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).