poseidon

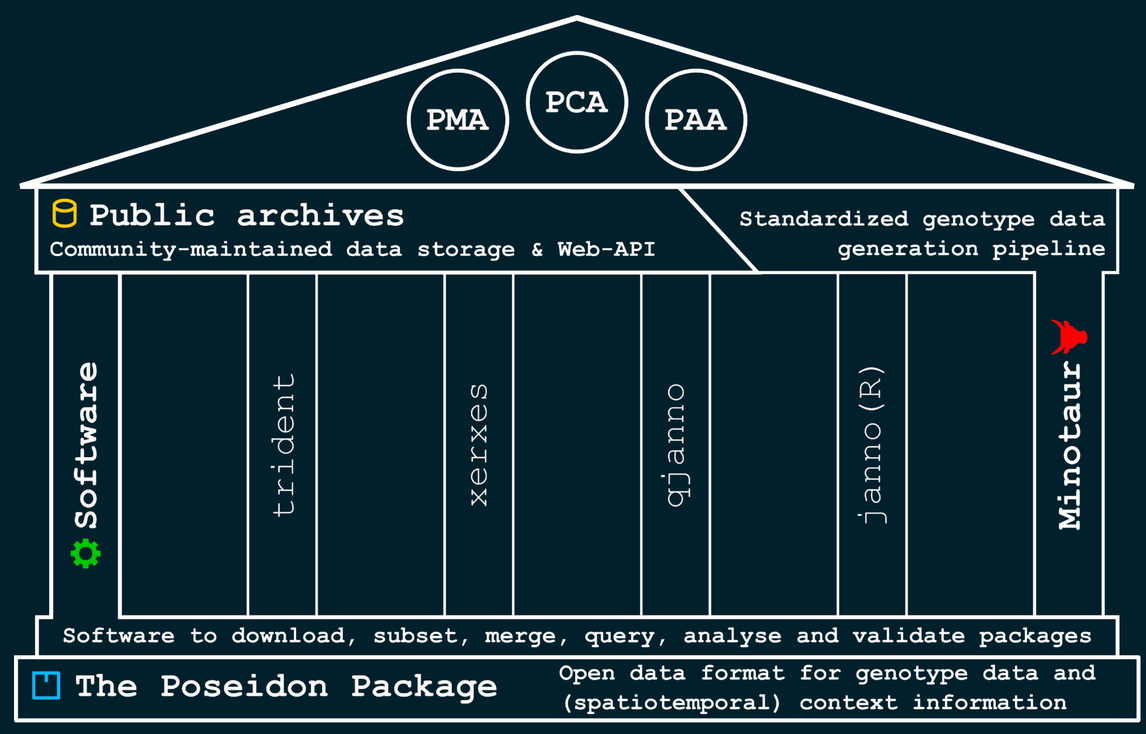
Powerful and FAIR archaeogenetic genotype data management

Clemens Schmid[1,2], Thiseas C. Lamnidis[1], Ayshin Ghalichi[1], Dhananjaya B. A. Mudiyanselage[1,3], Wolfgang Haak[1] and Stephan Schiffels[1]

1: Max Planck Institute for Evolutionary Anthropology, Department of Archaeogenetics, Leipzig, Germany; 2: Max Planck Institute for Geoanthropology, International Max Planck Research School for the Science of Human History, Jena, Germany; 3: Universität des Saarlandes, Saarbrücken, Germany

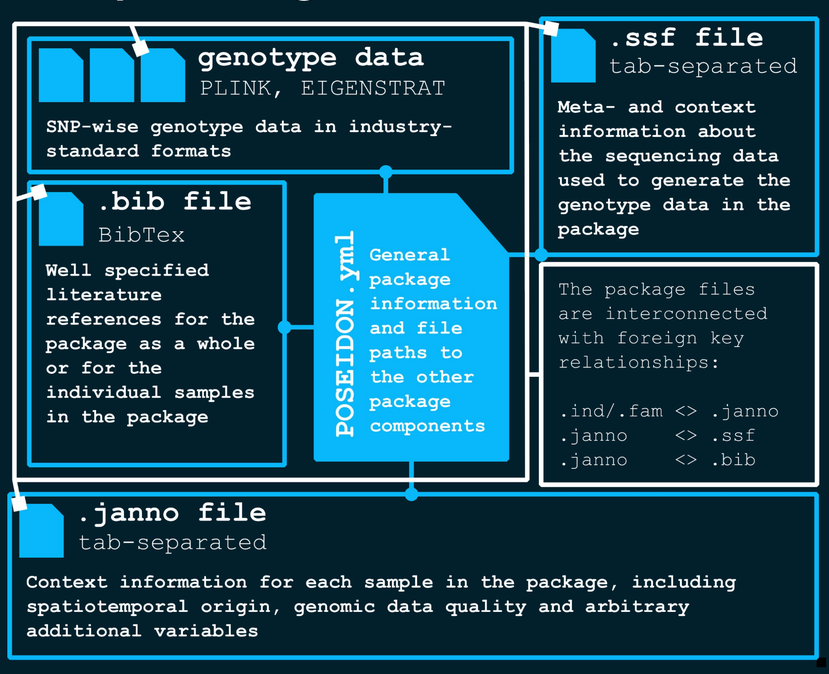
What is Poseidon?

Human archaeogenetics is a fast accelerating field, with new data being published faster than individual researchers can keep track of and co-analyze. Recently, the threshold of genome-wide data for 10,000 ancient human individuals was surpassed[1], with many of these samples featuring rich metadata ranging from archaeological field observations to radiocarbon dating. **Poseidon is an open computational framework to enable standardised and FAIR[2] handling of genotypes with this highly relevant context information.** It includes a well-specified data format, advanced software tools, and public, community-maintained archives to support the entire archaeogenetic research cycle, from data acquisition to management, analysis and publication.



The Poseidon package

A Poseidon package bundles genotype data in EIGENSTRAT / PLINK format with human- and machine-readable meta-data. This includes sample-wise context like spatio-temporal origin and genetic data quality in the .janno, literature in the .bib, and pointers to sequencing data in the .ssf file. .janno and .ssf have predefined variables, but can store arbitrary additional information.



The public archives

We created public archives for Poseidon packages to establish a central, community-maintained access point for published data in Poseidon format. The data is versioned with Git and hosted on GitHub for easy co-editing and automatic structural validation. It can be accessed through a Web-API with various endpoints at `server.poseidon-adna.org`.

Minotaur Archive
Per-paper packages with genotype data reprocessed by the Minotaur workflow

Community Archive
Per-paper packages with published genotype data, partially pre-populated from the AADR[4]

AADR Archive
Complete and structurally unified AADR releases repackaged in the Poseidon package format

The software tools

trident is a command line software tool to create, download, inspect and merge Poseidon packages – and therefore the central tool of the Poseidon framework. The `init` subcommand creates new packages from genotype data, `fetch` downloads them from the public archives through the Web-API, and `forge` merges and subsets them as specified. `list` gives an overview over entities in a set of packages and `validate` confirms their structural integrity.

```
trident list --remote --individuals --raw | grep Finland
trident fetch -d . -f "*2018_Lamnidis_Fennoscandia*"
trident forge -d . -d ~ -f "Finland_Levanluhta,-<JK1963>,<Ind>" -o mix
```

Code: A basic `trident` workflow to explore the public data archive, download relevant packages and create a new package from the downloaded and a local data collection.

xerxes is derived from `trident` and allows to directly perform various basic and experimental genomic data analyses on Poseidon packages. It implements allele sharing statistics (F_2, F_3, F_4, F_{ST}) with a flexible permutation interface.

janno is an R package to simplify reading .janno files into R and the popular tidyverse[3] ecosystem. It provides an S3 class `janno` that inherits from `tibble`.

qjanno is another command line tool to perform SQL queries on .janno files. On startup it creates a database in memory and reads .janno files into it. It then sends any user-provided SQL query to the database server and forwards its output.

```
qjanno " SELECT Country, COUNT (*) AS n
         FROM '2018_Lamnidis_Fennoscandia.janno'
         WHERE Date_Type <> 'modern' GROUP BY Country "
```

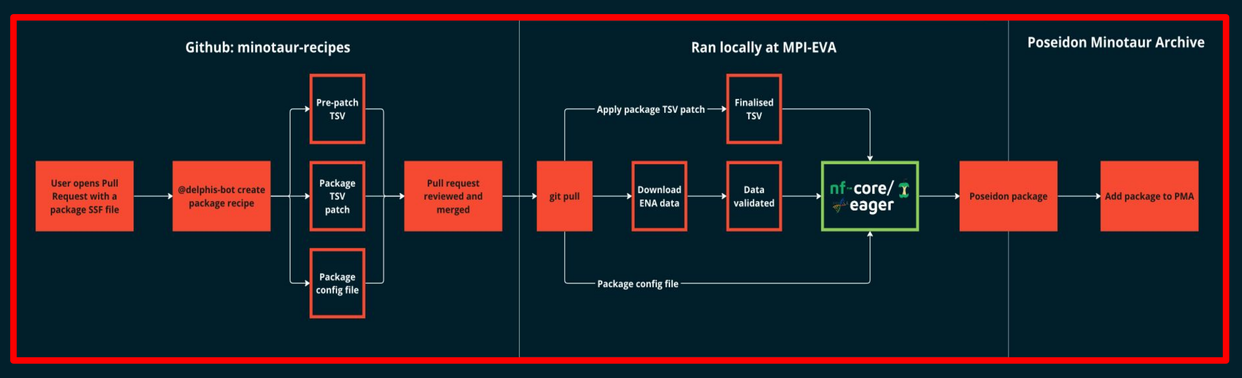
```
.---------.---.
| Country | n |
:=========:===:
| Finland | 7 |
| Russia  | 8 |
'---------'---'
```

Code: Using `qjanno` to load a .janno file, remove modern data, group by the country-of-origin variable and then return a table of countries and their number of samples.

The Minotaur workflow

We have put together a semi-automatic workflow to reproducibly process published sequencing data from the International Nucleotide Sequence Database Collaboration archives into Poseidon packages. Community members can request new packages by submitting a build recipe as a Pull Request against a dedicated GitHub repository. This recipe is created from a Sequencing Source File (.ssf), describing the sequencing data for the package and where it can be downloaded.

Using the recipe, the sequencing data gets processed through nf-core/eager[5] on computational infrastructure of MPI-EVA, using a standardised, yet flexible, set of parameters. The generated genotypes, together with descriptive statistics of the sequencing data (Endogenous, Damage, Nr_SNPs, Contamination), are compiled into a Poseidon package, and made available to users in the Minotaur Archive.



1: Ewen Callaway. "'Truly gobsmacked': Ancient-human genome count surpasses 10, 000". In: Nature 617.7959 (Apr. 2023). doi: 10.1038/d41586-023-01403-4
2: Mark D Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: Sci Data 3, 160018 (Mar. 2016). doi: 10.1038/sdata.2016.18
3: Hadley Wickham et al. "Welcome to the Tidyverse". In: JOSS 4.43 (Nov. 2019). doi: 10.21105/joss.01686
4: Swapan Mallick et al. "The Allen Ancient DNA Resource (AADR): A curated compendium of ancient human genomes". In: bioRxiv (Apr. 2023). doi: 10.1101/2023.04.06.535797
5: James A Fellows Yates et al. "Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager". In: PeerJ 9:e10947 (Mar. 2021). doi: 10.7717/peerj.10947

WEBSITE

poseidon-adna.org

MAX PLANCK INSTITUTE OF GEOANTHROPOLOGY

MAX PLANCK INSTITUTE FOR EVOLUTIONARY ANTHROPOLOGY

HAAM microscope

European Research Council
Established by the European Commission