South Africa

Pillay, Deenan, Africa Health Research Institute

PANGEA1 Demographic and Clinical Data.2019.v1

Study Documentation

April 18, 2019

Metadata Production

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

Table of Contents

PANGEA1 Demographic and Clinical Data.2019.v1

Overview

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

Abstract
PANGEA is an international consortium of researchers based in Africa, the US and the UK studying transmission dynamics in HIV epidemics in sub-Saharan Africa. The goal of the consortium is to analyse HIV phylogenetic and demographic data to identify individual and population-level factors that drive the epidemic, analyse the dynamics of the epidemic and translate these findings into information that can be used to more effectively target interventions. PANGEA encompasses four analysis themes: (1) molecular epidemiology and mathematical modelling, (2) phylodynamics, (3) mobility and migration, and (4) clinical science, drug resistance and ethics. Major variables within the dataset to answer this question are: (1) the genotype; (2) the participant's geolocation from which the genotype was derived; and (3) the date of sampling, (4) all relevant clinical data and (5) all relevant demographic data.

Kind of Data	Clinical and Demographic data
Unit of Analysis	Each sequences derived from a single specimen, but some sequences may have been repeated on the same specimen, or on another time point, depending on the quality of the sequence.

Scope & Coverage

Keywords	HIV full-length sequences; date of sampling; phylodynamics, phylogeny, HIV-1, ART history
Topics	HIV-1; Incidence; Phylogeny; Epidemics; Population Surveillance; Rural Population; HIV Infections; Africa
Time Period(s)	2010-2016
Countries	South Africa

Geographic Coverage
South Africa

Universe
HIV genome extracted and sequenced from participants infected with HIV

Producers & Sponsors

Primary Investigator(s)	Pillay, Deenan, Africa Health Research Institute
Other Producer(s)	Africa Health Research Institute (AHRI)
Funding Agency/ies	PANGEA consortium, funded by the Bill & Melinda Gates Foundation (BMGF) , Genotyping funding source
Other Acknowledgment(s)	Derache, Anne , Cleaned, aligned and help analyse the sequence data. , Africa Health Research Insitute

Sampling

Sampling Procedure
HIV positive individuals within the demographic surveillance area of the Africa Health Research Insititue from 2010 to 2016. Full-length HIV deep sequencing was attempted on samples that:
A) Had a blood sample taken (not DBS)

B) Had a viral load >1000 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 2016-12-04

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.Data.2019.v1

Files Description

Dataset contains 3 file(s)

AHRI.PANGEA1.ART.2019.v1	
# Cases	71421
# Variable(s)	7

AHRI.PANGEA1.INDIVIDUALS.2019.v1	
# Cases	3890
# Variable(s)	15

AHRI.PANGEA1.LAB.2019.v1	
# Cases	19197
# Variable(s)	6

Variables List

Dataset contains 28 variable(s)

File AHRI.PANGEA1.ART.2019.v1

#	Name	Label	Type	Format	Valid	Invalid	Question
1	PangeaId	Pangea unique Identifier of Individual	discrete	character-20	71421	0	-
2	ARTStart ..	Date when antiretroviral therapy was initiated	discrete	character-11	70901	-	-
3	ARTEncou ..	ART encounter or visit date	discrete	character-11	70901	-	-
4	ARTRegimen	ART regimen	discrete	character-244	63949	-	-
5	ARTEndEv ..	Date of ART end event	discrete	character-11	6440	-	-
6	ARTEndEv ..	Reasons for ART end event occuring	discrete	character-244	93	-	-
7	ARTEndEv ..	ARTEndEventType	discrete	numeric-12.0	5874	65547	-

File AHRI.PANGEA1.INDIVIDUALS.2019.v1

#	Name	Label	Type	Format	Valid	Invalid	Question
1	PangeaId	Pangea unique Identifier of Individual	discrete	character-244	3890	-	-
2	SampleId	Sample identifier	discrete	character-244	3890	-	-
3	SampleSo ..	A study a sample was taken from	discrete	numeric-12.0	3890	0	-
4	Source	-	discrete	character-244	3890	-	-
5	DoB	Individual's Date of Birth	discrete	character-11	3890	-	-
6	Age	Individual's age	continuous	numeric-12.0	3890	0	-
7	Sex	Gender	discrete	numeric-12.0	3889	1	-
8	SampleDate	Date when sample was taken	discrete	character-11	3890	-	-
9	ACDIS_Id	Surveillance unique Identifier of Individual	continuous	numeric-12.0	2736	1154	-
10	ARTEMIS_Id	ARTemis unique Identifier of Individual	continuous	numeric-12.0	1343	2547	-
11	ACCDB_Id	Tier.Net unique Identifier of Individual	continuous	numeric-12.0	2574	1316	-
12	TasP_Id	TasP unique Identifier of Individual	continuous	numeric-12.0	2538	1352	-
13	LatestNe ..	Date of last negative	discrete	character-11	304	-	-
14	Earliest ..	Date of earliest positive	discrete	character-11	3879	-	-
15	Earliest ..	Date of earliest known antiretroviral therapy	discrete	character-11	3585	-	-

File AHRI.PANGEA1.LAB.2019.v1

#	Name	Label	Type	Format	Valid	Invalid	Question
1	PangeaId	Pangea unique Identifier of Individual	discrete	character-20	19197	0	-

File AHRI.PANGEA1.LAB.2019.v1

#	Name	Label	Type	Format	Valid	Invalid	Question
2	SampleId	Sample identifier	discrete	character-20	19197	0	-
3	ResultDate	Date when sample was processed	discrete	character-11	19197	-	-
4	CD4Count	CD4 Count value	continuous	numeric-10.0	13822	5375	-
5	CD4Perce ..	CD4 Percentage value	continuous	numeric-10.0	4934	14263	-
6	Viralload	Viral load value	continuous	numeric-10.0	15189	4008	-

Variables Description

Dataset contains 28 variable(s)

File : AHRI.PANGEA1.ART.2019.v1

PangeaId: Pangea unique Identifier of Individual

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=71421 /-] [Invalid=0 /-]

ARTStartDate: Date when antiretroviral therapy was initiated

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=70901 /-]

ARTEncounterDate: ART encounter or visit date

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=70901 /-]

ARTRegimen: ART regimen

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=63949 /-]

ARTEndEventDate: Date of ART end event

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=6440 /-]

ARTEndEventReasons: Reasons for ART end event occuring

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=93 /-]

Value	Label	Cases	Percentage
Entry Required		5	5.4%
Lipodystrophy		7	7.5%
Other		40	43.0%
Policy change		20	21.5%
Poor adherence		7	7.5%
Renal impairment		1	1.1%
TasP trial switch to Atripla		4	4.3%
Virological failure		9	9.7%

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.

ARTEndEventType: ARTEndEventType

Information	[Type= discrete] [Format=numeric] [Range= 1-5] [Missing=*]
Statistics [NW/ W]	[Valid=5874 /-] [Invalid=65547 /-]

Value	Label	Cases	Percentage
1	Care changed	781	13.3%
2	Care interrupted	5028	85.6%
3	Death	65	1.1%
4	Transfer out	0	
5	Lost to follow up	0	
Sysmiss		65547	

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.

File : AHRI.PANGEA1.INDIVIDUALS.2019.v1

PangeaId: Pangea unique Identifier of Individual

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3890 /-]

SampleId: Sample identifier

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3890 /-]

SampleSource: A study a sample was taken from

Information	[Type= discrete] [Format=numeric] [Range= 1-4] [Missing=*]
Statistics [NW/ W]	[Valid=3890 /-] [Invalid=0 /-]

Value	Label	Cases	Percentage
1		9	0.2%
2		2538	65.2%
3		793	20.4%
4		550	14.1%

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.

Source

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3890 /-]

Value	Label	Cases	Percentage
ACDIS		9	0.2%
CC		793	20.4%
RES		550	14.1%
TasP		2538	65.2%

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.

DoB: Individual's Date of Birth

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3890 /-]

Age: Individual's age

Information	[Type= continuous] [Format=numeric] [Range= 14-114] [Missing=*]
Statistics [NW/ W]	[Valid=3890 /-] [Invalid=0 /-] [Mean=36.354 /-] [StdDev=12.299 /-]

Sex: Gender

Information	[Type= discrete] [Format=numeric] [Range= 1-9] [Missing=*]
Statistics [NW/ W]	[Valid=3889 /-] [Invalid=1 /-]

Value	Label	Cases	Percentage
1	Male	1177	30.3%
2	Female	2712	69.7%
9	Unknown	0	
Sysmiss		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.

SampleDate: Date when sample was taken

Information	[Type= discrete] [Format=character] [Missing=*]

File : AHRI.PANGEA1.INDIVIDUALS.2019.v1

# SampleDate: Date when sample was taken	
Statistics [NW/ W]	[Valid=3890 /-]
# ACDIS_Id: Surveillance unique Identifier of Individual	
Information	[Type= continuous] [Format=numeric] [Range= 84-218561] [Missing=*]
Statistics [NW/ W]	[Valid=2736 /-] [Invalid=1154 /-]
# ARTEMIS_Id: ARTemis unique Identifier of Individual	
Information	[Type= continuous] [Format=numeric] [Range= 41-200695] [Missing=*]
Statistics [NW/ W]	[Valid=1343 /-] [Invalid=2547 /-]
# ACCDB_Id: Tier.Net unique Identifier of Individual	
Information	[Type= continuous] [Format=numeric] [Range= 4-56761] [Missing=*]
Statistics [NW/ W]	[Valid=2574 /-] [Invalid=1316 /-]
# TasP_Id: TasP unique Identifier of Individual	
Information	[Type= continuous] [Format=numeric] [Range= 9-100102] [Missing=*]
Statistics [NW/ W]	[Valid=2538 /-] [Invalid=1352 /-] [Mean=11738.361 /-] [StdDev=8344.069 /-]
# LatestNegative: Date of last negative	
Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=304 /-]
# EarliestPositive: Date of earliest positive	
Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3879 /-]
# EarliestKnownART: Date of earliest known antiretroviral therapy	
Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=3585 /-]

File : AHRI.PANGEA1.LAB.2019.v1

PangeaId: Pangea unique Identifier of Individual

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=19197 /-] [Invalid=0 /-]

SampleId: Sample identifier

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=19197 /-] [Invalid=0 /-]

ResultDate: Date when sample was processed

Information	[Type= discrete] [Format=character] [Missing=*]
Statistics [NW/ W]	[Valid=19197 /-]

CD4Count: CD4 Count value

Information	[Type= continuous] [Format=numeric] [Range= 0-4495] [Missing=*/10001]
Statistics [NW/ W]	[Valid=13822 /-] [Invalid=5375 /-]

Value	Label	Cases	Percentage
10001	..		

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.

CD4Percentage: CD4 Percentage value

Information	[Type= continuous] [Format=numeric] [Range= 0-66] [Missing=*/101]
Statistics [NW/ W]	[Valid=4934 /-] [Invalid=14263 /-]

Value	Label	Cases	Percentage
101	..		

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.

Viralload: Viral load value

Information	[Type= continuous] [Format=numeric] [Range= 0-11000000] [Missing=*/100000001]
Statistics [NW/ W]	[Valid=15189 /-] [Invalid=4008 /-]

Value	Label	Cases	Percentage
100000001	..		

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.