**Appendix**

Words count: 1367

1. **SUPPLEMENTARY METHODS**

**1.1 PHYLOGENETIC ANALYSIS**

All molecular phylogenetic analyses were performed using Molecular Evolutionary Genetics Analysis version 6 (MEGA6) software (reference n°13 in the main text).

**Maximum likelihood phylogenetic analysis**

**HIV**

HIV RNA amplifications were done according to nested RT-PCR procedures established by the Agence Nationale de Recherches sur le SIDA et les hépatites virales (ANRS) (see Supplementary Table 1). The evolutionary history was inferred on HIV-1 C2V3 *env* and concatenated PR-RT sequences by using the Maximum Likelihood (ML) method based on the GTR+Γ+I (General Time Reversible model with Gamma distributed among-site rate heterogeneity, plus Invariant sites) model of nucleotide substitution recommended by the Find Best DNA/Protein models program inserted in the MEGA6 software. In addition to newly generated 202 HIV-1 *env* C2V3 and 192 concatenated PR-RT sequences (see Supplementary Figure 1) obtained from HIV-1-infected subjects during the Roka outbreak, we used in the alignment two distinct HIV sequences sets. First, a total of 88 reference sequences were selected from the Los Alamos HIV sequences database (<http://www.hiv.lanl.gov>). These sequences included three subtypes B from Thailand, 19 circulating recombinant forms (CRF) 01\_AE from Thailand, and 66 CRF sequences deriving from CRF01\_AE as follows: CRF15\_01B from Thailand (5 sequences), CRF22\_01A1 from Cameroon (11), CRF33\_01B from Malaysia (5), CRF34\_01B from Thailand (3), CRF48\_01B from Malaysia (3), CRF51\_01B from Malaysia and Singapore (4), CRF52\_01B from Thailand and Malaysia (3), CRF53\_01B from Malaysia (4), CRF54\_01B from Malaysia (3), CRF55\_01B from China (5), CRF58\_01B from Malaysia (6), CRF59\_01B from China (6), CRF67\_01B from China (2), CRF68\_01B from China (3), CRF69\_01B from Japan (1), CRF74\_01B from Malaysia (1), and CRF78\_cpx from China (1) (Supplementary Table 2). Second, we also used samples from HIV-1-infected subjects living in different geographical provinces in Cambodia. The sample sizes for these sequences were 45 and 24 for HIV-1 *env* C2V3 and concatenated PR-RT sequences, respectively. HIV-1 *env* C2V3 and concatenated PR-RT nucleotide sequences obtained from the Roka outbreak were submitted to GenBank and were assigned accession numbers KY570019–KY570220 and KY569803–KY569994, respectively. Accession numbers for sequences from other Cambodian geographical locations were KY570221–KY570265 and KY569995–KY570018, respectively.

**HCV**

Preliminary phylogenetic analysis of NS5B HCV sequences obtained from the Roka outbreak was conducted with 161 reference sequences of all genotypes selected from the GenBank database. This initial analysis revealed that the Roka outbreak was associated with two genotypes of HCV, including 1b and 6e genotypes. Consequently, the ML trees of HCV genotype 1b and 6e sequences were separately constructed. For the phylogenetic analysis of the genotype 1b cluster, 29 reference sequences (Supplementary Table 3) were aligned with the outbreak sequences. For the genotype 6e tree, we used two subtype 6e sequences, as well as other 27 subtypes 6 (Supplementary Table 4). In addition, we used HCV NS5B sequences that were routinely amplified and sequenced from specimens collected among Cambodian HCV-infected outpatients (37 sequences for HCV 1b alignment, and 18 sequences for HCV 6e alignment). The evolutionary history of HCV genotype 1b and 6e were inferred using respectively the HKY+Γ (Hasegawa Kino Yano model with gamma distributed among-site rate heterogeneity) [1], and K2+Γ+I (Kimura 2-parameter model with gamma distributed among-site rate heterogeneity, plus invariant sites) model of nucleotide substitution [2], as recommended by the Find Best DNA/Protein models program inserted in the MEGA software. HCV nucleotide sequences obtained from the Roka outbreak were submitted to GenBank under the accession numbers KY569650–KY569688 for genotype 1b, and KY569734–KY569784 for genotype 6e. Accession numbers for sequences from other geographical locations were KY569689–KY569725 for genotype 1b, and KY569785–KY569802 for genotype 6e.

**HBV**

Maximum likelihood phylogenies for S-gene HBV were performed using the K2+Γ+I model of nucleotide substitution, as recommended by the Find Best DNA/Protein models program inserted in the MEGA software. The outbreak sequences were aligned with 8 sequences obtained from Cambodian HBV-infected patients and 33 reference sequences retrieved from GenBank database. HBV nucleotide sequences obtained from the Roka outbreak and from other Cambodian geographical locations were submitted to GenBank under the accession numbers MF133378–MF133393, and MF133394–MF133401, respectively.

**Genetic distance**

Using MEGA, genetic distances between Roka HIV *env* C2V3 sequences, and between Roka HCV NS5Bgenotype 1b, and 6e sequences were calculated with GTR+Γ+I, TN93+G (Tamura-Nei model with gamma distributed among-site rate heterogeneity) and K2+Γ models, respectively (see reference n°14 in the main text).

**1.2 EVOLUTIONARY AND DEMOGRAPHIC RECONSTRUCTIONS**

**Clock prior elicitation**

The earliest samples from the Roka HIV outbreak were obtained on December 1, 2014 and the latest samples were obtained on January 28, 2015. Because of this narrow sampling window we lacked sufficient temporal spread in the data to precisely estimate the evolutionary rate. This precision was necessary to accurately infer the time to the most recent common ancestor (tMRCA) of the outbreak. To improve our inference of the evolutionary rate, we used an informative prior on the molecular clock for HIV *env* C2V3 and HCV NS5B sequences. This prior was derived via Bayesian coalescent analysis in BEAST 1.8.3 (see reference n°15 in the main text). We used the posterior mean and standard deviation (SD) obtained from this analysis as prior for the molecular clock of the Roka sequences.

**HIV**

Thirty HIV *env* C2V3 sequences unrelated to the Roka outbreak but sampled in Cambodia between 1997 and 2014 were included for the clock prior elicitation (accession numbers, KY570225-KY570232, KY570235, KY570238-KY570239, and KY570266-KY570284). This prior was a normal distribution with a mean of 4.1x10-3 substitutions per site per year (95% Highest Probability Density [HPD] confidence limits, 2.4x10-3 to 6.0x10-3) and a SD of 9.30x10-4 substitutions per site per year.

**HCV**

The HCV evolution rate was estimated at 2.4x10-3 substitutions per site per year (95% HPD confidence limits, 0.7x10-3 to 4.3x10-3) with a SD of 9.21x10-4 substitutions per site per year using 34 sequences collected between 2002 and 2015 from Cambodian HCV-infected patients (accession numbers, KY569689-KY569690, KY569692, KY569697-KY569700, KY569704-KY569706, KY569709-KY569714, and KY569716-KY569733). Our estimated substitution rate was similar to value that was previously described for the HCV NS5B region by others.[3]

**Coalescent analysis and skyline reconstruction**

**HIV**

We inferred posterior distributions of coalescent trees for the 198 Roka HIV *env* C2V3 sequences of the Roka cluster with the use of BEAST 1.8.3. We modeled the evolutionary process with an uncorrelated relaxed lognormal molecular clock and a GTR+Γ4+I nucleotide substitution model, as reported by Drummond et al [4]. The tMRCA and dynamics of the effective population size (*Ne*) over time were initially estimated by a non-parametric Bayesian skyline plot model (BSP) [5], as coalescent tree prior. Parametric estimates of growth rates were obtained under two demographic models: expansion and exponential growth [6, 7], in which adjustment to the data were assessed using the log marginal likelihood estimation (MLE) based on path sampling (PS) and stepping-stone sampling (SS) strategies.[8] We decided to present the results of the expansion model in the main text because Bayes Factors (BF) were <3 for PS and SS, when comparing expansion and exponential models. We ran Markov chain Monte Carlo (MCMC) for 100 million generations considering 10% of states as burn-in. We assessed MCMC convergence by looking at trace plots. All parameter estimates were taken from runs with ESS values of 200 or higher. We reconstructed viral population size over time and lineage through time (LTT) plot using demographic reconstruction and LTT analyses implemented in Tracer 1.6.0 [9], and exported the data to GraphPad Prism version 5.01 for Windows (GraphPad Software, san Diego, California, USA). Maximum Clade Credibility (MCC) trees were summarized using TreeAnnotator v1.6 and visualized using FigTree v1.4.3.

**HCV**

The tMRCAs were separately estimated for each HCV cluster that were identified in the Roka outbreak using BEAST 1.8.3. Analyses were conducted under the HKY for HCV 1b and HKY+I models of nucleotide substitution for HCV 6e [1]. Relaxed uncorrelated lognormal molecular clock models, combined with expansion and exponential growth models, were used as coalescent tree prior [6, 7]. Data adjustment was assessed using the log MLE based on PS and SS strategies [8]. For HCV 1b, results were similar between the two models (BF < 3). In contrast, for HCV 6e strains, the expansion growth model fitted better than the exponential growth model (BF>3). Thus, we decided to select the expansion model for the estimation of HCV 1b and 6e tMRCAs.

**2. SUPPLEMENTARY REFERENCES**

1. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol**1985**; 22:160-74.

2. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol**1980**; 16:111-20.

3. Gray RR, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus OG. The mode and tempo of hepatitis C virus evolution within and among hosts. BMC Evol Biol **2011**; 11:131.

4. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol**2006**; 4:e88.

5. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol**2005**; 22:1185-92.

6. Griffiths RC, Tavare S. Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci**1994**; 344:403-10.

7. Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics**2002**; 161:1307-20.

8. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol Biol Evol**2012**; 29:2157-67.

9. Rambaut A, Suchard MA, Xie D, Drummond AJ. Tracer 1.6. *Available at:* <http://beast.bio.ed.ac.uk/Tracer>2014.

**3. SUPPLEMENTARY TABLES**

**Supplementary Table 1.** **Details of HIV-1 primers for each subgenomic region**

The design of the primers was done by the ANRS coordinated action N°11 Resistance Study Group.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Viral region | PCR round | Primer name | Primer positions | Primer sequence |
| C2V3 ENV | FirstSecondFirstSecond | Env31Env8Env7ED33Env341Env342Env343Env344 | 6955-69737522-75407008-70267360-73816863-68867757-77776874-68947522-7540 | 5’-CAGTACAATGTACACATGG-3’5’-ATGGGAGGGGCATACATTG-3’5’-AATGGCAGTCTAGCAGAAG-3’5’-TTACAGTAGAAAAATTCCCCTC-3’5’-TCCTATACATTATTGTRCTCCAGC-3’5’-ATAGCTCCTATTCCCACTGC-3’5’-ATTGTACTCCAGCTGGTTATG-3’5’-ATGGGAGGAGCATACATTG-3’ |
| RT | FirstSecondSequencing | MJ3MJ4A35NE135A20NE120 | 2480-25013399-34202530-25643300-33342545-25643315-3334 | 5’-AGTAGGACCTACACCTGTCA-3’5’-CTGTTAGTGCTTTGGTTCCTCT-3’5’-TTGGTTGCACTTTAAATTTTCCCATTAGTCCTATT-3’5’-CCTACTAACTTCTGTATGTCATTGACAGTCCAGCT-3’ 5’-ATTTTCCCATTAGTCCTATT-3’5’-ATGTCATTGACAGTCCAGCT-3’ |
| PR | FirstSecond | 5’PROT13’PROT15’PROT23’PROT2 | 2082-21082703-27342136-21632621-2650 | 5’-TAATTTTTTAGGGAAGATCTGGCCTTCC-3’5’-GCAAATACTGGAGTATTGTATGGATTTTCAGG-3’5’-TCAGAGCAGACCAGAGCCAACAGCCCCA-3’5’-AATGCTTTTATTTTTTCTTCTGTCAATGGC-3’ |

Inner primers were further used for direct sequencing for each studied subgenomic region except RT region.

**Supplementary Table 2. HIV-1 reference sequence dataset used in phylogenetic analyses**

|  |  |  |  |
| --- | --- | --- | --- |
| **HIV-1 subtype** | **Country** | **No.** | **GenBank accession number** |
| B | Thailand | 3 | JX446796, JX446805, and JX447156 |
| CRF01\_AE | Thailand | 19 | JX447018, JX447543, JX446710, JX448027, JX446666, JX447891, JX447725, JN860767, JX448031, JX447734, JX448067, AY358046, AY358047, AY358066, AF259954, AF259955, AY358048, AY358039, and AB220945 |
| CRF15\_01B | Thailand | 5 | DQ354120, AF516184, AF530576, AF529572, and AF529573 |
| CRF22\_01A1 | Cameroon | 11 | AY371159, GQ229529, JN864049, KF716460, KF716461, KF716462, KF716463, JN864050, JN864051, JN864058, and JN864059 |
| CRF33\_01B | Malaysia | 5 | DQ366659, DQ366660, DQ366661, DQ366662, and GQ277610 |
| CRF34\_01B | Thailand | 3 | EF165539, EF165540, and EF165541 |
| CRF48\_01B | Malaysia | 3 | GQ175881, GQ175882, and GQ175883 |
| CRF51\_01B | Malaysia | 2 | KJ485698 and KJ485697 |
| CRF51\_01B | Singapore | 2 | JN029801 and JN029803 |
| CRF52\_01B | Thailand | 2 | DQ354113 and AY945734 |
| CRF52\_01B | Malaysia | 1 | DQ366664 |
| CRF53\_01B | Malaysia | 4 | DQ366663, JX390612, JX390611, and JX390610 |
| CRF54\_01B | Malaysia | 3 | EU031915, JX390977, and JX390976 |
| CRF55\_01B | China | 5 | JX574661, KF927151, KF927150, JX574663, and JX574662 |
| CRF58\_01B | Malaysia | 6 | KC522031, KC522035, KF425293, KC522034, KC522033, and KC522032 |
| CRF59\_01B | China | 6 | JX960635, KC462191, KJ484433, KJ484434, KC462190, and KJ484435 |
| CRF67\_01B | China | 2 | KC183779 and KC183780 |
| CRF68\_01B | China | 3 | KC183782, KC183783, and KF758551 |
| CRF69\_01B | Japan | 1 | LC027100 |
| CRF74\_01B | Malaysia | 1 | KR019771 |
| CRF78\_cpx | China | 1 | KU161145 |

**Supplementary Table 3. HCV reference sequence dataset used in phylogenetic analysis of the HCV genotype 1b outbreak-associated clusters**

|  |  |  |  |
| --- | --- | --- | --- |
| **HCV Genotype** | **Country** | **No.** | **GenBank accession number** |
| 1a | US | 1 | NC\_004102 |
| 1a | Viet Nam | 2 | AB523220; DQ155535 |
| 1b | China | 3 | AY587016; EU155364; GU451220 |
| 1b | Hong Kong | 1 | HM009071 |
| 1b | Japan | 3 | D89815; AB429050; M58335 |
| 1b | Laos | 1 | HE580375 |
| 1b | Thailand | 1 | KP323875 |
| 1b | Taiwan | 1 | AM494937 |
| 1b | Viet Nam | 1 | KJ470251 |
| 1b/2b | Japan | 1 | AB622121 |
| 1b/2k | Russia | 1 | AY587845 |
| 1c | China | 1 | KC844047 |
| 1e | United Kingdom | 1 | KC248194 |
| 1g | Spain | 1 | AM910652 |
| 1h | Cameroon | 1 | KC248199 |
| 1l | Africa | 1 | KC248193 |
| 2a | China | 1 | KF676351 |
| 2n/1b | Philippines | 1 | DQ364460 |
| 3a | China | 1 | HQ912953 |
| 4a | Japan | 1 | AB795432 |
| 5a | China | 1 | KC844046 |
| 6a | China | 1 | KC844038 |
| 6b | Thailand | 1 | D84262 |
| 6e | Viet Nam | 1 | EU246940 |

**Supplementary Table 4. HCV reference sequence dataset used in phylogenetic analysis of the HCV genotype 6e outbreak-associated clusters**

|  |  |  |  |
| --- | --- | --- | --- |
| **HCV genotype** | **Country** | **No.** | **GenBank accession number** |
| 1a | US | 1 | NC004102 |
| 2a | China | 1 | KF676351 |
| 3a | China | 1 | HQ912953 |
| 4a | Japan | 1 | AB795432 |
| 5a | China | 1 | KC844046 |
| 6a | China | 1 | KC844038 |
| 6b | Thailand | 1 | D84262 |
| 6c | Thailand | 1 | KP324209 |
| 6d | Viet Nam | 1 | D84263 |
| 6e | Viet Nam | 1 | EU246940 |
| 6e | China | 1 | DQ314805 |
| 6f | Thailand | 1 | DQ835760 |
| 6g | Hong Kong | 1 | DQ314806 |
| 6h | Hong Kong | 1 | HM009308 |
| 6i | Thailand | 1 | DQ835770 |
| 6j | Thailand | 1 | DQ835769 |
| 6k | China | 1 | AY878651 |
| 6l | US | 1 | EF424628 |
| 6m | Thailand | 1 | DQ835767 |
| 6n | Thailand | 1 | DQ835768 |
| 6o | Canada | 1 | EF424627 |
| 6p | Canada | 1 | EF424626 |
| 6q | Canada | 1 | EF116191 |
| 6r | Viet Nam | 1 | KM252786 |
| 6s | Canada | 1 | EF116177 |
| 6t | Viet Nam | 1 | EF632071 |
| 6u | China | 1 | EU408330 |
| 6v | China | 1 | EU158186 |
| 6w | China | 1 | DQ278892 |

**Supplementary Figure 1. Maximum-likelihood phylogenies of concatenated HIV-1 PR-RT sequences – Roka, Cambodia, 2014–2015**

Phylogeny branches and sequence names are colored as follows: 192 Roka sequences (red), 24 local reference sequences (blue), and 88 GenBank reference sequences (black). Numbers on branches indicate bootstrap support values based on 1,000 replicates. The phylogenetic analysis identified 189 sequences that formed a cluster (red-filled triangle). No Cambodian or GenBank reference sequence was closely related to the Roka HIV-1 cluster. The mean nucleotide identity ± SD was 99.7%±0.2 for the PR-RT cluster. Three infections sampled from Roka did not group with the outbreak-associated samples. These were NCHADS171, NCHADS184, and NCHADS185, the same as obtained with the HIV C2V3 *env* phylogeny. There were no PR-RT sequence available for NCHADS116. The branch lengths represent the number of substitutions per site, as indicated by the scale.