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 but future-proof
- complete and up-to-date
- with meta- and context information
- well-structured
- with good interfaces and software
- community-maintained

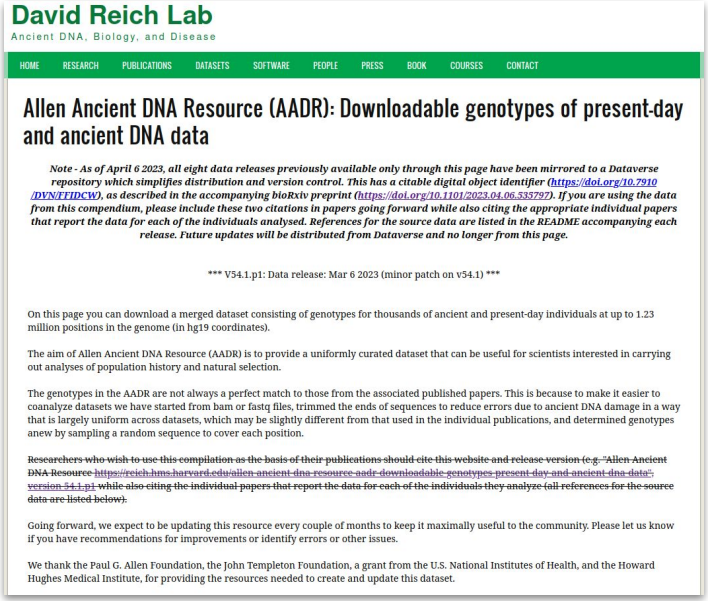
Motivation

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** but future-proof
- **complete** and up-to-date
- with meta- and **context information**
- well-structured
- with good interfaces and software
- community-maintained



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/FFIDCW), as described in the accompanying bioRxiv preprint (https://doi.org/10.1101/2023.04.06.535797). 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 coanalyze 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 columns

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 columns

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 columns

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`

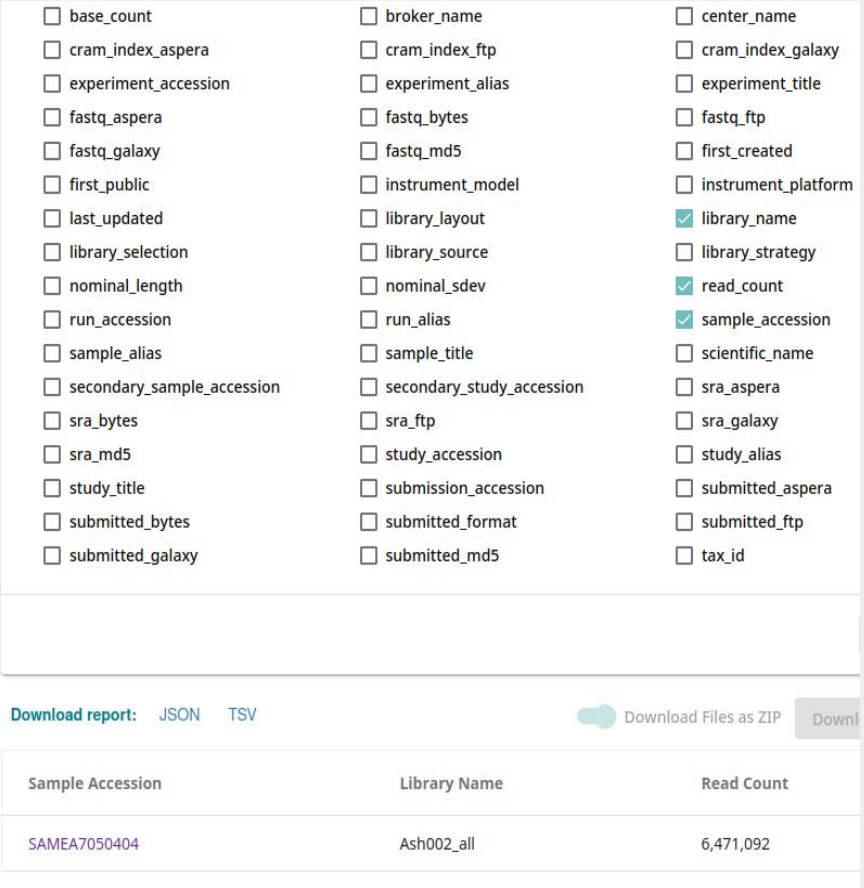
BibTeX file with all literature references that are relevant for a Poseidon package

- 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



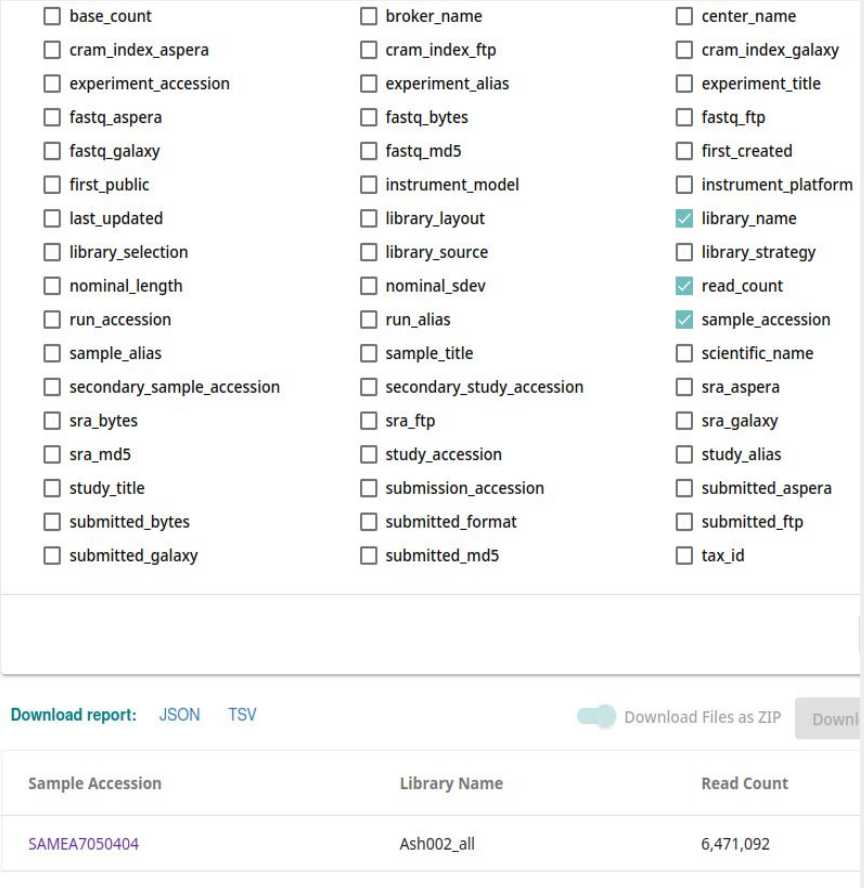
.ssf columns

```
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



.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
- controlled by the community
- Git 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

`.ssf`

Added in v2.7.0 (March 2023)

.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
- controlled by the community
- Git versioned
- accessible from an open webserver
- uniformly processed genotype data

PCA

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

Public data repositories

PMA

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

PCA

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

PAA

- Poseidonized version of the AADR

Public data repositories

PMA

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

PCA

- Poseidon packages for published papers
- controlled by the community
- Git versioned
- accessible from an open webserver
- ~~uniformly processed~~ author-submitted 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

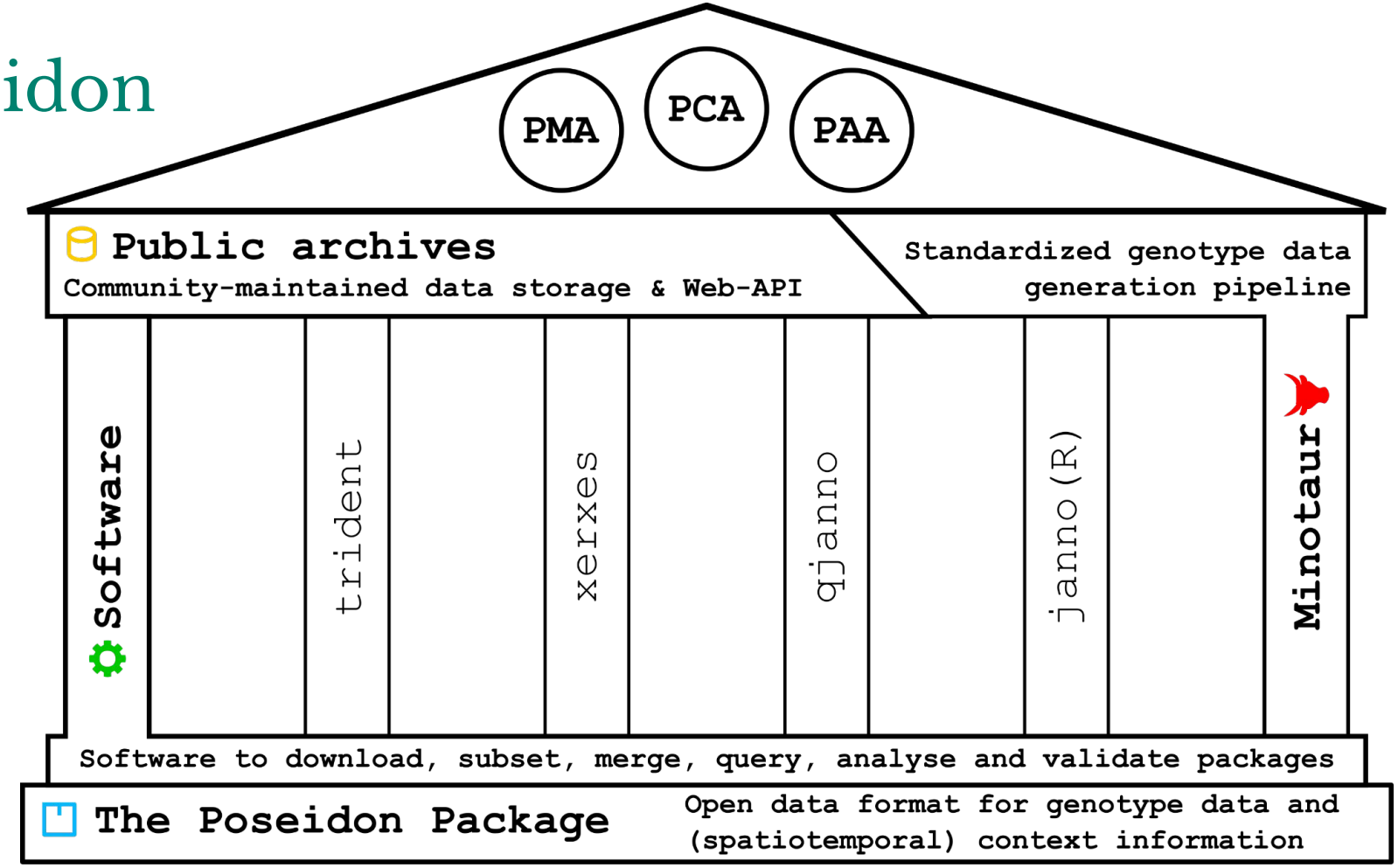
xerxes
Data analysis

janno
R interface

Poseidon

- Open package format specification
- Open data archives with community-based curation and web API
- Open source software tools precompiled for all major OSs
- Communication via GitHub, Mastodon and a blog
- Extensive documentation on poseidon-adna.org

Poseidon


PMA
PCA
PAA
Public archives
Community-maintained data storage & Web-API
Standardized genotype data generation pipeline
Software
trident
xerxes
qjanno
janno(R)
Minotaur
Software to download, subset, merge, query, analyse and validate packages
The Poseidon Package
Open data format for genotype data and (spatiotemporal) context information