South Africa

De Oliveira, Tulio, KwaZulu-Natal Research Innovation Sequencing Platform
Pillay, Deenan, Africa Health Research Institute

PANGEA1 HIV Clinical Data

Study Documentation

February 21, 2019

Metadata Production

Metadata Producer(s)	Africa Health Research Institute (AHRI)
Identification	DDI.AHRI.PANGEA1.HIV.Clinical.Data.2019.v1

Table of Contents

PANGEA1 HIV Clinical Data

Overview

Identification	AHRI.PANGEA1.HIV.Clinical.Data.2019.v1
Version	V1.0.0

Abstract

The purpose of this study was address one key question: What is the contribution of external sources of HIV in driving the epidemic within a community. High rates of cross-community transmission calls into question the efficacy of increasing local coverage of antiretroviral therapy to prevent new infections. To address this question, 1068 polymerase sequences were derived from sampling conducted within the Africa Health Research Insitute's demographic surveillance area between 2010 and 2013. AHRI genotypes were analysed against a large national dataset of homologous sequences in a novel phylodynamic framework to infer the overall transmission dynamics within the community. This allowed us to enumerate the relative contribution of local transmission versus external introductions to the overall HIV incidence in the community. Major variables within the dataset to answer this question are: (1) the genotype; (2) the community's location from which the genotype was derived; and (3) the date of sampling.

Kind of Data	HIV Genomic Data
Unit of Analysis	Each sequences derived from a single specimen.

Scope & Coverage

Keywords	HIV polymerase sequence; date of sampling; phylodynamics, phylogeny, HIV-1
Topics	HIV-1; Incidence; Phylogeny; Epidemics; Population Surveillance; Rural Population; HIV Infections; Africa
Time Period(s)	2010-2013
Countries	South Africa

Geographic Coverage

Treatment as Prevention trial area of the Africa Health Research Institute.

Universe

HIV positive participants enrolled in the Treatment as Prevention 12449 ANRS trial, who linked to care

Producers & Sponsors

Primary Investigator(s)	De Oliveira, Tulio, KwaZulu-Natal Research Innovation Sequencing Platform Pillay, Deenan, Africa Health Research Institute
Other Producer(s)	Africa Health Research Institute (AHRI)
Funding Agency/ies	South African Medical Research Council (SAMRC) , Genotyping funding source
Other Acknowledgment(s)	Wilkinson, Eduan , Cleaned, aligned and help analyse the sequence data. , KwaZulu-Natal Research Innovation Sequencing Platform

Sampling

Sampling Procedure

HIV positive individuals within the demographic surveillance area of the Africa Health Research Insititue from 2010 to 2013.
Individuals blood spots were only attempted sequencing if they:
A) Had a viral load done
B) Viral load was above 1,550 copies/ml

C) The laboratory got approximately 60% success rate on sequencing them, they should have the list that they attempt to sequence and the ones that were successful.

Data Collection

Data Collection Dates	start 2010-01-01 end 2013-12-04

Data Processing & Appraisal

Data Editing

Sequences were generated by the Durban based laboratory of AHRI. Sequences were aligned with one another in ClustalW and are presented in the standard fasta file format.

Accessibility

Access Conditions

1. The representative of the Receiving Organization agrees to comply with the following conditions:
2. Access to the restricted data will be limited to the Lead Researcher and other members of the research team listed in this request.
3. Copies of the restricted data or any data created on the basis of the original data will not be copied or made available to anyone other than those mentioned in this Data Access Agreement, unless formally authorized by the Data Archive.
4. The data will only be processed for the stated statistical and research purpose. They will be used for solely for reporting of aggregated information, and not for investigation of specific individuals or organizations. Data will not in any way be used for any administrative, proprietary or law enforcement purposes.
5. The Lead Researcher must state if it is their intention to match the restricted microdata with any other micro-dataset. If any matching is to take place, details must be provided of the datasets to be matched and of the reasons for the matching. Any datasets created as a result of matching will be considered to be restricted and must comply with the terms of this Data Access Agreement.
6. The Lead Researcher undertakes that no attempt will be made to identify any individual person, family, business, enterprise or organization. If such a unique disclosure is made inadvertently, no use will be made of the identity of any person or establishment discovered and full details will be reported to the Data Archive. The identification will not be revealed to any other person not included in the Data Access Agreement.
7. The Lead Researcher will implement security measures to prevent unauthorized access to licensed microdata acquired from the Data Archive. The microdata must be destroyed upon the completion of this research, unless the Data Archive obtains satisfactory guarantee that the data can be secured and provides written authorization to the Receiving Organization to retain them. Destruction of the microdata will be confirmed in writing by the Lead Researcher to the Data Archive.
8. Any books, articles, conference papers, theses, dissertations, reports, or other publications that employ data obtained from the Data Archive will cite the source of data in accordance with the citation requirement provided with the dataset.
9. An electronic copy of all reports and publications based on the requested data will be sent to the Data Archive.
10. The original collector of the data, the Data Archive, and the relevant funding agencies bear no responsibility for use of the data or for interpretations or inferences based upon such uses.
11. This agreement will come into force on the date that approval is given for access to the restricted dataset and remain in force until the completion date of the project or an earlier date if the project is completed ahead of time.
If there are any changes to the project specification, security arrangements, personnel or organization detailed in this application form, it is the responsibility of the Lead Researcher to seek the agreement of the Data Archive to these changes. Where there is a change to the employer organization of the Lead Researcher this will involve a new application being made and termination of the original project.
12. Breaches of the agreement will be taken seriously and the Data Archive will take action against those responsible for the lapse if willful or accidental. Failure to comply with the directions of the Data Archive will be deemed to be a major breach of the agreement and may involve recourse to legal proceedings. The Data Archive will maintain and share with partner data archives a register of those individuals and organizations which are responsible for breaching the terms of the Data Access Agreement and will impose sanctions on release of future data to these parties.

Citation Requirements

https://doi.org/10.23664/AHRI.PANGEA1.HIV.Clinical.Data.2019.v1

Files Description

Dataset contains 2 file(s)

Hexagons	
# Cases	1021
# Variable(s)	5

PangeaIndividualsFinalTasP	
# Cases	3966
# Variable(s)	24

Variables List

Dataset contains 29 variable(s)

File Hexagons

#	Name	Label	Type	Format	Valid	Invalid	Question
1	Id	-	continuous	numeric-12.0	1021	0	-
2	Uid	-	discrete	character-36	1021	0	-
3	Perimeter	-	discrete	character-244	1021	-	-
4	Centroid	-	discrete	character-244	1021	-	-
5	t	-	discrete	character-244	1021	-	-

File PangeaIndividualsFinalTasP

#	Name	Label	Type	Format	Valid	Invalid	Question
1	Individu ..	Unique Individual Identifier	continuous	numeric-12.0	3966	0	-
2	SampleId	Unique Sample Identifier	discrete	character-244	3966	-	-
3	Source	Source Project	discrete	character-244	3966	-	-
4	pangea_id	PANGEA Unique Identifier	discrete	character-244	3966	-	-
5	dob	Date of Birth	discrete	character-11	3966	-	-
6	Age	Age	continuous	numeric-12.0	3966	0	-
7	Gender	Gender	discrete	character-244	2590	-	-
8	sample_d ..	Sample Collection Date	discrete	character-11	3966	-	-
9	ACDIS_Id	ACDIS Unique Identifier	continuous	numeric-12.0	2785	1181	-
10	ARTEMIS_Id	ARTEMIS Unique Identifier	continuous	numeric-12.0	1367	2599	-
11	ACCDB_Id	ACCDB Unique Identifier	continuous	numeric-12.0	2626	1340	-
12	TasP_Id	TasP Unique Identifier	continuous	numeric-12.0	2590	1376	-
13	LatestNe ..	Last HIV Negative Date	discrete	character-11	304	-	-
14	Earliest ..	Earliest HIV Positive Date	discrete	character-11	3955	-	-
15	Earliest ..	Earliest Known ART Start Date	discrete	character-11	3661	-	-
16	LastKnow ..	Last Known ART Start Date	discrete	character-11	2256	-	-
17	LastKnow ..	Last Known ART Regimen	discrete	character-244	2256	-	-
18	LastKnow ..	Last Known ART Adherence VAS Score	continuous	numeric-12.0	2102	1864	-
19	ViralLoad	Viral Load	continuous	numeric-12.0	2527	1439	-
20	ViralLoa ..	Was Viral Load below LDL?	discrete	numeric-2.0	2526	1440	-
21	CD4Count	CD4 Count - Analysed by LAB on Sample	continuous	numeric-12.0	170	3796	-
22	CD4Count ..	CD4 Count - Analysed by Clinic Staff on PIMA	continuous	numeric-12.0	2128	1838	-
23	CD4Pct	CD4 Percentage - Analysed by LAB on Sample	continuous	numeric-12.0	170	3796	-
24	HexagonId	Hexagon Unique Identifier	continuous	numeric-12.0	2589	1377	-

Variables Description

Dataset contains 29 variable(s)

File : Hexagons

# Id	
Information	[Type= continuous] [Format=numeric] [Range= 1-1021] [Missing=*]
Statistics [NW/ W]	[Valid=1021 /-] [Invalid=0 /-] [Mean=511 /-] [StdDev=294.882 /-]
# Uid	
Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=1021 /-] [Invalid=0 /-]
# Perimeter	
Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=1021 /-]
# Centroid	
Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=1021 /-]
# t	
Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=1021 /-]

File : PangeaIndividualsFinalTasP

# IndividualId: Unique Individual Identifier

Information	[Type= continuous] [Format=numeric] [Range= 1-200695] [Missing=*]
Statistics [NW/ W]	[Valid=3966 /-] [Invalid=0 /-] [Mean=35127.359 /-] [StdDev=58416.372 /-]

# SampleId: Unique Sample Identifier

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3966 /-]

# Source: Source Project

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3966 /-]

Value	Label	Cases	Percentage
ACDIS		9	0.2%
CC		794	20.0%
RES		573	14.4%
TasP		2590	65.3%

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

# pangea_id: PANGEA Unique Identifier

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3966 /-]

# dob: Date of Birth

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3966 /-]

# Age: Age

Information	[Type= continuous] [Format=numeric] [Range= 14-114] [Missing=*]
Statistics [NW/ W]	[Valid=3966 /-] [Invalid=0 /-] [Mean=36.425 /-] [StdDev=12.314 /-]

# Gender: Gender

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=2590 /-]

Value	Label	Cases	Percentage
1		761	29.4%
2		1829	70.6%

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

# sample_date: Sample Collection Date

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3966 /-]

# ACDIS_Id: ACDIS Unique Identifier

Information	[Type= continuous] [Format=numeric] [Range= 84-218561] [Missing=*]
Statistics [NW/ W]	[Valid=2785 /-] [Invalid=1181 /-] [Mean=160690.758 /-] [StdDev=53859.445 /-]

# ARTEMIS_Id: ARTEMIS Unique Identifier

Information	[Type= continuous] [Format=numeric] [Range= 41-200695] [Missing=*]
Statistics [NW/ W]	[Valid=1367 /-] [Invalid=2599 /-] [Mean=92656.543 /-] [StdDev=74427.215 /-]

File : PangeaIndividualsFinalTasP

# ACCDB_Id: ACCDB Unique Identifier

Information	[Type= continuous] [Format=numeric] [Range= 4-56761] [Missing=*]
Statistics [NW/ W]	[Valid=2626 /-] [Invalid=1340 /-] [Mean=32619.001 /-] [StdDev=18233.265 /-]

# TasP_Id: TasP Unique Identifier

Information	[Type= continuous] [Format=numeric] [Range= 9-100102] [Missing=*]
Statistics [NW/ W]	[Valid=2590 /-] [Invalid=1376 /-] [Mean=11710.937 /-] [StdDev=8333.553 /-]

# LatestNegative: Last HIV Negative Date

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=304 /-]

# EarliestPositive: Earliest HIV Positive Date

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3955 /-]

# EarliestKnownART: Earliest Known ART Start Date

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3661 /-]

# LastKnownART: Last Known ART Start Date

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=2256 /-]

# LastKnownARTRegimen: Last Known ART Regimen

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=2256 /-]

Value	Label	Cases	Percentage
Abacavir,Lamivuc		13	0.6%
Abacavir,Lamivuc		1	0.0%
Dideoxyinosine, Didanosine,Lamiv		1	0.0%
Missing (iDART),Missing (iDART),Missing (iDART)		8	0.4%
No drug,No drug,Lopinavir		5	0.2%
Stavudine,Lamivu		21	0.9%
Stavudine,Lamivu		6	0.3%
Tenofovir,Emtrici		5	0.2%
Tenofovir,Emtrici		1878	83.2%
Tenofovir,Emtrici		18	0.8%
Tenofovir,Emtrici		1	0.0%
Tenofovir,Lamivu		63	2.8%
Tenofovir,Lamivu		63	2.8%
Tenofovir,Lamivu		4	0.2%
Tenofovir,Lamivu		3	0.1%
Tenofovir,Lamivu		1	0.0%

File : PangeaIndividualsFinalTasP

# LastKnownARTRegimen: Last Known ART Regimen

Value	Label	Cases	Percentage
Tenofovir,Missing (iDART),Efaviren		2	0.1%
Truvada,Etravirin		3	0.1%
Zidovudine,Dideo Didanosine,Lopin		2	0.1%
Zidovudine,Lamiv		81	3.6%
Zidovudine,Lamiv		75	3.3%
Zidovudine,Lamiv		2	0.1%

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

# LastKnownAdhereScore: Last Known ART Adherence VAS Score

Information	[Type= continuous] [Format=numeric] [Range= 1-100] [Missing=*]
Statistics [NW/ W]	[Valid=2102 /-] [Invalid=1864 /-] [Mean=96.829 /-] [StdDev=8.922 /-]

# ViralLoad: Viral Load

Information	[Type= continuous] [Format=numeric] [Range= 1-6277000] [Missing=*]
Statistics [NW/ W]	[Valid=2527 /-] [Invalid=1439 /-] [Mean=165876.255 /-] [StdDev=438237.754 /-]

# ViralLoadBelowLDL: Was Viral Load below LDL?

Information	[Type= discrete] [Format=numeric] [Range= 0-1] [Missing=*]
Statistics [NW/ W]	[Valid=2526 /-] [Invalid=1440 /-]

Value	Label	Cases	Percentage
0	No	2517	99.6%
1	Yes	9	0.4%
Sysmiss		1440	

Warning: these figures indicate the number of cases found in the data file. They cannot be interpreted as summary statistics of the population of interest.

# CD4Count: CD4 Count - Analysed by LAB on Sample

Information	[Type= continuous] [Format=numeric] [Range= 2-782] [Missing=*]
Statistics [NW/ W]	[Valid=170 /-] [Invalid=3796 /-] [Mean=270.529 /-] [StdDev=225.791 /-]

# CD4Count_PIMA: CD4 Count - Analysed by Clinic Staff on PIMA

Information	[Type= continuous] [Format=numeric] [Range= 4-3134] [Missing=*]
Statistics [NW/ W]	[Valid=2128 /-] [Invalid=1838 /-] [Mean=416.874 /-] [StdDev=253.064 /-]

# CD4Pct: CD4 Percentage - Analysed by LAB on Sample

Information	[Type= continuous] [Format=numeric] [Range= 2-37] [Missing=*]
Statistics [NW/ W]	[Valid=170 /-] [Invalid=3796 /-] [Mean=19.465 /-] [StdDev=7.962 /-]

# HexagonId: Hexagon Unique Identifier

Information	[Type= continuous] [Format=numeric] [Range= 412-1013] [Missing=*]
Statistics [NW/ W]	[Valid=2589 /-] [Invalid=1377 /-] [Mean=846.473 /-] [StdDev=103.569 /-]