University of Fribourg / Faculty of Science and Medicine / Department of Biology

Towards reliable, flexible and reproducible processing of low-depth and ancient sequencing data

Ilektra Schulz

When studying populations, the changes in genotype frequency over time, or the genetic differences between populations are investigated. This also means, the specific genotypes of individuals are often of less interest than overall population-scale differences. To capture these differences within a higher variety of samples, it has hence become common practice in population genetics to sequence a higher number of samples at a lower depth for the same budget as sequencing a lower number of samples at a higher depth. Additionally, when sequencing ancient DNA (aDNA), the fraction of endogenous DNA is usually low and the sequencing effort is mostly spent on environmental contaminants, also resulting in a low depth.

While a low sequencing depth does not allow for accurate genotype calls, it can be used to infer genetic diversity probabilistically, provided that the uncertainty in genotypes is properly accounted for. This uncertainty arises from systematic errors such as sequencing errors, which are not accurate and need to be recalibrated. When dealing with aDNA, there are additional post-mortem damages (PMD) in form of base misincorporations present in the data. With the ATLAS tool, we can reflect these uncertainties with a genotype likelihood model, that takes the error rate and PMD into account.

However, data curation, quality assessment, and data analysis, especially of numerous individuals, also goes along with a high computational effort and bioinformatic expertise. The numerous choices of methods, tools, and parameters has measurable impact on downstream analysis. Further, workflow reproducibility, portability, and consistency are main requirements to perform resourceful research.

This thesis presents a user-friendly and ready-to-use Next-Generation-Sequencing pipeline, aiming at the analysis of low-depth and aDNA samples.

The ATLAS-Pipeline covers the whole bioinformatical workflow of a project: From raw-data analysis (FASTQ to BAM file) over the calculation of reliable genotype likelihoods up to a principal component analysis (PCA). What makes this pipeline stand out against other standard pipelines in the field is that it infers population genetic quantities from genotype likelihoods directly, and that it accounts for genotype uncertainty without prior knowledge on variant sites, making it particularly applicable to non-model organisms.

Using this pipeline, we have answered questions about the Neolithisation of Europe. With demogenomic modeling, we discovered a series of previously undetected admixture events between the hunter gatherers and the ancestors of western early farmers shortly after the last glacial maximum. We further investigated an early contact zone between hunter-gatherers and early farmers in the Danube Gorges in modern-day Serbia. We found that this region, in specific the settlement of Lepenski Vir, was co-occupied by people from both genetic origins, and we could further identify multiple early-generation admixed individuals.

Jury:
Prof. Dr. Daniel Wegmann (thesis supervisor)
Dr. Benjamin Peter (external co-examiner)
Prof. Dr. Thomas Flatt (internal co-examiner)
Prof. Dr. Laure Weisskopf (president of the jury)