



News from the archaeogenetic genotype data management system Poseidon

Clemens Schmid



MAX PLANCK INSTITUTE
FOR EVOLUTIONARY ANTHROPOLOGY



MAX PLANCK INSTITUTE
FOR GEOANTHROPOLOGY

Who is Poseidon?



Ayshin Ghalichi



Wolfgang Haak



Stephan Schiffels



Thiseas Lamnidis



Dhananjaya
Athanayaka



Clemens Schmid

Who is Poseidon?



Ayshin Ghalichi



Wolfgang Haak



Stephan Schiffels



Thiseas Lamnidis



Dhananjaya
Athanayaka



Clemens Schmid

+ Power users, who

- ...report issues
- ...suggest features
- ...prepare packages
- ...share knowledge
- ...

What is Poseidon?

Data management system to handle genotype data with context information



What is Poseidon?

Data management system to handle genotype data with context information

1. A data format: The Poseidon package



What is Poseidon?

Data management system to handle genotype data with context information

1. A data format: The Poseidon package
2. Central repositories for published data



What is Poseidon?

Data management system to handle genotype data with context information

1. A data format: The Poseidon package
2. Central repositories for published data
3. Software to manage and analyse Poseidon packages



Motivation

New archaeogenetics projects require heaps of published genotype data

Motivation

New archaeogenetics projects require heaps of published genotype data

We don't want to collect this data from scratch for every project

Motivation

New archaeogenetics projects require heaps of published genotype data

We don't want to collect this data from scratch for every project

Instead:

- in one place
- complete and up-to-date
- with meta- and context information
- well-structured
- with good interfaces and software

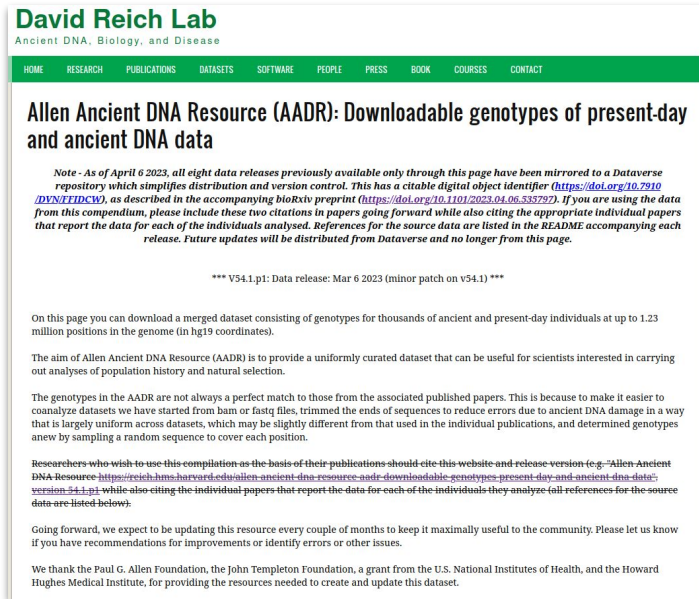
Motivation

New archaeogenetics projects require heaps of published genotype data

We don't want to collect this data from scratch for every project

Instead:

- **in one place**
- **complete** and **up-to-date**
- with meta- and **context information**
- well-structured
- with good interfaces and software



The screenshot shows the David Reich Lab website with a green header. The main heading is "Allen Ancient DNA Resource (AADR): Downloadable genotypes of present-day and ancient DNA data". Below this is a note about data releases being mirrored to a Dataverse repository. A status line indicates a data release on March 6, 2023. The text describes the dataset as a merged set of genotypes for thousands of individuals, covering up to 1.23 million positions in the genome. It states the aim is to provide a uniformly curated dataset for population history and natural selection analyses. It also mentions that the genotypes are not always a perfect match to published papers due to trimming for damage and sampling. A citation instruction is provided for researchers. A future update plan is mentioned, and the page concludes with acknowledgments to funding sources.

David Reich Lab
Ancient DNA, Biology, and Disease

HOME RESEARCH PUBLICATIONS DATASETS SOFTWARE PEOPLE PRESS BOOK COURSES CONTACT

Allen Ancient DNA Resource (AADR): Downloadable genotypes of present-day and ancient DNA data

Note - As of April 6 2023, all eight data releases previously available only through this page have been mirrored to a Dataverse repository which simplifies distribution and version control. This has a citable digital object identifier (<https://doi.org/10.7910/DVN/ZTHDCW>), as described in the accompanying bioRxiv preprint (<https://doi.org/10.1101/2023.04.06.535792>). If you are using the data from this compendium, please include these two citations in papers going forward while also citing the appropriate individual papers that report the data for each of the individuals analysed. References for the source data are listed in the README accompanying each release. Future updates will be distributed from Dataverse and no longer from this page.

*** V54.1.p1: Data release: Mar 6 2023 (minor patch on v54.1) ***

On this page you can download a merged dataset consisting of genotypes for thousands of ancient and present-day individuals at up to 1.23 million positions in the genome (in hg19 coordinates).

The aim of Allen Ancient DNA Resource (AADR) is to provide a uniformly curated dataset that can be useful for scientists interested in carrying out analyses of population history and natural selection.

The genotypes in the AADR are not always a perfect match to those from the associated published papers. This is because to make it easier to co-analyze datasets we have started from bam or fastq files, trimmed the ends of sequences to reduce errors due to ancient DNA damage in a way that is largely uniform across datasets, which may be slightly different from that used in the individual publications, and determined genotypes anew by sampling a random sequence to cover each position.

Researchers who wish to use this compilation as the basis of their publications should cite this website and release version (e.g. "Allen Ancient DNA Resource" <https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data>, version 54.1.p1) while also citing the individual papers that report the data for each of the individuals they analyze (all references for the source data are listed below).

Going forward, we expect to be updating this resource every couple of months to keep it maximally useful to the community. Please let us know if you have recommendations for improvements or identify errors or other issues.

We thank the Paul G. Allen Foundation, the John Templeton Foundation, a grant from the U.S. National Institutes of Health, and the Howard Hughes Medical Institute, for providing the resources needed to create and update this dataset.

The Poseidon package

POSEIDON.yml

```
title: ...  
version: ...  
maintainer: ...
```

The Poseidon package

POSEIDON.yml

```
title: ...  
version: ...  
maintainer: ...  
file paths:  
  - Genotype data
```

.geno

.snp

.ind

The Poseidon package

POSEIDON.yml

```
title: ...  
version: ...  
maintainer: ...  
file paths:  
  - Genotype data  
  - Context info  
  - Bibliography  
  - Sequencing  
    info
```

.geno

.snp

.ind

.janno

.bib

.ssf

.janno

.bib

.ssf

.janno

.tsv file with meta- and context data for each individual in a Poseidon package

- Each row features information for one individual
- The columns are defined as a well-specified set of variables

.bib

.ssf

.janno

Identifiers

Poseidon_ID
Alternative_IDs
Collection_ID
Group_Name

Bio. relatedness

Relation_To
Relation_Degree
Relation_Type
Relation_Note

Spatial position

Country
Country_ISO
Location
Site
Latitude
Longitude

Temporal position

Date_Type
Date_C14_Labnr
Date_C14_Uncal_BP
Date_C14_Uncal_BP_Err
Date_BC_AD_Start
Date_BC_AD_Median
Date_BC_AD_Stop
Date_Note

Individual properties

Genetic_Sex
MT_Haplogroup
Y_Haplogroup

Library properties

Source_Tissue
Nr_Libraries
Capture_Type
UDG
Library_Names
Library_Built
Genotype_Ploidy
Data_Preparation_Pipeline_URL

Data yield

Endogenous
Nr_SNPs
Coverage_on_
Target_SNPs

Data quality

Damage
Contamination
Contamination_Err
Contamination_Meas
Contamination_Note

Context information

Genetic_Source_
Accession_IDs
Primary_Contact
Publication
Note
Keywords

Arbitrary, additional columns

.janno

Identifiers

Poseidon_ID
Alternative_IDs
Collection_ID
Group_Name

Bio. relatedness

Relation_To
Relation_Degree
Relation_Type
Relation_Note

Spatial position

Country
Country_ISO
Location
Site
Latitude
Longitude

Temporal position

Date_Type
Date_C14_Labnr
Date_C14_Uncal_BP
Date_C14_Uncal_BP_Err
Date_BC_AD_Start
Date_BC_AD_Median
Date_BC_AD_Stop
Date_Note

Individual properties

Genetic_Sex
MT_Haplogroup
Y_Haplogroup

Library properties

Source_Tissue
Nr_Libraries
Capture_Type
UDG
Library_Names
Library_Built
Genotype_Ploidy
Data_Preparation_Pipeline_URL

Data yield

Endogenous
Nr_SNPs
Coverage_on_
Target_SNPs

Data quality

Damage
Contamination
Contamination_Err
Contamination_Meas
Contamination_Note

Context information

Genetic_Source_
Accession_IDs
Primary_Contact
Publication
Note
Keywords

Arbitrary, additional columns

.janno

Identifiers

Poseidon_ID
Alternative_IDs
Collection_ID
Group_Name

Bio. relatedness

Relation_To
Relation_Degree
Relation_Type
Relation_Note

Spatial position

Country
Country_ISO
Location
Site
Latitude
Longitude

Temporal position

Date_Type
Date_C14_Labnr
Date_C14_Uncal_BP
Date_C14_Uncal_BP_Err
Date_BC_AD_Start
Date_BC_AD_Median
Date_BC_AD_Stop
Date_Note

Individual properties

Genetic_Sex
MT_Haplogroup
Y_Haplogroup

Library properties

Source_Tissue
Nr_Libraries
Capture_Type
UDG
Library_Names
Library_Built
Genotype_Ploidy
Data_Preparation_Pipeline_URL

Data yield

Endogenous
Nr_SNPs
Coverage_on_
Target_SNPs

Data quality

Damage
Contamination
Contamination_Err
Contamination_Meas
Contamination_Note

Context information

Genetic_Source_
Accession_IDs
Primary_Contact
Publication
Note
Keywords

Arbitrary, additional columns

.janno

.tsv file with meta- and context data for each individual in a Poseidon package

- Each row features information for one individual
- The columns are defined as a well-specified set of variables
- Human- and machine-readable
- Software validation

.bib

.ssf

.janno

.tsv file with meta- and context data for each individual in a Poseidon package

- Each row features information for one individual
- The columns are defined as a well-specified set of variables
- Human- and machine-readable
- Software validation

.bib

BibTeX file with all literature references that are relevant for a Poseidon package

- Each entry describes one paper in a standard format
- Can be rendered to any citation style

.ssf

.janno

.tsv file with meta- and context data for each individual in a Poseidon package

- Each row features information for one individual
- The columns are defined as a well-specified set of variables
- Human- and machine-readable
- Software validation

.bib

BibTeX file with all literature references that are relevant for a Poseidon package

- Each entry describes one paper in a standard format
- Can be rendered to any citation style
- Linked to the individuals in the .janno file with a 1:n relationship
- Carried along upon merging/subsetting with trident

.ssf

.janno

.tsv file with meta- and context data for each individual in a Poseidon package

- Each row features information for one individual
- The columns are defined as a well-specified set of variables
- Human- and machine-readable
- Software validation

.bib

BibTeX file with all literature references that are relevant for a Poseidon package

- Each entry describes one paper in a standard format
- Can be rendered to any citation style
- Linked to the individuals in the .janno file with a 1:n relationship
- Carried along upon merging/subsetting with trident

.ssf

.tsv file with metadata for raw sequencing data of individuals in a Poseidon package

.janno

.tsv file with meta- and context data for each individual in a Poseidon package

- Each row features information for one individual
- The columns are defined as a well-specified set of variables
- Human- and machine-readable
- Software validation

.bib

BibTeX file with all literature references that are relevant for a Poseidon package

- Each entry describes one paper in a standard format
- Can be rendered to any citation style
- Linked to the individuals in the .janno file with a 1:n relationship
- Carried along upon merging/subsetting with trident

.ssf

.tsv file with metadata for raw sequencing data of individuals in a Poseidon package

- Each row represents a sequencing entity (library)
- The columns feature processing- and download info

<input type="checkbox"/> base_count	<input type="checkbox"/> broker_name	<input type="checkbox"/> center_name
<input type="checkbox"/> cram_index_aspera	<input type="checkbox"/> cram_index_ftp	<input type="checkbox"/> cram_index_galaxy
<input type="checkbox"/> experiment_accession	<input type="checkbox"/> experiment_alias	<input type="checkbox"/> experiment_title
<input type="checkbox"/> fastq_aspera	<input type="checkbox"/> fastq_bytes	<input type="checkbox"/> fastq_ftp
<input type="checkbox"/> fastq_galaxy	<input type="checkbox"/> fastq_md5	<input type="checkbox"/> first_created
<input type="checkbox"/> first_public	<input type="checkbox"/> instrument_model	<input type="checkbox"/> instrument_platform
<input type="checkbox"/> last_updated	<input type="checkbox"/> library_layout	<input checked="" type="checkbox"/> library_name
<input type="checkbox"/> library_selection	<input type="checkbox"/> library_source	<input type="checkbox"/> library_strategy
<input type="checkbox"/> nominal_length	<input type="checkbox"/> nominal_sdev	<input checked="" type="checkbox"/> read_count
<input type="checkbox"/> run_accession	<input type="checkbox"/> run_alias	<input checked="" type="checkbox"/> sample_accession
<input type="checkbox"/> sample_alias	<input type="checkbox"/> sample_title	<input type="checkbox"/> scientific_name
<input type="checkbox"/> secondary_sample_accession	<input type="checkbox"/> secondary_study_accession	<input type="checkbox"/> sra_aspera
<input type="checkbox"/> sra_bytes	<input type="checkbox"/> sra_ftp	<input type="checkbox"/> sra_galaxy
<input type="checkbox"/> sra_md5	<input type="checkbox"/> study_accession	<input type="checkbox"/> study_alias
<input type="checkbox"/> study_title	<input type="checkbox"/> submission_accession	<input type="checkbox"/> submitted_aspera
<input type="checkbox"/> submitted_bytes	<input type="checkbox"/> submitted_format	<input type="checkbox"/> submitted_ftp
<input type="checkbox"/> submitted_galaxy	<input type="checkbox"/> submitted_md5	<input type="checkbox"/> tax_id

Download report: [JSON](#) [TSV](#)

 Download Files as ZIP [Download](#)

Sample Accession	Library Name	Read Count
SAMEA7050404	Ash002_all	6,471,092

.ssf columns

poseidon_IDS
 udg
 library_built
 sample_accession
 study_accession
 run_accession
 sample_alias
 secondary_sample_
 accession
 first_public
 last_updated
 instrument_model
 library_layout
 library_source
 instrument_platform
 library_name
 library_strategy
 fastq_ftp
 fastq_aspera
 fastq_bytes
 fastq_md5
 read_count
 submitted_ftp

.ssf

.tsv file with metadata for raw sequencing data of individuals in a Poseidon package

- Each row represents a sequencing entity (library)
- The columns feature processing- and download info

<input type="checkbox"/> base_count	<input type="checkbox"/> broker_name	<input type="checkbox"/> center_name
<input type="checkbox"/> cram_index_aspera	<input type="checkbox"/> cram_index_ftp	<input type="checkbox"/> cram_index_galaxy
<input type="checkbox"/> experiment_accession	<input type="checkbox"/> experiment_alias	<input type="checkbox"/> experiment_title
<input type="checkbox"/> fastq_aspera	<input type="checkbox"/> fastq_bytes	<input type="checkbox"/> fastq_ftp
<input type="checkbox"/> fastq_galaxy	<input type="checkbox"/> fastq_md5	<input type="checkbox"/> first_created
<input type="checkbox"/> first_public	<input type="checkbox"/> instrument_model	<input type="checkbox"/> instrument_platform
<input type="checkbox"/> last_updated	<input type="checkbox"/> library_layout	<input checked="" type="checkbox"/> library_name
<input type="checkbox"/> library_selection	<input type="checkbox"/> library_source	<input type="checkbox"/> library_strategy
<input type="checkbox"/> nominal_length	<input type="checkbox"/> nominal_sdev	<input checked="" type="checkbox"/> read_count
<input type="checkbox"/> run_accession	<input type="checkbox"/> run_alias	<input checked="" type="checkbox"/> sample_accession
<input type="checkbox"/> sample_alias	<input type="checkbox"/> sample_title	<input type="checkbox"/> scientific_name
<input type="checkbox"/> secondary_sample_accession	<input type="checkbox"/> secondary_study_accession	<input type="checkbox"/> sra_aspera
<input type="checkbox"/> sra_bytes	<input type="checkbox"/> sra_ftp	<input type="checkbox"/> sra_galaxy
<input type="checkbox"/> sra_md5	<input type="checkbox"/> study_accession	<input type="checkbox"/> study_alias
<input type="checkbox"/> study_title	<input type="checkbox"/> submission_accession	<input type="checkbox"/> submitted_aspera
<input type="checkbox"/> submitted_bytes	<input type="checkbox"/> submitted_format	<input type="checkbox"/> submitted_ftp
<input type="checkbox"/> submitted_galaxy	<input type="checkbox"/> submitted_md5	<input type="checkbox"/> tax_id

Download report: [JSON](#) [TSV](#)

 Download Files as ZIP [Download](#)

Sample Accession	Library Name	Read Count
SAMEA7050404	Ash002_all	6,471,092

.ssf columns

poseidon_IDs
 udg
 library_built
 sample_accession
 study_accession
 run_accession
 sample_alias
 secondary_sample_
 accession
 first_public
 last_updated
 instrument_model
 library_layout
 library_source
 instrument_platform
 library_name
 library_strategy
 fastq_ftp
 fastq_aspera
 fastq_bytes
 fastq_md5
 read_count
 submitted_ftp

.ssf

.tsv file with metadata for raw sequencing data of individuals in a Poseidon package

- Each row represents a sequencing entity (library)
- The columns feature processing- and download info
- Linked to individuals with an n:n foreign-key relationship



.ssf

.tsv file with metadata for raw sequencing data of individuals in a Poseidon package

- Each row represents a sequencing entity (library)
- The columns feature processing- and download info
- Linked to individuals with an n:n foreign-key relationship
- Lays the foundation for the **Minotaur pipeline**



Pipeline for reproducible
genotype generation from
ENA/SRA entries with a
semi-automatic interface
on GitHub

.ssf

.tsv file with metadata for raw
sequencing data of individuals
in a Poseidon package

- Each row represents a sequencing entity (library)
- The columns feature processing- and download info
- Linked to individuals with an n:n foreign-key relationship
- Lays the foundation for the **Minotaur pipeline**

Public data repositories

PMA

- Poseidon packages for published papers
- FAIR
- controlled by the community
- versioned
- accessible from an open webserver
- uniformly processed genotype data



Pipeline for reproducible genotype generation from ENA/SRA entries with a semi-automatic interface on GitHub

Added in v2.7.0
(March 2023)

.ssf

.tsv file with metadata for raw sequencing data of individuals in a Poseidon package

- Each row represents a sequencing entity (library)
- The columns feature processing- and download info
- Linked to individuals with an n:n foreign-key relationship
- Lays the foundation for the **Minotaur pipeline**

Public data repositories

PMA

- Poseidon packages for published papers
- FAIR
- controlled by the community
- versioned
- accessible from an open webserver
- uniformly processed genotype data

PCA



- Poseidon packages for published papers
- FAIR
- controlled by the community
- versioned
- accessible from an open webserver
- ~~- uniformly processed~~ genotype data

Public data repositories

PMA

- Poseidon packages for published papers
- FAIR
- controlled by the community
- versioned
- accessible from an open webserver
- uniformly processed genotype data

PCA



- Poseidon packages for published papers
- FAIR
- controlled by the community
- versioned
- accessible from an open webserver
- ~~- uniformly processed~~ genotype data

PAA

- Poseidonized version of the AADR

Public data repositories

PMA

- Poseidon packages for published papers
- FAIR
- controlled by the community
- versioned
- accessible from an open webserver
- uniformly processed genotype data

PCA



- Poseidon packages for published papers
- FAIR
- controlled by the community
- versioned
- accessible from an open webserver
- ~~- uniformly processed~~ genotype data

PAA

- Poseidonized version of the AADR

Features:

- Fixed inconsistencies
- Machine-readable ^{14}C dates
- BibTeX entries for each individual
- Compatible with Poseidon software

Software tools



qjanno

Data querying



trident

Data handling



xerxes

Data analysis



janno

R interface

Current main projects

- Adding hundreds of .ssf files to the community-archive repository
- Implementing and testing the Minotaur pipeline
- Writing the Poseidon paper



Poseidon



Format definition and code openly available



Software tools precompiled for all major OSs



Open data API and community-based data curation



Communication via Mastodon and a blog



Extensive documentation on poseidon-adna.org