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

New methods to detect divergent selection and their application to domesticated species

Marco Galimberti

Humans have always been interested in the process of domestication, as the several domesticated species can proof. Even nowadays animals and plants are being domesticated, showing that this interest is not over. In the act of understanding this process, we have developed two different methods to infer selection using whole genome sequencing (WGS) data, and we have investigated the history of date palms.
In Chapter 1 we use WGS data to support the hypothesis emerged from the analysis of seed morphology and microsatellites of the discovery of a wild population of date palms (Phoenix dactylifera L.) in Oman. This sheds light on a complex domestication history regarding the African and the Middle Eastern cultivated date palms, supporting the scenario where previously domesticated individuals were imported from the Middle East and then crossed with local wild individuals in North Africa, referred as a secondary domestication event.
Chapter 2 introduces a method for genome-wide inference of local population mixtures. Whereas several methods have been developed to build phylogenetic trees and to infer population splits and gene flow, very little is known about the estimation of mixture proportions along the genome. A multiple-population genome-wide allele frequency dataset is provided as an input, and a Hidden Markov Model (HMM) is implemented in order to account for linkage. The hidden state of the HMM are representative of the magnitude of the mixture in a specific locus. We validate the goodness of our approach using simulations. We have implemented the method in a software package which is called TreeSwirl, which is available online with its wiki.
Inferring the strength of selection is the main topic of Chapter 3, where we detect divergent and balancing selection from linked sites. An F-model is used to describe the allele frequencies differences among populations, which are obtained from SNP data. The F-model permits to summarize these differences in allele frequencies to reflect locus and population specific contributions, which are seen as a measure of selection and drift, respectively. The introduction of linkage is provided by the implementation of a HMM, approach which is also used in Chapter 2. However, here the states of the HMM represent the strength of selection involved in the specific locus. Taking linkage into account has increased the statistical power of inferring loci under selection, when compared to previous methods that assumes independent sites. The proposed improvements were implemented in a software package, called Flink, which is available online with its wiki. We have applied it on data from the Human Genome Diversity Project, emphasizing particular attention on the lactase gene.

Jury:
Prof. Daniel Wegmann (thesis supervisor)
Prof. Jérôme Goudet (external co-examiner)
Prof. Thomas Flatt (internal co-examiner)
Prof. Jörn Dengjel (president of the jury)