Poseidon - A toolbox for archaeogenetic data management

Clemens Schmid, Max Planck Institute for the Science of Human History (MPI-SHH), schmid@shh.mpg.de
Ayshin Ghalichi, Max Planck Institute for the Science of Human History (MPI-SHH), ghalichi@shh.mpg.de
Wolfgang Haak, Max Planck Institute for the Science of Human History (MPI-SHH), haak@shh.mpg.de
Stephan Schiffels, Max Planck Institute for the Science of Human History (MPI-SHH), schiffels@shh.mpg.de

The recent increase in openly available ancient human DNA samples demands for new software solutions to store, distribute and analyse both genomic as well as archaeological context data. In this paper we present Poseidon, a computational framework including an open data format, software, and a public online repository to enable convenient, reliable, and FAIR access to genotype data from all around the world.

Archaeogenetics has become a fast accelerating field, with new data coming out faster than people can co-analyse (Orlando et al. 2021). If one considers samples currently being processed in the world's largest laboratories, we are now quickly approaching genome-wide data for 10,000 ancient human individuals. In addition, emergent fields such as ancient metagenomics and paleo-proteomics are adding complexity to a data landscape that already hosts traditional archaeological data with contributions from long-established non-genetic technologies like radiocarbon dating and stable isotope analyses.

Data from genetic analyses in academic papers is usually shared by releasing raw sequencing output into public repositories like the European Nucleotide Archive (ENA). Archaeological context information for individual samples is documented in supplementary tables attached to the publications. So while most data in the field is technically accessible, it does not satisfy the FAIR principles of open data: Findability, Accessibility, Interoperability, and Reproducibility (Wilkinson et al. 2016)

Beyond ethical concerns, this raises a number of concrete practical issues:

- Intermediate data such as genotypes or metagenomic profiles are often not released at all, making it hard to reproduce specific results.
- The connection between individuals, contextual information and genetic data is challenging to maintain across very different repositories and sources.
- Meta-analyses spanning datasets require enormous amounts of work on data collection and curation.
- Incrementally produced data, for example by adding new data to previously published individuals, cannot be easily connected to the same individuals.

To mitigate some of these problems, we propose a package data format which bundles the big genotype data in industry-standard formats (EIGENSTRAT, PLINK) with a flat context data file: The .janno file. For each sample, this

human- and machine-readable .tsv file format stores metadata (e.g. publication, keywords) together with information about spatiotemporal origin (e.g. coordinates, radiocarbon dates) and genetic data preparation context (e.g. shotgun sequencing vs. target enrichment) as well as quality markers (e.g. number of autosomal SNPs on the 1240k array). Genotype data and .janno file are supplemented with a BibTeX file for the relevant citations plus optional metadata files (README, CHANGELOG) to make the package complete.

The main workhorse for operations on these Poseidon packages is the command-line software trident. It is written in the functional programming language Haskell, builds on strong type-safety and clean interfaces, and provides modules for package creation, inspection, validation and analysis. trident handles both context data as well as the large genotype data files – the latter via stream processing. Sample entities are internally represented with specific data types, which ensures strict parsing constraints to maintain structural correctness and machine readability for all data in Poseidon. trident also serves as a command line client to download already available packages from a central online repository we host and maintain. We implemented the respective webserver relying on the same Haskell infrastructure as trident.

For integration of Poseidon packages with an R data analysis pipeline, we provide the poseidonR package to load .janno files into a tidyverse-compatible, tibble-derived S3 object. poseidonR also provides functions for bulk radiocarbon date calibration and age sampling on these janno objects.

In the spirit of the session, our presentation will highlight multiple aspects of our learning experience when developing Poseidon. This includes...

- the value we saw in a clear file format definition for the .janno file
- the advantages of a strictly typed programming language like Haskell for parsing operations
- the challenges of command line interfaces as the true lingua franca of scientific computing (Prasad et al. 2020)
- the necessity for code review, unit tests and manual alpha testing

Poseidon is a community project. Code and data are open. All documentation for the data format as well as the software tools are available online. Both for software development and for (context) data keeping we rely on version control with Git/Github.

References

Orlando, Ludovic, Robin Allaby, Pontus Skoglund, Clio Der Sarkissian, Philipp W. Stockhammer, María C. Ávila-Arcos, Qiaomei Fu, et al. 2021. "Ancient DNA Analysis." Nature Reviews Methods Primers 1 (1): 14. https://doi.org/10.1038/s43586-020-00011-0.
Wilkinson, Mark D., Michel Dumontier, I. Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR

Guiding Principles for Scientific Data Management and Stewardship." Scientific Data 3 (March): 160018. https://doi.org/10.1038/sdata.2016.18.
Aanand, Prasad, Firshman Ben, Tashian Carl, and Parish Eva. 2020. "Command Line Interface Guidelines." December 23, 2020. https://clig.dev.