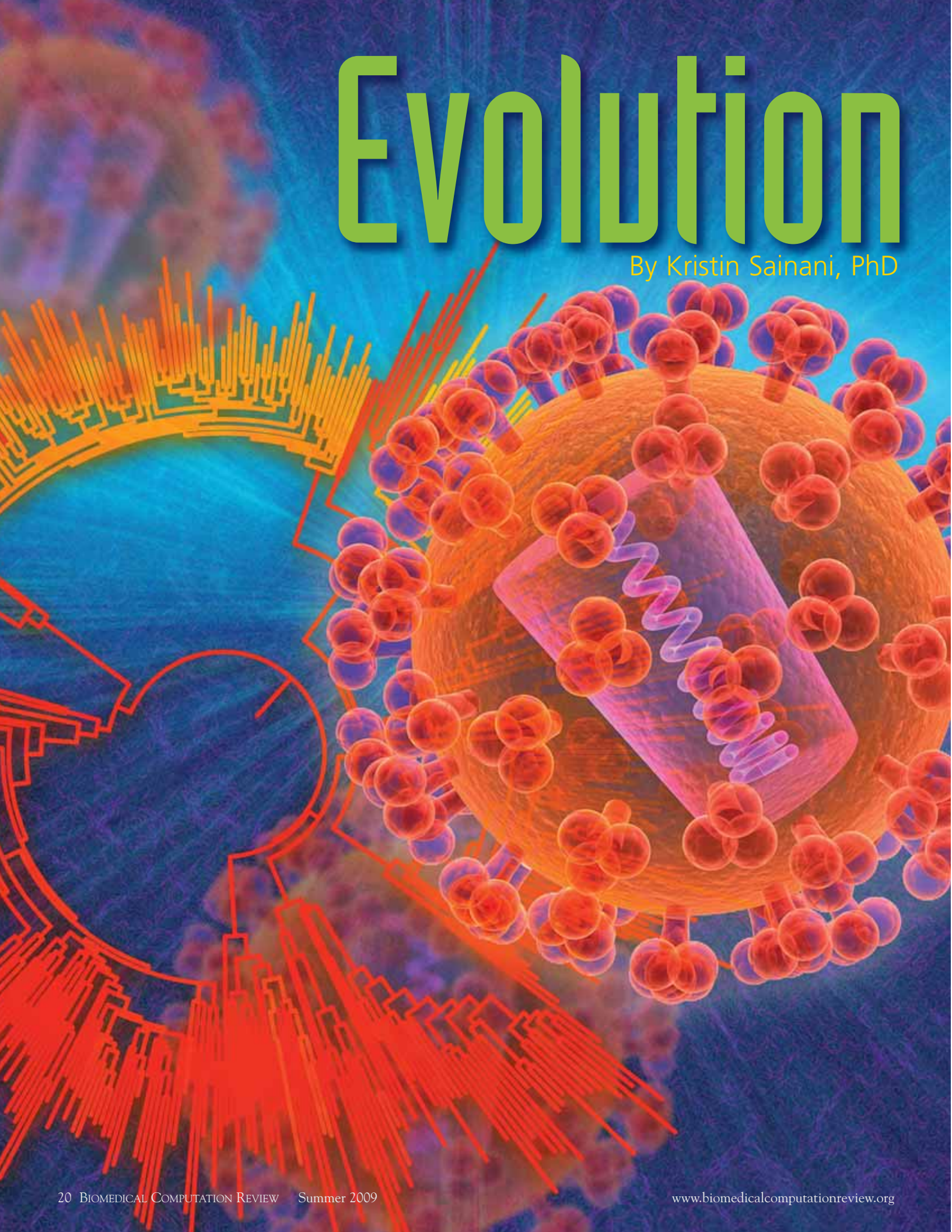
Evolution

By Kristin Sainani, PhD



and HIV:

Using computational phylogenetics to close in on a killer

When Darwin published *On the Origin of Species* in 1859, it would be decades before HIV would jump from monkeys to humans and set off a devastating worldwide pandemic. But evolution is at the heart of HIV's biology, and Darwin would no doubt have marveled at the virus's evolutionary prowess. The virus evolves a million times faster than humans do. So fast, in fact, that a 2007 *Proceedings of the National Academy of Sciences* paper estimated that HIV and other retroviruses live just beneath the "evolutionary speed limit"—a notch faster and they would mutate themselves into oblivion.

Speedy evolution is HIV's secret weapon—allowing it to evade the immune system, resist drug treatment, and, thus far, remain impervious to vaccines. But it may also be the key to disarming the virus. The better scientists understand HIV evolution, the better they can contain the pandemic, improve treatments, design vaccines, and devise novel ways to fight the virus.

Scientists study HIV evolution at multiple scales—globally, regionally, locally, and even within a single host. Using computational methods developed for the study of phylogenetics—the study of how organisms are genetically related to one another—they can date the age of ancestral viral sequences; unravel the virus's travels across nations and among different populations; reconstruct networks showing which individuals transmitted HIV to one another; and identify genes under selective pressure. The methods are the same whether applied at the population or host level—for example, migration patterns between tissues in a single host are resolved in the same way as migration patterns between countries. "This is a remarkable thing. We use the same underlying statistical and computational framework to tackle really quite different biological questions," says **Oliver Pybus**, **PhD**, a research fellow in the department of zoology at Oxford University.

"Lots of tools have been developed to do evolutionary analysis of gene sequences. And a lot of these tools have cut their teeth on HIV," says Eddie Holmes.

The study of HIV evolution is not only critical to fighting the virus; it has also driven advances in the computational tools used to study evolution in general. "Lots of tools have been developed to do evolutionary analysis of gene sequences. And a lot of these tools have cut their teeth on HIV," says **Eddie Holmes**, **PhD**, professor of biology at Penn State University. "It's like the space program; it's this kind of glittery prize of modern science. Some of the smartest people have worked on HIV to try and make these techniques."

Phylogenetics in the era of HIV

At the heart of the study of evolution is the phylogenetic tree. Scientists align a group of sequences (either of the whole genome or of a particular gene) and compare the nucleotides at every position to establish how genetically distant the strains are. These genetic distances define the phylogenetic tree: the order of sequences in the tree as well as the branch lengths.

Computationally, it's what's known as an "NP-hard" problem. "Which means, as your dataset gets bigger, your solution space gets ridiculous," says **Keith A. Crandall**, **PhD**, professor of biology at Brigham Young University. For example, the number of possible trees that you can build out of just 50 or 60 sequences exceeds the number of particles in the universe, says **Alexei Drummond**, **PhD**, associate professor of computer science at the University of Auckland.

Finding clever ways to search the tree space "is where a lot of the action is in phylogenetics," Crandall says. There's been a lot of advancement in this area in the past decade using Bayesian statistics, he says. For example, the Bayesian Markov chain Monte Carlo (MCMC) method is implemented in the popular program BEAST (Bayesian Evolutionary Analysis Sampling Trees, co-created by Drummond, http://beast.bio.ed.ac.uk/). "It doesn't attempt to find a single best answer. It tries to give you a set of trees that are representative, that are plausible, given your data and the model," Drummond says. Generating a set of trees has an added advantage—it contains inherent information about phylogenetic uncertainty. If 95 percent of the trees contain a particular feature, you can have 95 percent confidence in this feature.

Tree reconstruction assumes an underlying evolutionary model—which specifies, for example, whether A to G and C to T substitutions occur at the same or different rates. In the past, these models were over-simplified, says **Spencer Muse**, **PhD**, associate professor of statistics at North Carolina State University. They assumed that changes at one nucleotide position were independent of changes in other positions, which is unlikely to be true within a codon; they also ignored evolutionary constraints imposed by quirks of viral biology such as overlapping reading frames (where multiple genes with different starting points overlap the same sequence). But "there's a much richer class of models available now," Muse says. He and **Sergei Kosakovsky Pond**, **PhD**, developed the popular program HyPhy (Hypothesis Testing Using Phylogenies, http://www.hyphy.org/), which among other features, allows users to flexibly specify evolutionary models. "If you can write the model down, you can put it in the package," says Kosakovsky Pond, who is an assistant adjunct professor of medicine at the University of California, San Diego.

Many inferences can be gleaned from evolutionary trees once they're built, as each unique pattern of evolution leaves a unique signature in the tree. For example, within a host, HIV evolution is primarily driven by natural selection (immune or drug pressures); one lineage survives at a time, and this gives rise to a tree with a single diverging branch. In contrast, at the population level, HIV primarily evolves by random mutations (genetic drift)—and this results in dense trees with lots of branches at each time point. In the past decade, important advances have been made in the computational and statistical techniques that are used to make inferences from evolutionary trees. These are highlighted in the examples that follow.

The global level: dating HIV's origins

Scientists have used phylogenetic analysis to detail the history of the HIV pandemic—including when, where, and how it got into humans, as well as when and how it spread throughout the world. "Computational analysis has been absolutely fundamental in understanding the origins of the virus. And it's been a real success story," Holmes says.

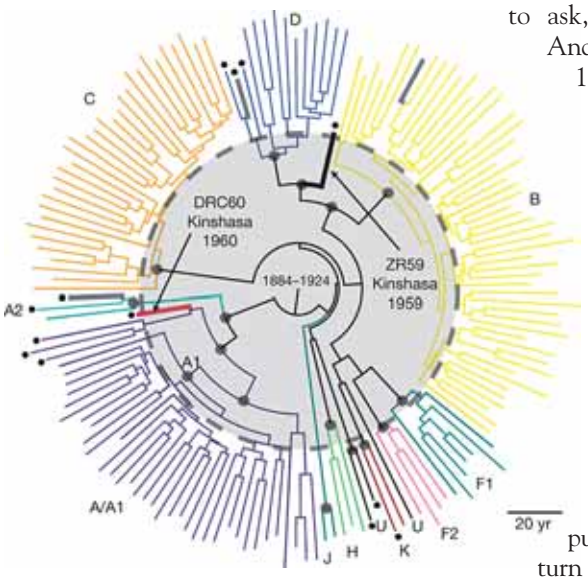
HIV can be divided into two types (HIV-1 and HIV-2) and three groups within HIV-1 (M, N, and O), but HIV-

1 M is the strain that predominates in the global pandemic. This strain descended from viruses found in chimpanzees in Eastern Cameroon, and appears to have first gained a foothold in what is now the city of Kinshasa in the Democratic Republic of the Congo. The two oldest known HIV sequences were unearthed there—one from a stored 1959 blood sample and another from a 1960 tissue sample, which was just discovered last year.

It's been hotly debated as to when HIV-1 M first crossed into humans. To date such events, scientists must convert from units of genetic distance on an evolutionary tree to units of time. Initially, they did this by assuming that that all lineages of the tree evolved at the same rate, but this was biologically unrealistic.

"Computational analysis has been absolutely fundamental in understanding the origins of the virus. And it's been a real success story," Holmes says.

The development of "relaxed" molecular clock models—which relax that strict assumption—has improved the accuracy of dating. This model is implemented in programs such as BEAST. How accurate is it? To check this, you can pretend that you don't know the ages of the oldest sequences and use the tree to date them, Pybus says. "We use the rest of the data to ask, exactly how old are they? And the result comes out bang on, 1959 and 1960."



Dating the Birth of HIV. Scientists used a representative sample of modern sequences and the two oldest known sequences (from Kinshasa, 1959 and 1960) to date the origins of the virus to somewhere between 1884 and 1924, most likely 1908. Branch lengths are depicted in units of time in years. Different colors represent different subtypes of HIV. Reprinted by permission from MacMillan Publishers, Ltd.,* Nature *455: 661-664 (2 October 2008).

In a 2008 *Nature* paper, scientists showed that HIV-1 M entered humans decades earlier than had previously been thought. Using BEAST software, they built an evolutionary tree from the 1959 and 1960 sequences and a sample of modern sequences, and then dated the root of the tree. Whereas earlier studies had pinpointed the date at around 1930, the new study put the estimate closer to the turn of the century (between 1884 to 1924, most likely 1908).

The *Nature* paper clearly refutes the contentious theory that HIV was introduced to humans during mass polio vaccination campaigns in Africa in the late 1950s. HIV was circulating in humans long before then. "So a lot of these evolutionary techniques pretty much put the squash on that hypothesis," Crandall says.

From Kinshasa, HIV-1 M made its way out to the rest of the world, and diverged into ten unique subtypes (A, B, C, D, E, F, G, H, J, and K). A 2007 paper in *PNAS* showed that HIV travelled from Africa to Haiti in 1966 (likely in a single host); and then from Haiti to the U.S. in 1969 (also in a single host); these "founder events" gave rise to the B subtype which now predominates in North America and Europe.

The regional level: monitoring national epidemics

At the population level, HIV evolution is driven primarily by random mutation (genetic drift), rather than particular selective pressures. So, how HIV evolves depends on how many people it infects and where it happens to spread—and the evolutionary trees reflect these factors. Thus, scientists can work backward from the trees to unravel the virus's demographic history in a particular region (called "coalescent theory"), as well as its migration patterns (aptly named "phylogeography"). These computational techniques can complement or even stand in for traditional epidemiology, and can help guide intervention strategies.

Coalescent theory reveals how quickly a virus was sweeping through a particular region just from the shape of the evolutionary tree. "Different rates of transmission give rise to different shaped trees, and coalescent theory is an explicit mathematical formulation of that," Pybus says. This mathematical framework is implemented in BEAST.

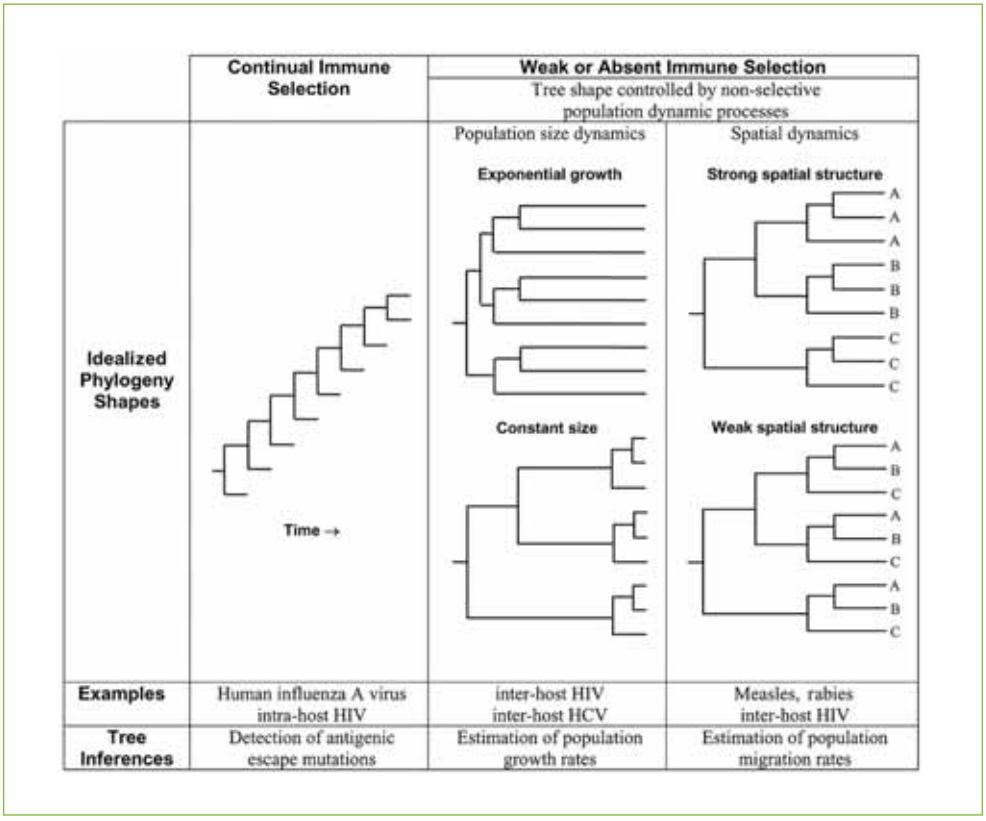
"You can find signatures in the tree

"Different rates of transmission give rise to different shaped trees, and coalescent theory is an explicit mathematical formulation of that," Pybus says.

for not only what the size of the population was, but whether or not it was changing through time. So, exponentially growing populations give a different signature in the shape of the tree than populations with a constant size. That has been used in HIV to look at how rapidly the HIV pandemic has expanded throughout various parts of the world," Drummond explains.

Using phylogenetics techniques including coalescent theory, Pybus and his colleagues showed that there were six independent introductions of HIV into gay men in the UK in the early 1980s; and each of those strains spread rapidly until the mid-1990s and then trailed off, corresponding to the introduction of effective combination therapy against HIV. Surveillance data from the UK showed similar overall patterns, but missed the underlying genetic structure of the epidemic.

"Sometimes these methods are more accurate for tracking prevalence than surveillance can be—especially in places where surveillance is very limited or governments have reason to hide information," says **Sudeb Dalai**, **MS**, an MD/PhD student at Stanford. "The genotypes are going to reflect the truth regardless of whether the surveillance actually does or not," he says. Dalai's team reconstructed the HIV epidemic in Zimbabwe—a place where surveillance data are shaky—and showed that the epidemic grew exponentially in the 1980s, correlating with political change and instability in Zimbabwe, but reached a plateau by 1991, possibly reflecting effective intervention campaigns. When they back-calculated HIV incidence from mortality statistics, they got a similar trajectory for the epidemic.



Signatures of Evolution. When strong selective pressures (such as immune pressures) are driving evolution, less fit lineages die out and the most fit lineages survive, giving rise to a characteristic tree pattern (left panel). In the absence of strong selective pressures (such as at the population level for HIV), multiple lineages exist at once (center and right panels). Different growth rates for the virus also give rise to different tree patterns; the top tree in the center panel reflects exponential growth and the bottom tree in the center panel reflects constant growth. From Grenfell,* et al.*, Unifying the Epidemiological and Evolutionary Dynamics of Pathogens.* Science *303: 327-331 (16 January 2004). Reprinted with permission from AAAS.

"Sometimes these methods are more accurate for tracking prevalence than surveillance can be—especially in places where surveillance is very limited or governments have reason to hide information," says Sudeb Dalai.

With phylogeography, scientists blend phylogenetic information with geographical information to evaluate how the virus travels in space. One can count migration events off an evolutionary tree as follows: if an ancestral sequence was sampled from region A and a direct descendent was sampled from region B, you can infer an A to B migration, explains **Marco Salemi**, **PhD**, assistant professor of pathology, immunology, and laboratory medicine at the University of Florida.

Using MacClade software, a popular

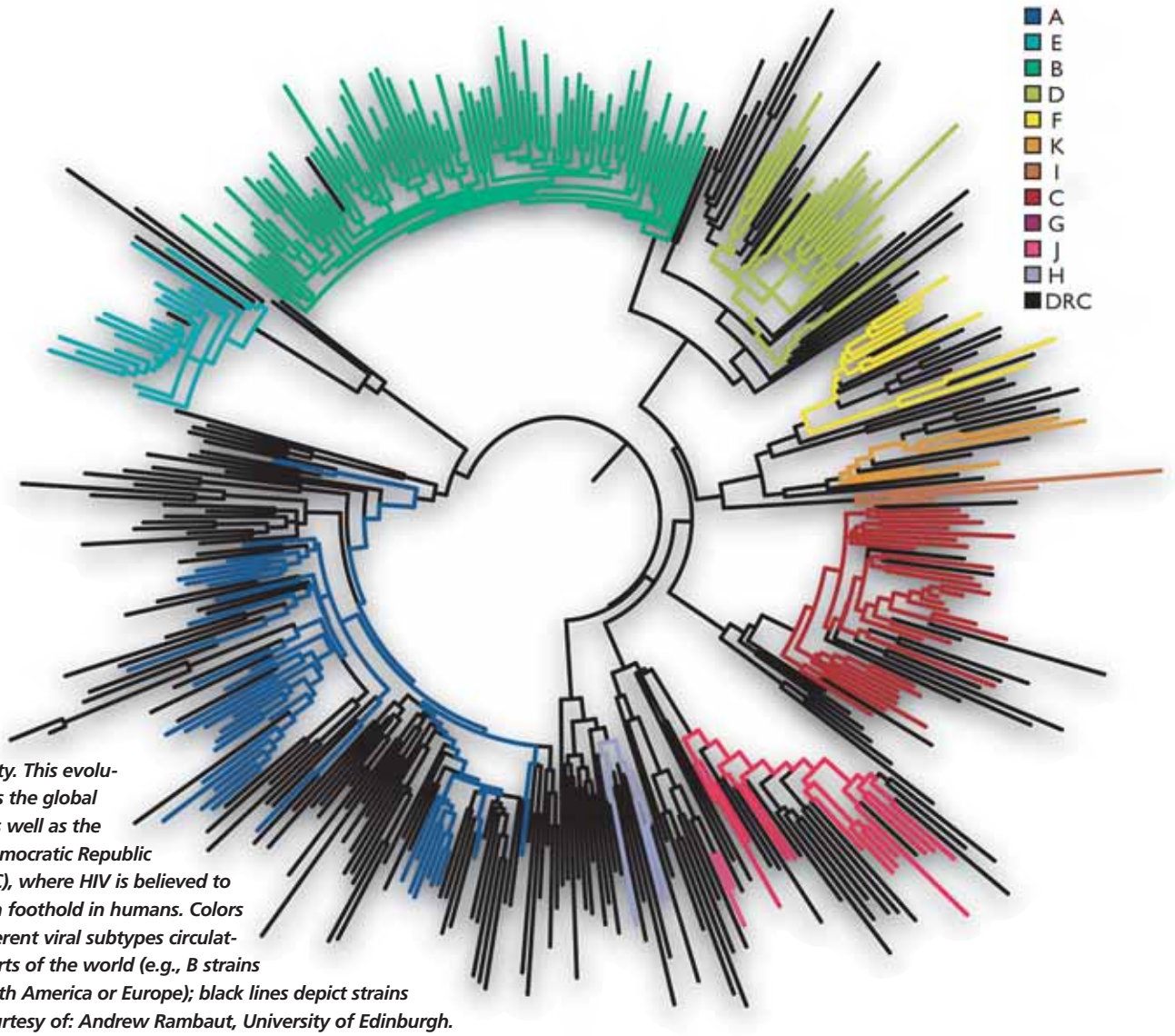
program for doing phylogeography, Salemi and his colleagues studied the HIV epidemic in Albania and Bulgaria, two countries where traditional epidemiologic data are lacking. Both countries—which were part of the Soviet Bloc during the Cold War—had explosive HIV epidemics during the early 1990s, likely related to the end of communism and the turmoil caused by nearby wars. The epidemics are dominated by subtype A, which is also prevalent in Russia and the Ukraine. But, surprisingly, Salemi and his team traced the source of the epidemics to Africa, not to Russia or the Ukraine. In Albania, HIV was introduced in the center of the country (in the capital) and then slowly spread to the periphery of the country, including its ports, which are just two or three hours to Italy by boat. "So it has the potential to modify the epidemic that is ongoing in Italy, in France, in Western Europe, and later on eventually in the U.S.," Salemi says. In the last four years, the Italian authorities have found that 30 percent of new HIV infections are with non-B subtypes, up from only 5 percent before, he says. These findings may have implications for where to target interventions.

The local level: linking infected individuals

Phylogenetics can also identify transmission networks—people who likely infected one another—and thus may help direct local public health efforts as well as help prove guilt or innocence in criminal cases involving HIV transmission.

For example, a landmark 2009 paper in the journal *AIDS* describes an effort to use sequence data for local public health surveillance in San Diego. "We're very interested in finding hotspots of HIV transmission—people who transmit to quite a few people, the nodes of the network so to speak," says lead author **Davey Smith, MD**, assistant professor of medicine at the University of California, San Diego. "Then the next step would be to intervene on those individuals, or at least to figure out what characteristics are common to them." This approach is routinely used to help control other reportable diseases such as syphilis and tuberculosis, but this is one of the first attempts to adapt it to HIV.

If you take HIV virus from two random HIV-positive people in San Diego,



***Circulating Diversity.** This evolutionary tree shows the global diversity of HIV, as well as the diversity in the Democratic Republic of the Congo (DRC), where HIV is believed to have first gained a foothold in humans. Colors represent the different viral subtypes circulating in different parts of the world (e.g., B strains sampled from North America or Europe); black lines depict strains from the DRC. Courtesy of: Andrew Rambaut, University of Edinburgh.*

their HIV polymerase gene sequences will be about 5 percent different, says co-author Sergei Kosakovsky Pond (also of the University of California, San Diego). If they are less than 1 percent apart, then they are almost certainly linked—either through direct transmission within the pair or through transmission from a common partner. In their study of 637 individuals, they found that 25 percent were linked; the largest cluster comprised 12 individuals. Next, you can draw diagrams showing how all the sequences in the clusters connect, as well as incorporate information on "people factors"—for example, data on the patients' sexual partners—to build a comprehensive computational model of the local transmission networks, Kosakovsky Pond says.

The UCSD researchers also track the transmission of drug resistant strains. About 20 percent of new HIV cases in San Diego are infected with a drug resistant strain, Smith says. This is a major problem because it takes three drugs to control the infection, and if a person is already resistant to just one of these drugs, they may quickly develop resistance to the other two. "So then we've just blown three drugs for this person," Smith says. In a 2008 paper in the *Open AIDS* journal, Smith's team showed that methamphetamine users in San Diego have a high frequency of transmitted drug resistance.

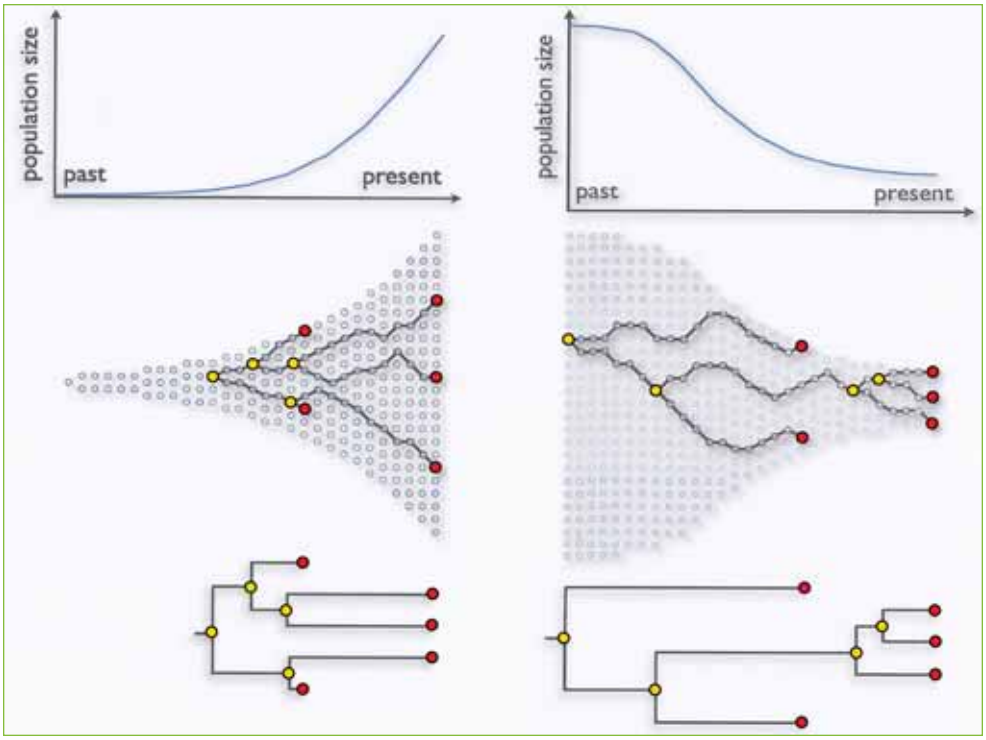
Phylogenetic data have also been used as evidence in HIV criminal trials. For example, in a highly publicized case in Libya, six international medical workers were sentenced to death for allegedly infecting hundreds of children in a Libyan hospital with HIV. A month before the final appeal hearing in 2006, Pybus and his colleagues were asked to analyze viral sequences from the infected children. "That basically gave us a few weeks to actually do the analysis and write it up and get it published before this trial. We knuckled down and did the analysis in about seven days. We didn't get much sleep," Pybus says. Using BEAST software on a 60-processor computer cluster at Oxford University (which they tied up for the week), they built evolutionary trees and dated the most recent common ancestor of the outbreak. The paper was published in *Nature* with just days to spare before the hearing. Their findings unequivocally exonerated the medical workers: The outbreak arose from a single ancestor that predated the arrival of the medical staff to the hospital (March of 1998); and 40 percent of the diversity in the circulating strains was already present when the staff arrived. The epidemic was probably due to a long-standing infection control problem at the hospital, Pybus says.

Though the paper was ignored at the 2006 trial and the medical workers were again sentenced to death, it apparently had an impact behind the scenes, Pybus says. "From what I've heard, our analysis did help in that it changed the tone of the diplomatic negotiations afterward." Six months later, the medical workers were released. "That was enough politics for me for one lifetime," Pybus says.

The host level: glimpsing natural selection in real-time

Scientists use the same techniques to study within-host evolution as they use to study population-level evolution. There's one difference, however: natural selection plays a much bigger role in driving evolution within an individual, as the virus attempts to escape specific immune and drug pressures. "It's beautiful natural selection, just like Darwin explained," Crandall says. Identifying these escape mutations presents an additional challenge for modelers.

Evolutionary studies show that when HIV is transmitted to a new host, a single virus is often responsible for seeding the infection. A few weeks into the infection all the viruses have a single common ancestor that dates to the start of the infection. "HIV goes through a really severe evolutionary bottleneck when it gets transmitted from one person to another," says **Bette Korber, PhD**, a laboratory fellow in the theoretical biology and biophysics group at the Los Alamos National Laboratory. The virus that is successfully transmitted may have unique characteristics, and could be specifically targeted by vaccines or early drug treatment.

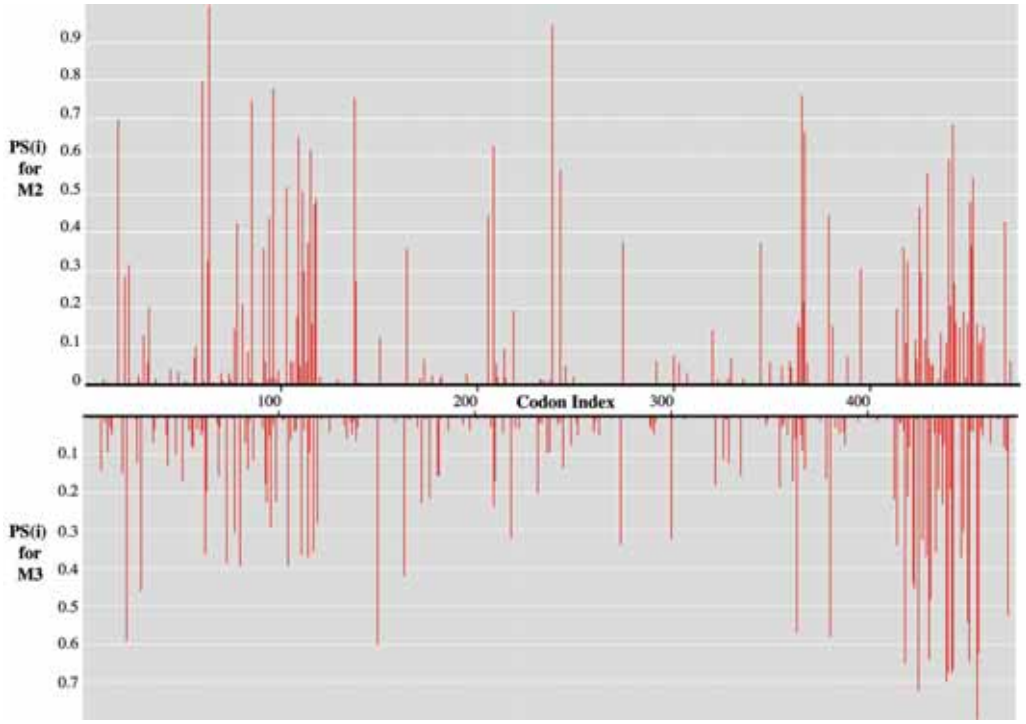


Coalescent Theory Explained. The shape of an evolutionary tree reflects the underlying population dynamics of the virus. The left panel illustrates the characteristic tree for exponential growth, whereas the right panel illustrates the characteristic tree for exponential decline. When the population size is small (e.g., early in the left example; later in the right example), branching events are more common. Courtesy of: Andrew Rambaut, University of Edinburgh.

"You can actually see, if you look at the evolutionary trees, lineages keep dying out and only one survives; and that process occurs over and over, across five to ten years of infection," Pybus says. "This arms race goes on and on; the only problem is, the immune system always seems to lose."

After transmission, HIV undergoes weeks of rapid replication (acute infection) until the immune system finally wakes up and starts fighting back. In the absence of treatment, the virus and the immune system settle into "an evolutionary arms race," Pybus says. "You can actually see, if you look at the evolutionary trees, lineages keep dying out and only one survives; and that process occurs over and over, across five to ten years of infection," he says. "This arms race goes on and on; the only problem is, the immune system always seems to lose."

The virus evolves to escape two different immune pressures—antibodies and killer T cells (cell-mediated immunity). To escape antibodies, HIV changes the shape of its envelope (surface) proteins. To escape cell-mediated immunity, HIV switches amino acids in epitopes, which are short snippets of viral protein that are displayed on the surface of HIV-infected host cells to alert killer T cells. Scientists can identify these escape mutations (using programs such as HyPhy) because genes undergoing positive selection leave a classic genetic signature—non-synonymous mutations (mutations that change the amino acid) occur more frequently than synonymous mutations (mutations that preserve the amino acid). Understanding these



Selection Detection. Scientists use computational methods to detect areas of the HIV genome that are evolving under positive selection (where nucleotide changes that alter the amino acid occur more frequently than changes that preserve the amino acid). Here, for a 500-codon stretch of the HIV genome, red bars indicate the probability that each codon is undergoing positive selection for each of two models of evolution—an over-simplified model (above) and a more biologically realistic model (below). The inferences from the two models differ considerably at several sites (for example, codons 65, 120, 230, and 470). Courtesy: Spencer Muse, North Carolina State University.

escape routes informs vaccine design (see sidebar on vaccine design).

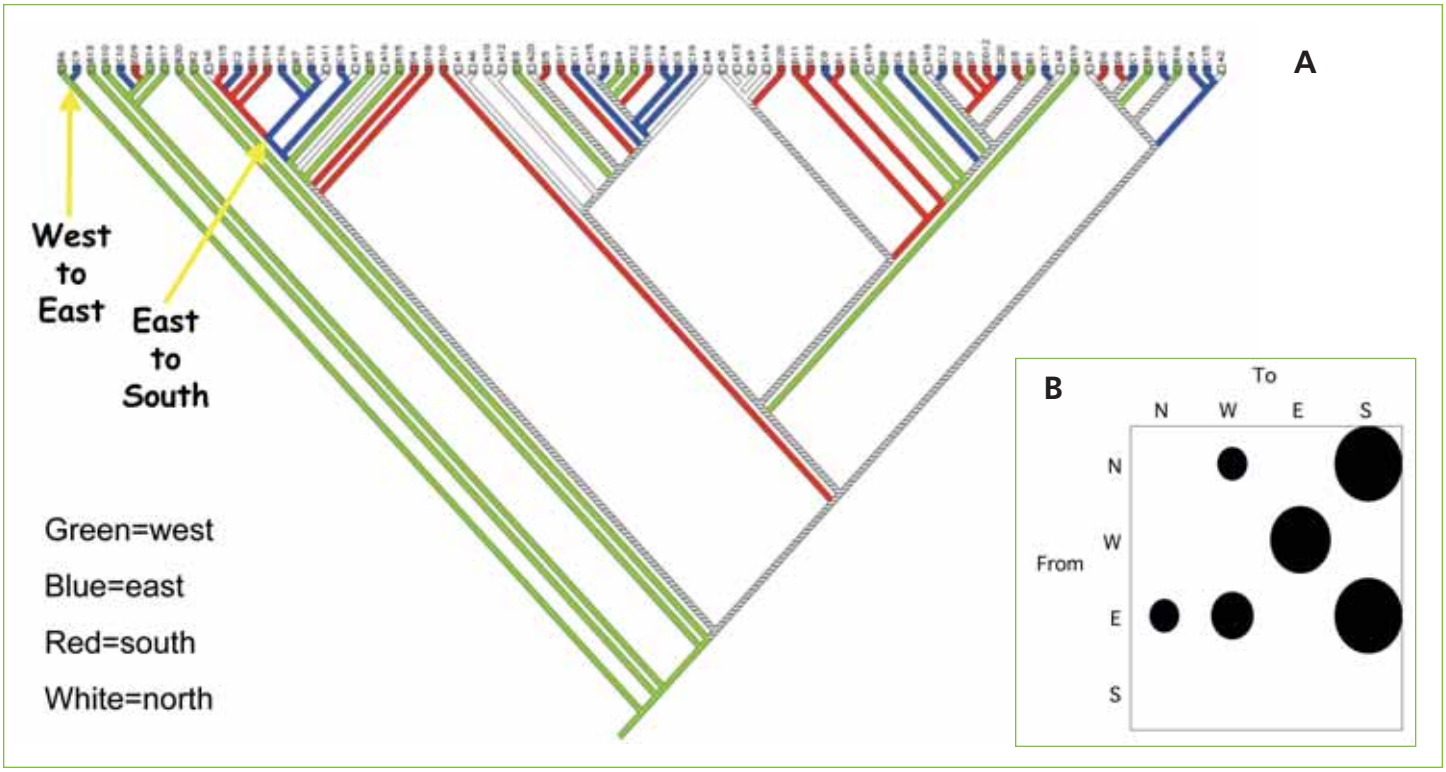
HIV infects many different cell types in the body (not just immune system cells). As a result, the virus may become isolated in particular tissues and evolve independently from viruses in the rest of the body, giving rise to tissue-specific strains. For example, about 70 percent of patients exhibit near-complete phylogenetic segregation between sequences in the brain and in the blood, says **Satish K. Pillai, PhD**, assistant professor of medicine at the University of California, San Francisco.

In addition to the establishment of latent infection (non-productive infection of long-lived resting cells), this "compartmentalization" effect helps explain why we can control HIV infection with drugs but we can't eradicate it, Pillai says. Drugs can push the virus to basically undetectable levels in the blood, but the virus may continue to thrive elsewhere. "The virus has a peaceful little sanctuary site that it can hide out in," Pillai says. If scientists could figure out where the virus is hiding and continuing to replicate, they could design specific treatments to target those cells.

Understanding how the virus evolves in different parts of the body has other clinical implications as well. "It's not purely an academic pursuit," Pillai says. Evolutionary pressures differ between tissues, which may cause HIV to evolve in specific, predictable ways. In a February 2005 *Journal of Virology* paper, Pillai and his colleagues documented unique genetic signatures associated with viruses in the male genital tract. This could open the door to vaccines or microbicides that specifically target genetic variants that reside here, Pillai says.

In the brain, viral evolution may be related to HIV-associated dementia—a debilitating condition that occurs even in those on effective drug treatment, Salemi says. In a September 2005 paper in the *Journal of Virology*, Salemi and his colleagues used phylogeography to track the migration of HIV through the brain of a patient who had severe HIV-associated dementia at death. They found that the virus entered the brain through the meninges and then spread to other brain areas, including the temporal lobe. Viruses in the temporal lobe were evolving at a faster rate than elsewhere in the brain, which could be a clue to the cause of dementia, Salemi says.

Pillai's team also studies HIV evolution in the brain, using virus sampled from the cerebrospinal fluid of living patients. They are hunting for genetic signatures of brain-associated HIV and for mutations that correlate with cognitive impairment. For example, using machine learning techniques to correlate particular mutations with scores on a cognitive deficit test, they identified a serine residue in a particular loop of the envelope protein that is "very significantly correlated with severe cognitive impairment," Pillai says. It may be possible to design a therapeutic vaccine to steer HIV evolution away from acquiring such harmful mutations during the course of infection, Pillai says. "That would be really cool. We don't have that technology yet, but I



Phylogeography Explained. By constructing an evolutionary tree from HIV sequences collected from different geographical regions, scientists can determine the pattern of gene flow from one region to another. In panel A, colors represent strains from different geographical regions (green=west, blue=east, red=south, white=north); migration events can be directly counted off the tree as indicated. This information is compiled computationally and translated into a bubble plot (panel B) which quantifies the gene flow between different regions. Pictures were generated using MacClade software. Courtesy of: Marco Salemi, University of Florida

Vaccines get an evolution lesson

The hunt for an HIV vaccine has been marked by highly publicized failures and enormous disappointments. HIV presents an immense challenge to vaccine designers because the organism is so diverse. As a benchmark, consider the flu vaccine—it must be reformulated annually because flu strains diverge by about 1 to 2 percent per year in the population. In comparison, HIV mutates by about 1 percent per year within just a single person; and, across the globe, different subtypes of HIV differ by up to 35 percent. Designing a vaccine to cover all (or even a useful fraction) of this diversity has turned out to be, thus far, insurmountable.

To meet this challenge, evolutionary scientists are designing viral proteins *de novo* in the computer that attempt to summarize HIV's variation. Thus far, they can make a "consensus" sequence by determining the most common amino acid at each position from a broad sample of sequences; reconstruct HIV's ancestral sequence (which is genetically between all modern strains); or build a "center of the tree" sequence, which has the lowest total genetic distance to all other sequences in an evolutionary tree. The resulting computer-generated proteins serve as immunogens in vaccines.

"It took a while to get everyone accustomed to the idea that you might want to artificially design a protein on the computer rather than use a natural protein. People didn't know if it would fold properly, if it would be antigenic, or if it would have the same sites that were relevant for an immune response as a natural strain. As it turns out, it does," Bette Korber says. Korber runs the Los Alamos National Laboratory HIV Sequence Database (http://www.hiv.lanl.gov), which provides the datasets often used for these approaches. In a 2008 paper in *PNAS*, her group showed that a consensus envelope protein stimulated three-to-four fold higher cell-mediated immune responses in monkeys than a natural envelope protein.

Another tactic is to design a vaccine from a diverse set of viral epitopes—the small fragments (typically nine amino acids) of virus recognized by killer T cells. Korber is piecing together the most common epitopes (cataloged in the Los Alamos database) into a small set of composite proteins. "I create sort of little Frankenstein proteins that look and feel like HIV proteins but they don't exist in nature," Korber says. So far, the proteins are showing good immunogenicity in animals, she says. "We're getting really good signals in the mice and the monkeys—which makes us delighted; they're doing a lot better than just natural proteins."

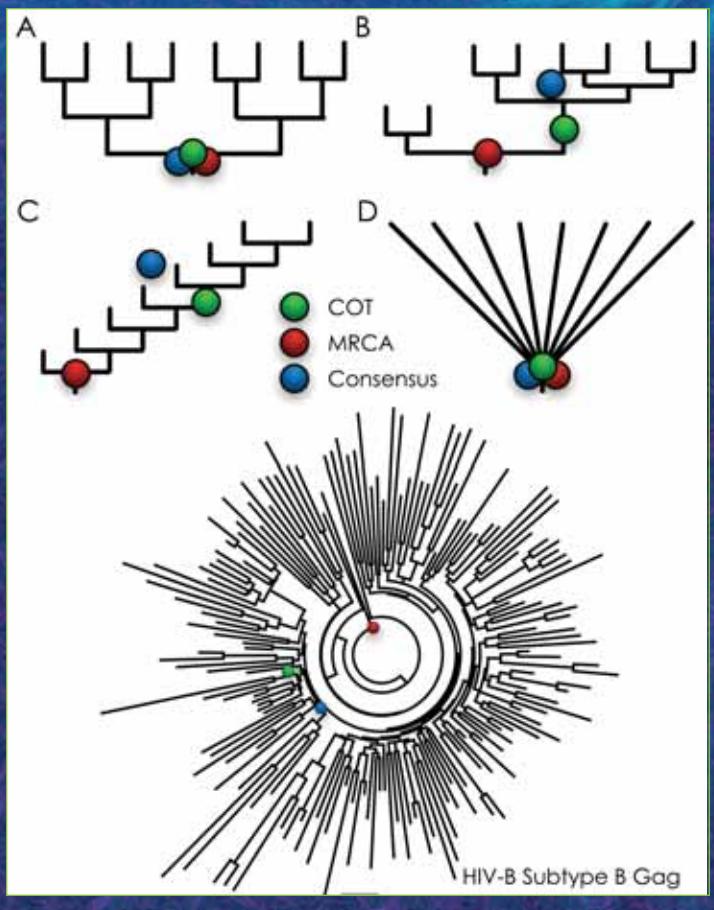
"It took a while to get everyone accustomed to the idea that you might want to artificially design a protein on the computer rather than use a natural protein," Bette Korber says.

Despite Korber's success in animal models, other groups report poor results with similar strategies. Both the center of the tree and a "center of the tree plus" method (which added diverse epitopes) came up negative, says **David Nickle, PhD**, a senior research biologist at Rosetta Bioinformatics. "I've become convinced that none of those [approaches] are going to be adequate," agrees James Mullins, who collaborates with Nickle.

Mullins and Nickle believe that HIV tricks the immune system into mounting an initial response against "decoy elements"—pieces of the virus that are easily mutated. The immune system then becomes trapped by its initial response (a phenomenon called "original antigenic sin"), and ends up focusing only on these variable regions, to HIV's benefit. "My prediction is that all vaccine approaches will fail until we take into account and remove these decoy elements," Mullins says.

They are designing a "conserved elements" vaccine that contains only segments of HIV that are highly conserved—regions that don't evolve much and may not tolerate variation. "The more conserved an amino acid is in viral evolution, the more likely it is that it plays a critical role in the function of the virus," Mullins says.

Getting the immune system to attack these areas first may be the key to crippling the virus, they believe. "If you let the immune system choose what to mount an immune response to, maybe it chooses badly. So we want to redirect the immune system to mount a response to these conserved regions, even though it may be harder," Nickle says. The approach is still in the early stages of development.



Computational Vaccine Design. To tackle HIV's diversity, evolutionary scientists are designing artificial "summary" HIV proteins in the computer that may stimulate a broader immune response than natural HIV proteins. This picture compares three approaches—center of the tree (COT, green), ancestral (MRCA, red), and consensus (blue). The center of the tree approach constructs a sequence with the lowest total genetic distance to all variants in the tree; the ancestral approach reconstructs the sequence of the most recent common ancestor of the tree; and the consensus approach chooses the most common amino acid at each position from all variants in the tree. Depending on the evolutionary history, the three approaches may yield very similar or very different proteins (upper panels A-D). The bottom panel shows results from the three approaches for the HIV gag protein (lower panel). Courtesy of: David Nickle, Rosetta Bioinformatics.

think it's a possibility."

Unlike the immune system, which eventually loses out to HIV evolution, drug treatment can keep the virus in check indefinitely. But keeping ahead of drug resistance is a major undertaking. "The virus is fantastically plastic and adaptable. It can evolve resistance to all these drugs that we've developed against it. And it pretty much tends to evolve resistance to them in clinical trials—before they even get put on the market," Pybus says. HIV-1 protease (one of the nine viral proteins) has 99 amino acids, and in very heavily treated people as many as 25 of the positions can be mutated, says **Robert Shafer, MD**, associate professor of medicine and pathology in the division of infectious diseases at Stanford University; Shafer runs the Stanford University HIV drug resistance database (http://hivdb.stanford.edu/). "That never ceases to amaze me: that one quarter of the amino acids can be mutated—and it still functions," he says.

The virus mutates in predictable ways in response to particular drugs, so the challenge is to document and keep track of these mutations to help physicians, epidemiologists, and drug designers. The Stanford database contains about 100,000 viral sequences, linked to data on *in vitro* and *in vivo* drug resistance, which researchers use to identify partial and full resistance mutations. Physicians can also enter sequence data and retrieve detailed information about their patients' mutations. "Helping clinicians interpret drug resistance tests is what's given the database the most recognition," Shafer says.

Computational scientists are working on providing new tools for physicians—for example, algorithms that predict the optimal drug regimen for a patient based on sequence data. "We infer rules for selecting optimal treatments from clinical databases; it's a big machine learning problem," says **Niko Beerenwinkel, PhD**, assistant professor of computational biology at the Swiss Federal Institute of Technology Zürich.

Beerenwinkel and his colleagues try to predict not only current drug resistance (both to single drugs and drug combinations), but also the potential for the virus to develop resistance in the future. To do this, they reconstruct the typical evolutionary paths of HIV under certain drug pressures—the sequence of mutations that may develop and the likely time frame. The resulting "mutagenic trees" are incorporated into their genotype-to-phenotype prediction algorithm

"That never ceases to amaze me: that one quarter of the amino acids can be mutated—and it still functions," Robert Shafer says.

Snapshots of a Pandemic. Evolutionary trees were created from all the HIV whole genome sequences available in 1993 (n=15, at top), 2003 (n=397, at center), and 2009 (n=1885, at bottom) from the Los Alamos HIV database (http://www.hiv.lanl.gov) and GenBank. Different colors depict different subtypes and recombinants of HIV—which is the most sequenced organism ever. This picture shows the increasing availability of whole genome sequences as well as HIV's increasing diversity. Courtesy of: Keith Crandall and Matthew Bendall, Brigham Young University

(http://www.geno2pheno.org). If the virus only needs one mutation to escape drug regimen A and five to escape B, we can predict that B will suppress the virus for a longer period, Beerenwinkel explains.

Evolving phylogenetics

Despite the progress made in HIV evolution and phylogenetics, some challenges remain. One issue is how to deal with recombination—where two different viral strains infect the same cell and exchange genetic material, so-called "viral sex." HIV actually evolves more rapidly by recombination than by point mutation, says **James Mullins, PhD**, professor of microbiology and of medicine at the University of Washington. But most tree-building programs don't account for recombination—which can lead to mistakes (especially when dealing with whole genome sequences) since a recombinant sequence actually has two separate lineages. "No one's attacked that problem really effectively in phylogenetics; I would say that's an understatement," Mullins says.

Several programs identify recombinant sequences and remove them prior to tree building. But what's really needed is a program that can detect recombination, figure out the breakpoints, and incorporate that history into the tree. This is a difficult task, because it greatly increases the number of possible trees. "If the phylogeny problem is NP-hard, one could say the recombination problem is NP-harder," Pybus quips. Nobody has solved the problem adequately yet, but BEAST developers are working on it, Drummond says.

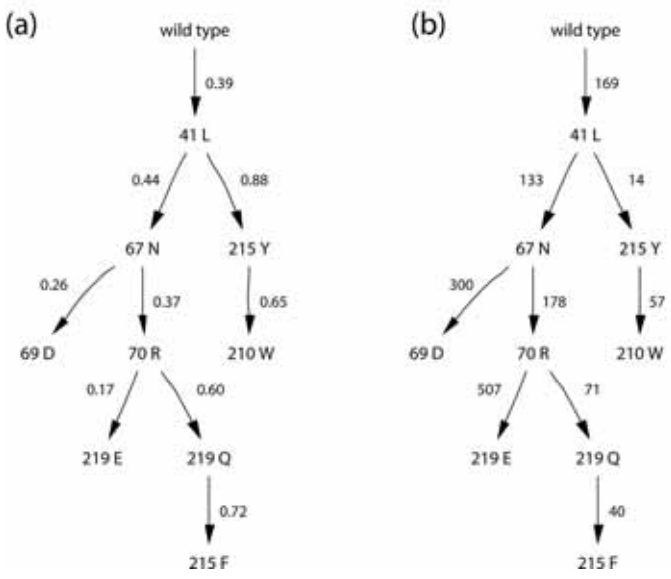
Another challenge is the rise of next generation sequencing platforms, such as 454 pyrosequencing, which increase the speed of sequencing by orders of magnitude. Besides providing a wealth of data for building evolutionary trees, the technology allows "deep sequencing," the ability to detect viral variants within a single patient that are present at very low levels—including low-level drug resistant variants—rather than just the dominant clones. This information may improve our ability to predict drug failure.

"But there's a lag between the sequencing technology and our methodology that processes these sequences," Kosakovsky Pond says. Current phylogenetics programs can handle hundreds of sequences, but next generation sequencing may provide thousands or tens of thousands of complete HIV genomes at once. "It's going to be a bit of a tidal wave for those of us who do the analysis," Pybus says.

Besides the sheer volume of data, the technologies present new bioinformatics problems, says **Allen Rodrigo, PhD**, professor of computational biology and bioinformatics and director of the Bioinformatics Institute at the University of Auckland in New Zealand. They yield short reading lengths, which have to be assembled; and they also have high error rates—which means it can be difficult to differentiate technical errors from real mutations in HIV. "This is going to open up a whole new set of computational challenges that we're just starting to look at," Rodrigo says.

> "If the phylogeny problem is NP-hard, one could say the recombination problem is NP-harder," Pybus says.

Researchers in HIV are once again at the forefront, driving forward these advancements in the study of evolution. Hopefully, these tools will, in turn, help drive HIV into extinction. □



***Predicting Evolution.** HIV develops resistance mutations to particular drugs in predictable ways. By linking genotypic and phenotypic data from large HIV databases, researchers can tease out these mutation pathways (called "mutagenic trees"). This picture illustrates the typical amino acid changes that HIV may undergo to develop resistance to the drug AZT. In these panels, the numbers shown along the arrows indicate (a) the probability of a particular mutation and (b) the average number of days it takes for each such mutation to occur. These mutagenic trees are incorporated into algorithms that predict optimal drug combinations for patients based on their viral genotypes (www.geno2pheno.org). Courtesy of: Niko Beerenwinkel, Swiss Federal Institute of Technology Zürich.*