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

Statistical inference of genetic diversity: computational tools for low-depth and ancient sequencing data

Vivian Link

Genetic diversity is the raw material of evolution. By elucidating how diversity changes over space and time, we can learn about the evolutionary history of living beings. The information about genetic diversity is contained within the genotypes. The genotypes cannot be observed, but we can produce sequencing data in the form of sequencing reads. Producing sequencing reads can be seen as sampling from the genotypes. However, this sampling process is not perfect: there are sequencing errors, biases specific to the sequencing technique, and due to cost or physical constraints, we sometimes only obtain very few reads. In the case of ancient DNA, inferring the true genotype is further complicated by post-mortem damage, which introduces fake mutations after the death of the individual. Sequencing data may thus provide false or insufficient information about the true genotypes, which is why it is called "noisy". However, since we are not interested in the genotypes themselves, but in a global parameter of diversity, we do not need to explicitly determine what the genotypes are. Instead, we can account for the genotype uncertainty by integrating over the genotype likelihoods. This statistical framework allows the sequencing data from many sites to directly be used to infer our parameter of genetic diversity.

The goal of my PhD project was to develop genetic diversity inference methods for noisy data. Many of these methods are based on the statistical framework described above. Most are implemented in a C++ program called Analysis Tools for Low Depth and Ancient Samples (ATLAS). The genotype uncertainty can only properly be accounted for if it is correctly modeled. ATLAS ensures this by estimating post-mortem damage in ancient samples and accounting for it directly in the genotype likelihoods. Further, it estimates recalibrated sequencing error rates, which are also accounted for in the genotype likelihoods. With the help of ATLAS, we answered questions about human history. We compared the ancient genomes of early European farmers to those of early farmers from the Aegean and the eastern Fertile Crescent, now Iran. We found that the farming lifestyle spread from the Aegean to Europe due to migration, but between the Aegean and Iran farming was probably also transmitted culturally. In another project we analyzed human samples from a Bronze Age battlefield and found that the frequency of lactase persistence (LP), which is the capacity to digest milk as an adult, was low (7%). This only left 3000 years for this allele to reach the modern frequency (around 80%). We estimated a positive selection coefficient of 6%, which is strong and could be explained by other factors besides milk being a calorie supplement. Finally, we also developed a C++ program called Tools to Incorporate Genotyping ERrors (TIGER). This tool estimates the genotyping error in Restriction site Associated DNA (RAD) sequencing data based on different types of external information. It is known that RAD sequencing causes many biases in genotyping and TIGER helps researcher overcome them. ATLAS and TIGER are well- documented, publicly available and compatible with existing tools, and will help researchers answer diverse evolutionary questions on the basis of genetic diversity.

Jury:
Prof. Daniel Wegmann (thesis supervisor)
Prof. Anders Albrechsen (external co-examinator)
Prof. Thomas Flatt (internal co-examinator)
Prof. Dominique Glauser (president of the jury)