Bayesian maximum a posteriori algorithms for modern and ancient DNA

Dissertation, 2016

Gabriel Renaud

When DNA is sequenced, nucleotide calls are produced along with their individual error probabilities, which are usually reported in the form of a per-base quality score. However, these quality scores have not generally been incorporated into probabilistic models as there is typically a poor correlation between the predicted and observed error rates. Computational tools aimed at sequence analysis have therefore used arbitrary cutoffs on quality scores which often unnecessarily reduce the amount of data that can be analyzed. A different approach involves recalibration of those quality scores using known genomic variants to measure empirical error rates. However, for this heuristic to work, an adequate characterization of the variants present in a population must be available -which means that this approach is not possible for a wide range of species.

This thesis develops methods to directly produce error probabilities that are representative of their empirical error rates for raw sequencing data. These can then be incorporated into Bayesian maximum a posteriori algorithms to make highly accurate inferences about the likelihood of the model that gave rise to this observed data. First, an algorithm to produce highly accurate nucleotide basecalls along with calibrated error probabilities is presented. Using the resulting data, individual reads can be robustly assigned to their samples of origin and ancient DNA fragments can be inferred even at high error rates. For archaic hominin samples, the number of DNA fragments from present-day human contamination can also be accurately quantified.

The novel algorithms developed during the course of this thesis provide an alternative approach to working with Illumina sequence data. They also provide a demonstrable improvement over existing computational methods for basecalling, inferring ancient DNA fragments, demultiplexing, and estimating present-day human contamination along with reconstruction of mitochondrial genomes in ancient hominins.