Progress towards developing the 'pathogen toolkit'

NIJ Grant 2011-DN-BX-K534

Priyanka Kshatriya,[1] Vinson Doyle,[3] Bradley J. Nelson,[3] Xiang, Qin,[1] John Andersen,[3] Jeremy M Brown,[3] and Michael L. Metzker[1,2,*]

[1]Human Genome Sequencing Center and [2]Department of Molecular & Human Genetics, Baylor College of Medicine, One Baylor Plaza, N1409, Houston, TX, 77030; [3]Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, USA

Abstract

Microbial forensics is an emerging field that presents enormous challenges. Unlike human forensic analysis, disease-causing microbial pathogens of humans exhibit remarkable genomic diversity generated through a number of elaborate mechanisms, including high mutation and recombination rates as well as diverse responses to selection. One major goal of microbial forensics is to use this genetic diversity to identify the source of a pathogen used to commit a crime. While phylogenetic analysis of nucleotide variation within a small number of genes has been used in past forensic studies to assess relationships among pathogens, a large fraction of those genomes remain uncharacterized, ignoring useful information contained in the presence or absence of different genes. Additionally, complex evolutionary processes that generate variation in phylogenetic signal across genomes, such as host-specific responses, recombination, and convergent selection, have not been accounted for in current forensic studies. Recent advances in next-generation sequencing (NGS) technologies and phylogenetic analysis of complete genomes (phylogenomics) have the potential to significantly alter the technological approaches used in characterizing case samples. HIV transmission is an excellent test case for microbial forensics due to the virus' extremely high mutation rate, rapid response to selection, and strict dependence on human hosts.

In this study, we have demonstrated feasibility of full-length genome sequence production from individuals infected with HIV using NGS technologies and phylogenomic analysis. Samples from the Texas HIV transmission case (CC samples) were used in this study. For a subset of full-length HIV genomes, we have demonstrated initial phylogenetic estimates from six non-overlapping genic regions showing (*i*) the monophyly of clones sampled from the same recipient individuals, (*ii*) variation in relationships amongst groups of clones from different individuals, and (*iii*) variation in rates of evolution across different genomic regions. These results are significant in demonstrating the integrity of the phylogenetic signal in these assembled CC molecular clone sequences as the extrinsic evidence, revealed during criminal proceedings, strongly suggested that each individual included in our initial analyses only acted as a recipient within the Texas transmission cluster. Distinct relationships among clones from different individuals inferred from numerous genic regions are strong evidence of independent evolutionary histories, thus providing greater confidence in forensic conclusions. Several technical challenges are currently being addressed that include (*i*) optimizing *de novo* assemblies that will produce larger numbers of successful full-length genome sequences with low SNP counts and (*ii*) understanding the high estimates of human sequences in the subset of HIV molecular clone reads. Overcoming these challenges will lead into the development of a robust 'pathogen toolkit' that can provide a detailed roadmap for future forensics studies.

Table of Contents

Executive Summary:

Introduction – Problem and Purpose

Microbial forensics is an emerging field that is very important for public safety, yet it presents enormous challenges to both the scientific and legal communities. Microbes have developed a number of mechanisms for generating natural genetic variation, such as high mutation and high recombination rates. Recombinant DNA technologies can also be employed to rapidly combine novel genetic elements conferring new biological, and presumably pathogenic, properties to microbial 'super bugs' with potentially devastating effects. *One major goal of microbial forensics is to use this genetic variation to identify the source of a pathogen used to commit crimes or acts of bioterrorism.*

Rapid pathogen evolution invalidates the use of DNA 'fingerprinting' techniques. Instead, phylogenetic methods are ideally suited because they can explicitly model genetic change through time in order to identify pathogen source populations. While phylogenetic analysis of nucleotide variation within individual genes has been used in previous forensic studies, several drawbacks exist. First, much genomic information is ignored, including gene presence or absence, insertion and deletion events, and structural rearrangements. Second, forensic studies based on the phylogenetic analysis of only a few genes may not reflect the true evolutionary history of that organism. Differences between gene trees and true organismal trees can result from biological processes such as recombination, laboratory engineering through recombinant techniques, and/or analytical difficulties (e.g., misleading signal due to convergent evolution).

Next-generation sequencing (NGS) technologies have changed the way we think about scientific approaches in basic, applied, and clinical research. The major advance offered by NGS is the ability to produce an enormous volume of data cheaply, in some cases, in excess of one billion short reads per instrument run. Whole-genome sequencing of many related organisms allows large-scale comparative and evolutionary studies to be performed that were unimagined just a few years ago. However, NGS has not yet been utilized for forensic studies of HIV transmission. In this report, we described our progress in characterizing molecular clones derived from long-range PCR amplicons of individuals infected with HIV involved in a criminal case. The methods being developed here will have broad impact in the forensics field as a pathogen toolkit for future studies.

Research Design and Methods

Long-range PCR: To minimize false positives due to contamination, standard operating procedures employed in previous microbial forensics studies were used at all times. Genomic DNA for Texas case samples CC01-CC08, stored in a minus 80°C freezer since 2009, were individually retrieved and handled to avoid sample mix-up. Samples were processed in reverse order starting with CC08 and finishing with CC01. Along with each genomic DNA sample, the PCR setup included a negative control (molecular grade water) in order to check for contamination. For each CC sample, ten genomic DNA and two blank-water control reactions were PCR amplified as a set. Except for sample CC08, all other CC samples revealed a positive 9-kb PCR product in

the first PCR round. For CC08, 2 µL of either the first genomic DNA-amplified product or blank water control was used as template for the second round of PCR.

Cloning 9-kb fragments: For gel purification of 9-kb PCR products, each 9-kb band was excised, purified, and cloned. Gel-purified products were used for ligation reactions into the PCR XL TOPO vector. Ligation reactions were transformed into electrocompetent cells using electroporation. The transformed cell solution was then plated onto selective LB agar plates (containing zeocin and kanamycin) and incubated overnight. Approximately 30-50 colonies for each CC sample were selected as clones, followed by mini-preps and the preparation of glycerol stocks. To confirm the presence of a 9-kb insert, clones were digested with *Eco*RI and/or *Bam*HI, separately. Digested plasmids were analyzed on an agarose gel and clones that exhibited a restriction enzyme digestion pattern with DNA fragments that summed to approximately 12.5-kb in size were qualified initially as "full-length" HIV genome clones.

NGS library generation: Initially, 160 molecular clones (i.e., eight CC samples each with 20 full-length molecular clones verified by restriction enzyme analysis) were submitted to the Human Genome Sequencing Center's (HGSC) Library QC and Library Automation groups. Quality control tests were conducted on the 160 clones to quantitate DNA concentrations and verify clone sizes. Of this initial set, 156 molecular clones passed both tests and were further processed by the NGS library production group. Each HGSC-qualified molecular clone (qMC) was constructed into individually bar-coded Illumina paired-end (PE) libraries. Briefly, DNA from each molecular clone was sheared into ~550 bp fragments. A series of molecular biology techniques was then implemented including DNA end-repair, A-tailing, and Illumina adapter ligation with each step being followed by a purification step. DNA fragments were amplified to generate bar-coded NGS libraries. PCR products were then purified and quantified, and their size distribution was analyzed.

NGS sequencing: Aliquots of the libraries were prepared and combined into two pools (Pool 1: 76 samples and Pool 2: 80 samples), for which molecular template clusters were amplified in two different lanes on separate Illumina flowcells. The two flowcells contained amplified HIV libraries were sequenced using the Illumina TruSeq.v3 chemistry. Each flowcell was run on different HiSeq 2000 instruments, yielding a total of 759 million reads with an average of 4.87 ± 0.76 million reads per library sample. The minimum number of reads was 3.11 million (CC04-19) and the maximum was 7.20 million reads (CC02-05).

Read pre-processing before assembly: Illumina sequencing data was processed to generate raw read sequences and adapter sequences were removed. Base-calls at the end of reads with Illumina quality scores of ≤2 were trimmed (referred to as "trimmed reads" from here on). Attempts were made to map trimmed reads to the human reference genome. Reads mapped to the human genome sequence were removed, with the remaining reads then screened for *E. coli*, ΦX174 (Illumina's internal sequencing control) and the PCR XL TOPO cloning vector sequence. Illumina reads that did not map to any of the reference sequences and had lengths of ≥30 bp were subsequently used in *de novo* assemblies (referred to as "screened reads" from here on).

Assembly: Velvet, Fermi, MIRA, Newbler, and Phrap assembly programs were evaluated for *de novo* assembly quality for a subset of samples using all screened reads as input. With the exception of Phrap, none of the programs listed above were able to assemble the screened reads into a single HIV contig using the parameters tested. Due to high read-depth coverage, only 50,000 reads were randomly extracted from screened reads (all screened reads if the total number was <50,000 reads) and used for Phrap assembly. If the HIV contig from the trimmed read assembly was >8.9-kb, the contig was subsequently used in the multiple sequence alignment and phylogenetic analysis. If assembly produced a contig <8.9-kb, it was considered as an initially failed product and was not included in the analysis of this report. These assemblies have been earmarked for more depth analysis to understand the reason(s) for the failure. The Los Alamos National Laboratory (LANL) curated HIV database was used to compare all *de novo* assemblies from the CC clones to the HXB2 genome to verify expected gene organization.

Outgroup selection: In order to assess the monophyly of HIV sequences obtained from case samples, we selected sequences from the LANL curated HIV database to include in all subsequent analyses. Each set of intra-individual sequences was aligned and pairwise distances between all sequences were computed. The two most divergent sequences within each set were then chosen to query the LANL HIV database using BLAST. The top match for each query, excluding sequences flagged as problematic, were added to the set of control sequences. In addition to these control sequences, the standard HIV-1 reference genome (HXB2) was also included.

Genome partitioning: A significant body of research has led to a detailed annotation of the HIV-1 genome. These annotations provided the genome coordinates used to extract regions of interest from the reference genome and orthologous sequences from the case and control sequences. Several regions of the HIV-1 genome were selected for analysis in order to evaluate variation in evolutionary rate and topological signal across the HIV-1 genome. We selected only non-overlapping protein coding regions (plus the LTR region) in order to avoid analyzing the same genomic regions in separate analyses.

Phylogenetic analysis: Phylogenetic inferences were drawn from multiple sequence alignments after masking sites whose alignment was uncertain. Filtering these sites is expected to help reduce spurious phylogenetic inferences resulting from alignment uncertainty. For each genomic region, the best-fit model of sequence evolution was chosen from among 88 candidate models and used to find the maximum-likelihood (ML) topology. In addition to assemblies comprised solely of newly generated Illumina sequences, we also aligned relevant subsets of the new sequences to previously generated Sanger sequences of *env* and *pol* from independent CC clones. Alignment, model selection, and ML phylogenetic inference were all performed as outlined above in order to verify the integrity of Illumina sequences compared with Sanger sequences from the same CC samples.

Modifications to the original research design: As noted in our March 30, 2013 progress report, the specific aims of the proposal remained the same as those originally funded, although the work would now be primarily focused on Texas case samples, and with time permitting, the

Washington and Louisiana case samples would then be processed. Since that report, our efforts have focused exclusively on the Texas case samples.

Phylogenetic methods to account for unmodeled processes: In addition to generating and analyzing whole viral genome sequences, we also developed a new analytical approach to detect the action of viral evolutionary processes not accounted for by 'standard' phylogenetic models (1) and applied it to an existing Sanger sequence data set of *env* sequences from the Texas case (2). More specifically, we examined the potential roles of convergent evolution and recombination in driving spurious phylogenetic conclusions. To test for the influence of recombination, we employed a hidden Markov model (3) that allows for variation in topology across sites. To compare models with different numbers of topologies, we developed an approximate AIC approach and validated it through simulations. To detect the action of selection, we employed codon-position-specific phylogenetic analyses and codon-based models that allow us to detect sites with elevated non-synonymous (dN) to synonymous (dS) rate ratios (dN/dS). Lastly, we developed a site-specific-likelihood profile by calculating site-specific likelihoods using the overall maximum likelihood (ML) tree and the best tree consistent with particular transmission hypotheses. The difference between these site-specific likelihoods then acts as a measure of site-specific support for or against the transmission hypothesis. Using permutation tests, we can assess whether sites in particular codon positions exhibit more or less support for the transmission hypothesis (or the ML tree) than expected by chance. We can also visualize the distribution of support across an alignment to see if sites supporting particular hypotheses are spatially clustered.

Findings

Long-range PCR: As we described in our original proposal, the technique of long-range PCR was employed to amplify ~9-kb of the HIV-1 genome. In our July 30, 2012 progress report, we reported successful long-range PCR for the clone pNL4-3 and case sample CC08. As noted in that report, however, we obtained several major, non-specific bands, which reduced the yield of the desired 9-kb product. Attempts to improve product specificity by altering temperature conditions were met with little success. We, therefore, proceeded by purifying the 9-kb band from the CC08 sample and cloned it using the pCR-XL-TOPO cloning vector system. Our preliminary results suggested the 20 molecular clones analyzed by the restriction enzyme *Eco*RI revealed that we had cloned the 9-kb PCR product. This result, however, could not be confirmed when using the alternative restriction enzyme, *Bam*HI. Moreover, Sanger sequencing of the ends of the 20 molecular clones did not confirm that we had cloned full-length HIV genomes (data not shown).

Identification of robust long-range PCR primers: Our original nested PCR primers were amplifying genomic DNA in a non-specific manner. After an exhaustive search, we settled on a new set of nested primer sequences. For the majority of CC samples, these new primers produced a single PCR product band of ~9-kb without any noticeable non-specific bands. With the exception of CC08, all Texas case samples were amplified with just the outer PCR primers. All PCR products were, nonetheless, purified by agarose gel electrophoresis.

Cloning 9-kb fragments: After gel purification, the 9-kb fragments were cloned to resolve the

population of HIV genomes into individual isolates. In general, the cloning efficiency was sufficiently high to produce hundreds of colonies per plate with low background. Restriction enzyme analysis suggested that 63% of colonies selected contained the full-length insert. For each CC sample, 20 molecular clones containing full-length inserts were selected for NGS library preparation.

NGS library generation and sequencing: Molecular clones were further quality-controlled for quantity and size, with 97.5% (156/160) passing this second QC criterion. These "HGSC-qualified molecular clones" were then constructed into individually bar-coded PE libraries. The individual PE libraries were then combined into two pools, which were amplified as molecular clusters in different lanes on two separate flowcells and sequenced on an Illumina HiSeq 2000 on different runs.

De novo assembly: As the *de novo* assembly of Illumina paired-end reads from full-length molecular clones has not been described previously, we initially tested several genome assemblers. We found by using the default settings for each of the assemblers, only Phrap gave single contigs of approximately 9-kb in size for the majority of CC molecular clone sequence reads. We used Phrap for the results reported here, although this program can only handle up to approximately 50,000 reads and it does not take advantage of Illumina PE reads. Nonetheless, as an initial assessment of the sequencing data, we assembled HIV genomes with Phrap using up to 50,000 reads that were randomly selected from screened reads.

De novo assembly analysis: For *de* novo assemblies, we used an initial acceptability criterion of one contig between 8,900 and 9,400 bp. Approximately 76% (119/156) of the CC molecular clones met this criterion. *De novo* sequence assemblies were then characterized by alignment to the HXB2 genome using the LANL sequence locator tool to determine if the gene order was consistent with the HXB2 reference. All 119 assemblies satisfied this condition.

Discovery of alternative haplotype reads in *de novo* assemblies: Further examination of the quality of the *de novo* assemblies resulted in the finding of two distinct types: (*I*) those with low estimates of human sequence and the expected clonal sequence alignments and (*II*) those with high estimates of human sequence (~60%) and mixed haplotype reads in the alignment. To determine the number of type II assembly errors, we ran SNP detection analyses on the 119 *de novo* assemblies. These results were plotted with the percentage of human sequences identified during the pre-processing step. While the phenomenon is not understood at this time, we observed a strong correlation between those clonal samples with a high percentage of human reads and high SNP counts. From this analysis, the number of type I and type II *de novo* assemblies were 41 and 78, respectively, with the latter removed from further analysis pending causal understanding of the correlation noted above. In order to estimate variation in evolutionary rate and phylogenetic signal across the HIV-1 genome, we conservatively included only those sequences with zero SNPs for all phylogenetic analyses in this report.

Preliminary analysis of HIV reads mapping to the human genome: We note that the CC samples are human genomic DNA containing HIV that exists as integrated proviruses. If human sequence contamination did occur in our CC samples, we would expect reads mapping all across the human

genome. We observed, however, that the HIV reads mapped to 25 localized regions in the human genome. The mapped regions ranged in size from 4,978 bp to 11,456 bp, with an average of ~8-kb. The majority of reads from the CC molecular clones map to either Chr9 or Chr22, both of which contain at least one human endogenous retroviral (ERV) sequence. We note that human ERVs are widely spread throughout the genome, constituting approximately 8% of all sequences in humans. While HIV whole-genome strategies have reported high host (human) contamination in HIV sequencing efforts, our finding represents an unexplained phenomenon. First, due to the size of the mapped regions in the human genome, this observation suggests non-specific amplification of human sequences roughly the size of our target: 9-kb. While it is not uncommon for PCR to amplify non-specific regions, the cloning process that we have employed here should have resolved any mixture into individual sequences – being either human or HIV, but not both. We are currently investigating this observation by a variety of methods.

Phylogenetic analyses: Initial phylogenetic estimates of type I *de novo* assemblies from six non-overlapping regions of the HIV-1 genome demonstrate (*i*) the monophyly of clones sampled from the same recipient individuals, (*ii*) variation in relationships among groups of clones from different individuals, and (*iii*) variation in rates of evolution across different genomic regions. Result (*i*) is important for demonstrating the integrity of the phylogenetic signal in these assembled clones, since external evidence strongly suggests that each of the individuals included in these initial analyses only acted as a recipient within this transmission cluster. If strongly supported non-monophyly of clones from the same individual were found, it would indicate some problem with the *de novo* assembly, with the fit of the phylogenetic model to the data, or with the presumption that a very stong bottleneck in viral populations occurs at the time of transmission. Further verification of the phylogenetic signal in the Illumina data was achieved by aligning the relevant data subsets from *env* and *pol* to previously generated Sanger sequence data and performing phylogenetic inference on the combined data. Assembled Illumina clone sequences were all inferred to be closely related to Sanger-sequenced clones from the same individuals.

Results (*ii*) and (*iii*) derived from type I assemblies indicate the potential for whole genome sequences to provide more highly resolved and robust forensic conclusions from phylogenetic evidence. Distinct relationships among clones from different individuals inferred from numerous genomic regions is strong evidence of independent evolutionary histories for these regions caused by recombination. Increased sampling of independent genomic regions provides much greater confidence in forensic conclusions. In particular, repeated observations of (*i*) overall ingroup monophyly, (*ii*) monophyly of clones from individual recipients, and (*iii*) paraphyly of clones from source individuals across different genomic regions should greatly increase confidence in inferences of source-recipient relationships. Variation in evolutionary rates across different genomic regions also indicates the potential for these regions to provide resolution of transmission events at different timescales. Our preliminary evidence indicates that the regions we and others have previously sequenced for forensic inference (partial sequences of *pol* and *env*) may sit at the extremes of evolutionary rate in the HIV-1 genome. Sequences from other regions may provide greater resolution (i.e., be more likely to retain source paraphyly) for transmission events at timescales intermediate between those resolvable by *env* and *pol*.

Further, the six non-overlapping regions that we have investigated thus far based on Illumina data are fairly large. Investigation of targeted subsets of these regions that vary even more in evolutionary rate may be able to provide resolution of transmission events younger or older than those previously investigated.

Phylogenetic methods to account for unmodeled processes: With our new analytical approach, we were able to demonstrate that convergent evolution has likely influenced the distribution and strength of phylogenetic support across at least one HIV gene region (*env*), leading to certain spurious phylogenetic conclusions using standard phylogenetic models. Initial *env*-based phylogenetic estimates from the Texas case (2) suggested paraphyly of the viral lineages from one individual (CC07) who was thought to only be a recipient of HIV-1. Since paraphyly is a signature of transmission to other individuals, this result was initially puzzling. The series of analyses that we employed strongly implicate convergent evolution as the cause of this puzzling pattern. We were able to demonstrate that 1st and 2nd codon positions harbor more support for CC07 paraphyly than do 3rd codon positions. We were also able to show that phylogenetic signal varies across different sections of the *env* gene, but that such variation is not likely driven by recombination. Rather, this variation is better explained as a by-product of convergent evolution occurring at only a small number of spatially restricted sites.

Conclusions

In this study, we have demonstrated feasibility of full-length genome sequence production from individuals infected with HIV using NGS technologies and phylogenomic analysis. For a subset of full-length HIV genomes, we have demonstrated initial phylogenetic estimates from six non-overlapping genic regions showing (*i*) the monophyly of clones sampled from the same recipient individuals, (*ii*) variation in relationships amongst groups of clones from different individuals, and (*iii*) variation in rates of evolution across different genomic regions. These results are significant in demonstrating the integrity of the phylogenetic signal in these assembled CC molecular clone sequences as the extrinsic evidence, revealed during criminal proceedings, strongly suggested that each individual included in our initial analyses only acted as a recipient within the Texas transmission cluster. Our results also suggest that whole-genome sequences provide more highly resolved and robust forensic conclusions from phylogenomic evidence. Distinct relationships among clones from different individuals inferred from numerous genic regions are strong evidence of independent evolutionary histories, thus providing greater confidence in forensic conclusions. Several technical challenges are currently being addressed that include (*i*) optimizing *de novo* assemblies that will produce larger numbers of successful full-length genome sequences with low SNP counts and (*ii*) understanding the high estimates of human sequences in the subset of HIV molecular clone reads. Overcoming these challenges will lead into the development of a robust 'pathogen toolkit' that can provide a detailed roadmap for future forensics studies.

We believe this work can be expanded to the more general field of microbial forensics to aid in solving crimes involving a range of pathogens, as well as developing measures to enhance public safety. Examples of microbial forensic studies that will benefit from our study include characterization of vaginal swabs to test for sexually transmitted microbes in suspected sexual

assault cases involving minors, identification of introduced pathogenic microbes that contaminate food (e.g., *Salmonella* and *Shigella*) or water supplies (e.g., the recent cholera outbreak in Haiti), non-curable diseases transmitted with the intent to cause bodily harm (e.g., those associated with HIV and hepatitis C infections), or anthrax exposure with the intent to cause death. While we believe that policy recommendations may be premature, demonstration projects involving other microbial forensic studies that employ NGS with phylogenomic approaches will shed light on this topic, providing more informed decisions on public policy issues.

Main Body of the Draft Technical Report

Introduction

Statement of the problem: Microbial forensics is an emerging field that is very important for public safety, yet it presents enormous challenges to both the scientific and legal communities. Microbial pathogens comprise a highly diverse set of organisms with at least 1,400 or more known to cause disease in humans (4). These pathogenic agents have been used in acts of biocrime, bioterrorism, and military operations over the course of human history (5). Microbes have developed a number of mechanisms for generating natural genetic variation, such as high mutation (6-8) and high recombination rates (9-11). Recombinant DNA technologies can also be employed to rapidly combine novel genetic elements conferring new biological, and presumably pathogenic, properties to microbial 'super bugs' with potentially devastating effects. *One major goal of microbial forensics is to use this genetic variation to identify the source of a pathogen used to commit crimes or acts of bioterrorism.*

Our previous work focused on characterizing two separate regions, *pol* and *env*, of the HIV genome (2, 12), with the latter study centered specifically on inferring direction of transmission between case samples. These genes were selected because they exhibit different biological functions, are targeted by different selective pressures (i.e., drug versus host, respectively), and are known to exhibit different rates of evolution and degrees of genetic diversity. Factors contributing to the creation and maintenance of genetic diversity across viral lineages are high mutation (6-8) and recombination rates (9-11) coupled with an estimated replicative production of 10^8 to 10^{10} virions per day (13-15). This expansion is offset by lineage extinction from the production of defective, non-replicating virions (16), the effectiveness of the host's immune system, and the efficacy of highly active antiretroviral therapies (17). While many of these biological processes are accounted for in traditional phylogenetic analysis, certain complexities are often ignored for the sake of computational simplicity. Unmodeled complexities relevant to HIV evolution include intragenic recombination (17-20), convergent evolution due to selection (17, 21-23), site-specific shifts in the rate of evolution across different viral lineages (24), and unforeseen heterogeneity in the evolutionary process across sites (25, 26).

Statement of hypothesis or rationale for the research: It is well established that many of these processes occur frequently in HIV evolution and can affect the results of phylogenetic analyses (17, 21, 27). In cases of suspected HIV transmission, use of more realistic models of sequence

evolution and inclusion of genes beyond just *pol* and *env* should provide more accurate estimates of true paraphyletic relationship among case samples. Thus, we hypothesized that the application of long-range PCR technology to amplify the entire HIV genome, which will then be sequenced with NGS technologies and characterized by phylogenomic methods would incorporate many additional complexities of the evolutionary processes relevant to HIV.

Methods

PCR precautionary steps and laboratory protocol between samples: To minimize false positives due to contamination, standard operating procedures were used at all times (2, 12). For example, all of the reagents were prepared using sterile autoclaved water and stored at appropriate temperatures. The PCR setup, the addition of genomic and PCR template DNAs, and PCR cloning were carried out at separate stations to avoid carryover contamination. Dedicated pipets and filter barrier tips were used for at step involving PCR setup, addition of DNA templates, and plasmid cloning. All bench tops, including the PCR setup station, DNA addition station, cloning station, plasmid prep station, and equipment, including the agarose gel running apparatus, Safe Imager Blue-Light Transilluminator, water-bath, and pipets were cleaned with 2% (vol/vol) Clorox bleach solution, and then with 70% (vol/vol) ethanol solution. The Clorox-Ethanol wash down was carried out after every major protocol step as well as at the beginning of each new case sample. All overnight cultures were soaked in 2% Clorox bleach solution for up to 12 hours. The biohazard material was disposed of appropriately in the biohazard bin. Disposal of other waste material and solutions was carried out according to the Baylor College of Medicine safety policy. All cloning reagents (requiring minus 80^0C storage) and plasmid DNAs of each sample were then stored in a minus 80^0C freezer housed in a separate room from the Metzker Laboratory. Antibiotic selective plates were prepared and stored at 4^0C for up to three weeks after which fresh stock of LB agar plates was made.

Long-range PCR: Genomic DNA for Texas case samples CC01-CC08, stored in a minus 80°C freezer since 2009, were individually retrieved and handled to avoid sample mix-up. Samples were processed in reverse order starting with CC08 and finishing with CC01. The genomic DNA concentration ranged from 15-143 ng/µL. FIG. 1 shows the primer binding sites that targeted ~9-kb of the HIV genome by employing the long-range, nested PCR technique.[1] The following primers were used as described by Salvi et al. (28): 626s 5'-TCTCTAGCAGTGGCGCCCGAACAGGG, 691s 5'-GCAGGACTCGGCTTGCTGAAGC, 9614a 5'-GGCAAGCTTTATTGAGGCTTAAG, and 9680a 5'-GGTCTGAGGGATCTCTAGTTACCAGAGTC.[2] Master mixes for both PCR rounds were prepared using the SequalPrep Long Range PCR kit (Life Technologies), which is a blend of Platinum *Taq* DNA polymerase and *Pyrococcus* species GB-polymerase (a proof-reading enzyme). We note that pooling samples (see below) and using a proof-reading polymerase can help reduce PCR-recombination effects and increase amplicon fidelity (29), although polymerase blends are necessary in achieving large amplicons (30). Along with each genomic DNA sample, the PCR setup

[1] The outer primers yielded a target amplicon size of 9,054 bp and the inner primers yielded a target amplicon size of 8,923 bp.

[2] Salvi et al. noted that each primer was designated by the nucleotide position relative to NL4-3 genome followed by the letter "s" or "a" denoting the sense or antisense strand, respectively.

included a negative control (molecular grade water) in order to check for contamination. For each CC sample, ten genomic DNA and two blank water control reactions were PCR amplified as a set. Cycle conditions for 1st (626s/9680a) and 2nd (691s/9614a) PCR rounds were as follows: 2 min at 94°C, and then 10 cycles at 94°C for 10 sec, 60°C for 30 sec, 68°C for 9 min followed by 22 cycles at 94°C for 10 sec, 60°C for 30 sec, 68°C for 9 min + 20 sec added accumulatively with each cycle. The final extension step was 72°C for 5 min. The cycle numbers were kept relatively low in an effort to minimize PCR recombination effects (29).

First PCR-round products were pooled and analyzed on a 0.5% agarose gel and electrophoresed at 45 V for 6 hrs. Except for sample CC08, all other CC samples revealed a positive 9-kb PCR product (see FIG. 2). For CC08, 2 µL of either the first genomic DNA-amplified product or blank water control was used as template for the second PCR-round, for which products were pooled and run on 0.5% agarose gel using the conditions described above.

Cloning 9-kb fragments: For gel purification of 9-kb PCR products, the SYBR DNA Gel Stain and SYBR Safe Transilluminator (Life Technologies) were used. Each 9-kb band was excised using a sterile razor, purified using the S.N.A.P gel purification kit, and cloned using the pCR-XL Topo Cloning kit (both kits from Life Technologies). The concentration of gel-purified 9-kb PCR products was determined using a Nanodrop device, which ranged from 4-14 ng/µL. Approximately four volumes (i.e., 16-56 ng) of gel-purified products were used for ligation reactions into 10 ng of the PCR XL TOPO vector. Ligation reactions were then transformed into TOP10 electrocompetent cells (Life Technologies) using electroporation (Bio-Rad), according to the manufacturer's protocol. Two hundred µL of the transformed cell solution were then plated onto the selective LB agar plates (zeocin 25 µg/mL and kanamycin 50 µg/mL) and incubated overnight at 37°C. On average, >100 to >200 colonies were obtained on the LB plates with approximately 30-50 colonies for each CC sample being selected as clones. A summary of the cloning results is shown in TABLE 1.[3] Mini-preps were performed using the Zyppy Plasmid Miniprep kit (Zymo Research). Glycerol stocks were prepared by transferring 200 µL of the overnight culture to 800 µL of the 80% glycerol stock. To confirm the presence of a 9-kb insert, clones were digested with 15U/µg *Eco*RI and/or *Bam*HI[4], separately (Life Technologies), using the following conditions: *Eco*RI at 37°C for 4 hrs; 65°C for 20 mins and *Bam*HI 30°C for 4 hrs; 65°C for 20 mins. Approximately 0.1-0.3 µg of the digested plasmids were analyzed on a 0.5% agarose gel and electrophoresed at 45 V for 5 hrs. Clones that exhibited a restriction enzyme digestion

[3] Although the cloning results are summarized in a single table, each CC sample was handled separately in the laboratory and characterized through the steps of identifying 20 or more full-length molecular clones.

[4] For CC08 and CC07, both *Eco*RI and *Bam*HI restriction enzyme digest analysis was performed to verify the cloning of a 9-kb product. Thereafter, *Eco*RI (CC06, CC05, CC04, and CC01) or *Bam*HI (CC03 and CC02) restriction enzyme digest analysis was performed on the remaining CC molecular clones to verify the cloning of a 9-kb product.

pattern with DNA fragments that summed to approximately 12.5-kb in size were qualified initially as "full-length" HIV genome clones.[5]

NGS library generation: Initially, 160 molecular clones (i.e., eight CC samples each with 20 full-length molecular clones verified by restriction enzyme analysis) were submitted to the Human Genome Sequencing Center's (HGSC) Library QC and Library Automation groups. Quality control tests were conducted on the 160 clones using the PicoGreen assay to quantitate DNA concentrations, and flash gels using Agilent's Bioanalyzer to verify clone sizes. Of this initial set, 156 molecular clones passed both tests and were further processed by the NGS library production group.[6]

Each HGSC-qualified molecular clone (qMC) was constructed into individually bar-coded Illumina paired-end (PE) libraries using the BCM-HGSC optimized protocol performed on Beck-man Coulter Biomek robots (FIG. 3). Briefly, 0.5 µg DNA from each molecular clone was sheared into ~550 bp fragments using the E210 system (Covaris, Inc.). A series of molecular biology techniques was then implemented including DNA end-repair, A-tailing, and Illumina adapter ligation with each step being followed by a purification step using Agencourt XP Beads (Beckman Coulter Genomics, Inc.). DNA fragments were then amplified using Illumina Index PE primers to generate bar-coded NGS libraries. PCR products were then purified and quantified, and their size distribution was analyzed using the LabChip GX electrophoresis system (Perkin Elmer).

NGS sequencing: Aliquots (10 nM) of the libraries were prepared and combined into two pools (Pool 1: 76 samples and Pool 2: 80 samples), for which molecular template clusters were amplified in two different lanes on separate flowcells using the cBOT cluster station. The two flowcells contained amplified HIV libraries were sequenced using the Illumina TruSeq.v3 chemistry as 2x100 base reads with 10 additional cycles to read the barcodes. Each flowcell was run on different HiSeq 2000 instruments, yielding a total of 759 million reads with an average of 4.87 ± 0.76 million reads per library sample. The minimum number of reads was 3.11 million (CC04-19) and the maximum was 7.20 million reads (CC02-05).

Read pre-processing before assembly: Illumina sequencing data was processed using CASAVA 1.8.2 (Illumina) to generate raw read fastq formatted read sequences. Adapter sequences in Illumina reads were removed using the SeqPrep program[7] with a minimum read length of 50 bp after adapter trimming. Base-calls at the end of reads with Illumina quality scores of ≤2 were trimmed (referred to as "trimmed reads" from here on) using a BCM-HGSC in-house script. The trimmed reads were mapped to the human reference genome using BWA (31) with default

[5] We note that clones digested with the same restriction enzyme sometimes gave different fragment sizes, which we attributed to nucleotide sequence variation within distinct HIV sequences. Including the vector of 3.5-kb, the expect size of the digested plasmid sum to ~12.5-kb.

[6] Molecular clones CC01-09, CC08-07, CC08-08, and CC08-11 failed the flash gel assay exhibiting fragment sizes of approximately 5-kb.

[7] *See* https://github.com/jstjohn/SeqPrep

parameters and BMTagger.[8] Reads mapped to the human genome sequence by BWA or BMTagger were removed, with the remaining reads then being screened for *E. coli*, ΦX174 (Illumina's internal sequencing control) and the PCR XL TOPO cloning vector sequence using the cross_match program.[9] Illumina reads that did not map to any of the reference sequences and had lengths of ≥30 bp were subsequently used in *de novo* assemblies (referred to as "screened reads" from here on).

Evaluation of other assembly programs: Velvet (32), Fermi (33), MIRA,[10] Newbler from Roche 454, and Phrap[9] assembly programs were evaluated for *de novo* assembly quality for a subset of samples using all screened reads as input. The parameters used for the evaluation were Velvet with *k*-mer size of 29 and 31, Fermi with default parameters, MIRA with default parameters for Illumina paired reads, and Newbler with default parameters. With the exception of Phrap (see below), none of the programs listed above were able to assemble the screened reads into a single HIV contig using the parameters tested.

Phrap assembly: Due to high read depth coverage, only 50,000 reads were randomly extracted from screened reads (all screened reads if the total number was <50,000 reads) and used for Phrap assembly.[9] If the HIV contig from the trimmed read assembly was >8.9-kb, the contig was subsequently used in the multiple sequence alignment and phylogenetic analysis. If assembly produced a contig <8.9-kb, it was considered as an initially failed product and was not included in the analysis of this report. These assemblies have been earmarked for more depth analysis to understand the reason(s) for the failure.

De novo assembly analysis: The Los Alamos National Laboratory (LANL) curated HIV database[11] was used to compare all *de novo* assemblies from the CC clones to the HXB2 genome to assess the overall gene organization of the assemblies. This comparison was accomplished using the sequence locator tool, which generates a graphical report that displays gene products, nucleotide coordinates, start and stop codon positions, deletions or frame-shifts and other HIV genomic structural elements present in the CC clone *de novo* assemblies.[11]

Control sequence (Outgroup) selection: In order to assess the monophyly of HIV sequences obtained from the Texas case samples, we selected sequences from the LANL curated HIV database[12] to include in all subsequent analyses. All sequences in the LANL database are also

[8] *See* ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/

[9] *See* http://www.phrap.org/phredphrapconsed.html. Phrap parameters used in the *de novo* assembly of CC sample clone were minmatch = 10, minscore = 24, and forcelevel = 10.

[10] *See* http://mira-assembler.sourceforge.net/

[11] *See* http://www.hiv.lanl.gov/content/index

[12] *See* http://www.hiv.lanl.gov

found in Genbank. The LANL data, however, have been screened for problematic sequences, such as hypermutant, contaminant, or synthetic sequences.[13]

To obtain the multiple sequence alignment, each set of intra-individual sequences were aligned using the Auto option in MAFFT v. 7 program (34-36) and pairwise distances (uncorrected p-distance) were computed among all intra-individual genome sequences using PAUP* (37). The two most divergent sequences within each set were then chosen to query the LANL HIV database using BLAST (38). The top match for each query, excluding sequences flagged as problematic, were added to the set of control sequences. Sequences were not added if a sequence from the same study was already present among the set of control sequences. In addition to these control sequences, the standard HIV-1 reference genome (HXB2, Genbank accession number K03455) was also included in this study.

Genome partitioning: A significant body of research has led to a detailed annotation of the HIV-1 genome regarding gene boundaries, as exemplified by the HXB2 reference. These annotations (see FIG. 5 legend) provided the genome coordinates used to extract the region of interest from the reference genome and orthologous sequences from the case and control sequences using a custom python script. The extracted sequence from HXB2 was locally BLASTed against each case and control genome sequence in order to extract homologous sequences. Each set of orthologous sequences was saved to a separate file. These regions of the HIV-1 genome were selected for analysis in order to evaluate variation in evolutionary rate and topological signal across the HIV-1 genome. We selected only non-overlapping protein coding regions (plus the LTR region) in order to avoid analyzing the same genomic regions in separate analyses.

Alignment and alignment filtering: Base nucleotide alignments were estimated using MAFFT v. 7 and assessed for alignment uncertainty using Guidance v. 1.3.3 (39) using default settings. Phylogenetic inferences were made from multiple sequence alignments after masking individual sites whose placement in the alignment was sensitive to the choice of guide tree. Sites with a confidence score below 0.93 (default) were masked. Filtering these sites is expected to help reduce spurious phylogenetic inferences resulting from alignment uncertainty.

Phylogenetic analysis: For each *de novo* assembly derived from Illumina read data, the best-fit model of sequence evolution was chosen using AICc in jModelTest 2.1.4 (40, 41) from among 88 candidate models. We used Garli 2.0 (42) to find the maximum-likelihood (ML) topology from five search replicates under the best-fit model of sequence evolution.

In addition to assemblies comprised solely of newly generated Illumina sequences, we also aligned the relevant subsets of the new sequences to previously generated Sanger sequences of *env* and *pol* from independent CC clones (2). Alignment, model selection, and ML phylogenetic inference were all performed as outlined above in order to verify the integrity of Illumina sequences compared with Sanger sequences from the same CC samples.

[13] *See* http://www.hiv.lanl.gov/components/sequence/HIV/search/help.html#bad_seq

Phylogenetic methods to account for unmodeled processes: In addition to generating and analyzing whole viral genome sequences, we also developed a new analytical approach to detect the action of viral evolutionary processes not accounted for by 'standard' phylogenetic models (1) and applied it to an existing Sanger sequence data set of *env* sequences from the Texas case (2). More specifically, we examined the potential roles of convergent evolution and recombination in driving spurious phylogenetic conclusions. To test for the influence of recombination, we employed a hidden Markov model (3) that allows for variation in topology across sites. To compare models with different numbers of topologies, we developed an approximate AIC approach and validated it through simulations. To detect the action of selection, we employed codon-position-specific phylogenetic analyses and codon-based models that allow us to detect sites with elevated non-synonymous (dN) to synonymous (dS) rate ratios (dN/dS). Lastly, we developed a site-specific-likelihood profile by calculating site-specific likelihoods using the overall maximum likelihood (ML) tree and the best tree consistent with particular transmission hypotheses. The difference between these site-specific likelihoods then acts as a measure of site-specific support for or against the transmission hypothesis. Using permutation tests, we can assess whether sites in particular codon positions exhibit more or less support for the transmission hypothesis (or the ML tree) than expected by chance. We can also visualize the distribution of support across an alignment to see if sites supporting particular hypotheses are spatially clustered.

Results

Statement of Results:

Long-range PCR: As we described in our original proposal, the technique of long-range PCR was employed to amplify ~9-kb of the HIV-1 genome (FIG. 1). In our initial progress report, we described successful long-range PCR for the clone pNL4-3 and case sample CC08.[14] As noted in that report, however, we obtained several major, non-specific bands, which reduced the yield of the desired 9-kb product.[15] Attempts to improve product specificity by altering temperature conditions were met with little success. We, therefore, proceeded by purifying the 9-kb band from the CC08 sample and cloned it using the pCR-XL-TOPO cloning vector system.[16] Our preliminary results suggested the 20 molecular clones analyzed by the restriction enzyme *Eco*RI revealed that we had cloned the 9-kb PCR product.[17] This result, however, could not be confirmed when using the alternative restriction enzyme, *Bam*HI. Moreover, Sanger sequencing of the ends of the 20 molecular clones did not confirm that we had cloned full-length HIV genomes (data not shown).

Identification of robust long-range PCR primers: We originally developed our nested PCR primers based on sequences described by Gao *et al.* (43) and Salminen *et al.* (44). Our results, however,

[14] *See* July 30, 2012 Progress Report: Fig. 2B for pNL4-3 and Fig. 3 for CC08.

[15] *Id.*, p. 2.

[16] *Id.*, p. 3.

[17] *Id.*, p. 3.

suggested that these primer sets were amplifying genomic DNA in a non-specific manner. After an exhaustive search, we settled on a new set of nested primer sequences, described in Salvi *et al.* (28). FIG. 2 illustrates the results of long-range PCR for all eight CC samples, which worked robustly for all case samples.[18] For the majority of CC samples, a single PCR product band of ~9-kb was obtained without any noticeable non-specific bands. With the exception of CC08,[19] all Texas case samples were amplified with just the outer PCR primers to yield a positive signal. All PCR products were, nonetheless, purified by agarose gel electrophoresis.

Cloning 9-kb fragments: After gel purification, the 9-kb fragments were then cloned to resolve the population of HIV genomes into individual isolates. As noted in our initial progress report, we employed the pCR-XL-TOPO cloning vector system, which has been optimized for larger inserts.[20] For each CC sample, the gel-purified 9-kb insert was ligated into the pCR-XL-TOPO cloning vector. The ligation product was transformed into ultra-competent cells and plated for overnight growth. In general, the cloning efficiency was sufficiently high to produce hundreds of colonies per plate with low background (see Methods section).

To test for molecular clones containing full-length HIV genomes, restriction enzyme analysis was performed with *Eco*R1 and/or *Bam*HI. Not all clones tested contained a full-length, 9-kb insert, and overall, 63% of clones selected contained the full-length insert. For each CC sample, 20 molecular clones containing full-length inserts were then selected for NGS library preparation.

NGS library generation and sequencing: The molecular clones were further quality controlled for quantity and size using the PicoGreen assay and flash gels using the Agilent Bioanalyzer. Of these clones, 97.5% (156/160) passed this QC criterion. These "HGSC-qualified molecular clones" (qMC) were then constructed into individually bar-coded PE libraries using automated protocols developed at the BCM-HGSC (see FIG. 3). The individual libraries were then combined into two pools: Pool 1 contained 76 libraries from CC02, CC03, CC04, and CC05 and Pool 2 contained 80 libraries from CC01, CC06, CC07, and CC08. The two pools were amplified as molecular clusters in different lanes on two separate flowcells and sequenced on an Illumina HiSeq 2000 on different runs. Fastq files were created from Illumina read data using the CASAVA 1.8.2 program. The Illumina reads were then screened to remove the adapter sequences and trimmed to remove low quality bases at the ends of the reads. These "trimmed reads" were then mapped to the human genome using BWA or BMTagger to remove human sequence contamination. Additional sequence screening was performed to remove control and vector sequences to produce "screened reads", see Methods section for details.

---

[18] Although the results of long-range PCR are illustrated in a single figure, each CC sample was handled separately in the laboratory and characterized through the steps of identifying 20 or more full-length molecular clones before starting long-range PCR on the next CC sample. The order of processing the CC samples started at CC08 and proceeded in reverse numerical order to CC01.

[19] Isolation of genomic DNA from PBMCs from CC08 yielded the lowest concentration of 15 ng/µl. All other CC samples ranged from 40 to 143 ng/µl.

[20] *See* July 30, 2012 Progress Report, p. 3.

De novo assembly: As the *de novo* assembly of Illumina PE reads from full-length molecular clones has not been described previously, we initially tested several genome assemblers including Velvet (32), Fermi (33), MIRA,[21] Newbler from Roche 454, and Phrap.[22] We found by using the default settings for each of the assemblers, only Phrap gave single contigs of approximately 9-kb in size for the majority of CC molecular clone sequence reads. We therefore settled for this report on the Phrap program developed by Ewing and Green during the Human Genome Project (45), although a limitation of the program is that it can only handle up to approximately 50,000 reads. Another limitation described below is that Phrap does not take advantage of Illumina PE reads. Nonetheless, as an initial assessment of the sequencing data, we assembled HIV genomes with Phrap using up to 50,000 reads that were randomly selected from screened reads.[23]

De novo assembly analysis: We used an initial criterion that *de novo* assemblies were acceptable if they produced one contig in the size range of ≥8,900 bp and ≤9,400 bp. Approximately 76% (119/156) of the CC molecular clones met this criterion.[24] *De novo* sequence assemblies were then characterized by alignment to the HXB2 genome using the LANL sequence locator tool to determine if the gene order was consistent with the HXB2 reference. All 119 assemblies satisfied this condition.

Discovery of alternative haplotype reads in *de novo* assemblies: Further examination of the quality of the *de novo* assemblies resulted in the finding of two distinct types: (*I*) those with low estimates of human sequence and the expected clonal sequence alignments (not shown) and (*II*) those with high estimates of human sequence (~60%) and mixed haplotype reads in the alignment (FIG. 4A). In the analysis of sequence variation, tools such as Atlas-SNP (46) have been used to call single nucleotide polymorphisms (SNPs) in re-sequencing efforts (47). To determine the number of type II assembly errors, we ran the Atlas-SNP2 program on the 119 *de novo* assemblies to count the number of SNPs. These data were plotted along with the percentage of human sequences identified during the pre-processing step. While the phenomenon is not understood at this time, we observed a strong correlation between those clonal samples with a high percentage of human reads and high SNP counts (FIG. 4B). From this analysis, the number of type I and type II *de novo* assemblies were 41 and 78, respectively,[25] with the latter removed from further analysis pending causal understanding of the correlation noted above. In order to estimate variation in evolutionary rate and phylogenetic signal across the HIV-1 genome, we conservatively included only those sequences with zero SNPs for all phylogenetic analyses in this report.

---

[21] *See* http://mira-assembler.sourceforge.net/

[22] *See* http://www.phrap.org/phredphrapconsed.html

[23] Some CC case samples had fewer than 50k reads after human removal, for which all reads were used by Phrap for *de novo* assemblies.

[24] The remaining 37 assemblies will not be discussed further in this report as they are currently being investigated.

[25] The type I cut-off was human sequence estimates ≤1% and SNP counts ≤ 10.

Preliminary analysis of HIV reads mapping to the human genome: We note that the CC samples are human genomic DNA containing HIV that exists as integrated proviruses. The expected result if human sequence contamination did occur in our CC samples would be reads mapping all across the human genome. We observed, however, that the HIV reads mapped to 25 localized regions in the human genome, see TABLE 2. The mapped regions ranged in size from 4,978 bp to 11,456 bp, with an average of ~8-kb. The majority of reads from the CC molecular clones map to either Chr9 or Chr22, both of which contain at least one human endogenous retroviral (ERV) sequence. We note that human ERVs are widely spread throughout the genome, constituting approximately 8% of all sequences in humans (48). While HIV whole-genome strategies have reported high host (human) contamination in HIV sequencing efforts (49), our finding represents an unexplained phenomenon. First, due to the size of the mapped regions in the human genome, this observation suggests non-specific amplification of human sequences roughly the size of our target: 9-kb. While it is not uncommon for PCR to amplify non-specific regions, the cloning process that we have employed here should have resolved any mixture into individual sequences – being either human or HIV, but not both. We are currently investigating this observation by a variety of methods.

Phylogenetic analyses: Initial phylogenetic estimates of type I *de novo* assemblies from six non-overlapping regions of the HIV-1 genome (FIG. 5) demonstrate (*i*) the monophyly of clones sampled from the same recipient individuals, (*ii*) variation in relationships among groups of clones from different individuals, and (*iii*) variation in rates of evolution across different genomic regions. Result (*i*) is important for demonstrating the integrity of the phylogenetic signal in these assembled clones, since external evidence strongly suggests that each of the individuals included in these initial analyses only acted as a recipient within this transmission cluster (2). If strongly supported non-monophyly of clones from the same individual were found, it would indicate some problem with the *de novo* assembly, with the fit of the phylogenetic model to the data, or with the presumption that a very stong bottleneck in viral populations occurs at the time of transmission. Further verification of the phylogenetic signal in the Illumina data was achieved by aligning the relevant data subsets from *env* and *pol* to previously generated Sanger sequence data and performing phylogenetic inference on the combined data. Assembled Illumina clone sequences were all inferred to be closely related to Sanger-sequenced clones from the same individuals (FIG. 6).

Results (*ii*) and (*iii*) derived from type I assemblies indicate the potential for whole genome sequences to provide more highly resolved and robust forensic conclusions from phylogenetic evidence. Distinct relationships among clones from different individuals inferred from numerous genomic regions is strong evidence of independent evolutionary histories for these regions caused by recombination. Increased sampling of independent genomic regions provides much greater confidence in forensic conclusions. In particular, repeated observations of (*i*) overall ingroup monophyly, (*ii*) monophyly of clones from individual recipients, and (*iii*) paraphyly of clones from source individuals across different genomic regions should greatly increase confidence in inferences of source-recipient relationships. Variation in evolutionary rates across different genomic regions also indicates the potential for these regions to provide resolution of transmission events at different timescales. Our preliminary evidence (FIG. 5) indicates that the

regions we and others have previously sequenced for forensic inference (partial sequences of *pol* and *env*) may sit at the extremes of evolutionary rate in the HIV-1 genome. Sequences from other regions may provide greater resolution (i.e., be more likely to retain source paraphyly) for transmission events at timescales intermediate between those resolvable by *env* and *pol*. Further, the six non-overlapping regions that we have investigated thus far based on Illumina data are fairly large. Investigation of targeted subsets of these regions that vary even more in evolutionary rate may be able to provide resolution of transmission events younger or older than those previously investigated.

Phylogenetic methods to account for unmodeled processes: With our new analytical approach, we were able to demonstrate that convergent evolution has likely influenced the distribution and strength of phylogenetic support across at least one HIV gene region (*env*), leading to certain spurious phylogenetic conclusions using standard phylogenetic models. Initial *env*-based phylogenetic estimates from the Texas case (2) suggested paraphyly of the viral lineages from one individual (CC07) who was thought to only be a recipient of HIV-1. Since paraphyly is a signature of transmission to other individuals, this result was initially puzzling. The series of analyses that we employed strongly implicate convergent evolution as the cause of this puzzling pattern. We were able to demonstrate that 1st and 2nd codon positions harbor more support for CC07 paraphyly than do 3rd codon positions. We were also able to show that phylogenetic signal varies across different sections of the *env* gene, but that such variation is not likely driven by recombination. Rather, this variation is better explained as a by-product of convergent evolution occurring at only a small number of spatially restricted sites.

Tables:

TABLE 1. Summary of gDNA concentrations and cloning results for the Texas case samples.

		Number of colonies		Number of clones	
sample ID	gDNA conc (ng/µl)	Test Plate	Negative control plate	Clones tested	Clones with 9-kb Insert
CC08	15	200+	2	29	21
CC07	143	100+	3	32	23
CC06	81	100+	3	35	22
CC05	81	200+	2	48	20
CC04	103	100+	3	30	27
CC03	54	200+	2	35	23
CC02	101	200+	0	35	23
CC01	40	100+	0	40	20

TABLE 2. Summary of human mapping data.

Chromosome No: Start-Stop position	Length of mapping region (bp)	No. of HIV assemblies mapping to regions	No. of ERVs* in mapped regions
1: 204,466,150 to 204,475,734	9,585	2	2
2: 9,249,662 to 9,258,679	9,018	1	1
2: 242,468,648 to 242,473,822	5,175	1	3
3: 31,929,386 to 31,936,644	7,259	2	2
3: 48,057,678 to 48,064,153	6,476	3	1
4: 2,387,560 to 2,393,653	6,094	9	3
4: 25,714,412 to 25,720,591	6,180	2	0
4: 41,674,103 to 41,685,383	11,281	4	0
4: 159,743,681 to 159,749,745	6,065	3	1
5: 42,940,423 to 42,947,444	7,022	5	1
5: 176,199,843 to 176,204,820	4,978	3	6
8: 86,488,827 to 86,499,764	10,938	3	0
9: 125,643,494 to 125,654,398	10,905	5	0
9: 137,566,165 to 137,574,051	7,887	118	1
12: 122,811,675 to 122,822,082	10,408	22	0
13: 82,868,175 to 82,874,688	6,514	4	1
16: 19,269,679 to 19,276,437	6,759	14	7
16: 27,406,602 to 27,416,514	9,913	7	0
16: 31,393,589 to 31,400,015	6,427	1	1
16: 57,163,447 to 57,174,902	11,456	4	1
16: 71,556,230 to 71,564,448	8,219	2	3
16: 74,900,631 to 74,907,735	7,105	3	5
17: 49,628,344 to 49,635,982	7,639	7	1
19: 41,955,072 to 41,964,604	9,533	2	16
22: 45,791,458 to 45,798,451	6,994	48	2

* Human EVRs include ERV1, ERL, and ERVL-MaLR

Figures:

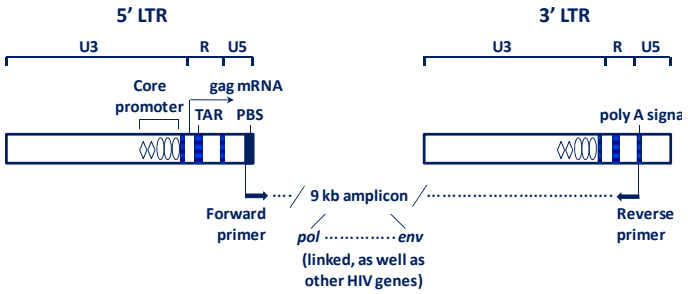

FIG. 1. HIV genome cartoon, illustrating the location of long-range PCR primers. Figure adapted from ref. (44).

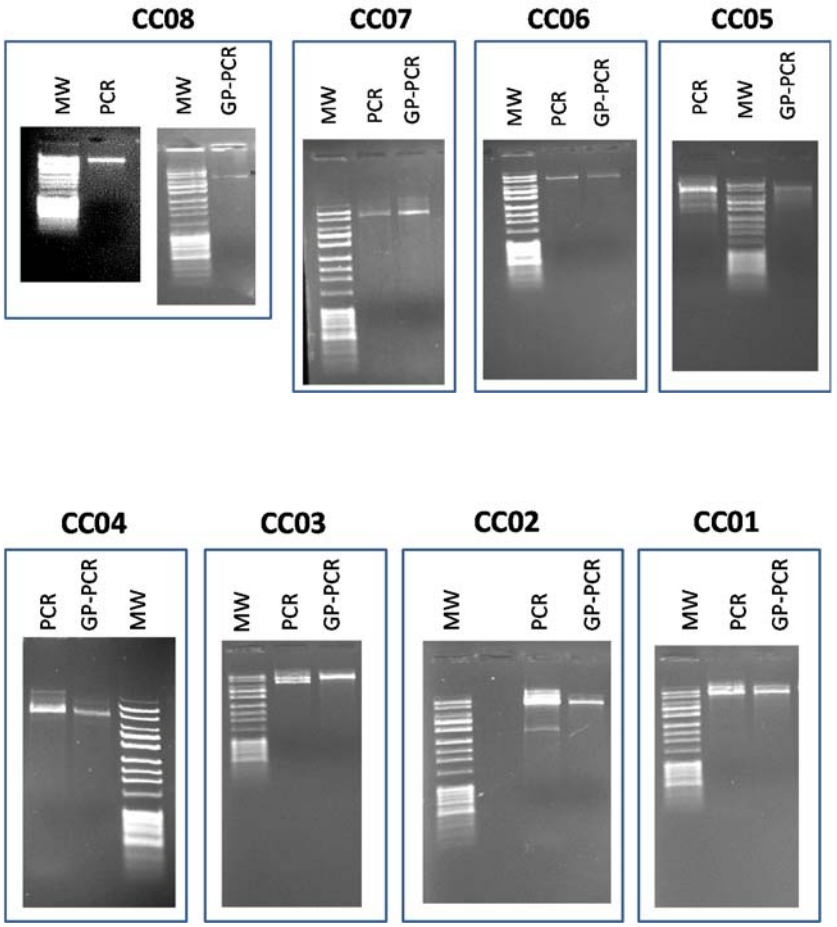

FIG. 2. Long-range PCR results from Texas case samples. MW = 10-kb molecular weight marker; PCR = original PCR products; GP-PCR = gel-purified PCR product.

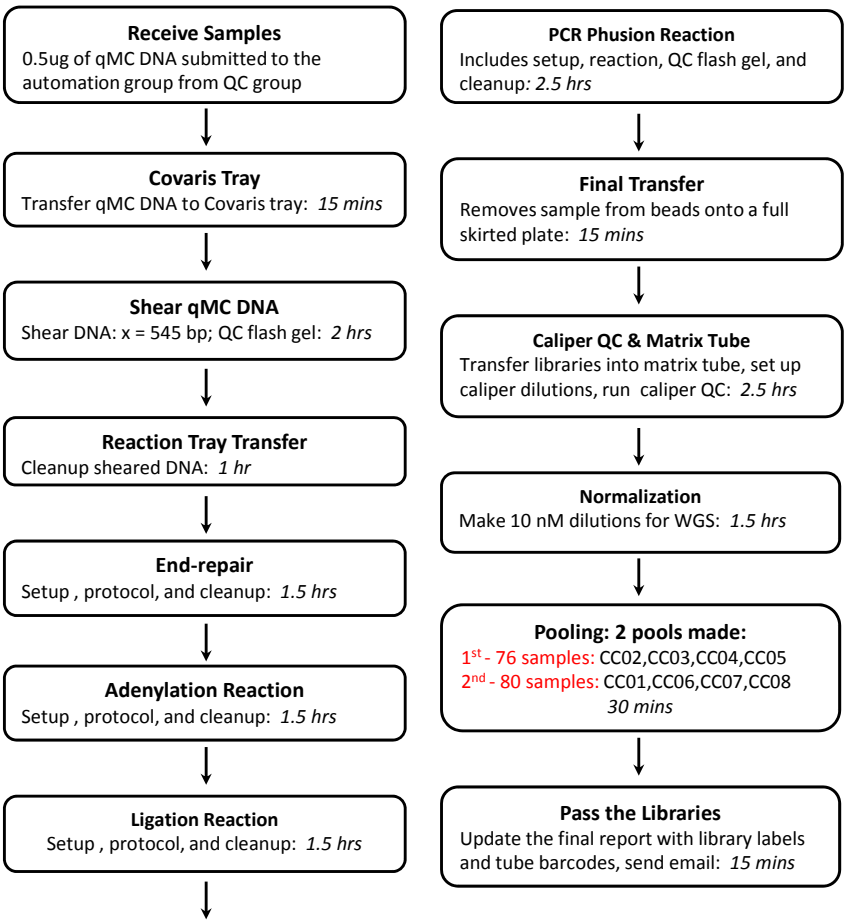

FIG. 3. Automation workflow protocol in preparing NGS libraries of 9-kb HIV genome clones.

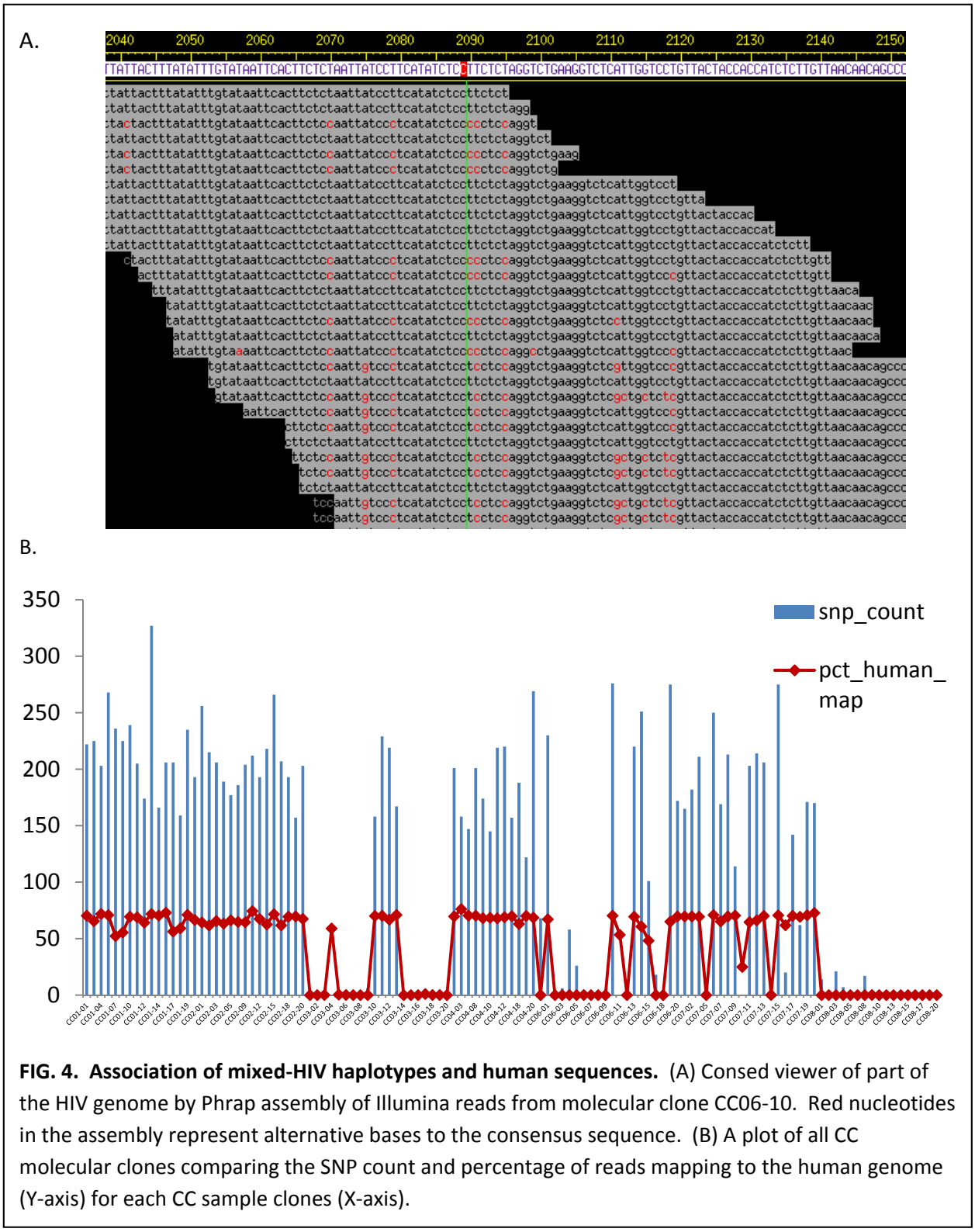

FIG. 4. Association of mixed-HIV haplotypes and human sequences. (A) Consed viewer of part of the HIV genome by Phrap assembly of Illumina reads from molecular clone CC06-10. Red nucleotides in the assembly represent alternative bases to the consensus sequence. (B) A plot of all CC molecular clones comparing the SNP count and percentage of reads mapping to the human genome (Y-axis) for each CC sample clones (X-axis).

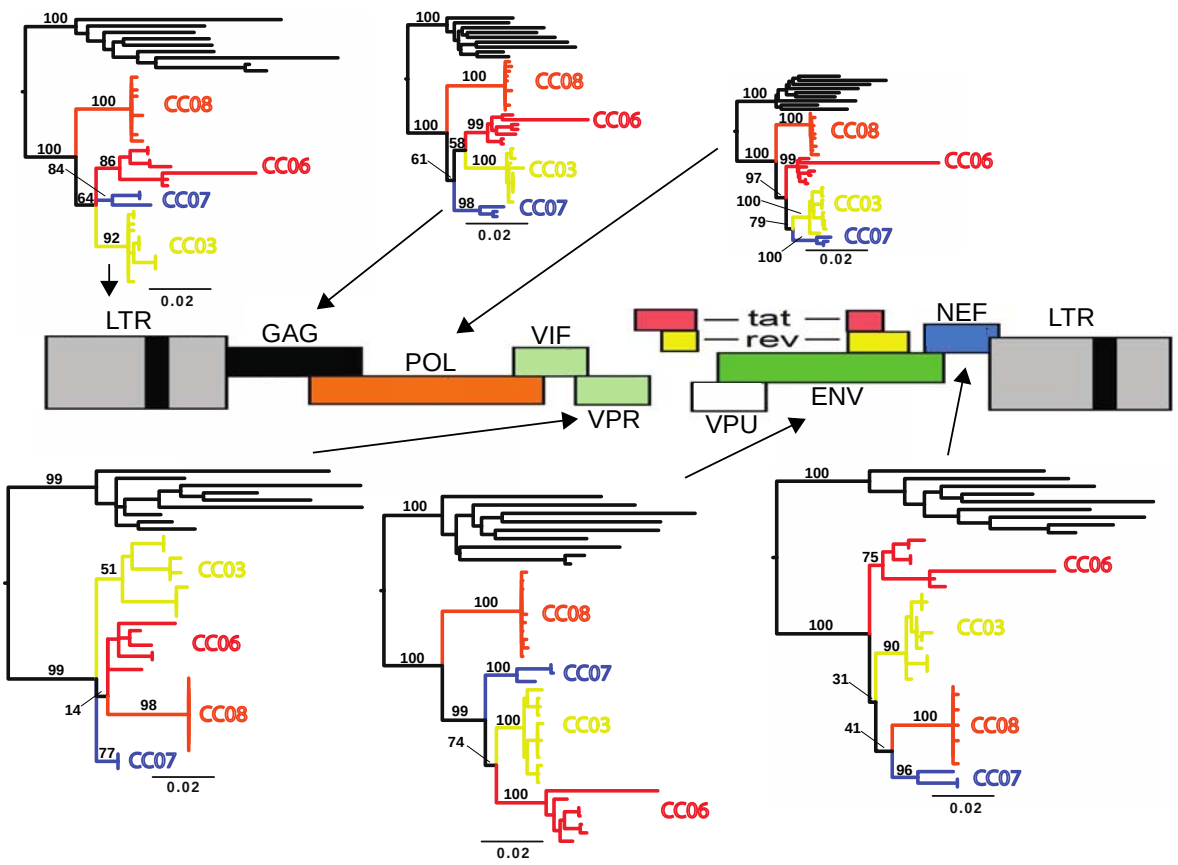

FIG. 5 Maximum-likelihood gene trees inferred from six non-overlapping regions of the HIV-1 genome. Each tree has been scaled such that the branch-length scales are the same among individual gene trees. The variation in tree size illustrates the variation in evolutionary rate across the HIV-1 genome. Branch support values are bootstrap proportions from 200 pseudoreplicates. The following start and stop coordinates from the HXB2 genome were used to retrieve the homologous sequences from assembled case and control sequences: LTR – 1, 634; GAG – 790, 2292; POL – 2358, 5096; VPR – 5559, 5850; ENV – 6225, 8795; NEF – 8797, 9417.

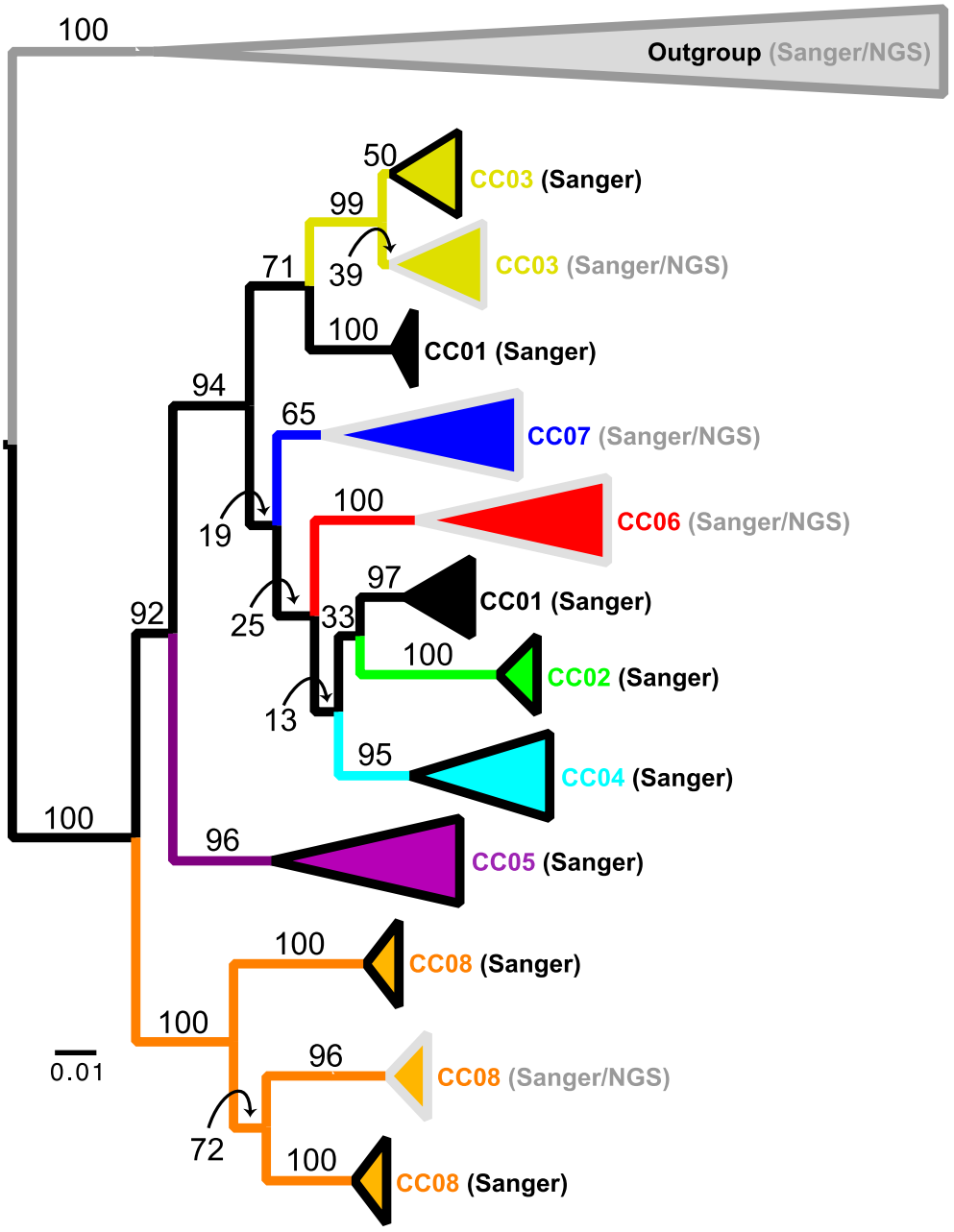

FIG. 6 Combined phylogenetic tree of Sanger and Illumina data for the *env* gene region. Branch support values are bootstrap proportions from 200 pseudoreplicates.

Conclusions

Discussion of findings: In this study, we have demonstrated feasibility of full-length genome sequence production from individuals infected with HIV using NGS technologies and phylogenomic analysis. For a subset of full-length HIV genomes, we have demonstrated initial phylogenetic estimates from six non-overlapping genic regions showing (*i*) the monophyly of clones sampled from the same recipient individuals, (*ii*) variation in relationships amongst groups of clones from different individuals, and (*iii*) variation in rates of evolution across different genomic regions. These results are significant in demonstrating the integrity of the phylogenetic signal in these assembled CC molecular clone sequences as the extrinsic evidence, revealed during criminal proceedings, strongly suggested that each individual included in our initial analyses only acted as a recipient within the Texas transmission cluster (2). Distinct relationships among clones from different individuals inferred from numerous genic regions are strong evidence of independent evolutionary histories, thus providing greater confidence in forensic conclusions. We have also demonstrated that complex evolutionary processes, like convergent evolution, can affect 'standard' phylogenetic results and provided an analytical framework to detect the spurious signal produced by such processes. Several technical challenges are currently being addressed that include (*i*) optimizing *de novo* assemblies that will produce larger numbers of successful full-length genome sequences with low SNP counts and (*ii*) understanding the high estimates of human sequences in the subset of HIV molecular clone reads. Overcoming these challenges will lead into the development of a robust 'pathogen toolkit' that can provide a detailed roadmap for future forensics studies.

Implications for policy and practice: As noted above, several technical challenges exist in the current study. Assuming these issues are resolved, we believe this work could have a significant impact in terms of public practice. Examples of microbial forensic studies that will benefit from our study include characterization of vaginal swabs to test for sexually transmitted microbes in suspected sexual assault cases involving minors (50), pathogenic microbes intentionally introduced to contaminate food supplies, such as *Salmonella* (51) and *Shigella* (52), non-curable diseases transmitted with the intent to cause bodily harm, such as those associated with HIV (2, 12, 53, 54) and hepatitis C (12, 54) infections, or anthrax exposure with the intent to cause death (55). Although the recent cholera outbreak in Haiti does not appear intentional (56), the occurrence highlights the potential for deliberate contamination of public water supplies. While we believe that policy recommendations may be premature, demonstration projects involving other microbial forensic studies that employ NGS with phylogenomic approaches will shed light on this topic, providing more informed decisions on public policy issues.

Implications for further research: We believe this work can be expanded to the more general field of microbial forensics to aid in solving crimes involving a range of pathogens, as well as developing measures to enhance public safety. Because microbes have developed a number of elaborate mechanisms for generating tremendous genetic diversity, the demonstration of NGS technologies coupled with advanced phylogenetic methods are being developed into a 'pathogen toolkit'. We believe these advanced 'tools' will provide (*i)* identification of the most informative sources of genomic variation across different pathogen species, (*ii*) characterization of the standing variation in large isolate sets from model species and localities, (*iii*) tests for the action

of complex evolutionary processes and their effects on phylogenetic inference in model pathogens using newly developed phylogenetic models, and (*iv*) empirical benchmarking of whole-genome sequences and phylogenomics to accurately recover externally verified transmission events.

References

1. Doyle, V. P., Andersen, J. J., Nelson, B. J., Metzker, M. L., and Brown, J. M. (2014) Untangling the influences of unmodeled evolutionary processes on phylogenetic signal in a forensically important HIV-1 transmission cluster. *Mol. Phylogenet. Evol.* **75**, 126-137
2. Scaduto, D. I., Brown, J. M., Haaland, W. C., Zwickl, D. J., Hillis, D. M., and Metzker, M. L. (2010) Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proc. Natl. Acad. Sci. USA* **107**, 21242-21247
3. Boussau, B., Guéguen, L., and Gouy, M. (2009) A mixture model and a hidden Markov model to simultaneously detect recombination breakpoints and reconstruct phylogenies. *Evol. Bioinformatics* **5**, 67
4. Taylor, L. H., Latham, S. M., and woolhouse, M. E. J. (2001) Risk factors for human disease emergence. *Phil. Trans. R. Soc. Lond. B* **356**, 983-989
5. Budowle, B., Murch, R., and Chakraborty, R. (2005) Microbial forensics: the next forensic challenge. *Int. J. Legal Med.* **119**, 317-330
6. Preston, B. D., Poiesz, B. J., and Loeb, L. A. (1988) Fidelity of HIV-1 reverse transcriptase. *Science* **242**, 1168-1171
7. Boyer, J. C., Bebenek, K., and Kunkel, T. A. (1992) Unequal human immunodeficiency virus type 1 reverse transcriptase error rates with RNA and DNA templates. *Proc. Natl. Acad. Sci. USA* **89**, 6919-6923
8. Mansky, L. M., and Temin, H. M. (1995) Lower *in vivo* mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.* **69**, 5087-5094
9. Zhuang, J., Jetzt, A. E., Sun, G., Yu, H., Klarmann, G., Ron, Y., Preston, B. D., and Dougherty, J. P. (2002) Human immunodeficiency virus Type 1 recombination: Rate, fidelity, and putative hot spots. *J. Virol.* **76**, 11273-11282
10. Jung, A., Maier, R., Vartanian, J.-P., Bocharov, G., Jung, V., Fischer, U., Meese, E., Wain-Hobson, S., and Meyerhans, A. (2002) Recombination: Multiply infected spleen cells in HIV patients. *Nature* **418**, 144-144
11. Rhodes, T., Wargo, H., and Hu, W.-S. (2003) High Rates of human immunodeficiency virus type 1 recombination: Near-random segregation of markers one kilobase apart in one round of viral replication. *J. Virol.* **77**, 11193-11200
12. Metzker, M. L., Mindell, D. P., Liu, X.-M., Ptak, R. G., Gibbs, R. A., and Hillis, D. M. (2002) Molecular evidence of HIV-1 transmission in a criminal case. *Proc. Natl. Acad. Sci. USA* **99**, 14292-14297
13. Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifson, J. D., Bonhoeffer, S., Nowak, M. A., Hahn, B. H., Saag, M. S., and Shaw, G. M. (1995) Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* **373**, 117-122
14. Ho, D. D., Neumann, A. U., Perelson, A. S., Chen, W., Leonard, J. M., and Markowitz, M. (1995) Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature* **373**, 123-126
15. Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M., and Ho, D. D. (1996) HIV-1 dynamics in vivo: Virion clearance rate, infected cell life-span, and viral generation time. *Science* **271**, 1582-1586

16. Coffin, J. M. (1995) HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science* **267**, 483-489
17. Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. (2004) The causes and consequences of HIV evolution. *Nature Rev. Genet.* **5**, 52-61
18. Minin, V. N., Dorman, K. S., Fang, F., and Suchard, M. A. (2005) Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics* **21**, 3034-3042
19. Ané, C., Larget, B., Baum, D. A., Smith, S. D., and Rokas, A. (2007) Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.* **24**, 412-426
20. Ané, C. (2011) Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. *Genome Biol. Evol.* **3**, 246-258
21. Lemey, P., Rambaut, A., and Pybus, O. G. (2006) HIV evolutionary dynamics within and among hosts. *AIDS Reviews* **8**, 125-140
22. Guindon, S., Rodrigo, A. G., Dyer, K. A., and Huelsenbeck, J. P. (2004) Modeling the site-specific variation of selection patterns along lineages. *Proc. Natl Acad. Sci. USA* **101**, 12957-12962
23. Huelsenbeck, J. P., Jain, S., Frost, S. W. D., and Pond, S. L. K. (2006) A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc. Natl Acad. Sci. USA* **103**, 6263-6268
24. Penn, O., Stern, A., Rubinstein, N. D., Dutheil, J., Bacharach, E., Galtier, N., and Pupko, T. (2008) Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. *PLoS Comput. Biol.* **4**, e1000214
25. Moore, B., McGuire, J., Ronquist, F., and Huelsenbeck, J. P. (2011) Bayesian analysis of partitioned data. *Syst Biol, Accepted*
26. Huelsenbeck, J. P., and Suchard, M. A. (2007) A nonparametric method for accommodating and testing across-site rate variation. *Syst Biol* **56**, 975-987
27. Lemey, P., Derdelinckx, I., Rambaut, A., Van Laethem, K., Dumont, S., Vermeulen, S., Van Wijngaerden, E., and Vandamme, A.-M. (2005) Molecular footprint of drug-selective pressure in a human immunodeficiency virus transmission chain. *J. Virol.* **79**, 11981-11989
28. Salvi, R., Garbuglia, A. R., Di Caro, A., Pulciani, S., Montella, F., and Benedetto, A. (1998) Grossly defective *nef* gene sequences in a human immunodeficiency virus type 1-seropositive long-term nonprogressor. *J. Virol.* **72**, 3646-3657
29. Lahr, D. J. G., and Katz, L. A. (2009) Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques* **47**, 857–866
30. Barnes, W. M. (1994) PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc. Natl. Acad. Sci. USA* **91**, 2216-2220
31. Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760
32. Zerbino, D. R., and Birney, E. (2008) Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821-829
33. Li, H. (2012) Exploring single-sample SNP and INDEL calling with whole-genome *de novo* assembly. *Bioinformatics* **28**, 1838-1844
34. Katoh, K., Kuma, K.-i., Toh, H., and Miyata, T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511-518
35. Katoh, K., Misawa, K., Kuma, K. i., and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059-3066
36. Katoh, K., and Toh, H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics* **9**, 286-298

37. Swofford, D. L. (2000) PAUP*. Phylogenetic Analysis Using Parsimony (*and other substitutions). Version 4. Sinauer Associates, Sunderland, MA
38. Altschul, S. F., Gish, W., Miller, W., Meyers, E. W., and Lipman, D. J. (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.* **215**, 403-410
39. Penn, O., Privman, E., Landan, G., Graur, D., and Pupko, T. (2010) An alignment confidence score capturing robustness to guide tree uncertainty. *Mol. Biol. Evol.* **27**, 1759-1767
40. Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Meth.* **9**, 772-772
41. Guindon, S. p., and Gascuel, O. (2003) A Simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696-704
42. Zwickl, D. J. (2006) Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin. The Garli program is available from http://garli.nescent.org.
43. Gao, F., Robertson, D. L., Carruthers, C. D., Morrison, S. G., Jian, B., Chen, Y., Barre-Sinoussi, F., Girard, M., Srinivasan, A., Abimiku, A. l. G., Shaw, G. M., Sharp, P. M., and Hahn, B. H. (1998) A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. *J. Virol.* **72**, 5680-5698
44. Salminen, M. O., Koch, C., Sanders-Buell, E., Ehrenberg, P. K., Michael, N. L., Carr, J. K., Burke, D. S., and McCutchan, F. E. (1995) Recovery of virtually full-length HIV-1 provirus of diverse subtypes from primary virus cultures using the polymerase chain reaction. *Virology* **213**, 80-86
45. Ewing, B., and Green, P. (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8**, 186-194
46. Shen, Y., Wan, Z., Coarfa, C., Drabek, R., Chen, L., Ostrowski, E. A., Liu, Y., Weinstock, G. M., Wheeler, D. A., Gibbs, R. A., and Yu, F. (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res.* **20**, 273-280
47. The 1000 Genomes Project Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073
48. International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921
49. Malboeuf, C. M., Yang, X., Charlebois, P., Qu, J., Berlin, A. M., Casali, M., Pesko, K. N., Boutwell, C. L., DeVincenzo, J. P., Ebel, G. D., Allen, T. M., Zody, M. C., Henn, M. R., and Levin, J. Z. (2013) Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification. *Nucleic Acids Res.* **41**, e13
50. Centers for Disease Control and Prevention. (2002) Sexual assault and STDs. Sexually transmitted diseases treatment guidelines. *MMWR Recomm. Rep.* **51 (RR-6)**, 69-74
51. Török, T. J., Tauxe, R. V., Wise, R. P., Livengood, J. R., Sokolow, R., Mauvais, S., Birkness, K. A., Skeels, M. R., Horan, J. M., and Foster, L. R. (1997) A large community outbreak of salmonellosis caused by intentional contamination of restaurant salad bars. *JAMA* **278**, 389-395
52. Kolavic, S. A., Kimura, A., Simons, S. L., Slutsker, L., Barth, S., and Haley, C. E. (1997) An outbreak of *Shigella dysenteriae* type 2 among laboratory workers due to intentional food contamination. *JAMA* **278**, 396-398
53. Albert, J., Wahlberg, J., Leitner, T., Escanilla, D., and Uhlén, M. (1994) Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 pol and gag genes. *J. Virol.* **68**, 5918-5924
54. de Oliveira, T., Pybus, O. G., Rambaut, A., Salemi, M., Cassol, S., Ciccozzi, M., Rezza, G., Gattinara, G. C., D'Arrigo, R., Amicosante, M., Perrin, L., Colizzi, V., Perno, C. F., and Benghazi Study, G. (2006) Molecular epidemiology: HIV-1 and HCV sequences from Libyan outbreak. *Nature* **444**, 836-837

55. Read, T. D., Salzberg, S. L., Pop, M., Shumway, M., Umayam, L., Jiang, L., Holtzapple, E., Busch, J. D., Smith, K. L., Schupp, J. M., Solomon, D., Keim, P., and Fraser, C. M. (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* **296**, 2028-2033

56. Chin, C.-S., Sorenson, J., Harris, J. B., Robins, W. P., Charles, R. C., Jean-Charles, R. R., Bullard, J., Webster, D. R., Kasarskis, A., Peluso, P., Paxinos, E. E., Yamaichi, Y., Calderwood, S. B., Mekalanos, J. J., Schadt, E. E., and Waldor, M. K. (2011) The Origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364**, 33-42

Dissemination of Research Findings

Publications: The manuscript by Doyle et al. entitled, "Untangling the influences of unmodeled evolutionary processes on phylogenetic signal in a forensically important HIV-1 transmission cluster" has now been published, see reference 1 above.

Presentations: Drs. Metzker and Brown have made a number of presentations involving research derived from our NIJ grant. Those meetings are listed below:

- Dr. Metzker attended the American Society of Microbiology in June 2012 and gave a lecture on the use of NGS in pathogen forensics. A portion of that talk focused on progress being made under the NIJ grant.

- Dr. Brown attended the annual meeting of the Society for Molecular Biology and Evolution, held at the Convention Centre in Dublin, Ireland from June 23rd-27th, 2012.

- Dr. Brown attended the First Joint Congress on Evolution, sponsored by five academic societies[26] and held at the Convention Centre in Ottawa, Ontario, Canada from July 6th-10th, 2012.

- Dr. Metzker was invited to (and Dr. Steve Scherer at BCM spoke at) the ABRF 2013 Satellite Workshop in Palm Springs, CA on March 2, 2013. The title of the talk was "*Application of NGS in HIV Forensics.*" A portion of that talk focused on progress being made under the NIJ grant.

- Dr. Doyle, John Andersen, and Brad Nelson attended MEEGID XI, the 11th International Conference on Molecular Epidemiology and Evolutionary Genetics of Infectious Diseases, held in New Orleans, LA from Oct. 30th – Nov. 2nd, 2012.

[26] American Society of Naturalists, Canadian Society for Ecology and Evolution, European Society for Evolutionary Biology, Society for the Study of Evolution, and the Society of Systematic Biologists