RESEARCH ARTICLE

Deep sequencing of near full-length HIV-1 genomes from plasma identifies circulating subtype C and infrequent occurrence of AC recombinant form in Southern India

Shuba Varshini Alampalli[1☯], Michael M. Thomson[2☯], Raghavan Sampathkumar[1], Karthi Sivaraman[1], Anto Jesuraj U. K. J.[3], Chirag Dhar[3], George D. Souza[4], Neil Berry[5‡], Annapurna Vyakarnam[1,6‡*]

1 Centre for Infectious Disease Research (CIDR), Indian Institute of Science, Bengaluru, India, **2** Centro Nacional de Microbiología, Instituto de Salud Carlos III, Ctra. Majadahonda-Pozuelo, Majadahonda, Madrid, Spain, **3** Department of Infectious Diseases, St John's Research Institute, Bengaluru, India, **4** Department of Pulmonary Medicine & Department of Infectious Diseases, St John's Research Institute, Bengaluru, India, **5** Division of Virology, NIBSC, South Mimms, United Kingdom, **6** Department of Infectious Diseases, King's College London, London, United Kingdom

☯ These authors contributed equally to this work.
‡ These authors also contributed equally to this work.
* anna.vyakarnam@kcl.ac.uk, avyakarnam@cidr.iisc.ernet.in

Abstract

India has the third largest number of HIV-1-infected individuals accounting for approximately 2.1 million people, with a predominance of circulating subtype C strains and a low prevalence of subtype A and A1C and BC recombinant forms, identified over the past two decades. Recovery of near full-length HIV-1 genomes from a plasma source coupled with advances in next generation sequencing (NGS) technologies and development of universal methods for amplifying whole genomes of HIV-1 circulating in a target geography or population provides the opportunity for a detailed analysis of HIV-1 strain identification, evolution and dynamics. Here we describe the development and implementation of approaches for HIV-1 NGS analysis in a southern Indian cohort. Plasma samples ($n = 20$) were obtained from HIV-1-confirmed individuals living in and around the city of Bengaluru. Near full-length genome recovery was obtained for 9 Indian HIV-1 patients, with recovery of full-length *gag* and *env* genes for 10 and 2 additional subjects, respectively. Phylogenetic analyses indicate the majority of sequences to be represented by subtype C viruses branching within a monophyletic clade, comprising viruses from India, Nepal, Myanmar and China and closely related to a southern African cluster, with a low prevalence of the A1C recombinant form also present. Development of algorithms for bespoke recovery and analysis at a local level will further aid clinical management of HIV-1 infected Indian subjects and delineate the progress of the HIV-1 pandemic in this and other geographical regions.

OPEN ACCESS

Citation: Alampalli SV, Thomson MM, Sampathkumar R, Sivaraman K, U. K. J. AJ, Dhar C, et al. (2017) Deep sequencing of near full-length HIV-1 genomes from plasma identifies circulating subtype C and infrequent occurrence of AC recombinant form in Southern India. PLoS ONE 12 (12): e0188603. https://doi.org/10.1371/journal.pone.0188603

Editor: William A. Paxton, University of Liverpool Institute of Infection and Global Health, UNITED KINGDOM

Received: September 8, 2017

Accepted: November 9, 2017

Published: December 8, 2017

Data Availability Statement: NCBI GenBank accession numbers for the sequences reported here are: KY713228 (SC007); KY713229 (SC008); KY713230 (SC013); KY713231 (SC015); KY713232 (SC017); KY713233 (SC022); KY713234 (SC062); KY713235 (SC085); KY713236 (SHE001); KY713237 (gag gene SC012); KY713238 (gag gene SC018); KY713239 (gag gene SC061); KY713240 (gag gene SC063);

KY713241 (gag gene SC064); KY713242 (gag gene SC065); KY713243 (gag gene SC072); KY713244 (gag gene SC073); KY713245 (gag gene SC019); KY713246 (env gene SC019); KY713247 (gag gene SC023); KY713248 (env gene SC023).

Funding: This work was funded partly by a EC FP7 EURIPRED (FP7-INFRA-2012 Grant Agreement No. 312661) to AV, MMT, GDS and a Centre for Excellence award from the Department of Biotechnology (DBT), India to AV (BT/01/CEIB/12/III/09). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Characterising the genetic composition of HIV-1 circulating in different geographical locales represents a significant undertaking, with implications for vaccine design, treatment efficacy regimes and understanding the molecular epidemiology of transmitted variants in target populations, with group M subtypes (A-K) and circulating recombinant forms (CRFs) comprising the majority of the genotypes in global circulation. HIV-1 clade C infections represent approximately half of the total HIV-1 infections worldwide, being the majority variant in Southern Africa and the Indian subcontinent [1]. HIV-1 clade C therefore represents a significant target for vaccine efforts. We were interested in determining the specific variants of HIV-1 clade C circulating in a cohort of HIV-1-infected individuals in Bengaluru, southern India, using deep sequencing approaches. Previously, whole genome recovery of HIV-1 RNA from plasma using a pan-HIV-1 amplification strategy and next generation sequencing (NGS) and bioinformatics analyses has been reported [2–4], providing fresh insight on HIV-1 variation by taking into account many under-represented HIV-1 variants and recombinant forms in a given host.

In this study, we describe modifications of the whole genome amplification protocol described by Gall *et al* [4] applied to clinically-relevant plasma viruses and a bioinformatics framework in the form of a bespoke pipeline, to amplify HIV-1 genomes from Indian clinical isolates and assemble them reliably post-NGS sequencing. Further, we describe a phylogenetic analysis of these Indian HIV-1 sequences using this pipeline for the first time, as either whole genome sequences or *gag* genes to determine the relationship with the diversity of HIV-1 clade C infections worldwide.

Materials and methods

Clinical details

This study was conducted according to the principles expressed in the Declaration of Helsinki. This study was approved by the Ethical Review Committee of St. John's Medical College Hospital, Bengaluru, India (Ref no: 55/2015). Patients were included in this study after obtaining written consent. Relevant clinical information was documented in a pro forma and 30 ml of EDTA-anticoagulated peripheral blood was collected by venipuncture. A total of twenty subjects confirmed HIV seropositive (by the standard HIV I & II ELISA test and western blot) with known CD4$^+$ T-cell counts and plasma HIV viral loads (Roche diagnostics, Germany) were studied (see Table 1). Sixteen out of twenty subjects with no prior history of treatment with antiretroviral drugs [antiretroviral treatment (ART)-naïve] or post-exposure prophylaxis were included and were confirmed to be TB negative. Four out of the twenty HIV positive subjects (SHE001, SC007, SC019 and SC062) were diagnosed with pulmonary and/or extrapulmonary tuberculosis (TB). A diagnosis of pulmonary TB (SHE001) was ascertained by sputum smear microscopy and culture. Standard smear grading of 1+, 2+ and 3+ was used to ascertain the bacterial burden. Smear-negative cases of pulmonary TB cases were diagnosed and classified by the treating clinician from plain chest radiographs as per the Revised National Tuberculosis Control Program (RNTCP) standards. A diagnosis of extra-pulmonary TB (SC007, SC019 and SC062) was established from tissue specimens by Ziehl-Neelsen staining for detection of acid-fast bacilli in tissue samples (obtained as surgical specimens/biopsies/fine-needle aspiration specimens). The RNTCP-approved Nucleic-Acid Amplification Test (NAAT) GeneXpert MTB/Rif of tissue sample was used for confirmation of this diagnosis. Subjects SC007 and SC019 had received treatment for either TB and /or HIV. SC007 was treated for TB (Anti-Tubercular/Tuberculosis Therapy/Treatment (ATT) but ART-naïve) for 52 days, whereas SC019 had completed 6 months of treatment for both TB and HIV prior to blood collection

Table 1. Clinical data.

S. No.	Sample	Age	Sex	Patient Category and Therapy	Type and site of TB	Date of viral RNA isolation	CD4 count (cells/mm^3)	Viral Load (copies/ml)
1	SHE001	45	M	HIV+TB+ (Treatment Naïve)	Pulmonary TB, suspected intestinal TB	16-04-2015	773	138000
2	SC007	26	M	HIV+TB+ (ATT 52 days and ART-naïve)	Extra pulmonary TB, TB lymphadenitis	05-05-2015	510	725000
3	SC008	49	M	HIV+ ART-naïve	None	20-05-2015	389	2980000
4	SC012	34	F	HIV+ ART-naïve	None	25-05-2015	582	14100
5	SC013	30	F	HIV+ ART-naïve	None	28-05-2015	510	120000
6	SC015	30	M	HIV+ ART-naïve	None	29-05-2015	415	63000
7	SC017	45	M	HIV+ ART-naïve	None	01-06-2015	475	108000
8	SC018	31	M	HIV+ ART-naïve	None	02-06-2015	382	9010
9	SC019	34	M	HIV+TB+ (Treated for both for 6 months)	Extra pulmonary TB, Abdominal TB	06-02-2015	267	347000
10	SC022	42	F	HIV+ ART-naïve	None	08-06-2015	853	39900
11	SC023	49	M	HIV+ ART-naïve	None	08-06-2015	583	98600
12	SC061	30	F	HIV+ ART-naïve	None	31-08-2015	452	80900
13	SC062	51	M	HIV+TB+ (Treatment Naïve)	Extra pulmonary TB, Lymph node	03-09-2015	167	752000
14	SC063	28	F	HIV+ ART-naïve	None	07-09-2015	158	81600
15	SC064	35	M	HIV+ ART-naïve	None	07-09-2015	109	3330000
16	SC065	24	M	HIV+ ART-naïve	None	08-09-2015	497	10800
17	SC072	32	M	HIV+ ART-naïve	None	29-09-2015	230	5970
18	SC073	27	M	HIV+ ART-naïve	None	29-09-2015	581	131000
19	SC078	32	F	HIV+ ART-naïve	None	06-10-2015	755	13300
20	SC085	42	F	HIV+ ART-naïve	None	03-11-2015	99	368000

https://doi.org/10.1371/journal.pone.0188603.t001

for this study. Of the 20 samples studied, 65% were male; the median age was 33 years (range 24–51); median CD4$^+$ T-cell count at the time of study was 463.5 cells/mm^3 (range 99–853); and the median viral load was 103,300 copies/ml (range 5,978–3,330,000 copies/ml). The correlation between CD4$^+$ T-cell count and viral load for each sample is depicted in Figure A in S1 File.

Viral RNA isolation

Viral RNA was purified from 1 ml of EDTA-treated plasma using the QIAamp viral RNA mini-kits (Qiagen, Cat # 52906) as per manufacturer protocol with some modifications. Instead of the recommended 140 μl, 1 ml of EDTA-treated plasma was processed in a single QIAamp Mini column and RNA was eluted in two batches of 40 μl of elution buffer each. 10 μl aliquots of purified plasma RNA were stored at -20˚C for immediate use (less than 2 hours); long-term storage of RNA was at -80˚C.

Primer design and one-step RT-PCR

The strategy adopted was to develop the pan-HIV-1 primer approach for the putative amplification of HIV-1 genomes of all major groups (M, N and O), their subtypes and recombinants as previously reported [4]. Initial analysis of field samples from the Indian cohort suggested modifications of primers to improve amplification efficiency and coverage of the *gag-pol* region were required. Further modifications were introduced based on consensus C-clade genome sequences retrieved from the Los Alamos HIV Sequence Database and an extended

Table 2. Primers used in this study.

Amplicon Name	Primer Sequence	Coordinates on HXB2 genome	Reference
Amp1	Pan-HIV-1F: AGCCYGGGAGCTCTCTG Pan-HIV-1R: CCTCCAATTCCYCCTATCATTTT	26–1953 (1928 bp)	Gall et al., JCM 2012
Amp2 (1F&1R)	HIV-A2-SE-F1: AgTATgggCAAgCAgggAgCT HIV-A2-SE-R1: TgTCCTTCCTTTCCACATTTCC	891-2051(1160 bp; overlap with Amp1: 1062 bp)	This study
Amp3 (2F&3R)	HIV-A2-SE-F2: gATgACAgCATgTCAgggAgT HIV-A2-SE-R3: TATAggCTgTACTgTCCATTA	1827–3281 (1454 bp; overlap with Amp2: 224 bp)	This study
Amp4 (4F&4R)	HIV-A2-SE-F4: CTTCCACAgggATggAAggAT HIV-A2-SE-R4: CTgCCATTTgTACTgCTgTCTT	2994–4767 (1773 bp; overlap with Amp3: 287 bp)	This study
Amp5	Pan-HIV-3F: TTAAAAGAAAAGGGGGGATTGGG Pan-HIV-3R: TGGCYTGTACCGTCAGCG	4329–7394 (3066 bp; overlap with Amp4: 438 bp)	Gall et al., JCM 2012
Amp6	Pan-HIV-4F: CCTATGGCAGGAAGAAGCG Pan-HIV-4R: CTTWTATGCAGCWTCTGAGGG	5513–9063 (3551 bp; overlap with Amp5:1881 bp)	Gall et al., JCM 2012

https://doi.org/10.1371/journal.pone.0188603.t002

amplicon strategy adopted (summarized in Figure B in S1 File). Using the primers shown in Table 2, one-step RT-PCRs were performed to derive overlapping amplicons of 1.9 kb, 1.1 kb, 1.4 kb, 1.7 kb, 3 kb, and 3.5 kb using SuperScript III One-Step RT-PCR with Platinum *Taq* DNA High Fidelity Polymerase kits (Invitrogen, Cat # 12574–035). Each 25 μl reaction mixture contained 12.5 μl reaction mix (2 x), 4.5 μl RNase-free water, 1 μl each of each primer (conc. 20 nmol/μl), 1 μl SuperScript III RT/ Platinum Taq High Fidelity mix and 5 μl of template RNA. Cycling conditions were 50˚C for 30 min; 94˚C for 2 min; 35 cycles of 94˚C for 15 sec, 58˚C for 30 sec, and 68˚C for 4 min 30 sec; and, finally, 68˚C for 10 min. Amplicon quality was verified by agarose gel electrophoresis.

Illumina sequencing

Massive parallel sequencing of the pooled amplicons was carried out on an Illumina NextSeq 500 analyser. Library preparation was performed at Genotypic Technology's Genomics facility following Nextera XT DNA Library Preparation protocol (Cat #FC-131-1024). 1 ng of Qubit-quantified genomic DNA was tagmented (fragmented and tagged) using Amplicon Tagment Mix provided in the Nextera XT Kit. The adapter tagged DNA was subjected to 12 cycles of indexing-PCR [72˚C for 3 min followed by denaturation at 95˚C for 30 sec, cycling (95˚C for 10 sec, 55˚C for 30 sec, 72˚C for 30 sec) and 72˚C for 5 min] to enrich the adapter-tagged fragments. The PCR product was purified using HighPrep beads (Magbio Genomics, #AC-60050). Quantification and size distribution of the prepared libraries were determined using Qubit fluorometer and the Agilent D1000 TapeStation respectively according to the manufacturer's instructions.

Bioinformatics

QC and genome assembly.

Raw reads generated from the Illumina machine were subjected to quality assessment using FastQC [5] and Trimmomatic v0.36 [6]. Based on per base quality score and k-mer presence, sequences were trimmed 10 bases at the 5'-end and five bases at 3'-end. Trimmed reads were mapped using bowtie2 [7] (default parameters) onto 2870 complete HIV-1 genomes to identify the closest reference. HIV-1 mapped reads were extracted and *de novo* assembled using IVA (Iterative Virus Assembler) [8]. IVA-assembled contigs were used as input for PRICE (Paired-Read Iterative Contig Extension) [9], with default options. PRICE uses incomplete assemblies of the clinical isolate genome and extends the contigs to fill the gaps iteratively. We ran PRICE for 10 iterations, choosing all contigs above 1000 bases at each cycle as templates for the next cycle. In nine cases, we obtained near complete genome assemblies (size > 8.5 kb) using this strategy and the remaining eleven cases could be assembled to obtain the *gag* gene owing to inadequate coverage of reads. In two cases, we assembled both *gag* and *env* genes but did not have enough read support to complete the whole genome. The genomes were made into their own consensus sequence using samtools [10] and bcftools [11].

Phylogenetic analyses. HIV-1 reference sequences were downloaded from the Los Alamos HIV Sequence Database [12] for phylogenetic analyses. Multiple sequence alignment was performed using MUSCLE [13] (default parameters) and phylogenetic trees were generated via maximum likelihood (ML) under the general time-reversible with gamma-distributed rate heterogeneity among sites (GTR+Γ) evolutionary model using RAxML v8.2.10 [14]. Trees were visualized with MEGA 7 [15]. The presence of intersubtype recombination was initially checked with Recombination Identification Program (RIP) [16], with subsequent analysis of the mosaic structure with the bootscanning method implemented in Simplot v3.5.1 [17]. For SNP-based phylogenetic tree, the HIV-1 specific filter reads for each sample were trimmed (adaptor/transposon sequence, 5' < Q10 and 3' <Q20) and clipped for an average quality score of 20 using Trimmomatic v0.36. QC-ed reads were mapped using Bowtie2 and duplicates marked using Picard MarkDuplicates [18]. SNP/InDels were saved in VCF format using samtools and bcftools. In order to include multiple reference sequences for a SNP-based phylogenetic tree, multiple sequence alignment among the reference was carried out with MUSCLE and corresponding bases (indicated as SNP/InDel by vcftools [19]) against each reference were extracted and concatenated using in-house R-scripts. The SNPs and InDels were concatenated for each sample following the same order of sequences in each case.

Results

Amplicon recovery and protocol optimization

Adaptation of the 4-amplicon strategy to recover near full-length HIV-1 genomes to a 6-amplicon strategy (Figure B in S1 File) for large-scale HIV-1 genome sequencing improved recovery of HIV-1 gene fragments from the Bengaluru cohort. While providing enhanced recovery of viral strains circulating in this population, amplification efficiency remained variable with some isolates, which appeared independent of baseline viral load status. Amplicons for each primer-pair, pooled in equimolar concentrations based upon NanoDrop readings, resulted in more than 1 million reads for each sample (minimum 30,000 reads per amplicon).

Initial de novo assembly and extension

In order to assemble a viable HIV-1 genome and enhance the potentially unique nature of each clinical isolate, we followed a two-step process involving *de novo* assembly using IVA (step 1) and PRICE (step 2). Extracted reads were mapped onto a consensus C-clade genome with the resulting read coverage for each sample shown in Table 3. SC012, SC018, SC019, SC023, SC061, SC063, SC064, SC065, SC072, SC073 and SC078 were analyzed across *gag* as a result of low coverage downstream of the *pol* gene.

IVA uses an overlap consensus layout along with iteratively aligning reads to contigs so as to minimise sequencing errors and rare variations within the quasispecies. PRICE utilises the preassembled contigs from IVA and extends iteratively, until sufficiently long contigs are produced and gaps filled. In our runs, use of PRICE allowed us to merge multiple kilobase-sized contigs (Table 3) into a single genome-size assembly (>8500 bases) for nine strains.

Phylogenetic analysis of the clinical samples

Of the 20 Indian clinical samples, 9 yielded near full-length genomes (NFLG), with 10 further samples assembled across the complete *gag* gene. Sample SC078 had very few HIV-1 specific

Table 3. Summary of NGS results for each viral isolate.

Whole Genome

						Read Count for each primer used in this study					
Sl. No.	Sample ID	Total No. of paired-end reads	Percentage of HIV-1 specific reads	Contigs After PRICE	Length of the contig	Amp1	Amp2	Amp3	Amp4	Amp5	Amp6
1	SC007	12,048,590	45.86%	1	9041	1,081,560	839,188	2,560,787	2,406,795	38,537	17,069
2	SC008	3,456,568	16.75%	1	9084	82,876	58,111	193,065	312,744	34,558	16,670
3	SHE001	3,168,528	28.06%	1	9114	111,893	85,609	321,476	595,475	15,005	4,469
4	SC013	2,259,720	29.48%	1	9187	439,695	399,894	161,808	64,259	93,514	16,011
5	SC015	6,464,458	2.37%	1	9122	75,807	64,536	22,762	5,293	11,775	4,483
6	SC017	6,119,426	3.02%	1	9043	83,658	79,833	52,115	68,269	1,665	443
7	SC022	6,891,246	8.60%	1	8940	327,410	309,199	97,578	125,710	26,937	6,228
8	SC062	6,649,608	25.91%	1	9073	894,930	846,293	354,433	258,557	55,657	29,616
9	SC085	1,558,952	14.98%	1	9117	127,991	110,737	113,634	28,640	34,927	9,489

Partial Genome

						Read Count for each primer used in this study					
Sl. No.	Sample ID	Total No. of paired-end reads	Percentage of HIV-1 specific reads	Contigs After PRICE	Length of the contig	Amp1	Amp2	Amp3	Amp4	Amp5	Amp6
1	SC012	223,832	1.57%	1	1839	2,720	2,673	352	0	0	0
2	SC018	305,190	34.23%	1	2924	62,556	58,685	9,208	64	198	67
3	SC019	6,233,546	15.60%	2	4959;3621	602,209	569,824	112,558	22,808	34,062	10,188
4	SC023	6,185,790	9.91%	2	4841;4384	481,739	431,574	75,507	6,475	7,109	8,795
5	SC061	300,540	52.14%	2	3076;1324	88,149	81,933	14,422	203	360	1,014
6	SC063	1,720,104	2.33%	3	2948;2354;1919	47,377	42,021	7,144	788	515	1,394
7	SC064	168,998	63.92%	2	3592;3297	61,325	56,149	13,412	1,328	3,570	1,346
8	SC065	1,869,004	2.62%	1	1978	56,656	52,643	2,304	2	160	221
9	SC072	156,858	0.68%	1	2748	390	300	66	13	0	0
10	SC073	297,644	15.23%	1	2150	32,580	32,123	4,005	2	102	30

https://doi.org/10.1371/journal.pone.0188603.t003

reads and could not be assembled across the complete *gag* gene and was therefore removed from further analyses.

Since recombinant sequences can affect node support values and topologies in a phylogenetic tree, prior to tree construction, the newly obtained sequences were checked for the presence of intersubtype recombination using RIP. These analyses indicated that one virus, SC017, was A1C recombinant and all other sequences were of homogeneous subtypes (subtype C, except SC072 *gag* sequence, which was of subtype B). ML trees were constructed based on either whole-genome (Fig 1A) or *gag* alone (Fig 1B), excluding SC017 from the analyses. Multiple sequence alignments were generated with the HIV-1 subtype references downloaded from the Los Alamos HIV Sequence Database. In order to examine the phylogenetic relationship of the subtype C viruses from Bengaluru with other subtype C viruses from India and other countries, in the whole-genome tree, all subtype C NFLG sequences from India and other Asian countries (China, Myanmar and Nepal) available at the Los Alamos database, as well as references from the 11 subtype C clusters previously identified in Africa [20], were included in the analysis. The ML tree generated with this alignment (Fig 1A) shows that the Bengaluru sequences branch within a strongly supported monophyletic clade comprising all but one Asian subtype C sequences (the only exception being one sequence from India), within which sequences from Nepal, on the one hand, and China and Myanmar on the other, group in respective sub-clusters. The Asian subtype C clade is closely related to the previously identified African C6 cluster, comprising viruses from Botswana and South Africa, which in

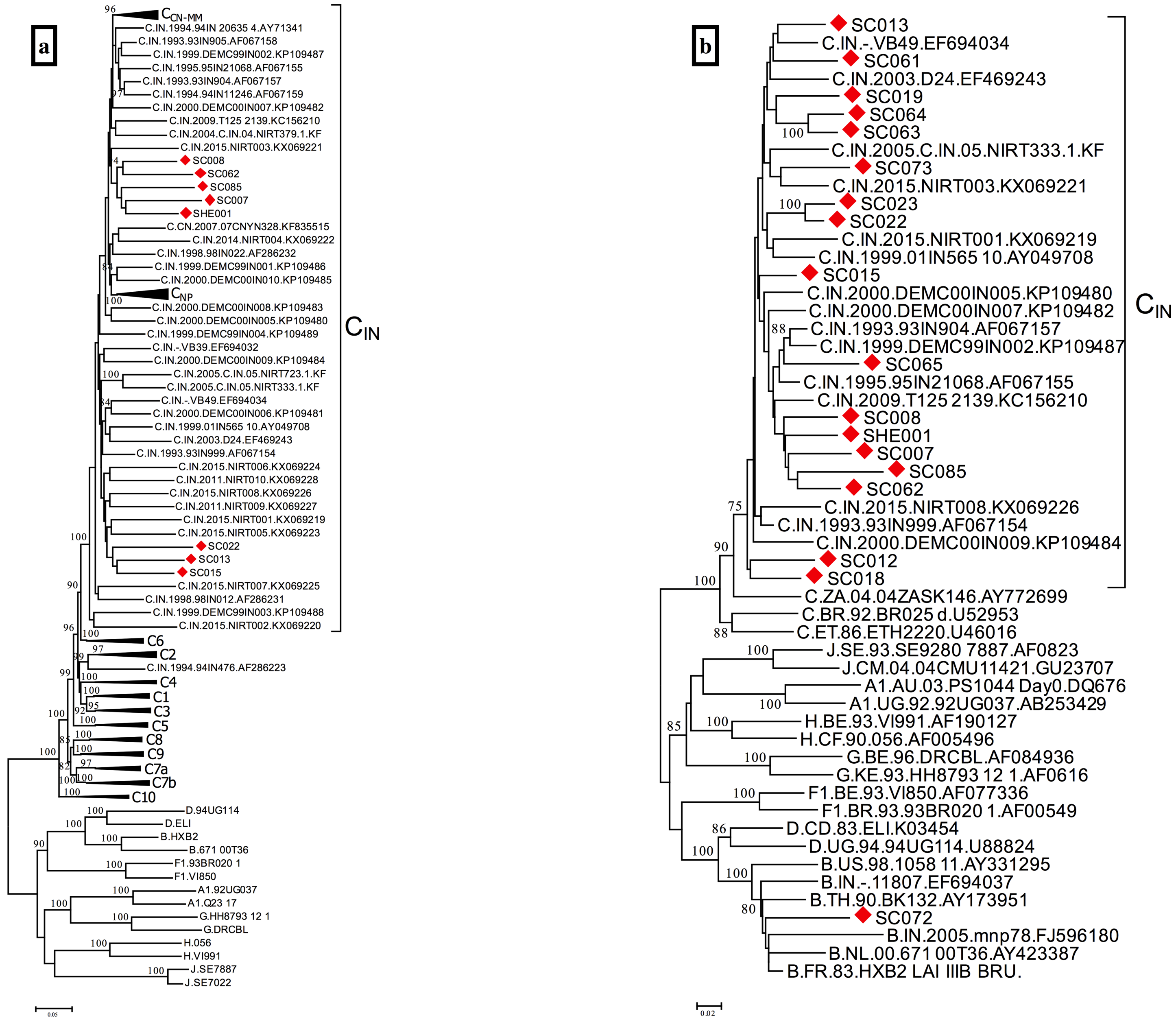
a
C CN-MM
C.IN.1994.94IN 20635 4.AY71341
C.IN.1993.93IN905.AF067158
C.IN.1999.DEMC99IN002.KP109487
C.IN.1995.95IN21068.AF067155
C.IN.1993.93IN904.AF067157
C.IN.1994.94IN11246.AF067159
C.IN.2000.DEMC00IN007.KP109482
C.IN.2009.T125 2139.KC156210
C.IN.2004.C.IN.04.NIRT379.1.KF
C.IN.2015.NIRT003.KX069221
SC008
SC062
SC085
SC007
SHE001
C.CN.2007.07CNYN328.KF835515
C.IN.2014.NIRT004.KX069222
C.IN.1998.98IN022.AF286232
C.IN.1999.DEMC99IN001.KP109486
C.IN.2000.DEMC00IN010.KP109485
C NP
C.IN.2000.DEMC00IN008.KP109483
C.IN.2000.DEMC00IN005.KP109480
C.IN.1999.DEMC99IN004.KP109489
C.IN.-.VB39.EF694032
C.IN.2000.DEMC00IN009.KP109484
C.IN.2005.C.IN.05.NIRT723.1.KF
C.IN.2005.C.IN.05.NIRT333.1.KF
C.IN.-.VB49.EF694034
C.IN.2000.DEMC00IN006.KP109481
C.IN.1999.01IN565 10.AY049708
C.IN.2003.D24.EF469243
C.IN.1993.93IN999.AF067154
C.IN.2015.NIRT006.KX069224
C.IN.2011.NIRT010.KX069228
C.IN.2015.NIRT008.KX069226
C.IN.2011.NIRT009.KX069227
C.IN.2015.NIRT001.KX069219
C.IN.2015.NIRT005.KX069223
SC022
SC013
SC015
C.IN.2015.NIRT007.KX069225
C.IN.1998.98IN012.AF286231
C.IN.1999.DEMC99IN003.KP109488
C.IN.2015.NIRT002.KX069220
C IN
C6
C2
C.IN.1994.94IN476.AF286223
C4
C1
C3
C5
C8
C9
C7a
C7b
C10
D.94UG114
D.ELI
B.HXB2
B.671 00T36
F1.93BR020 1
F1.VI850
A1.92UG037
A1.Q23 17
G.HH8793 12 1
G.DRCBL
H.056
H.VI991
J.SE7887
J.SE7022
0.05
b
SC013
C.IN.-.VB49.EF694034
SC061
C.IN.2003.D24.EF469243
SC019
SC064
SC063
C.IN.2005.C.IN.05.NIRT333.1.KF
SC073
C.IN.2015.NIRT003.KX069221
SC023
SC022
C.IN.2015.NIRT001.KX069219
C.IN.1999.01IN565 10.AY049708
SC015
C.IN.2000.DEMC00IN005.KP109480
C.IN.2000.DEMC00IN007.KP109482
C.IN.1993.93IN904.AF067157
C.IN.1999.DEMC99IN002.KP109487
SC065
C.IN.1995.95IN21068.AF067155
C.IN.2009.T125 2139.KC156210
SC008
SHE001
SC007
SC085
SC062
C.IN.2015.NIRT008.KX069226
C.IN.1993.93IN999.AF067154
C.IN.2000.DEMC00IN009.KP109484
SC012
SC018
C IN
C.ZA.04.04ZASK146.AY772699
C.BR.92.BR025 d.U52953
C.ET.86.ETH2220.U46016
J.SE.93.SE9280 7887.AF0823
J.CM.04.04CMU11421.GU23707
A1.AU.03.PS1044 Day0.DQ676
A1.UG.92.92UG037.AB253429
H.BE.93.VI991.AF190127
H.CF.90.056.AF005496
G.BE.96.DRCBL.AF084936
G.KE.93.HH8793 12 1.AF0616
F1.BE.93.VI850.AF077336
F1.BR.93.93BR020 1.AF00549
D.CD.83.ELI.K03454
D.UG.94.94UG114.U88824
B.US.98.1058 11.AY331295
B.IN.-.11807.EF694037
B.TH.90.BK132.AY173951
SC072
B.IN.2005.mnp78.FJ596180
B.NL.00.671 00T36.AY423387
B.FR.83.HXB2 LAI IIIB BRU.
0.02

Fig 1. Maximum likelihood phylogenetic trees of HIV-1 sequences from Bengaluru. Sequences from Bengaluru obtained in this study are labelled with red diamonds. Only bootstrap values ≥70% are shown. (a) Phylogenetic tree of NFLG sequences of subtype C. All NFLG subtype C sequences from Asia available at the Los Alamos HIV Sequence Database are included in the analysis. References of African subtype C clusters C1 through C10 [20] are also included and are shown compressed in triangles, as are clusters comprising sequences from China and Myanmar (C_{CN-MM}) and from Nepal (C_{NP}) (the last cluster comprising one sequence from India) nested within the Indian clade (C_{IN}). (b) Phylogenetic tree of *gag* sequences of subtypes C and B. Fifteen randomly selected subtype C sequences and the only two subtype B sequences from India available at the Los Alamos database were included in the analysis, together with subtype references.

https://doi.org/10.1371/journal.pone.0188603.g001

turn forms part of a supercluster comprising six clusters (C1 through C6) associated with southern Africa [20]. Within the Asian subtype C clade, the Bengaluru sequences failed to group with each other, except two sequences, SC008 and SC062, which joined in a cluster supported by a 94% bootstrap value. In the *gag* tree (Fig 1B), in which 15 other subtype C and the two subtype B full-length *gag* sequences from India available at the Los Alamos database were included, 18 sequences from Bengaluru were of subtype C, branching within the Indian/Asian clade, with only two pairs of sequences (SC022/SC023; SC063/SC064) grouping in well supported clusters, and one was of subtype B. Both viruses sequenced only in *env* were subtype C, branching within the C_{IN} clade, in agreement with the *gag* phylogenetic tree (results not shown).

SNP based phylogenetic tree was constructed in the same way and the orientation of the samples reflects the results of the previous trees. The SNPs and InDels in each sample annotated on consensus subtype-C genome are represented in Figure C in S1 File and many of the variations appear to be in the *env* gene.

The mosaic structure of SC017 was analyzed by bootscanning using Indian A1 and C subtype references, together with references of subtypes B and H used as outgroups. The resulting bootscan plot (Fig 2) shows a complex structure with multiple breakpoints along the genome. To examine whether SC017 is related to other Indian A1C recombinant viruses, an ML phylogenetic tree was constructed with all Indian A1C recombinant NFLG sequences (only one per patient) deposited at the Los Alamos database. The tree shows clustering (75% bootstrap support) of SC017 only with one A1C recombinant, 95IN21301 (Figure D in S1 File). However, its mosaic structure differs substantially from that of SC017 (Figure E in S1 File).

NCBI GenBank accession numbers for the sequences reported here are: KY713228 (SC007); KY713229 (SC008); KY713230 (SC013); KY713231 (SC015); KY713232 (SC017); KY713233 (SC022); KY713234 (SC062); KY713235 (SC085); KY713236 (SHE001); KY713237 (*gag* gene SC012); KY713238 (*gag* gene SC018); KY713239 (*gag* gene SC061); KY713240 (*gag* gene SC063); KY713241 (*gag* gene SC064); KY713242 (*gag* gene SC065); KY713243 (*gag* gene SC072); KY713244 (*gag* gene SC073); KY713245 (*gag* gene SC019); KY713246 (*env* gene SC019); KY713247 (*gag* gene SC023); KY713248 (*env* gene SC023).

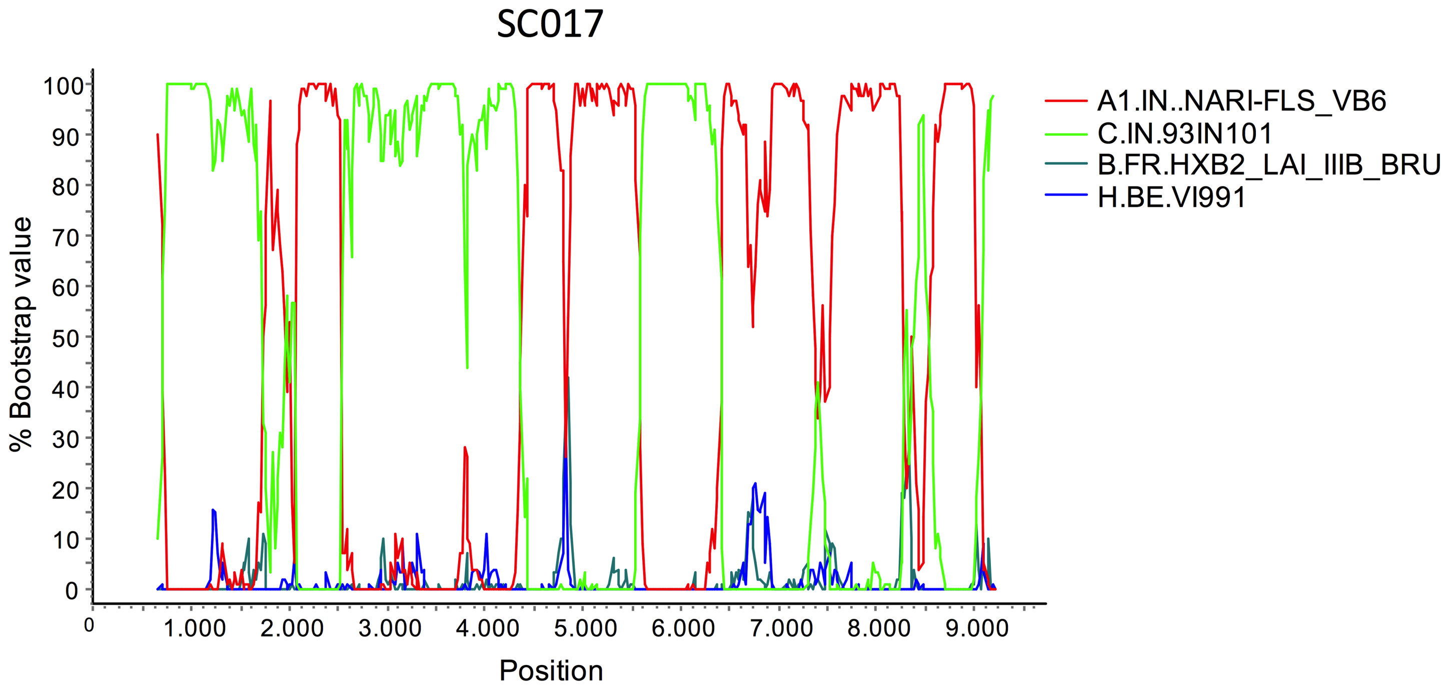

Fig 2. Bootscan analysis of the NFLG of SC017. The horizontal axis represents the position in the HXB2 genome and the vertical axis represents percent bootstrap values supporting clustering with reference sequences. Trees were constructed with the neighbor-joining method, using a window of 250 nt, sliding along the alignment in 20 nt steps.

https://doi.org/10.1371/journal.pone.0188603.g002

Discussion

Here, we describe near full length genome recovery and assembly of HIV-1 from a clinical cohort in southern India using plasma-derived viral RNA, comprising recently replicating viral variants circulating in 20 HIV-1-diagnosed individuals from Bengaluru city. Overlapping amplicons using RT-PCR-based protocols and Illumina NextSeq 150 technology were used to generate *de novo* genome assemblies, optimized to enhance sequencing data/flow-through and to reflect evolution of HIV-1 genomes using phylogenetic analyses. Several universal methods for amplifying and sequencing whole HIV-1 genomes have been reported [2–4, 21–24]. However, we adapted the universal amplification protocol of Gall *et al* [4], where oligonucleotides were originally selected based on the large number of collated HIV-1 sequences in the Los Alamos database minimising amplification bias within highly conserved regions using bioinformatics predictions of sequence fidelity. Originally performed on 10 WHO HIV-1 genotype reference strains and 15 HIV-1 plasma RNA samples of various subtypes and CRFs an overall sensitivity of detection of 3000 HIV-1 RNA copies/ml was achieved. However, direct application of the universal primers appeared suboptimal for many field strains circulating in India, with plasma virus representing the primary source of viral RNA and we were unable to improve on the 3000 RNA copy number detection threshold; an overall 70% amplification efficiency success rate was, however, consistent with previous findings [2]. Interestingly, the inability to amplify from a proportion of Indian HIV-1 strains appeared independent of viral load even with re-design of primers, for example, to the *gag/pol* region making further refinement and re-design of these protocols inevitable. This is perhaps due to lack of documented circulating strains from South India, which can be used as a template for more efficient primer design and our data emphasizes significant differences in the *gag-pol* region of the circulating strains compared to the available reference genomes.

While Illumina sequencing is a stochastic process which may not attain equal coverage of all amplicons in a given pool, this could potentially be overcome by preparing individual libraries for each amplicon of every sample. However, the significant associated cost escalation precluded using this option. At least a million reads of a single library were generated per sample (comprising an equimolar pool of all 6 amplicons), resulting in ~30,000 reads per amplicon which represented a mathematical minimum requirement for downstream *de novo* genome assembly. In addition, *de novo* assemblies are not robust in assembling whole genomes from short, paired-end data with inherent gaps in sequence coverage. Mindful of these limitations, in an Indian clade C dominant population, we manually assessed read quality, assembly construction and number of iterations during assembly extension for each sample, rather than use a standard reference-based assembly approach.

The data generated extends our knowledge of HIV-1 strain characterization of viruses circulating in India reported since the 1990s, confirming that in addition to the largely subtype C predominance, subtypes A1 and B are also circulating as minor clades giving rise to A1C and BC recombinant forms [25–35]. Although analyses have usually been restricted to single genes, typically *gag*, *pol*, or *env*, some authors have reported sequencing of full-length proviral DNA of HIV-1 in India from cultured peripheral blood mononuclear cells [17,33,35–38], representative of both archived and replicating viral genomes. Reports of HIV-1 whole genome sequencing from plasma RNA, to identify viruses that are currently replicating and circulating in India, is however limited to one publication [2], in which sequences from 10 field HIV-1 samples genotypically confirmed to be subtype C strains were reported.

Phylogenetic analyses of partial or full-length genomes have shown that the great majority of HIV-1 subtype C strains from India group in a monophyletic clade [17,25,26,30,36,38–41], designated C_{IN} [26], which, according to phylodynamic analyses, originated in the 1970s [41–

43], although sporadic cases of subtype C viruses branching outside of the C_{IN} clade have been reported [6,30,36,41]. Analyses of relationships with viruses from other countries indicate that the African viruses most closely related to C_{IN} are from the southern part of the continent [36,41], and that subtype C strains circulating in China [36,42,44], Myanmar [42], Nepal [45,46] and Bangladesh [47] derive from the Indian clade. Our ML phylogenetic trees confirm the monophyly of most Indian subtype C viruses, with all 8 NFLG and 18 *gag* subtype C viruses from Bengaluru branching within the C_{IN} clade, interspersed among other Indian viruses (Fig 1). Additionally, for the first time we examine the relationship of the C_{IN} clade with the previously identified African subtype C clusters [20], finding that it is closely related to a cluster (C6) comprising viruses from Botswana and South Africa (Fig 1A), which in turn groups in a supercluster (C1-C6) associated to Southern African countries, thus confirming previous reports indicating a close relationship between Indian and Southern African subtype C viruses.

Of the 18 viruses from Bengaluru characterized in the NFLG or *gag* or *env* genes, only two were of non-subtype C genetic forms: one A1C recombinant (SC017), characterized in the NFLG (Fig 2), and one of subtype B (SC072), characterized in *gag* (Fig 1B). These results are consistent with previous studies reporting subtype B and A1C viruses to be minor components of the Indian HIV-1 epidemic [17,27,28,31–33]. SC017 exhibited a complex mosaic structure, with multiple breakpoints (Fig 2). In a phylogenetic tree, it appeared to be related to another A1C recombinant from India (Figure D in S1 File), although their mosaic structures were clearly different (Figure E in S1 File), implying that SC017 represents a unique recombinant form, as do other A1C recombinants characterized in India, which fail to cluster with each other (Figure D in S1 File).

The extent of genetic mixing and its detailed nature in the context of the widespread distribution of HIV-1 subtype C in India will only become apparent by more systematic and detailed sequencing studies underscoring the utility of deep sequencing approaches to identify changes in the balance of epidemiologically relevant strains and the appearance of novel recombinants. These are most likely to be detected by direct sequence recovery from plasma samples, taken as evidence of productive, systemic infection and recent replication events in an individual where novel variants and recombinant forms will be detectable. Further analysis of the nine whole genomes and historical database isolates suggests limited evolution over time where the virus appears well-adapted to its host. Similar clustering patterns obtained with whole genome or *gag* gene suggests that overall recombination frequencies are also low. However, greater confidence in depicting the emergence of low frequency variants from whole genome recovery, (e.g. SC017, HIV-1 A1C) will be attained with the ability to fully map whole mosaic genomes. These will be further illuminated by refining and developing the types of approach described in this report, and elsewhere, given the urgent need for an effective HIV vaccine which is particularly able to target subtype C infections.

Conclusions

In conclusion, our data further indicate that Indian HIV-1 whole genome sequences converge to form a monophyletic lineage of subtype C, closely related to a southern African lineage, with sporadic cases of A1C recombinant forms exhibiting independent origins. While refinement of amplification protocols, sequencing platforms and bioinformatics tools for analysis of this kind of data are ongoing and likely to supersede the current study, application of techniques and approaches described here will benefit both the clinical management of HIV-1 patients, but will further allow a more precise description of molecular epidemiological trends and direction in specific geographical locales. Understanding complex interactions with coincident infections, such as tuberculosis or viral infections, will be facilitated by these types of approach. Such studies will benefit not only treatment approaches and regimens, but provide clearer data for vaccine design approaches when they become available.

Supporting information

S1 File. A) CD4$^+$ T-cell count and viral load correlation of clinical samples. B) Primer redesigning for amplicon recovery of samples involved in this study. C) SNPs and InDels annotated on consensus C genome for each sample. D) Maximum likelihood phylogenetic tree of NFLG sequences of SC017 and other A1C recombinant viruses from India. E) Bootscan analysis of the NFLG of 95IN21301.
(PDF)

Acknowledgments

This work was funded partly by a EC FP7 EURIPRED (FP7-INFRA-2012 Grant Agreement No. 312661) to AV, MMT, GDS and a Centre for Excellence award from the Department of Biotechnology (DBT), India to AV (BT/01/CEIB/12/III/09). We thank Mark Preston, NIBSC for discussions relating to some aspects of the study.

Author Contributions

Conceptualization: Annapurna Vyakarnam.

Data curation: Shuba Varshini Alampalli.

Formal analysis: Shuba Varshini Alampalli.

Investigation: Neil Berry, Annapurna Vyakarnam.

Methodology: Shuba Varshini Alampalli, Michael M. Thomson, Raghavan Sampathkumar, Karthi Sivaraman, Neil Berry, Annapurna Vyakarnam.

Resources: Anto Jesuraj U. K. J., Chirag Dhar, George D. Souza.

Validation: Shuba Varshini Alampalli.

Writing – original draft: Shuba Varshini Alampalli.

Writing – review & editing: Michael M. Thomson, Neil Berry, Annapurna Vyakarnam.

References

1. Hemelaar J, Gouws E, Ghys PD, Osmanov S. Global trends in molecular epidemiology of HIV-1 during 2000–2007. AIDS 2011; 25: 679–689. https://doi.org/10.1097/QAD.0b013e328342ff93 PMID: 21297424
2. Aralaguppe SG, Siddik AB, Manickam A, Ambikan AT, Kumar MM, Fernandes SJ, et al. Multiplexed next-generation sequencing and de novo assembly to obtain near full-length HIV-1 genome from plasma virus. J Virol Methods 2016; 236: 98–104. https://doi.org/10.1016/j.jviromet.2016.07.010 PMID: 27448822
3. Berg MG, Yamaguchi J, Alessandri-Gradt E, Tell RW, Plantier JC, Brennan CA. A Pan-HIV Strategy for Complete Genome Sequencing. J Clin Microbiol 2016; 54: 868–882. https://doi.org/10.1128/JCM.02479-15 PMID: 26699702
4. Gall A, Ferns B, Morris C, Watson S, Cotten M, Robinson M, et al. Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes. J Clin Microbiol 2012; 50: 3838–3844. https://doi.org/10.1128/JCM.01516-12 PMID: 22993180
5. Andrews S, Krueger F, Seconds-Pichon A, Biggins F, Wingett S. FastQC: a quality control for high throughput sequence data. 2017. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

6. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014; 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

7. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012; 9: 357–359. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

8. Hunt M, Gall A, Ong SH, Brener J, Ferns B, Goulder P, et al. IVA: accurate de novo assembly of RNA virus genomes. Bioinformatics 2015; 31: 2374–2376. https://doi.org/10.1093/bioinformatics/btv120 PMID: 25725497

9. Ruby JG, Bellare P, Derisi JL. PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3: Genes, Genomes, Genetics 2013; 3: 865–880.

10. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

11. Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics 2016; 32: 1749–1751. https://doi.org/10.1093/bioinformatics/btw044 PMID: 26826718

12. HIV Sequence Database 2014. Available: http://www.hiv.lanl.gov/content/sequence/HIV/mainpage. Accessed 29 July 2014.

13. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004; 32: 1792–1797. https://doi.org/10.1093/nar/gkh340 PMID: 15034147

14. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014; 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

15. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 2016; 33: 1870–1874. https://doi.org/10.1093/molbev/msw054 PMID: 27004904

16. Siepel AC, Halpern AL, Macken C, Korber BT. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. AIDS Res Hum Retroviruses 1995; 11: 1413–1416. https://doi.org/10.1089/aid.1995.11.1413 PMID: 8573400

17. Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS, Novak NG, et al. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J Virol 1999; 73: 152–160. PMID: 9847317

18. Tools, P. 2015. By Broad Institute.

19. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics 2011; 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522

20. Thomson MM, Fernández-García A. Phylogenetic structure in African HIV-1 subtype C revealed by selective sequential pruning. Virology 2011; 415: 30–38. https://doi.org/10.1016/j.virol.2011.03.021 PMID: 21507449

21. Wilkinson E, Holzmayer V, Jacobs GB, De Oliveira T, Brennan CA, Hackett J Jr., et al. Sequencing and phylogenetic analysis of near full-length HIV-1 subtypes A, B, G and unique recombinant AC and AD viral strains identified in South Africa. AIDS Res Hum Retroviruses 2015; 31: 412–420. https://doi.org/10.1089/AID.2014.0230 PMID: 25492033

22. Grossmann S, Nowak P, Neogi U. Subtype-independent near full-length HIV-1 genome sequencing and assembly to be used in large molecular epidemiological studies and clinical management. J Int AIDS Soc 2015; 18: 20035. https://doi.org/10.7448/IAS.18.1.20035 PMID: 26115688

23. Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, et al. Population genomics of intrapatient HIV-1 evolution. Elife 2015; 4.

24. Ode H, Matsuda M, Matsuoka K, Hachiya A, Hattori J, Kito Y. Quasispecies Analyses of the HIV-1 Near-full-length Genome With Illumina MiSeq. Front Microbiol 2015; 6: 1258. https://doi.org/10.3389/fmicb.2015.01258 PMID: 26617593

25. Grez M, Dietrich U, Balfe P, von Briesen H, Maniar JK, Mahambre G, et al. Genetic analysis of human immunodeficiency virus type 1 and 2 (HIV-1 and HIV-2) mixed infections in India reveals a recent spread of HIV-1 and HIV-2 from a single ancestor for each of these viruses. J Virol 1994; 68: 2161–2168. PMID: 8139000

26. Shankarappa R, Chatterjee R, Learn GH, Neogi D, Ding M, Roy P, et al. Human immunodeficiency virus type 1 env sequences from Calcutta in eastern India: identification of features that distinguish subtype C sequences in India from other subtype C sequences. J Virol 2001; 75: 10479–10487. https://doi.org/10.1128/JVI.75.21.10479-10487.2001 PMID: 11581417

27. Halani N, Wang B, Ge YC, Gharpure H, Hira S, Saksena NK. Changing epidemiology of HIV type 1 infections in India: evidence of subtype B introduction in Bombay from a common source. AIDS Res Hum Retroviruses 2001; 17: 637–642. https://doi.org/10.1089/088922201300119743 PMID: 11375060

28. Mandal D, Jana S, Bhattacharya SK, Chakrabarti S. HIV type 1 subtypes circulating in eastern and northeastern regions of India. AIDS Res Hum Retroviruses 2002; 18: 1219–1227. https://doi.org/10.1089/08892220260387968 PMID: 12494921

29. Siddappa NB, Dash PK, Mahadevan A, Jayasuryan N, Hu F, Dice B, et al. Identification of subtype C human immunodeficiency virus type 1 by subtype-specific PCR and its use in the characterization of viruses circulating in the southern parts of India. J Clin Microbiol 2004; 42: 2742–2751. https://doi.org/10.1128/JCM.42.6.2742-2751.2004 PMID: 15184461

30. Sengupta S, Khetawat D, Jana S, Sarkar K, Bhattacharya SK, Chakrabarti S. Polymorphism of HIV-1 gag (p17) gene from female sex workers in Calcutta, India. Arch Virol 2005; 150: 2117–2124. https://doi.org/10.1007/s00705-005-0562-5 PMID: 15959835

31. Tripathy SP, Kulkarni SS, Jadhav SD, Agnihotri KD, Jere AJ, Kurle SN, et al. Subtype B and subtype C HIV type 1 recombinants in the northeastern state of Manipur, India. AIDS Res Hum Retroviruses 2005; 21: 152–157. https://doi.org/10.1089/aid.2005.21.152 PMID: 15725754

32. Siddappa NB, Dash PK, Mahadevan A, Desai A, Jayasuryan N, Ravi V, et al. Identification of unique B/C recombinant strains of HIV-1 in the southern state of Karnataka, India. AIDS 2005; 19: 1426–1429. PMID: 16103776

33. Pandey S, Tripathy S, Paranjape R. Molecular characterization of unique intersubtype HIV type 1 A1/C recombinant strain circulating in Pune, India. AIDS Res Hum Retroviruses 2013; 29: 1245–1253. https://doi.org/10.1089/AID.2013.0150 PMID: 23742670

34. Azam M, Malik A, Rizvi M, Singh S, Gupta P, Rai A. Emergence of drug resistance-associated mutations in HIV-1 subtype C protease gene in north India. Virus Genes 2013; 47: 422–428. https://doi.org/10.1007/s11262-013-0961-8 PMID: 23888308

35. Pandey SS, Cherian S, Thakar M, Paranjape RS. Phylogenetic and Molecular Characterization of Six Full-Length HIV-1 Genomes from India Reveals a Monophyletic Lineage of Indian Sub-Subtype A1. AIDS Res Hum Retroviruses 2016; 32: 489–502. https://doi.org/10.1089/AID.2015.0207 PMID: 26756665

36. Rodenburg CM, Li Y, Trask SA, Chen Y, Decker J, Robertson DL, et al. Near full-length clones and reference sequences for subtype C isolates of HIV type 1 from three different continents. AIDS Res Hum Retroviruses 2001; 17: 161–168. https://doi.org/10.1089/08892220150217247 PMID: 11177395

37. Lakhashe S, Tripathy S, Paranjape R, Bhattacharya J. Evidence of a novel B/C recombinant exhibiting unique breakpoints of near full-length HIV type 1 genome from Northeastern India. AIDS Res Hum Retroviruses 2008; 24: 229–234. https://doi.org/10.1089/aid.2007.0229 PMID: 18284322

38. Hanna LE, Neogi U, Ranga U, Swaminathan S, Prasad VR. Phylogenetic characterization of six full-length HIV-1 subtype C molecular clones from three patients: identification of rare subtype C strains containing two NF-kappaB motifs in the long terminal repeat. AIDS Res Hum Retroviruses 2014; 30: 586–591. https://doi.org/10.1089/AID.2013.0275 PMID: 24387762

39. Novitsky VA, Montano MA, McLane MF, Renjifo B, Vannberg F, Foley BT, et al. Molecular cloning and phylogenetic analysis of human immunodeficiency virus type 1 subtype C: a set of 23 full-length clones from Botswana. J Virol 1999; 73: 4427–4432. PMID: 10196340

40. Kulkarni SS, Lapedes A, Tang H, Gnanakaran S, Daniels MG, Zhang M, et al. Highly complex neutralization determinants on a monophyletic lineage of newly transmitted subtype C HIV-1 Env clones from India. Virology 2009; 385: 505–520. https://doi.org/10.1016/j.virol.2008.12.032 PMID: 19167740

41. Shen C, Craigo J, Ding M, Chen Y, Gupta P. Origin and dynamics of HIV-1 subtype C infection in India. PLoS One 2011; 6: e25956. https://doi.org/10.1371/journal.pone.0025956 PMID: 22016790

42. Tee KK, Pybus OG, Li XJ, Han X, Shang H, Kamarulzaman A, et al. Temporal and spatial dynamics of human immunodeficiency virus type 1 circulating recombinant forms 08_BC and 07_BC in Asia. J Virol 2008; 82: 9206–9215. https://doi.org/10.1128/JVI.00399-08 PMID: 18596096

43. Neogi U, Bontell I, Shet A, De Costa A, Gupta S, Diwan V, et al. Molecular epidemiology of HIV-1 subtypes in India: origin and evolutionary history of the predominant subtype C. PLoS One 2012; 7: e39819. https://doi.org/10.1371/journal.pone.0039819 PMID: 22768132

44. Yu XF, Chen J, Shao Y, Beyrer C, Lai S. Two subtypes of HIV-1 among injection-drug users in southern China. Lancet 1998; 351: 1250. https://doi.org/10.1016/S0140-6736(05)79316-8 PMID: 9643749

45. Oelrichs RB, Shrestha IL, Anderson DA, Deacon NJ. The explosive human immunodeficiency virus type 1 epidemic among injecting drug users of Kathmandu, Nepal, is caused by a subtype C virus of restricted genetic diversity. J Virol 2000; 74: 1149–1157. PMID: 10627525

46. Bhusal N, Sutthent R, Horthongkham N, Athipanyasilp N, Kantakamalakul W. Prevalence of HIV-1 Subtypes and Antiretroviral Drug Resistance Mutations in Nepal. Curr HIV Res 2016; 14: 517–524. PMID: 27697032

47. Bontell I, Sarker MS, Rahman M, Afrad MH, Sonnerborg A, Azim T. Molecular dating of HIV-1 subtype C from Bangladesh. PLoS One 2013; 8: e79193. https://doi.org/10.1371/journal.pone.0079193 PMID: 24223905