UNAIDS 2024

BRIEF 1 of 3

The application of phylogenetics to HIV—insights into biology and epidemiology of HIV

Contents

Acknowledgments

UNAIDS gives special thanks to Lucie Abeler-Dörner, Ana Bulas Cruz, Josh Herbeck, Matthew Hall, Manon Ragonnet-Cronin, Chris Wymant, Katrina Lythgoe, Louise Karlsson, Christophe Fraser, for their contributions to this publication.

Introduction

Phylogenetics is the study of the *evolution* of organisms based on their genetic similarity. Phylogenetic techniques, laboratory methods to read genetic code present in all living organisms, can be used to compare different species or to compare different members of the same species. They can also be used to compare viruses like HIV. Phylogenetic techniques are used to detect the subtle changes that occur in the genetic code of each organism from one generation to the next. These changes are particularly pronounced in HIV making it a good candidate for phylogenetic studies.

This brief introduction to HIV phylogenetics is the first of three briefs on the phylogenetics of HIV. This brief introduces phylogenetics, explains how the genetic code can be read, and shows how the sequences can reveal insights into HIV biology and epidemiology. The second brief explores what we have learned from HIV phylogenetic studies and how HIV phylogenetics can and should be used to support public health programmes. The third brief discusses the ethical considerations of HIV phylogenetics, including the need for appropriate care and safeguards of sensitive information.

Genomes and genetics

Phylogenetics uses an organism's genetic information. All animals, plants, bacteria and viruses store this genetic information in their *genome*. The *genome* specifies how an organism develops, how it looks and how it reproduces. The genome is made either of *ribonucleic acid* (RNA) or *deoxyribonucleic acid* (DNA).

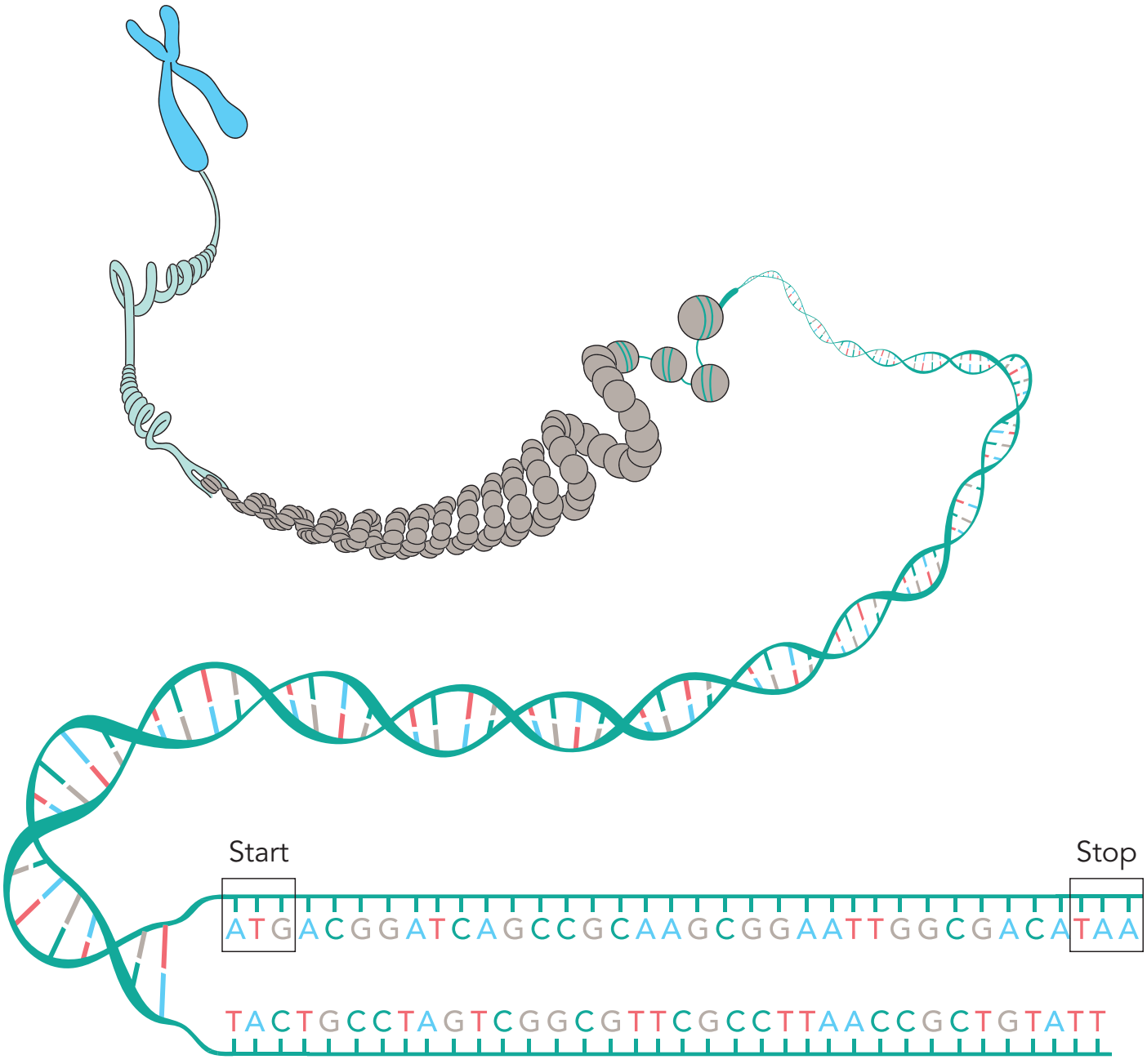
Both RNA and DNA are large molecules built from four smaller molecules called *nucleotides* or *bases*. In DNA, these nucleotides are adenosine (A), cytosine (C), guanine (G) and thymine (T). In RNA, thymine is replaced by a similar molecule called uracil (U). These nucleotides serve as the 'letters' of the genetic alphabet. As in a book, they are strung together in long lines to encode genetic information. The book, or the entirety of an organism's genetic material, is called the genome. RNA consists of a single linear string of A, C, G, and U. DNA consists of two strings of A, C, G, and T, which are arranged next to each other and connected by chemical bonds. A DNA molecule looks like a twisted ladder (the double helix) (Fig. 1). In this ladder, an A on one strand is always paired with a T on the other strand, while a C is always paired with a G. The information on both strands is therefore complementary and it is possible to reconstruct the second strand of a DNA molecule when the first strand is known.

Today, all animals, plants and bacteria use DNA for their genomes. But viruses can use either DNA or RNA. HIV has a genome made of RNA, and is an example of a retrovirus, which is why the medicines used to treat HIV are called *antiretrovirals*. This means that HIV can copy itself into a DNA form and then insert itself into the genome of infected cells. This ability of HIV to hide in the human genome is one of the reasons that an HIV infection is extremely difficult to cure. The HIV genome is made up of approximately 10 000 RNA nucleotides. For comparison, the human genome is much larger at over 3.2 billion DNA nucleotides.

Together with the natural and social environment we grow up in, our genetic make-up influences different aspects of our body—what we look like, how tall we are, how much we weigh and how susceptible

we are to different diseases. In a person, most of these traits are determined by many, sometimes hundreds of different sections of the genome. For viruses, the linkage between the genome and the characteristics of a given virus are often more straightforward as the genome is much smaller and one stretch of DNA or RNA is often responsible for one property of the virus.

Figure 1: DNA is usually arranged in a twisted ladder-like shape called the DNA double helix. The double helix is usually wound around circular proteins and packed tighter and tighter until it forms a large X-shaped structure called a chromosome.



Source: https://biologywarakwarak.wordpress.com/2012/01/15/the-3-magical-rules-to-determine-the-amino-acid-chain-from-a-dna-piece-without-error/

Sequencing

Genetic 'sequencing' means determining the sequence of A, C, G and T (or U) in a sample. Current sequencing techniques can only use DNA. This means that the genome of RNA viruses (such as HIV) must be translated into DNA before they can be sequenced. It's also important that there is enough DNA available for machine readers to measure, so a technique known as *polymerase chain reaction* (PCR), is used to generate enough DNA for sequencing. PCR duplicates DNA strands so that there is a measurable quantity of DNA.

The first sequencing technique to be developed (Sanger sequencing), works by analyzing all amplified DNA molecules at the same time. Only the most common nucleotide is recorded at each position, with rare variations not recorded. Thus, if two virus particles have the nucleotide A at position 1000 along the genome and one has a nucleotide T, an A would be recorded at this position. The result is a sequence which consists of the most common nucleotide at each position. This is called a *consensus sequence*. Because of technical limitations, Sanger sequencing produces sequences of around 500–800 nucleotides. To obtain an entire HIV genome of 10 000 nucleotides, it is necessary to line up overlapping sequences from the many different fragments.

'Next-generation' or 'deep' sequencing methods have been developed since 2005 and these have replaced Sanger sequencing in many applications, including in phylogenetic studies. As these methods require the DNA to be chopped up into fragments, assembling a complete genome after sequencing is more complex than in Sanger sequencing and needs greater computing power. This limitation can be partially overcome by newer, more advanced, protocols. Alternatively, entirely different sequencing technologies such as Oxford Nanopore sequencing can be used. Oxford Nanopore technology does not have a restriction on the length of the fragments that it can read. At least theoretically, it can read an entire HIV genome, or *haplotype*, in one step. At the moment, such systems are inefficient in analyzing large numbers of samples and are much more prone to errors than deep sequencing. However, the field is developing rapidly, and it is likely

that datasets generated by these methods will become available in the near future.

Deep sequencing has one great advantage: because the DNA fragments are arranged in a spatial pattern in a microchip-like surface during the sequencing process, the sequence of nucleotides from separate fragments can be recorded separately. Deep sequencing methods therefore enable us to not only study the diversity of viruses found in different people, but also the diversity of viruses found within a single individual. This allows the study of *minority variants*, nucleotide deviations from the consensus sequence.

Traditionally, HIV is sequenced from a person's serum (the acellular fraction of a blood sample). If people living with HIV are on effective antiretroviral treatment, the number of viral particles in serum is so low that there is not enough for sequencing. There is, however, an alternative. HIV can insert its genetic material into the human genome; these copies of the virus are still present in treated individuals and cause a resurgence of the infection when treatment is stopped. Methods to prepare human cells for sequencing of these proviruses exist but are more complicated to implement and less widely used than those using serum as the starting point.

Reading a phylogenetic tree

Phylogenetics is the study of the *evolution* of organisms based on their genetic similarity. Like family trees, phylogenetic trees depict the genetic relationship between different genetic sequences. Phylogenetics uses mathematical models to predict how genetic sequences that existed in the past could have evolved to give rise to the genetic sequences we see in a given dataset.

Phylogenetic trees go back to an origin—the predicted most recent common ancestor of all sequences in the dataset. The concept is analogous to a parent being the most recent common ancestor of a set of siblings, and a grandparent being the most recent common ancestor of all cousins in a family. An example could be an evolutionary tree that shows when birds and mammals split from reptiles. For viruses like HIV, phylogenetic trees can reconstruct how a virus spreads through a population.

Phylogenetic techniques make use of the fact that the genetic sequences of all living organisms change over time. Nucleotides are replaced by other nucleotides owing to imprecise copying; this change is called a *mutation*. Mutations accumulate over time and may be passed on to the next generation. This means sequences from closely related organisms are more similar (i.e., share more common nucleotides) than sequences from organisms that are more distantly related. Viruses accumulate mutations more quickly than bacteria, plants or animals, and HIV accumulates mutations more quickly than most other viruses. This makes phylogenetic analysis of HIV more informative than for many other viruses.

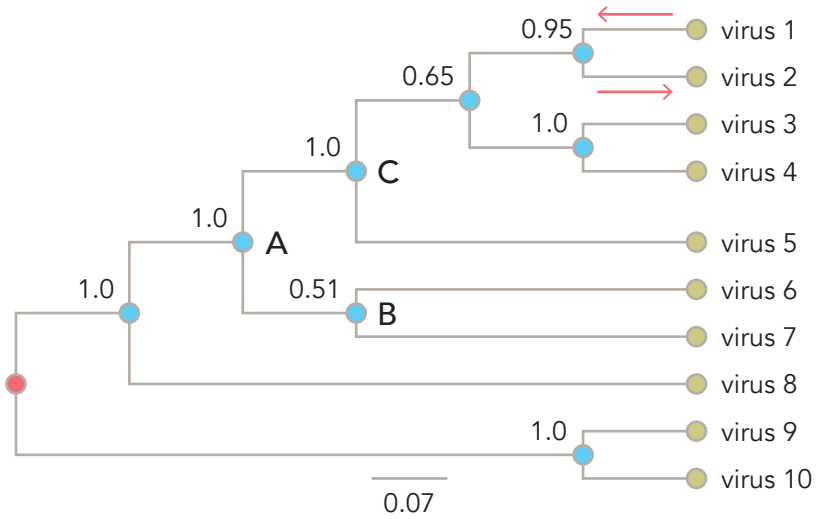


Figure 2: An example of a rooted phylogenetic tree. The red dot is the origin ancestor of all other viruses in the tree. This figure was taken from the ARTIC Network tutorial on how to read a phylogenetics tree (http://artic.network/how-to-read-a-tree.html). The scale bar marks the genetic distance in substitutions per nucleotide.

A schematic example of a viral phylogenetic tree is depicted in Figure 2. The horizontal lines are called branches and represent genetic change accumulating over successive generations of organisms from ancestors to descendants. The length of the branch is proportional to the amount of genetic change. The vertical lines in this kind of tree do not carry any information, they are simply used to spread out the branches and make the tree more readable.

The circles are called *nodes* and represent ancestors that mark the end of a branch. The green circles at the 'tips' of the tree represent the sequences present in the data set. Moving left from the tips corresponds to moving back in time through the evolutionary history of these samples. Every blue dot on the horizontal lines represents an organism that was not observed but inferred to have existed in the past. Once we move far enough back into evolutionary history, the sampled organisms begin to share history and share ancestors, in much the same way that cousins have different parents but shared grandparents (and great-grandparents etc.). For example, B is the most recent common ancestor of organisms 6 and 7. Any 'cut' in the phylogeny results in *clade*, an ancestral sequence and all its descendants. The internal node marked with a dark red circle is the 'root' of the tree, representing the most recent common ancestor of all the viruses in the tree.

In Figure 2 the genetic distance between viruses 1 and 2 is the sum of the lengths of the two branches marked with the red arrows. The legend at the bottom provides a scale for these lengths, in this case 0.07, in units of 'substitutions per nucleotide'. This means that a branch with the same length as the scale represents 7 mutations accumulating for each 100 nucleotides in the genome.

In some trees, the x axis is units of time rather than genetic distance. Mutations do not always happen at the same rate, as evolution might sometimes be fast and sometimes slow. It is difficult to convert a phylogenetic tree based on genetic distance into a phylogenetic tree based on time. A variety of mathematical models exist to perform this task.

As different techniques can be used to create phylogenetic trees, more than one tree can be created. A phylogenetic tree in a scientific publication will usually show the most likely tree identified by an algorithm. It may further include information to summarize the many possible trees that were generated during the analysis. A number near an internal node of the tree is sometimes used to indicate the fraction of generated trees for which the tips descended from this node group together. In Figure 2, the 1.0 next to internal node A indicates that a node splitting the two branches that gave rise to nodes B and C was present in all trees constructed. The 0.51 next to node B indicates that this node was only present in 51% of all trees constructed. When comparing longer sequences, e.g., the entire HIV genome, it is good practice to construct multiple trees from different parts of the genome to make sure the resulting trees are consistent.

Intra-host diversity

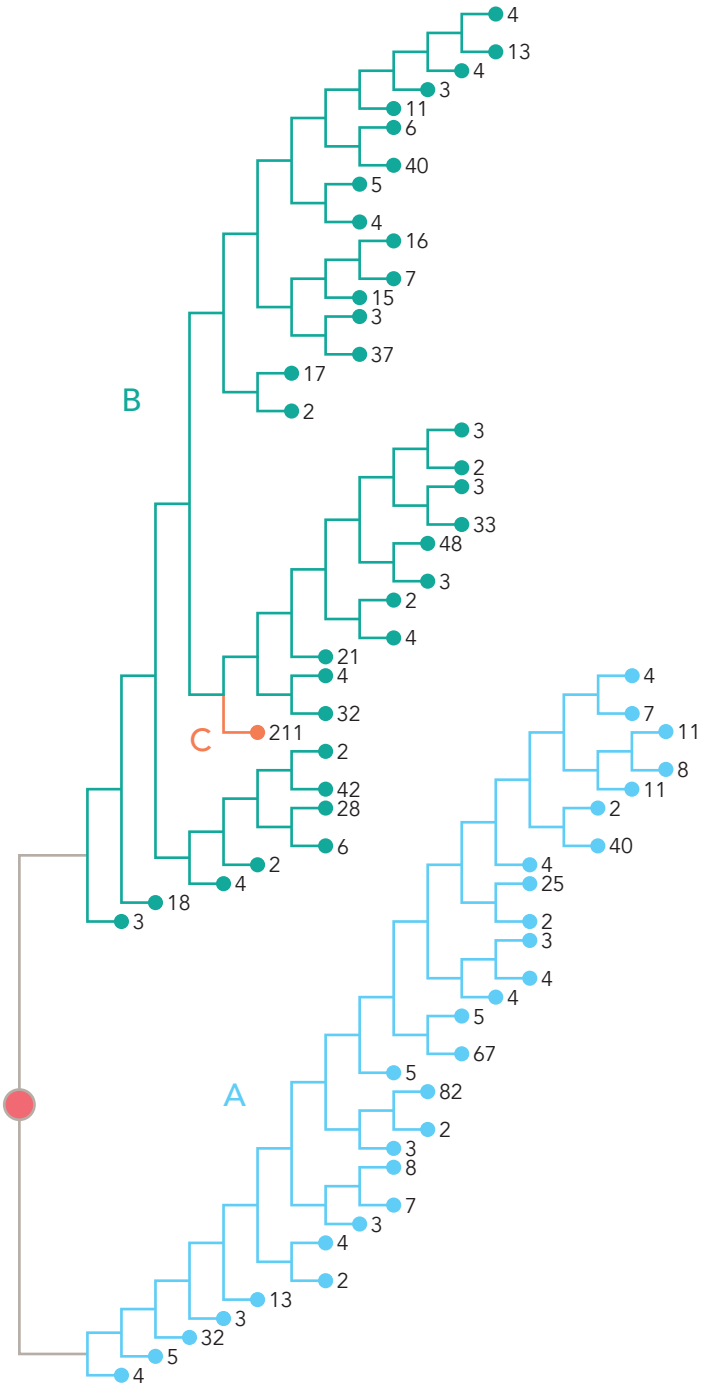


Figure 3: A phylogenetic tree depicting variation within the samples from one person (within one colour) as well as similarity between samples from two people (A-F, different colours). Numbers at each tip node indicate the number of identified copies of the same genome in a sample.

Deep sequencing has allowed researchers to go beyond the consensus sequence and consider the genetic variation found within the HIV sample of a single person, not just the variation between different people. Viruses accumulate mutations as they move through a population. Viruses that cause short infections which are quickly resolved, accumulate very few mutations in any given individual. However, because HIV causes a chronic infection, many mutations accumulate in each person until replication of HIV is stopped by effective antiretroviral treatment. HIV infections are usually caused by a single viral particle which means that all viruses in an individual usually originate from a common sequence. With each round of virus replication, mutations are introduced, and within a few weeks the viruses have changed so much that sequencing reveals a collection of closely related but not identical viruses. This collection is called a *quasispecies*. Many of these mutations will lead to viruses that are no longer infectious, but many will also lead to variants of the original HIV that are slightly different to the parent virus and in some cases may be able to function differently, for example by developing resistance to a particular antiretroviral treatment, or by escaping from some immune defenses of the human host. Since HIV causes a chronic infection, the quasispecies has time to gradually change and become more diverse. This diversity can also be shown in a phylogenetic tree (Figure 3). The sequences from the viruses found in one person (denoted in the same colour) are in most cases more closely related to each other than the sequences from viruses found in different persons (denoted by different colours). This variation among viruses from a single person is called *intra-host diversity*.

Intra-host diversity increases with time since infection. Methods have been developed to estimate the time since infection from the HIV genetic diversity. These methods work reasonably well at the population level but come with significant uncertainty at the individual level. They will be discussed in more detail in the second brief in this series, *Practical Uses of Phylogenetics in Public Health*.

Intra-host diversity can also be used to find pairs of individuals in a dataset that are likely direct transmission pairs. In Figure 3, sample C was found to have 211 identical sequences; such little intra-host diversity suggests the individual was sampled shortly after acquiring HIV. The 211 sequences are closely related to and are downstream in evolution, (evolved after,) from sequences of sample B, suggesting that participant B is upstream in the transmission chain, (infected prior,) that led to participant F acquiring HIV, most likely in a direct transmission. Viruses from sample A on the other hand are much less related to viruses from samples B and C and direct transmission is unlikely. Analyses of this kind only generate a probability that two sequences are directly linked. From a conceptual point of view, they can therefore never be used to prove transmission between two individuals. Transmission probabilities from these analyses are highly sensitive; the ethical implications are discussed in the third brief of this series, *Ethics Considerations for HIV Phylogenetic Analyses*

Until now, HIV transmission analyses have mostly been undertaken in research contexts. Public health agencies may opt to use cluster analyses instead of phylogenetic transmission analyses as they are faster and simpler to run and generate less sensitive data. Cluster analyses compare sequences based on their genetic similarity and group similar sequences into clusters. They make no assumptions about the direction of transmission. Cluster analyses have been used in public health contexts in various countries to find fast-growing sub-epidemics. The public health uses of phylogenetic and cluster-based approaches will be discussed in more detail in the second brief of this series, "Practical Uses of Phylogenetics in Public Health.

From sample to analysis

In a typical HIV phylogenetic study, blood samples are collected from study participants, spun in a centrifuge to remove the human blood cells, frozen and shipped to the laboratory. In the laboratory, RNA is extracted from the samples and converted into DNA. At this stage, the samples still contain a large amount of human DNA, so they are enriched for HIV-derived DNA. During these steps, the DNA is amplified by PCR to generate enough DNA for the sequencing process. The DNA is prepared for sequencing and then sequenced. Sequencing success is highly dependent on the *viral load* of the participant, the number of viral particles per unit blood. The sequencer provides raw data files which may contain hundreds or millions of sequence fragments, depending on how much HIV RNA was present in the original sample. The fragments are then aligned to a position in the HIV genome. The most common nucleotide at each position is determined to construct a consensus sequence. The data files with all sequencing reads are stored as well.

Most HIV phylogenetic analyses to date have used consensus sequences. Analyses can use the whole genome or just one part of it, for example the sequence of the polymerase gene (*pol*) which is routinely sequenced for drug resistance analysis in many countries. Analyses with consensus sequences have been used to reveal how HIV spread within Africa and beyond in the 20th century, to study the properties and evolution of different *HIV subtypes*, and to learn more about risks of transmission.

Consensus sequences are often made available to public databases, especially after the analyses are published in scientific journals. The sequences in public databases are free to be used by anyone, but generally only contain limited additional information, for example country and year of collection. Some studies also consider the information from the individual fragments, for example transmission analyses. These are generally not made publicly available as they contain more sensitive information and because the files are very large. Pathogen phylogenetics only use pathogen sequences. All traces of human genomes contained in the samples are usually deleted before analysis.

Benefits and risks of HIV phylogenetics

HIV phylogenetic analyses can be used to understand the dynamics of HIV epidemics, including how HIV has spread around the globe in the past 50 years, how different subtypes have emerged and developed, how migration influences these patterns, and how HIV transmission impacts different age groups today.

HIV phylogenetic analyses can generate highly sensitive data, which means that their use is not always warranted and generated data needs to be handled with great care. Most importantly, phylogenetics cannot definitively implicate an individual in a transmission chain; however, it can definitively exclude an individual in a transmission chain.

Limitations

Conducting phylogenetic studies has important capital and running costs. Working with deep-sequencing data can take a lot of computational time. Most analyses can be performed on a standard laptop computer, but genome assembly and phylogenetic analysis for a few thousand samples can take weeks on a high-performance computer cluster, depending on the research questions, the analyses performed, and the methods used.

Conclusion

Phylogenetics is a powerful tool to help answer virological, clinical and epidemiological research questions. To maximise the potential benefits of phylogenetics, it must be combined with other informative data sources. As the techniques improve and costs decrease, phylogenetics is likely to become a more common source of data in epidemiological and clinical studies. This brief has laid out the basic approaches currently being used to create and analyse sequence data from HIV. The second brief in this series describes examples of how these phylogenetic approaches have been and are being used for public health and for research that aims to improve HIV programme efficiency and effectiveness. The third brief explores the ethical considerations that such approaches use and considers how any risks can best be mitigated and harms avoided.

Glossary

Antiretrovirals

Drugs that interfere with the life cycle of HIV in the cell and stops the virus from generating new viral particles. A successful antiretroviral therapy or ART cocktail for HIV consists of three different drugs combined into a single pill that usually needs to be taken daily.

Base or base pair

The building block of DNA or RNA; a base is also called nucleotide.

Clade

One organism (or virus) and all its descendants.

Consensus sequence

A genetic sequence produced from multiple sequence fragments of the same sample, with the most common nucleotide being reported for each position. It is possible that no single molecule in the sample is identical to the consensus sequence, but it gives a useful overall summary of sequences present in the sample.

Deoxyribonucleic acid (DNA)

The molecule inside cells that contains the genetic information responsible for the development and function of an organism. DNA consists of the four bases adenine (A), cytosine (C), guanine (G) and thymine (T).

Evolution

Change in the heritable characteristics of biological populations over successive generations.

Genome

The entirety of all genetic information in an organism, encoded by DNA in most organisms and by DNA or RNA in viruses.

Haplotype

A group of genes inherited from a single parent. For a virus, the genetic sequence contained in a single virus particle. A quasispecies consists of many distinct haplotypes.

HIV subtype

Group of phylogenetically linked HIV variants that are more closely related to each other than to other subtypes. Genetic variation within a subtype can be 15–20%, whereas variation between subtypes is usually 25–35%.

Intra-host diversity

Genetic diversity which exists among viruses in a single person. Intra-host diversity develops as the originally transmitted viral particle accumulates mutations and forms a quasispecies.

Minority variant

A genetic variant, for example a single nucleotide polymorphism which is not represented in the consensus sequence.

Mutation

A change in the sequence of A, C, G and T/U in a genome. The change can be a replacement of one or more nucleotides by other nucleotides, a loss of one or more nucleotides (deletion), a gain of one or more nucleotides (insertion), or a combination of the three.

Node

The end of a branch in a phylogenetic tree, representing an observed or inferred genetic sequence. There are two types of nodes: the tips, which are the sequences present in the data set, and the internal nodes, which are the branching points in the tree. The root is the internal node that represents the most recent common ancestor of all sequences in the data set.

Nucleotide

Building block of DNA and RNA; also called base in a genetic context.

Polymerase chain reaction (PCR)

A fast and inexpensive technique to create copies of small fragments of DNA. Most genetic analyses require more DNA than can be found in the original sample and have only become possible since PCR was invented in 1985. The discovery was deemed so significant that it led to a Nobel prize in chemistry in 1993.

Provirus

Virus genome which is integrated into the DNA of a host cell.

Retrovirus

Member of a family of viruses that inserts a DNA copy of its RNA genome into the DNA of a host cell. This makes it virtually impossible for the immune system to completely clear a retroviral infection.

Ribonucleic acid (RNA)

A molecule inside cells that performs various roles in the translation of DNA into proteins. Some viruses use RNA rather than DNA to store their genetic information. RNA consists of the four bases adenine (A), cytosine (C), guanine (G) and uracil (U).

Quasispecies

Collection of slightly different viruses (or haplotypes) found in the same person.

Viral load

Density of virus particles in a fluid, often the number of virus particles per milliliter of blood.

Suggested citation. The application of phylogenetics to HIV—insights into biology and epidemiologyof HIV. Geneva: Joint United Nations Programme on HIV/AIDS; 2024. Licence: CC BY-NC-SA 3.0 IGO.

UNAIDS/JC3106aE

UNAIDS
Joint United Nations
Programme on HIV/AIDS

20 Avenue Appia
1211 Geneva 27
Switzerland

+41 22 791 3666

unaids.org