV-TRACE: a guide for public health departments to reconstructing HIV-1 transmission clusters,” courtesy of Joel Wertheim.

Why pairwise alignment?

Secure HIV-TRACE was designed to detect transmission clusters by analyzing the 1497 nucleotide region spanning the HIV-1 *pro/rt* region common in public health surveillance activities, drug-resistance screening, and research studies. This genomic region is from a conserved genomic region with very limited length variation (unlike, say, *env*) across all major HIV-1 subtypes and circulating recombinant forms. The rarity of insertions and deletions permits robust pairwise alignment to a reference sequence. This approach is a timesaving measure compared with the more computational intensive approach of multiple sequence alignment, because it has linear complexity in the number of sequences; popular multiple sequence alignment algorithms all have superlinear complexity. **Secure HIV-TRACE** uses a modified version of the Smith-Waterman algorithm, which aligns nucleotide sequences by considering amino-acid translations of constituent codons; this approach allows us to make full use of amino-acid conservation to preserve alignment accuracy for divergent sequences (e.g., those from different subtypes).

Why genetic distance?

Genetic distance provides a measure of epidemiological relatedness, because it increases as a function of time since transmission (in a linear fashion, as a first order approximation). This increase in genetic distance, due to an underlying molecular clock, provides a proxy for the amount of time that has passed since two viral strains diverged from one another. The molecular clock in HIV, however, is highly imprecise because of factors such as latency and natural selection due to immune escape and antiretroviral treatment. Furthermore, the virus evolves in both the donor and recipient, so the distance between two strains is not simply a multiplier for the time since transmission. However, genetic distance serves as a useful proxy for epidemiological relatedness.

Why use a fixed distance cutoff?

Our recent work in named partners in New York City has demonstrated that genetic distance alone provides better insight into who are potential transmission partners than partner tracing alone. The distribution of pairwise distances among a population of named partners in New York City has the characteristic property of resembling a mixture of two distributions (see FIGURE 24): a component near 0 (i.e., closely/recently related sequences) and a component near 0.06 (i.e., two random sequences of the same subtype). Distance cutoffs of 0.01 to 0.02 segregate the two components nicely.

What is TN93 genetic distance?

TN93 is the name of a nucleotide substitution model developed by Koichiro Tamura and Masatoshi Nei, published in 1993. Hence, TN93. Nucleotide substitution models are used in evolutionary analyses to correct for multiple substitutions (e.g., change from an A to a T then to C, before another genetic sequence has been sampled) and/or reversions (e.g., change from an A to a T back to an A, before another genetic sequence has been sampled) at a given site. Highly divergent sequences, with a greater number of substitutions separating them, are more likely to require complicated evolutionary models to properly estimate the level of divergence. The simplest evolutionary model, JC69, has a single parameter governing mutation rates among different nucleotides, and assumes equal frequencies for all nucleotides. In contrast, a more complex evolutionary model like general time reversible model with gamma rate variation (GTR+ Γ_4) allows all nucleotide substitutions to occur at a unique rate, unique equilibrium base frequencies, and rate variation across sites. Importantly, over relatively short evolutionary distances (i.e., <0.05 substitutions/site), GTR+ Γ_4 does not improve distance estimation accuracy for simpler models like JC69, because not enough time has elapsed for a substantial number of multiple substations and/or reversions. In basic calculus terms, most curves resemble straight lines if you zoom in closely enough.

For **Secure HIV-TRACE**, we wanted an evolutionary model that optimizes both realism and computational efficiency. Simple models like JC69 and K2P (Kimura 2-parameter) have obvious shortcomings when applied to HIV: these models do not permit unequal nucleotide base frequencies, and HIV has notorious high frequencies of adenine (A) and low frequencies of uracil/thymine (U/T). The TN93 substitution model allows for unequal base frequencies and three different rates of substitutions between nucleotide bases: transitions between purines (i.e., A and G), transitions between pyrimidines (i.e., C and U/T), and transversions between purines and pyrimidines (e.g., A or U/T to C or G). Furthermore, distances estimated under TN93 can be represented by a closed form solution (i.e., computed without numerical optimization, simply from pairwise differences in nucleotide counts), which permits rapid computation of pairwise distances. More complex models require relatively expensive numerical optimization, especially because it will have to be done hundreds of millions or billions of times, to find all relevant distances. Over relatively short evolutionary distances (i.e., <0.05 substitutions/site), more complex models do not improve distance estimation accuracy. Therefore, when using genetic distances that are expected to be between 0.01 and 0.02 substitutions/site divergent to identify potential transmission partners, a substitution model more complicated than TN93 is not needed, and there are no appreciable computational savings to be had by using cruder models. As an example, our implementation can compute approximately 10 million TN93 distances per second on a single server node.

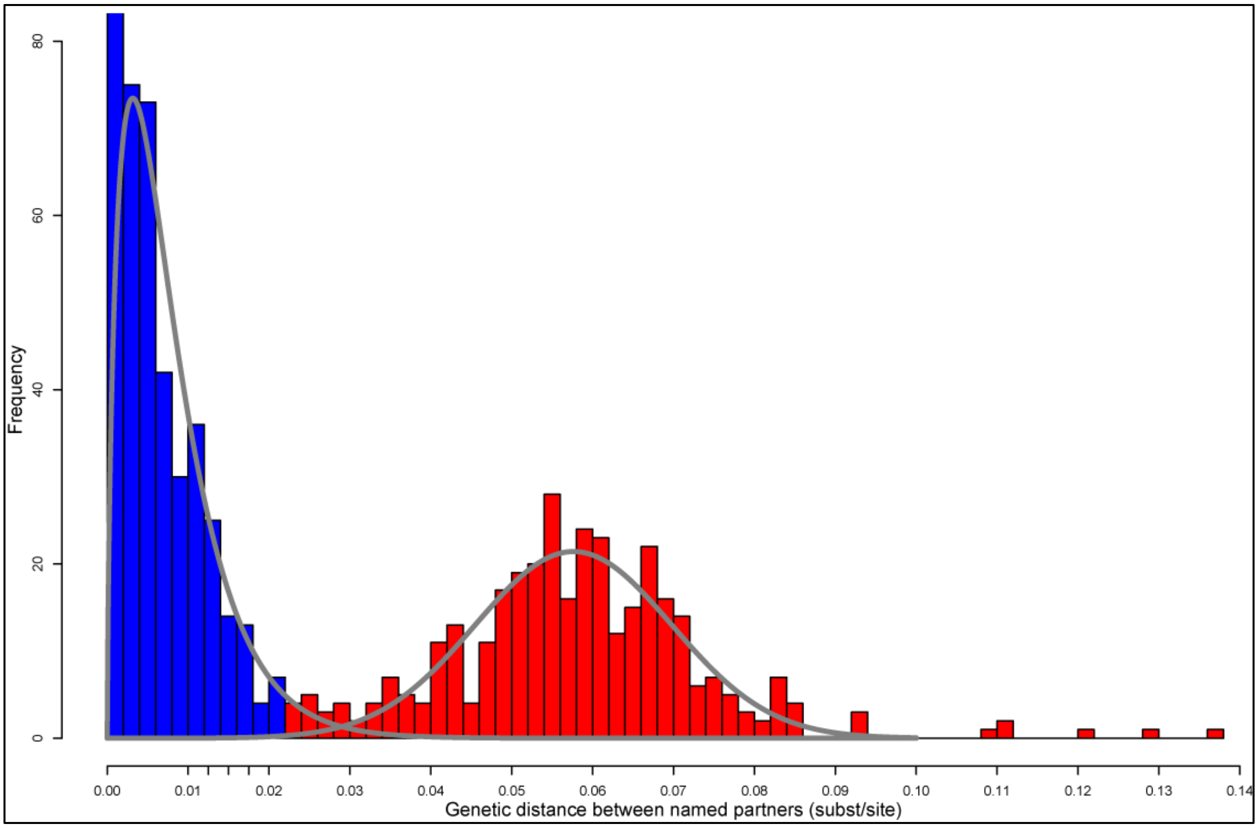
Why not phylogenetics?

Phylogenetics is an extraordinarily powerful tool for understanding viral evolutionary history and dynamics, but phylogenetics says little about whether the relatedness of viruses A and B is epidemiologically meaningful. (For example, to say that two randomly selected subtype B sequences have a meaningful epidemiological linkage would be to say that we care about events that happened more than 50 years ago.) In fact, many HIV transmission network studies that used phylogenies also needed a genetic distance component.

A major problem with relying on phylogenetics to define what can be in a single cluster is that the models are highly contingent on the data and can change in counterintuitive ways. When the goal is tracking transmission network growth over time while adding more and more sequence data, this is a highly undesirable feature. Sequences that are clustered when using Secure HIV-TRACE will always be clustered by Secure HIV-TRACE if the analysis parameters stay the same; adding more data can only increase the size of clusters, not break them apart.

Another issue with the phylogenetic approach is that it takes a lot of computational time, especially for big datasets with tens or hundreds of thousands of HIV sequences. Currently almost half a million sequences are contained in the U.S. National HIV Surveillance database. And unlike a phylogenetic approach which requires a complete re-analysis when a few new sequences are added, with our genetic distance approach, only the new sequences need to be considered, and all the previous computational work can be kept: like adding new pieces to a jigsaw puzzle.

Figure 24. Distribution of genetic distances separating named partners in New York City. Potential transmission clusters are shown in blue. Random, within-subtype variation is shown in red.



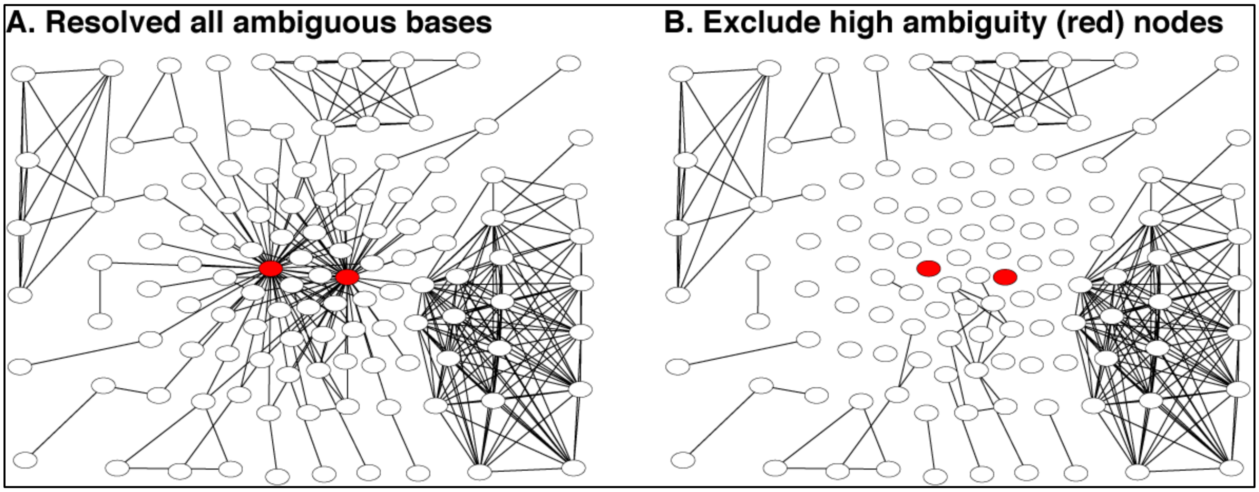
What are ambiguous nucleotides? Or ambiguities?

When HIV infects an individual, it forms genetically diverse and potentially complex populations within that person. Currently, the sequence data reported to the National HIV Surveillance System are produced by using bulk Sanger sequencing, which produces a single genetic sequence representing this circulating population. If, for example, a thymine (T) nucleotide is contained at a given sequence site in at least 80% of the intra-host population, Sanger sequencing typically identifies a T at that site. However, when some intermixing strains have one nucleotide at a position and others have a different nucleotide at the same position, Sanger sequencing typically reports diversity at polymorphic sites as common nucleotide IUPAC ambiguity codes (e.g., R [A or G], Y [C or T], N [any nucleotide]). In standard phylogenetic inference, nucleotide ambiguities are “partially missing data” (e.g., Y is either C or T, but not A or G). When using pairwise distances (as in **Secure HIV-TRACE**) to construct genetic transmission networks, these nucleotide ambiguities have the potential to greatly complicate inference (see **FIGURE 25A**). The most conservative approach is to average the distance between ambiguities and resolved bases (e.g., Y is 0.5 differences from either C or T). But averaging ambiguities in transmission network analysis decreases the probability that sequences from chronically infected individuals—who are likely to have a more diverse viral population—will cluster in the network. Therefore, **resolving ambiguities** (so that Y would be 0 differences from either C or T, and 1 difference from A or G) appears to be an attractive option. However, if we are too permissive in our tolerance of ambiguities, unrelated viruses can become connected in our network. Variable sites are not uniformly distributed across the HIV genome. As a result, if ambiguities are resolved in the genetic distance calculation for a highly polymporphic sequence, this highly polymorphic sequence is likely to link to many "unrelated" viruses. The result is a large transmission cluster in which most sequences are connected to the high ambiguity sequence, but not to each other.

For example, if sequences from two people differ at 5% of sites, their viruses represent random intra-subtype variation and are not likely potential transmission partners. However, if within one of these people, most of this variation is polymorphic, and ambiguities are resolved in the genetic distance calculation, the genetic distance separating these viruses may fall below the distance threshold. Since variable sites are not uniformly distributed across the HIV genome, the highly polymorphic sequence is also likely to link to many other "unrelated" viruses as well. The result is a large transmission cluster in which most sequences are connected to a hub (the high ambiguity sequence) but not to each other.

In an example from the San Diego Primary Infection Cohort (**FIGURE 25A**), the genetic transmission network is affected by handling of nucleotide ambiguities. When ambiguities are fully resolved, the largest cluster in this cohort contains 119 people. However, when this cluster was mapped onto a maximum likelihood phylogenetic tree, its members are dispersed across the tree, encompassing the genetic diversity of the entire city of San Diego. Furthermore, the majority of nodes in the cluster are connected via two nodes acting as hubs (highlighted in red in **FIGURE 25**) which have 5.8% and 7.6% ambiguities and represent the two highest degree nodes in the network. The nodes connected through the spokes on these hubs rarely share an edge with each other. This feature, along with the phylogenetic dispersion, suggests that this cluster is an artifact of nucleotide ambiguity resolution. When these two hubs are excluded from the analysis, the cluster breaks apart, resulting in several distinct clusters and unconnected nodes (**FIGURE 25B**).

Figure 25. Example in which two contaminant sequences with high numbers of nucleotide ambiguities (shown in red) can create artificial clustering among unlinked singletons and unrelated clusters. (A) The inferred cluster resolving all ambiguous nucleotides. (B) The same cluster where the two hubs (shown in red) are excluded from the analysis.


A. Resolved all ambiguous bases
B. Exclude high ambiguity (red) nodes

Clusters that resemble **FIGURE 25A** should be interpreted with extreme caution. They are almost always spurious and the result of erroneous inference due to high levels of nucleotide ambiguities (or contamination with "reference" strains). **Secure HIV-TRACE** has been developed to minimize the chance of this artifact occurring.

How does Secure HIV-TRACE handle ambiguous bases?

We recommend that nucleotide ambiguities be fully resolved when calculating genetic distance only when (i) the sequences have a low proportion of ambiguities or (ii) if the size of the dataset is small. When constructing a transmission network for datasets of thousands or tens of thousands of sequences, we recommend penalizing sequences with high levels of ambiguities. The "**Ambiguity Fraction**" parameter governs this penalty. The default "Ambiguity Fraction" value of 0.015 resolves the genetic distance between ambiguous nucleotides when calculating the distance between sequences with ≤1.5% ambiguities and averages the genetic distance between ambiguous nucleotides when calculating the distance between sequences with >1.5% ambiguities.

Although not currently implemented in Secure HIV-TRACE, future versions will identify sequences with >5% ambiguous nucleotides and flag them as problematic sequences and/or remove them from the analysis. This protocol follows the guide set forth by the Los Alamos National Laboratory (LANL) HIV Sequence Database (https://www.hiv.lanl.gov/components/sequence/HIV/search/help.html#bad_seq). Extremely high proportions of ambiguities can be the result of poor quality sequencing, contamination, or dual infection. Including these sequences can adversely affect the performance of **Secure HIV-TRACE**.

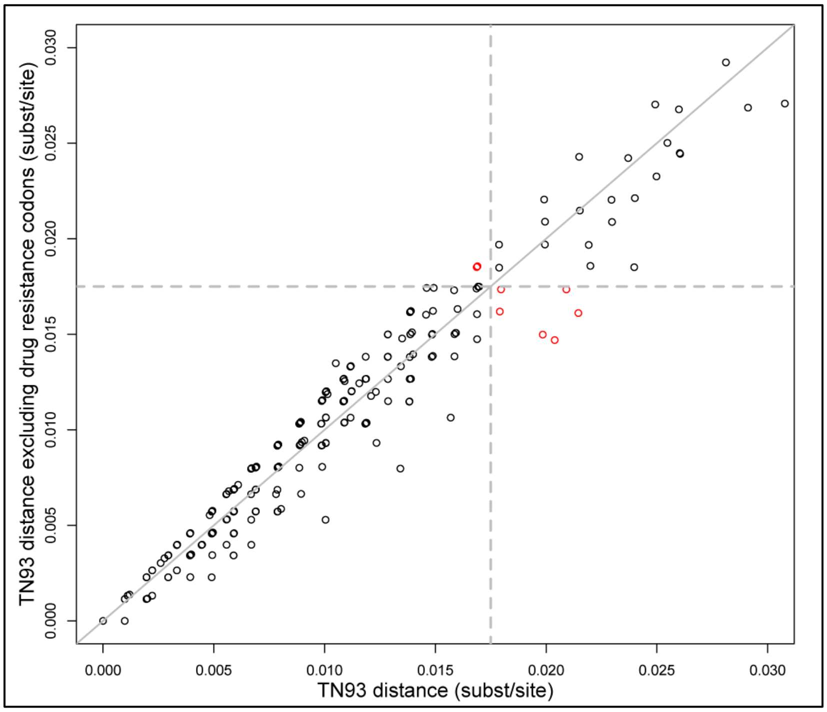
Why should I screen for laboratory contaminants?

Although the protocols for generating HIV-1 *pro/rt* genetic sequences are well validated, occasionally laboratory contamination with other genetic material is known to occur. This contamination is most often with the lab strain HXB2, but it can happen with any strain of HIV. Importantly, this contamination often results in a mixed sample where the resulting sequence is a combination of the isolate and the laboratory contaminant. This mixed sample often has high levels of ambiguous nucleotides and could compromise HIV-TRACE analysis if it were to be included, especially because mixing two unrelated strains will create ambiguities at many sites that tend to vary between strains, thereby enabling a "connection" through this sequence if ambiguous nucleotides are resolved (see above). Furthermore, if multiple contaminant sequences are included in the same analysis, they will erroneously be inferred to be part of the same cluster. Therefore, we screen every run for HXB2 linked sequences. Any sequence that links to HXB2 will be identified after the alignment phrase and excluded from further analysis.

What about drug-resistance–associated mutations (DRAMs)?

DRAMs often arise in HIV found in people taking antiretroviral therapy; they can be found in virus from both treatment-naive and treatment-experienced people who were initially infected with a drug-resistant virus. DRAMs typically occur at a select set of sites that are not polymorphic in the absence of prior antiviral therapy. This type of convergent evolution at the amino acid-level has the potential to negatively affect phylogenetic inference. The genetic distance separating two viruses that undergone convergent evolution will theoretically be lower than two viruses that have not experienced convergent evolution. In practice, however, we find little to no effect of excising DRAM sites from network inference. Specifically, transmission networks built at the city, national, global level are robust to inclusion of DRAM sites. For example, when analyzing a cohort of named partner pairs in New York City, only a small fraction of partners become either linked or unlinked when DRAMs are excluded (red in **FIGURE 26**). Therefore, we do not recommend excising DRAMs from transmission network analyses using HIV-TRACE. An exception to this recommendation is for studies focusing on the effect of DRAMs on network characteristics; in these instances, DRAM sites should be excised prior to network construction.

Figure 26. Genetic linkage including/excluding codons associated with drug-resistance mutations in a New York City surveillance cohort. Nodes in red change linkage depending on inclusion/exclusion of DRAMs.



Appendix D. Additional resources

Selected published articles that use HIV-TRACE analyses

- Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, Kosakovsky Pond SL. The global transmission network of HIV-1. *J Infect Dis* 2014;209(2):304–313.
- Little SJ, Kosakovsky Pond SL, Anderson CM, Young JA, Wertheim JO, Mehta SR, May S, Smith DM. Using HIV networks to inform real time prevention interventions. *PLoS One* 2014;9(6):e98443.
- Oster AM, France AM, Panneer N, Bañez Ocfemia MC, Campbell E, Dasgupta S, Switzer WM, Wertheim JO, Hernandez AL. Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data. *J Acquir Immune Defic Syndr* 2018. doi:10.1097/QAI.0000000000001856.
- Oster AM, Wertheim JO, Hernandez AL, Ocfemia MC, Saduvala N, Hall HI. Using molecular HIV surveillance data to understand transmission between subpopulations in the United States. *J Acquir Immune Defic Syndr* 2015;70(4):444–451.
- Whiteside YO, Song R, Wertheim JO, Oster AM. Molecular analysis allows inference into HIV transmission among young men who have sex with men in the United States. *AIDS* 2015;29(18):2517–2522.
- Wertheim JO, Oster AM, Hernandez AL, Saduvala N, Bañez Ocfemia MC, Hall HI. The international dimension of the U.S. HIV transmission network and onward transmission of HIV recently imported into the United States. *AIDS Res Hum Retroviruses* 2016;32(10–11):1046–1053.