

Phylogenetic Analysis of HIV-1 Shows Frequent Cross-Country Transmission and Local Population Expansions

Marc Bennedbæk, Anna Zhukova, Man-Hung Eric Tang, Jaclyn Bennet, Paula Munderi, Kiat Ruxrungtham, Magnus Gisslen, Michael Worobey, Jens Lundgren, Rasmus Marvig

▶ **To cite this version:**

Marc Bennedbæk, Anna Zhukova, Man-Hung Eric Tang, Jaclyn Bennet, Paula Munderi, et al.. Phylogenetic Analysis of HIV-1 Shows Frequent Cross-Country Transmission and Local Population Expansions. Virus Evolution, 2021, 7 (2), pp.veab055. 10.1093/ve/veab055 . pasteur-03266546

HAL Id: pasteur-03266546
https://pasteur.hal.science/pasteur-03266546

Submitted on 21 Jun 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Phylogenetic analysis of HIV-1 shows frequent cross-country transmission and local population expansions

Authors

Marc Bennedbæk[1,a], Anna Zhukova[2], Man-Hung Eric Tang[1,b], Jaclyn Bennet[3], Paula Munderi[4], Kiat Ruxrungtham[5], Magnus Gisslen[6,7], Michael Worobey[8], Jens D. Lundgren[1], Rasmus L. Marvig[9,a], for the INSIGHT START study group.

Author affiliations

[1] Centre of Excellence for Health, Immunity and Infection (CHIP), Department of Infectious Diseases, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark.

[2] Unité Bioinformatique Evolutive, Hub Bioinformatique et Biostatistique, USR3756 (C3BI//DBC), Institut Pasteur & CNRS, Paris, France.

[3] Clinical HIV Research Unit, Department of Internal Medicine, School of Clinical Medicine, Faculty of Health Sciences, University of Witwatersrand, Johannesburg, 2092 South Africa.

[4] MRC Uganda Research Unit on AIDS, Entebbe, Uganda

[5] HIV-NAT, Thai Red Cross AIDS Research Center, Department of medicine, Chulalongkorn University, Bangkok, Thailand.

[6] Department of Infectious Diseases, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden.

[7] Region Västra Götaland, Sahlgrenska University Hospital, Department of Infectious Diseases, Gothenburg, Sweden.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

© The Author(s) 2021. Published by Oxford University Press.

[8] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721, USA.

[9] Department of Genomic Medicine, Rigshospitalet, Copenhagen, Denmark.

Notes and correspondence

[a] Correspondence should be addressed to Marc Bennedbæk (marc.bennedbaek@regionh.dk) and Rasmus L. Marvig (rasmus.lykke.marvig@regionh.dk).

[b] Present address: Nordic Bioscience, ProScion, Herlev, Denmark.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Abstract

Understanding of pandemics depends on characterization of pathogen collections from well-defined and demographically diverse cohorts. Since its emergence in Congo almost a century ago, HIV-1 has geographically spread and genetically diversified into distinct viral subtypes. Phylogenetic analysis can be used to reconstruct the ancestry of the virus to inform on the origin and distribution of subtypes.

We sequenced two 3.6 kb amplicons of HIV-1 genomes from 3,197 participants in a clinical trial with consistent and uniform sampling at sites across 35 countries and analyzed our data with another 2,632 genomes that comprehensively reflects the HIV-1 genetic diversity. We used maximum likelihood phylogenetic analysis coupled with geographical information to infer the state of ancestors.

The majority of our sequenced genomes (n=2,501) were either pure subtypes (A-D, F, G) or CRF01_AE. The diversity and distribution of subtypes across geographical regions differed; United States showed the most homogenous subtype population, whereas African samples were most diverse. We delineated transmission of the four most prevalent subtypes in our dataset (A, B, C, and CRF01_AE), and our results suggest both continuous and frequent transmission of HIV-1 over country borders, as well as single transmission events being the seed of endemic population expansions.

Overall, we show that coupling of genetic and geographical information of HIV-1 can be used to understand origin and spread of pandemic pathogens.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Introduction

A high mutation rate, short generation time, lack of proofreading, and frequent recombination fuels HIV-1 molecular evolution with genetic variation (Castro-Nallar et al. 2012; Volz et al. 2013). This has resulted in diversification of HIV-1 group M into nine distinct subtypes (A-D, F-H, J, and K). Recombination events between multiple subtypes have also resulted in new circulating recombinant forms (CRFs) of the virus. The prevalence of the different subtypes and CRFs differ markedly, with some more prevalent than others in specific geographic regions (Hemelaar et al. 2011; Hemelaar et al. 2019). The divergence and expansion of the different subtypes of HIV-1 is elusive, as they are likely to have happened before the characterization of HIV-1 in 1983 (Gilbert et al. 2007).

Phylogenetics can be applied to infer the relatedness between HIV-1 sample genotypes, and geographical information contained in the genetic relationship can be used to infer the states of ancestral genotypes (Castro-Nallar et al. 2012). However, the retrospective nature of phylogenetics makes the approach dependent on the comprehensiveness of the sample collection, i.e. samples representing all genotypic states should ideally be included in the collection. In practice it is not possible to sample all genotypic states, and especially it is difficult to include samples of ancestral genotypes only available from historical sample collections. Nonetheless, a few available historical collections of HIV-1 covering decades have been successfully sequenced and used to date origins of types of HIV-1 (Bletsa et al. 2019; Gryseels et al. 2020; Worobey et al. 2016). Beyond the geographical information contained in phylogenies, they can be analyzed with respect to other sample information, e.g. temporal, epidemiological and demographic information, to further understand the observed genetic diversification and distribution of HIV-1 (Castro-Nallar et al. 2012).

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Here, we sequenced HIV-1 genomes sampled from participants in the Strategic Timing of AntiRetroviral Treatment (START) trial (Group et al. 2015), which constitute a well-defined and demographically diverse cohort of ART naïve HIV+ participants, and analyzed them together with a set of HIV-1 genomes defined by Los Alamos National Laboratory (LANL) to comprehensively reflect HIV-1 genetic diversity. With phylogenetic ancestry reconstruction, we used the sample collection to delineate and characterize transmission of subtypes across countries.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Methods

Description of sample collection from START clinical trial

All participants are from the START trial, which is managed by the International Network for Strategic Initiatives in Global HIV Trials (INSIGHT). The study design and data collection plan for START has previously been reported (Group et al. 2015). A total of 4,685 antiretroviral therapy (ART) naive participants from 35 countries (6 geographical regions) were included in the START study. The participants were enrolled between April 2009 and December 2013. Here, 3,785 participants with a viral load ≥1,000 cp/mL and with at least two baseline plasma samples available in the START biobank were included for HIV-1 genome sequencing.

Ethics

Samples included in this study were derived from participants who consented in the clinical trial, START (NCT00867048) (Group et al. 2015), run by INSIGHT. The study was approved by the institutional review board or ethics committee at each contributing center, and written informed consent was obtained from all participants. All informed consents were reviewed and approved by participant site ethics review committees.

Viral concentration and RNA extraction

Plasma samples were thawed at room temperature. Following this, 500µL plasma was transferred to new RNAse-free tubes and centrifuged at 2000x*g* for 15 minutes. The supernatant was then extracted and centrifuged at 21.000x*g* for 75 minutes and 360 µL of the top supernatant was discarded. Viral RNA was extracted using QIAamp viral RNA extraction kit (Qiagen) on a QIAcube robot according to the manufacturer's guidelines.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Reverse transcription and amplification of viral RNA

Reverse transcription polymerase chain reaction (RT-PCR) was used to amplify two amplicons of HIV-1 separately, covering positions 1485-5058 and 5967-9517 (later referred to as amplicons A and B) in the HXB2 genome sequence (GenBank accession number K03455), as described previously (primer sequences are listed in Supplementary Table 1) (Gall et al. 2012). The reverse-transcription and amplification were performed using SuperScript III One-Step RT-PCR System with Platinum Taq High Fidelity (Thermo Fisher Scientific) according to manufacturer's instructions with 10 µL viral RNA for each amplicon (Thermocycler settings are listed in SI, Table S7). The PCR products were purified using Ampure XP (Beckman Coulter) PCR purification according to the manufacturer's instructions. The two amplicons were pooled for each sample prior to DNA library preparation.

Amplicon DNA library preparation and sequencing

Libraries of DNA from pooled amplicons were prepared using a Nextera XT (Illumina, San Diego, CA, USA) sample preparation kit to target an insert size of 300-nt. A modified protocol was used, in which input DNA and reagent use was halved, except for normalization of libraries, where 1.5X magnetic normalization beads were used. DNA libraries were sequenced on an Illumina MiSeq machine using a MiSeq 150-cycle V3 reagent kit (Illumina), producing 75 nucleotide (nt) paired-end reads.

Removal of sequence readouts from human genome

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Sequencing reads from each sample were aligned against the human reference genome sequence Genome Reference Consortium Human Build 37 patch release 13 (GRCh37.p13) using Bowtie2 version 2.2.8 (Langmead and Salzberg 2012) (option "-X 1000") and only read pairs in which neither of the reads aligned to GRCh37.p13, were retained.

Alignment of reads against HXB2 reference to create consensus sequence

Cleaned reads from the previous step were aligned against the genome sequence of HIV-1 HXB2 (GenBank accession number K03455.1) with SSAHA2 version 2.5.3 (Ning et al. 2001) (option "-solexa" and "-pair 1,1000") and resulting sequence alignment map (SAM) files were converted to sorted binary alignment map (BAM) files using SAMtools 1.1 (Li et al. 2009). Insert size of mapped reads was on average 173 nt (interquartile range 123-216 nt) as determined with command "samtools stats [sorted BAM file]". Consensus sequences of reads aligned within regions targeted by amplicons A and B were generated from sorted BAM files with the command "samtools mpileup -d 1000000 -uf [HXB2 reference sequence] [sorted BAM file] | bcftools call -c | vcfutils.pl vcf2fq". Sample consensus sequences were trimmed to only contain the genomic position of the two amplicons, i.e. HXB2 reference genome positions 1485-5058 (amplicon length 3,574 nt) and 5967-9517 (amplicon length 3,551 nt). All samples (n=3,197) for which a consensus sequence was determined for either 90% of one of the amplicon regions or 50% of both amplicon regions were included for further subtype and phylogenetic analysis.

HIV-1 subtyping

The consensus sequence of each sample was analyzed with REGA HIV-1 Subtyping Tool version 3 (Pineda-Pena et al. 2013). The output was manually inspected to check for the presence

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

of subtype specific sequences within the given consensus sequence. Shannon-Weaver and 1 – Simpson diversity indexes were calculated using the vegan package in R (version 3.2.5). The diversity indexes were calculated based on the distribution of subtypes A, AB, B, BC, BF, C, and CRF01_AE across geographical regions.

Genome sequences from Los Alamos HIV Sequence Database

We downloaded publicly available HIV-1 genome sequences from Los Alamos HIV Sequence Database (LANL). We downloaded all genome sequences in LANL Filtered Web Alignment (https://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html) of subtypes A (A1-A6), B, C, D, F (F1,F2), G and CRF01_AE. The sequences were downloaded on 2020/11/23. The sequences were trimmed to match the START amplicons using MAFFT version 7.453 with the following command:

mafft --thread 26 --memsave --retree 1 --maxiterate 0 --add <LANL_filtered_web_aligment_fasta-file-with-sequences> --keeplength <K03455_subset_to_amplicon_A_and_B >

<LANL_filtered_web_alignment_aligned_and_subset_to_amplicon_A_and_B >. The aligned sequences were combined with the START sequences.

Maximum likelihood phylogenetic reconstruction

A maximum likelihood tree was constructed using ExaML (Exascale Maximum Likelihood) version 3.0.16 that uses the RAxML search algorithm (Kozlov et al. 2015). The sequences of pure subtypes A, B, C, D, F and G, as well as subtype CRF01_AE, were used for tree construction. Recombination hampers phylogenetic analysis, so we focused our analysis of

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

ancestry on subtypes A, B, C, and CRF01_AE, respectively (Felsenstein 2004). We used a generalized time reversible model (GTR) with a gamma distribution, and the analysis was parallelized using openmpi. The following three commands were used to produce the trees: (1) "raxmlHPC-AVX -y -m GTRGAMMA -p 12345 [sample consensus sequences in phylip format] -n StartingTree"; (2) "examl-OMP -s maximum_likelihood_tree.unpartitioned.binary -t RAxML_parsimonyTree.StartingTree -m GAMMA -n Tree1"; and (3) "examl-AVX -s maximum_likelihood_tree.unpartitioned.binary -t TreeSet -f E -m GAMMA -n T3uE". The maximum likelihood trees were visualized using the Interactive Tree of Life (iTol) online tool (Figure 1) (Letunic and Bork 2019).

LSD was used to extract monophyletic clades containing subtypes A, B, C, and CRF01_AE, respectively, and produce separate subtype specific rooted trees (To et al. 2016).

Reconstruction of ancestral states

Ancestral states were reconstructed using PastML 1.9.7 (Ishikawa et al. 2019). We used the rooted subtype specific trees produced by LSD as input for PastML. The tips in the subtype specific trees were annotated to country. Ancestral state reconstruction and visualization in PastML were run with the following command: "pastml --tree (International et al.) --data [metadata] --columns [metadata column] --name_column [metadata column] --tip_size_threshold 15 --html_compressed [compressed_tree] --html [uncompressed_tree] --prediction_method MPPA --model F81". We used the PastML utility script 'calculate_changes.py' (https://github.com/evolbioinfo/pastml/blob/master/pastml/utilities/calculate_changes.py) to count the inferred events of transmission between countries with the following command:

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

calculate_changes.py --tree (International et al.) --acr [file with combined ancestral states output from pastml] --columns Country --out_log [log file].

Ancestral states were reconstructed with PastML using an F81-like model, which generalizes states to the 4-state F81 model for nucleotide substitution (Felsenstein 1981). Under F81-like model, migration rate from a state i (e.g. location) to a different state j (i != j) is proportional to the equilibrium frequency of j, πj. In addition to the state equilibrium frequencies, PastML optimizes the rescaling factor (analogous to mutation rate under strict molecular clock), which is applied to all the tree branches and represents the average number of character changes per branch unit (e.g. year for dated trees).

We assessed the robustness of the ancestral state reconstruction by shuffling tip state annotations prior to PastML analysis. The random shuffling analysis was repeated three times for each subtype.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Results

Genotyping of HIV-1 from the START clinical trial

We obtained HIV+ plasma samples from 3,785 START trial participants with a viral load ≥1,000 copies/mL. For each sample, we sequenced two 3.6 kb amplicons of the HIV-1 genome that cover partial *gag* (54%) and *pol* (99%), and full the *env* gene, respectively. While we achieved an average sequencing coverage of the two HIV-1 genomic regions of 7,188-fold (median 5,334-fold) and 1,061-fold (median 303-fold), respectively, sequencing success varied both across samples and genomic regions. For further analysis, we only included those 3,197 samples for which sequence information was available for at least 90% of either of the genomic regions, or at least 50% of both genomic regions.

We analyzed each of the 3,197 genomes with REGA HIV-1 Subtyping Tool (Pineda-Pena et al.) to identify subtype specific sequences, and we assigned the samples to be either a pure subtype (n=2,354 samples distributed on subtypes A-D, F, and G), or a recombinant subtype, in cases where the genome showed presence of sequences specific to >1 pure subtypes (n=843 samples). We did not detail the recombinants down to specific CRF references, except for CRF01_AE (n=147). For example, if a genome showed to contain both sequence specific to C and D subtype, respectively, it was denoted as subtype 'CD' without further detailing of genomic mosaic structure defined by recombinant breakpoints.

Combined phylogenetic analysis of data from START and LANL

We combined our START sequences of pure subtypes (A, B, C, D, F, and G) and CRF01_AE (n=2,501 sequences) with all sequences of the same subtypes from the LANL Filtered Web

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Alignment (n=2,632) that consist of sequences that represent the fullest spectrum (diversity) of sequences in LANL (Table 1).

The LANL Filtered Web Alignment sequences spanned 72 countries (8 samples have no country information) of which 24 countries overlapped with the 35 countries in START (Supplementary Table 2-3). Sequences from START were sampled between April 2009 and December 2013 (Supplementary Figure 1). The majority (94%) of sequences in LANL Filtered Web Alignment were from samples after 1995, albeit the earliest samples dated back to 1979 (Supplementary Figure 1).

We constructed a maximum likelihood phylogeny of all samples in the combined dataset consisting of subtypes A, B, C, D, F, G, and CRF01_AE (n=5,133; Figure 1). The phylogeny confirmed that all samples of each subtype clustered as monophyletic groups, i.e. there was concordance between the assigned subtypes and the evolutionary relationship of samples. Subtypes A, B, C, and CRF01_AE were the most frequent subtypes in both the START and the LANL dataset, and our further analysis was focused on extracted monophyletic clades for each of these four subtypes (i.e. a rooted phylogenetic tree for each subtype; Supplementary Figure 2-5).

Origin and cross-country transmission of subtype B

Subtype B samples from LANL Filtered Web Alignment (n=1,196) encompassed 43 countries of which 15 countries overlapped with the 32 countries with subtype B samples from START (n=1,959; Supplementary Table 2-3).

We annotated the tips (samples) of the subtype B phylogenetic tree according to country of origin to infer the ancestral states (Figure 2). We found that our collection of subtype B samples

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

emerged in the United States (marginal probability 89%). To assess if our ancestral state estimates were a result of genuine structure in the data or due to collection skewness (subtype B samples were in general most often observed in United States (25%)).

We repeated the ancestral reconstruction with geographical annotation randomly shuffled between samples (Firth et al. 2010). Random shuffling of country annotation led to the estimate of ancestral state being unknown (Supplementary Figure 6). As such, the random shuffling of annotation supported the estimate that the ancestral state was in United States and was a result of genuine structure in the data.

We used ancestral state reconstruction to infer events of transmission between countries to explain the observed country of origin of our samples (Figure 2). We inferred 530 events of transmission of subtype B from the United States to Europe or Australia, and all transmission events led to clusters of 1-8 samples in the recipient country. We also inferred 61 transmissions of subtype B back again to the United States Europe or Australia.

While we also inferred 179 transmission events from the United States to Latin America leading to small clusters represented by 1-4 samples; six transmission events from the United States to Latin America distinguished themselves as they led to larger clusters of 19, 19, 22, 39, 106, and 166 samples, respectively (Figure 2). Five of these cluster included samples from both START and LANL dataset.

We found that transmission from the United States led to a large cluster represented by 27 samples from South Korea (Figure 2). Also, we found a cluster of 12 samples in Thailand to origin from the United States or Malaysia, and that the Thailand cluster seeded a large cluster in China represented by 39 samples (Figure 2).

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

While subtype B originally emerged in Central Africa (Gilbert et al. 2007), ancestral reconstruction estimates that all African subtype B samples in our collection are the results of introduction of subtype B from other geographical regions back to Africa.

Origin and cross-country transmission of subtype C

The second most prevalent subtype within our combined dataset was subtype C (n=949). Subtype C samples from LANL Filtered Web Alignment (n=720) encompassed 30 countries of which 10 countries overlapped with the 21 countries with subtype C samples from START (n=229) (Supplementary Table 2-3). Ancestral state reconstruction estimated the ancestor of subtype C samples to be in South Africa (marginal probability 99%) and root cluster encompassed 398 samples across both datasets (Figure 3). The country of the ancestor became unknown when the country annotation was randomly shuffled (Supplementary Figure 7). Accordingly, the annotation shuffling of supported the estimation of the subtype C ancestral state to be in South Africa and was the outcome of genuine structure in the data.

From the ancestral cluster in South Africa, we inferred four events of transmission that led to clusters represented by >10 samples: (a) transmission to a 39 sample cluster in India; (b) transmission to a 29 sample cluster in Tanzania with transmission onwards via Sweden (cluster with 29 samples) to Brazil (cluster with 20 samples) and from Tanzania back to South African cluster represented by 27 samples; (c) transmission to a 4 sample cluster in Belgium with transmission onwards to Brazil (cluster with 11 samples); and (d) transmission to a 13 sample cluster in Botswana with transmission onwards to an Indian cluster (42 samples) (Figure 3). The largest Brazilian cluster included samples from both START and LANL datasets.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Origin and cross-country transmission of subtype A

START subtype A samples (n=106) encompassed 12 countries, whereas LANL subtype A samples (n=231) encompassed 22 countries of which four countries overlapped with START (Supplementary Table 2-3). The ancestor of subtype A samples was estimated to be in Uganda (marginal probability 99.8%), and the root cluster of 66 samples encompassed samples from both datasets (Figure 4). Estimate of the ancestral state became unknown when the country annotation was randomly shuffled (Supplementary Figure 8); thus, the shuffling of country annotation supports that the estimation of subtype A ancestral state to be in Uganda was the outcome of genuine structure in the data. From the ancestral cluster in Uganda, we inferred up to 11 transmissions leading to clusters of sizes 1-3 and a transmission leading to a large cluster in Kenya (53 samples) with onwards transmission to clusters including a total of 184 samples (Figure 4).

The large Kenyan cluster was the source of two introductions to Europe that led to clusters represented by minimum four samples: That is, we delineated clusters of transmission from the Kenyan cluster to: (a) Tanzania (12 samples) to Cyprus (9 samples) to Greece (18 samples) to Portugal (4 samples); and (b) Congo (10 samples) with onwards transmission to Ukraine (13 samples) that split into a Russian (15 samples) and a Polish cluster (4 samples), respectively (Figure 4). While both START and LANL datasets included Ugandan samples, the other before-mentioned countries were either unique to LANL (Congo, Ukraine, Russia, Poland, Tanzania, Cyprus) or START (Greece, Portugal) dataset, respectively.

Origin and cross-country transmission of subtype CRF01_AE

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

For subtype CRF01_AE, the LANL dataset encompassed samples (n=292) from 14 countries of which three countries overlapped with the samples (n=147) from 12 countries encompassed by the START dataset (Supplementary Table 2-3).

Most (234 of 439 samples) of subtype CRF01_AE samples were from Thailand, but while the majority (81%) of subtype CRF01_AE samples in START dataset were from Thailand, subtype CRF01_AE samples in LANL dataset were more distributed on both China (42%) and Thailand (39%). The LANL subtype CRF01_AE dataset also included three samples taken in 1990 in the Central African Republic, and a cluster with these three samples placed origin of CRF01_AE to be Central African Republic (marginal probability 64%; Figure 5). Estimate of the ancestral state became unknown when the country annotation was randomly shuffled (Supplementary Figure 9); thus, the shuffling of country annotation supports that the estimation of subtype CRF01_AE ancestral state to be in Central African Republic was the outcome of genuine structure in the data.

A transmission cluster in Thailand (n=232 samples; including samples from both LANL and START datasets) was the largest cluster identified for CRF01_AE. From Thailand, we inferred 42 transmissions that led to small clusters outside Asia, whereas we found transmission from Thailand to other Asian countries that led to six large clusters of 7-27 samples (Figure 5).

Subtype distribution across regions for START clinical trial samples

Finally, we investigated the distribution of subtypes across geographical regions for START samples. Seven subtypes (A, AB, B, BC, BF, C, and CRF01_AE) were represented by at least 100 samples in the START collection (Table 2). The seven subtypes showed differences in distribution across geographical regions (Pearson's Chi-squared test with simulated P-value

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

based on 2,000 replicates; P-value=0.0005). Subtype B was the most dominant in Australia, Europe and Israel, Latin America, and United States; subtype C was most dominant in Africa; and subtype CRF01_AE was most dominant in Asia (Table 2). The population of HIV-1 in the United States was the most homogenous (Shannon-Weaver index 0.11), whereas the African population was the most diverse (Shannon-Weaver index 1.52; Table 2).

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Discussion

In this study, we sequenced HIV-1 genomes from 3,197 participants in a clinical trial sampling across 35 countries and analyzed our data with another 2,632 genomes that comprehensively reflects the HIV-1 genetic diversity.

We used the genetic information to delineate and characterize cross-country transmission of subtypes. Our transmission analysis of subtype B showed many parallel transmissions out from the United States that led to small transmission clusters in other regions (51 countries). We also found evidence for transmission back again to the United States from 29 different countries. We find that this supports a model of continuous and frequent transmission of subtype B over country borders.

While we in general found transmission to be identified by clusters represented by a few samples, transmission of subtype B from United States to Latin America showed examples of single transmissions that led to large clusters. We suggest that this is indicative that the HIV-1 population expansion within Latin America was local rather than being fueled by new transmissions from other regions. A similar indication of local clonal expansion was observed for subtype C for which an ancestral population in South Africa led to large clusters in India. This is in agreement with previous reports that the introduction of subtype C to India occurred a limited number of times (Neogi et al. 2012). We found the root of subtype CRF01_AE to be in central Africa which is also the known origin for this subtype (Gao et al. 1996; Murphy et al. 1993). Also, the distribution of subtype CRF01_AE in the START clinical trial confirms that subtype CRF01_AE is endemic in Asia (Angelis et al. 2015), and while we found frequent transmissions to Europe and Australia, our phylogenetic analysis shows that there is no expansion of subtype CRF01_AE outside Asia.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

A limitation of our analysis is that even though our sample collection includes 84 countries, many geographical regions are not well represented, and results from one country may not represent nearby countries, as there may exist large differences in subtype distribution between countries in the same region. This is illustrated by Greece that, unlike other European countries, had a high prevalence of subtype A in the START clinical trial. This agrees with a report on people living with HIV-1 in southwestern Greece that concludes subtype A has surpassed subtype B in new infections (Davanos et al. 2015), and we furthermore estimated that the Greek cluster of subtype A was transmitted from Uganda.

Also, we found that subtype B ancestral state was estimated to be in the United States; nonetheless, we note that, as we have no Caribbean samples in our collection, our analysis is not able to detect the suggested cryptic subtype B circulation in the Caribbean (Worobey et al. 2016).

We found it confirmatory that transmission clusters often included both START and LANL samples and as such both datasets both supported the existence of these clusters. While some countries were represented in both datasets, most countries were unique to each of the two datasets. In this regard, we found it confirmatory that we were able to delineate transmission over countries that were one after the other unique to one of the datasets. Finally, we also found it confirmatory that inferred cross-country transmission often followed the geographical position of countries, e.g. we observed a north-eastwards transmission of subtype C: South Africa to Botswana to India to China or Nepal. Nonetheless, while we found that the combined dataset showed to mutually support our findings, we note that analysis of combined datasets should be interpreted with caution as results may be confounded by general differences between the datasets, e.g. differences in sequence quality.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

We took advantage of the START sampling scheme to investigate the distribution of subtypes across geographical regions in years 2009-2013. While we identified the majority of START samples be pure subtypes (A-D, F, G), we found 26% recombinant subtypes including more than >100 samples of recombinants AB, BC, BF, and CRF01_AE, respectively. This is in line with that recombinant forms of HIV-1 have been estimated to account for 18-20% of infections worldwide (Buonaguro et al. 2007; Hemelaar et al. 2011; Hemelaar et al. 2019). It was beyond the scope of this study to perform a detailed characterization of the individual mosaics and their ancestry (except for CRF01_AE); nonetheless, we note that our observed subtype distributions confirms high prevalence of subtype BC and BF in Africa and Latin America, respectively (Carr et al. 2001; Melo et al. 2012). Also, we found subtype AB in Uganda, Thailand, and Greece, but our sample collection does not cover Eastern Europe where an AB-recombinant was first described in 1998 to circulate among injecting drug users in Eastern Europe (Liitsola et al. 1998), and further reports on distribution and transmission of AB-recombinants remains scarce (Neogi et al. 2017).

Like other studies using the same amplicon based sequencing approach (Cornelissen et al. 2017; Ratmann et al. 2017), we were limited by our ability to produce an evaluable result for all available START samples, and likewise we observe that failure increased for the 3' end of the genome and samples with low viral load. Also, our use of machine extraction as opposed to manual RNA extraction may affect our success rate (Cornelissen et al. 2017).

Finally, we note that our two datasets are fundamentally different in sense that the START samples represent HIV-1 from ART naïve clinical trial participants with CD4+ cell counts greater than 500 cells/mm^3 (Group et al. 2015), whereas LANL samples were selected to represent fullest spectrum of HIV-1 genetic diversity.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Overall, our data both extends current knowledge and exemplifies the limitations in phylogeographic analysis, where inference of transmission is dependent on both representative geographical and genotypic sampling, and as such the estimated countries of ancestors should only be regarded as the best proxies given not all countries are represented. Also, we note that our combined analysis encompasses 84 countries, but 44 of these are represented by less than 10 sequences.

In conclusion, we have used a combined dataset of HIV-1 genomes to present an analysis on the ancestry and transmission of HIV-1 across multiple viral subtypes and geographical regions. We envision that the presented data on HIV-1 genotypes can facilitate comparative genetic studies and be analyzed in concert with other information, e.g. markers of HIV pathogenesis, to further understand HIV-1.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Acknowledgements

The authors would like to specifically thank the participants in the START trial. See N Engl J Med 2015;373:795–807 for the complete list of START investigators. The authors would like to thank the reviewers and the handling editor for help to improve the manuscript.

Funding

The primary funder of the START trial was NIAID (grant numbers UM1-AI068641, UM1-AI120197 and 1U01-AI136780). This study was made possible through a grant from the Danish National Research Foundation (grant number 126). Additional support was obtained from the National Institute of Allergy and Infectious Diseases, National Institutes of Health Clinical Center, National Cancer Institute, National Heart, Lung, and Blood Institute, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institute of Mental Health, National Institute of Neurological Disorders and Stroke, National Institute of Arthritis and Musculoskeletal and Skin Diseases, Agence Nationale de Recherches sur le SIDA et les Hépatites Virales (France), National Health and Medical Research Council (Australia), Bundesministerium für Bildung und Forschung (Germany), European AIDS Treatment Network, Medical Research Council (United Kingdom), National Institute for Health Research, National Health Service (United Kingdom), and University of Minnesota. Antiretroviral drugs were donated to the central drug repository by AbbVie, Bristol-Myers Squibb, Gilead Sciences, GlaxoSmithKline/ViiV Healthcare, Janssen Scientific Affairs, and Merck. MW was supported by NIH/NIAID R01AI084691.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Author contributions

M.B, A.Z, J.D.L and R.L.M conceived the study; A.Z, J.D.L and R.L.M supervised the project; M.B, J.D.L and R.L.M designed the analysis plan; M.B performed experiments; M.B., A.Z, MH.T and R.L.M analyzed the data; M.B. drafted the manuscript; M.B, A.Z, MH.T, J.B, P.M, K.R, M.G, M.W, J.D.L and R.L.M revised the analysis plan and revised the manuscript.

Competing interests

The authors declare that they have no competing interests.

Data availability

All non-personally identifiable data will be made available to the readers of Virus Evolution, if the manuscript is accepted. Personal identifiable data can be accessed through agreement with the INSIGHT START Study Group to respect the donor consent of the participants in the START study.

References

Angelis, K., et al. (2015), 'Global Dispersal Pattern of HIV Type 1 Subtype CRF01_AE: A Genetic Trace of Human Mobility Related to Heterosexual Sexual Activities Centralized in Southeast Asia', *J Infect Dis,* 211 (11), 1735-44.

Bletsa, M., et al. (2019), 'Divergence dating using mixed effects clock modelling: An application to HIV-1', *Virus Evol,* 5 (2), vez036.

Buonaguro, L., Tornesello, M. L., and Buonaguro, F. M. (2007), 'Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications', *J Virol,* 81 (19), 10209-19.

Carr, J. K., et al. (2001), 'Diverse BF recombinants have spread widely since the introduction of HIV-1 into South America', *AIDS,* 15 (15), F41-7.

Castro-Nallar, E., et al. (2012), 'The evolution of HIV: inferences using phylogenetics', *Mol Phylogenet Evol,* 62 (2), 777-92.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Cornelissen, M., et al. (2017), 'From clinical sample to complete genome: Comparing methods for the extraction of HIV-1 RNA for high-throughput deep sequencing', *Virus Res,* 239, 10-16.
Davanos, N., et al. (2015), 'HIV-1 subtype characteristics of infected persons living in southwestern Greece', *HIV AIDS (Auckl),* 7, 277-83.
Felsenstein, J. (1981), 'Evolutionary trees from DNA sequences: a maximum likelihood approach', *J Mol Evol,* 17 (6), 368-76.
--- (2004), *Inferring Phylogenies* (Sunderland, MA: Sinauer).
Firth, C., et al. (2010), 'Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses', *Mol Biol Evol,* 27 (9), 2038-51.
Gall, A., et al. (2012), 'Universal amplification, next-generation sequencing, and assembly of HIV-1 genomes', *J Clin Microbiol,* 50 (12), 3838-44.
Gao, F., et al. (1996), 'The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin', *J Virol,* 70 (10), 7013-29.
Gilbert, M. T., et al. (2007), 'The emergence of HIV/AIDS in the Americas and beyond', *Proc Natl Acad Sci U S A,* 104 (47), 18566-70.
Group, Insight Start Study, et al. (2015), 'Initiation of Antiretroviral Therapy in Early Asymptomatic HIV Infection', *N Engl J Med,* 373 (9), 795-807.
Gryseels, S., et al. (2020), 'A near full-length HIV-1 genome from 1966 recovered from formalin-fixed paraffin-embedded tissue', *Proc Natl Acad Sci U S A,* 117 (22), 12222-29.
Hemelaar, J., et al. (2011), 'Global trends in molecular epidemiology of HIV-1 during 2000-2007', *AIDS,* 25 (5), 679-89.
Hemelaar, J., et al. (2019), 'Global and regional molecular epidemiology of HIV-1, 1990-2015: a systematic review, global survey, and trend analysis', *Lancet Infect Dis,* 19 (2), 143-55.
International, H. I. V. Controllers Study, et al. (2010), 'The major genetic determinants of HIV-1 control affect HLA class I peptide presentation', *Science,* 330 (6010), 1551-7.
Ishikawa, S. A., et al. (2019), 'A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios', *Mol Biol Evol*.
Kozlov, A. M., Aberer, A. J., and Stamatakis, A. (2015), 'ExaML version 3: a tool for phylogenomic analyses on supercomputers', *Bioinformatics,* 31 (15), 2577-9.
Langmead, B. and Salzberg, S. L. (2012), 'Fast gapped-read alignment with Bowtie 2', *Nat Methods,* 9 (4), 357-9.
Letunic, I. and Bork, P. (2019), 'Interactive Tree Of Life (iTOL) v4: recent updates and new developments', *Nucleic Acids Res,* 47 (W1), W256-W59.
Li, H., et al. (2009), 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics,* 25 (16), 2078-9.
Liitsola, K., et al. (1998), 'HIV-1 genetic subtype A/B recombinant strain causing an explosive epidemic in injecting drug users in Kaliningrad', *AIDS,* 12 (14), 1907-19.
Melo, F. L., Jamal, L. F., and Zanotto, P. M. (2012), 'Characterization of primary isolates of HIV type 1 CRF28_BF, CRF29_BF, and unique BF recombinants circulating in Sao Paulo, Brazil', *AIDS Res Hum Retroviruses,* 28 (9), 1082-8.
Murphy, E., et al. (1993), 'Diversity of V3 region sequences of human immunodeficiency viruses type 1 from the central African Republic', *AIDS Res Hum Retroviruses,* 9 (10), 997-1006.
Neogi, U., et al. (2017), 'Recent increased identification and transmission of HIV-1 unique recombinant forms in Sweden', *Sci Rep,* 7 (1), 6371.
Neogi, U., et al. (2012), 'Molecular epidemiology of HIV-1 subtypes in India: origin and evolutionary history of the predominant subtype C', *PLoS One,* 7 (6), e39819.
Ning, Z., Cox, A. J., and Mullikin, J. C. (2001), 'SSAHA: a fast search method for large DNA databases', *Genome Res,* 11 (10), 1725-9.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Pineda-Pena, A. C., et al. (2013), 'Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools', *Infect Genet Evol,* 19, 337-48.
Ratmann, O., et al. (2017), 'HIV-1 full-genome phylogenetics of generalized epidemics in sub-Saharan Africa: impact of missing nucleotide characters in next-generation sequences', *AIDS Res Hum Retroviruses,* 33 (11), 1083-98.
To, T. H., et al. (2016), 'Fast Dating Using Least-Squares Criteria and Algorithms', *Syst Biol,* 65 (1), 82-97.
Volz, E. M., Koelle, K., and Bedford, T. (2013), 'Viral phylodynamics', *PLoS Comput Biol,* 9 (3), e1002947.
Worobey, M., et al. (2016), '1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America', *Nature,* 539 (7627), 98-101.

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Tables and Figures

Table 1: Number of samples for subtypes A, B, C, D, F, G, and CRF01_AE in START and LANL dataset.

Subtype	*START*	*LANL*
A	106	231
B	1,959	1,196
C	229	720
D	35	71
F	20	46
G	5	76
CRF01_AE	147	292
Total	2,501	2,632

ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Table 2: Distribution of HIV-1 subtypes across geographical regions for START dataset. All subtypes found in less than 100 samples are grouped as 'Other'. Sequencing was not successful for 588 samples and these are shown as 'ND' (Not determined). The bottom two rows show the Shannon-Weaver and 1 – Simpson indexes of subtype diversity per region. Diversity indexes were calculated based in the distribution of subtypes A, AB, B, BC, BF, C, and CRF01_AE across geographical regions.

Subtype	***Africa***	***Asia***	***Australia***	***Europe and Israel***	***United States***	***Latin America***	***Total***
A	63 (9 %)	0	0	43 (3 %)	0	0	106 (3 %)
AB	43 (6 %)	19 (6 %)	0	44 (3 %)	2 (1 %)	2 (0 %)	110 (3 %)
CRF01_AE	0	121 (41 %)	5 (5 %)	21 (2 %)	0	0	147 (4 %)
B	42 (6 %)	23 (8 %)	74 (78 %)	888 (66 %)	338 (85 %)	594 (65 %)	1,959 (52 %)
BC	85 (12 %)	8 (3 %)	1 (1 %)	22 (2 %)	1 (0 %)	23 (3 %)	140 (4 %)
BF	0	0	0	26 (2 %)	1 (0 %)	132 (14	159 (4

ACCEPTED MANUSCRIPT

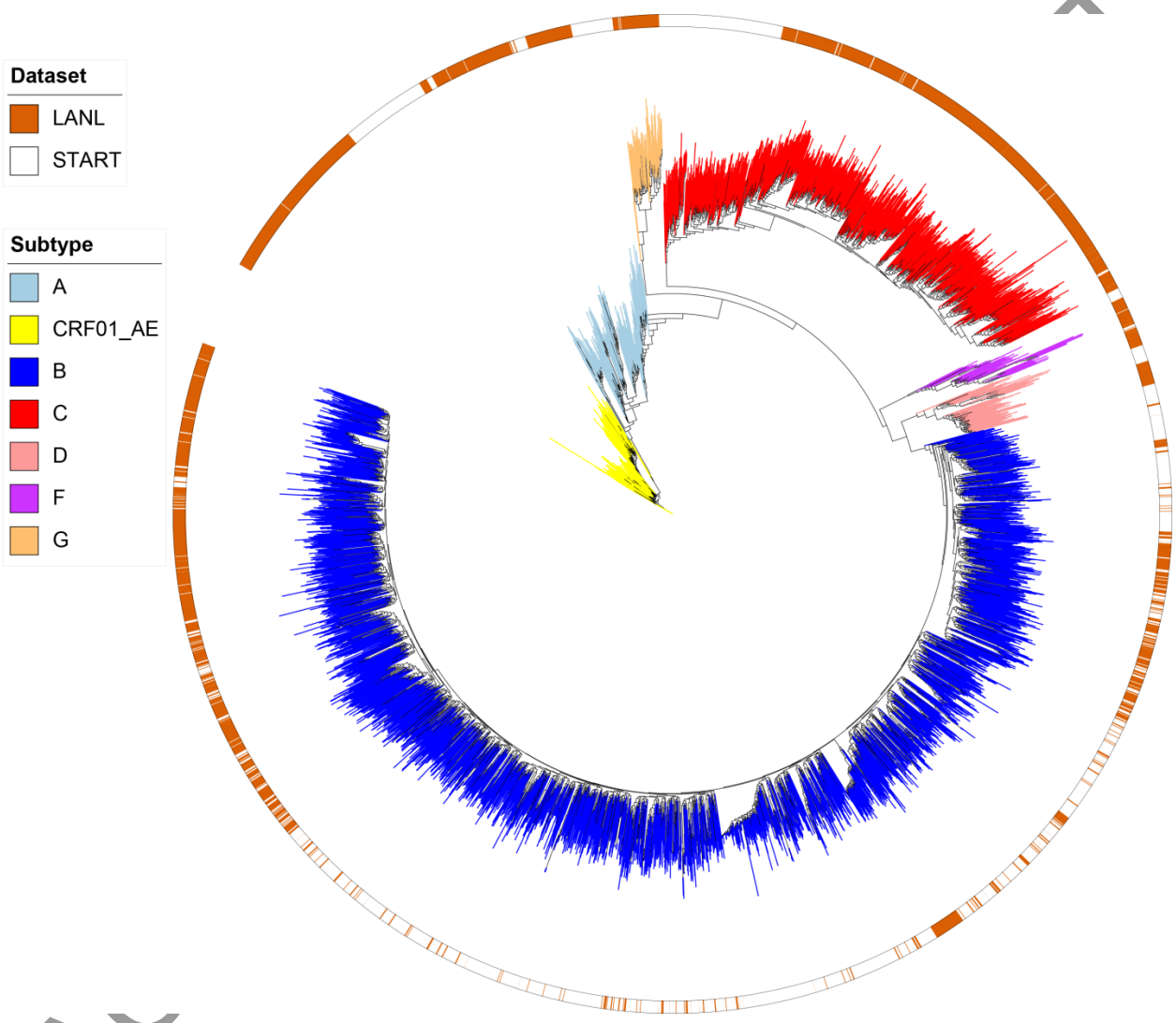
Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

						%)	%)
C	129 (18 %)	40 (13 %)	3 (3 %)	40 (3 %)	2 (1 %)	15 (2 %)	229 (6 %)
Other	189 (26 %)	19 (6 %)	0	109 (8 %)	5 (1 %)	25 (3 %)	347 (9 %)
ND	180 (25 %)	68 (23 %)	12 (13 %)	160 (12 %)	48 (12 %)	120 (13 %)	588 (16 %)
Total	731	298	95	1353	397	911	3,785
Shannon-Weaver Index	1.52	1.22	0.44	0.79	0.11	0.70	-
1 - Simpson index	0.76	0.61	0.20	0.32	0.03	0.37	-

ACCEPTED MANUSCRIPT

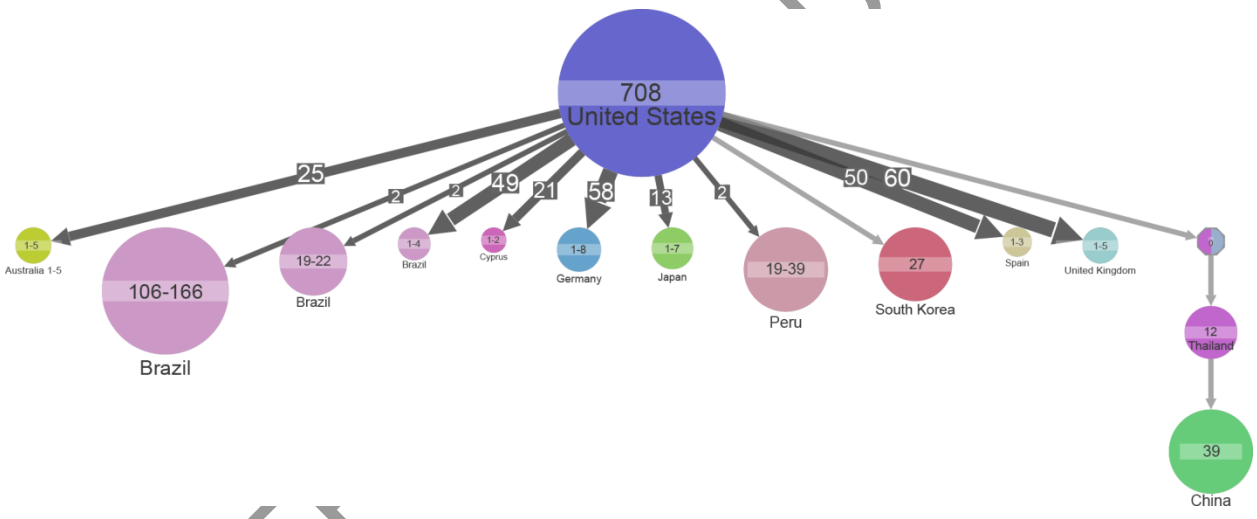
Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Figure 1: Maximum likelihood phylogenetic tree of all 5,133 samples in the combined START and LANL dataset that were defined as pure subtypes (A-D, F, G) or subtype CRF01_AE. The branches are colored according to subtype. The outer color-strip indicate if the samples are from either START or LANL Filtered Web alignment.



Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

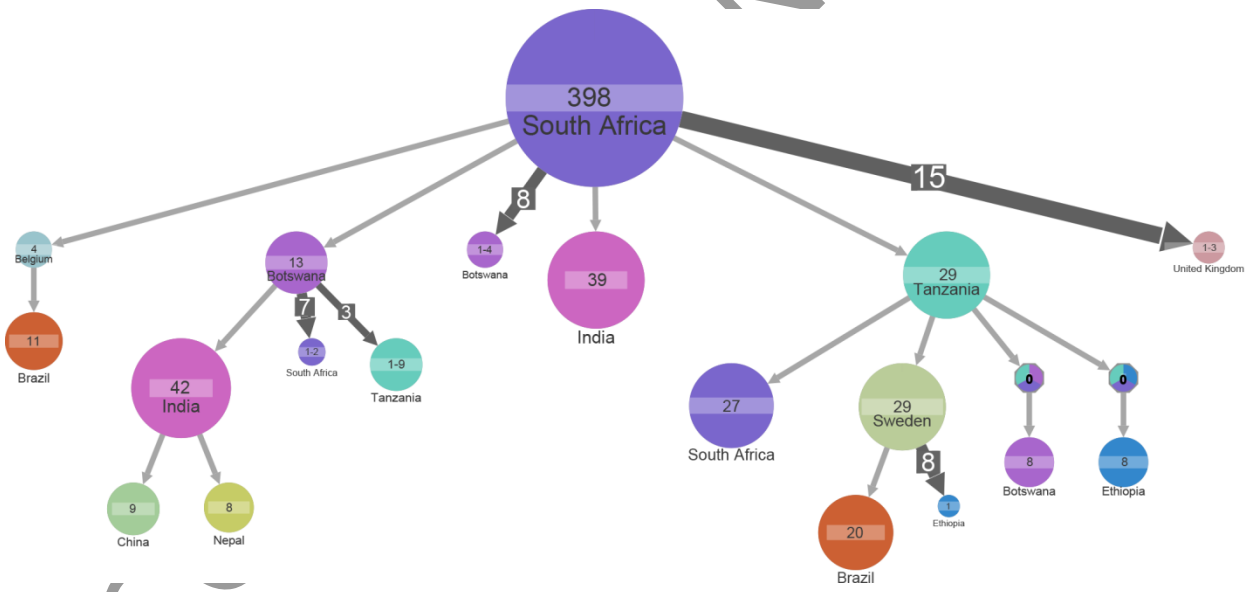
Figure 2: Ancestral state reconstruction of 3,155 subtype B samples. Ancestral state reconstruction of samples is shown for country origin. Circles denote genetic clusters of samples with the same state. The state and sample size of clusters are indicated for each circle. An arrow between two circles denotes events of transmission from the top cluster to the bottom cluster. The size and the number on top of the arrows indicate that the arrows represent multiple transmission events leading to clusters of similar sizes. Clusters with a "0" and multiple colors indicate that several corresponding states have similar marginal probabilities. The lowest marginal probability for resolved clusters shown is 74%. Clusters of size less than 19 are hidden to improve readability.



ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

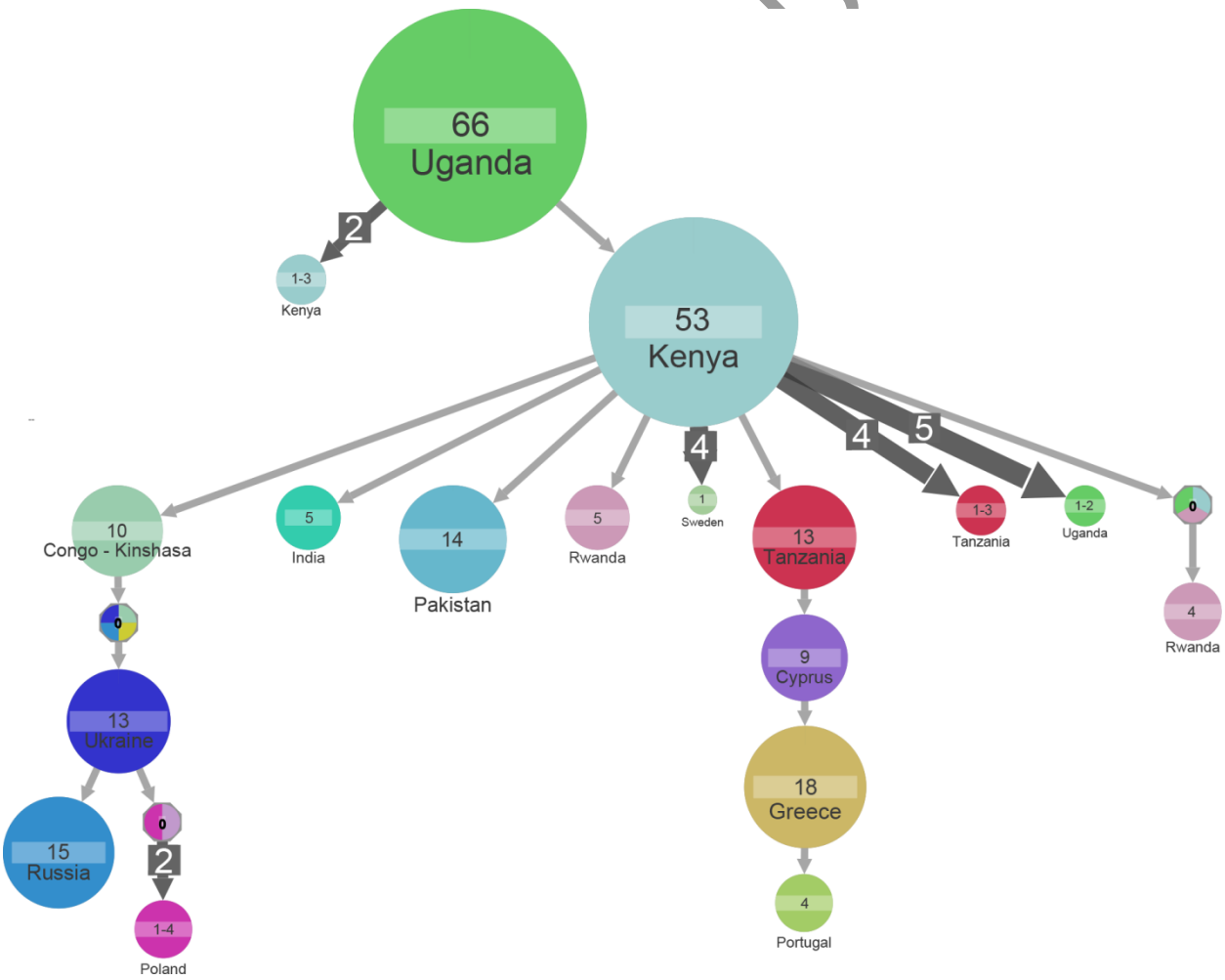
Figure 3: Ancestral state reconstruction of 949 subtype C samples. Ancestral state reconstruction of samples is shown for country origin. Circles denote genetic clusters of samples from the sample country. The country and sample size of clusters are indicated for each circle. An arrow between two circles denotes events of transmission from the top cluster to the bottom cluster. Clusters with a "0" and multiple colors indicate that several corresponding states have similar marginal probabilities. The lowest marginal probability for resolved clusters shown is 82%. The size and the number on top of the arrows indicate that the arrow represent multiple transmission events leading to clusters of similar sizes. Clusters of size less than 8 are hidden to improve readability.



MANUSCRIPT

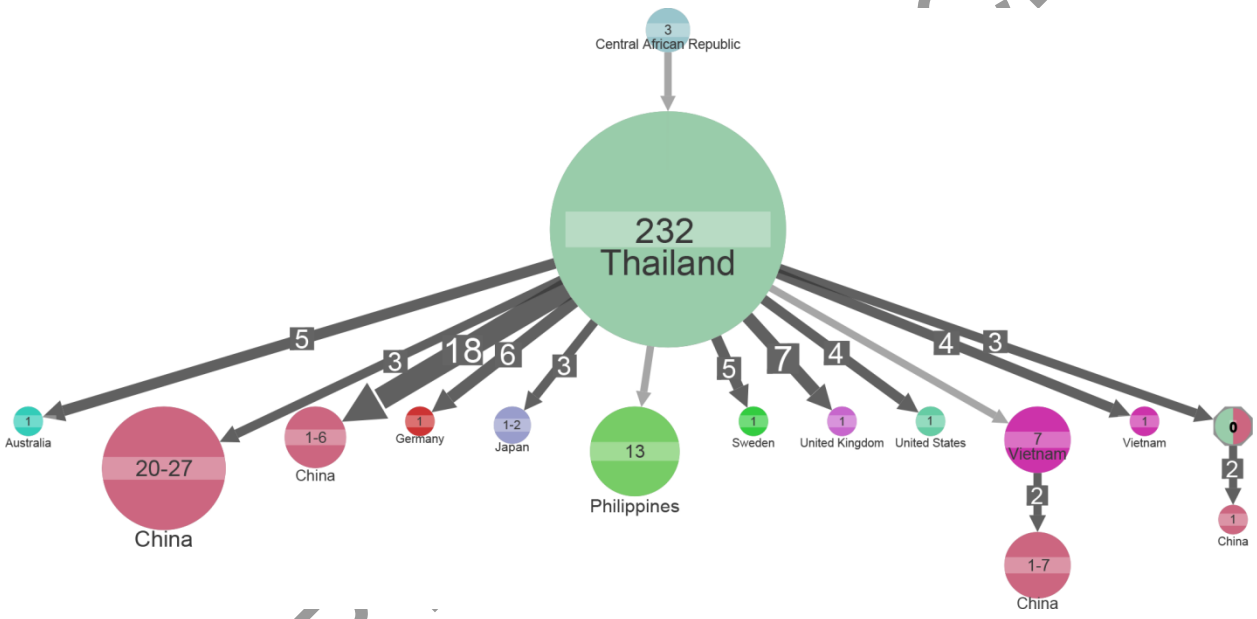
Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Figure 4: Ancestral state reconstruction of 337 subtype A samples. Ancestral state reconstruction of samples is shown for country origin. Circles denote genetic clusters of samples from the sample country. The country and sample size of clusters are indicated for each circle. An arrow between two circles denotes events of transmission from the top cluster to the bottom cluster. Clusters with a "0" and multiple colors indicate that several corresponding states have similar marginal probabilities. The lowest marginal probability for resolved clusters shown is 75%. The size and the number on top of the arrows indicate that the arrow represent multiple transmission events leading to clusters of similar sizes. Clusters of size less than 4 are hidden to improve readability.



Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021

Figure 5: Ancestral state reconstruction of 439 subtype CRF01_AE samples. Ancestral state reconstruction of samples is shown for country origin. Circles denote genetic clusters of samples from the sample country. The country and sample size of clusters are indicated for each circle. An arrow between two circles denotes events of transmission from the top cluster to the bottom cluster. The lowest marginal probability for resolved clusters shown is 64%. The size and the number on top of the arrows indicate that the arrow represent multiple transmission events leading to clusters of similar sizes. Clusters of size less than 4 are hidden to improve readability.



ACCEPTED MANUSCRIPT

Downloaded from https://academic.oup.com/ve/advance-article/doi/10.1093/ve/veab055/6295513 by Institut Pasteur user on 22 June 2021