Robust and Scalable Inference of Population History and Selection from Hundreds of Whole Genomes

Abstract: Demographic inference refers to the problem of inferring past population events (migrations, admixture, expansions, etc.) from patterns of mutations found in sampled DNA. Apart from intrinsic appeal of understanding, for example, how we originated and peopled our planet, this type of analysis is useful for forming a null model of human evolution, departures from which signal the presence of natural selection, population structure, and other interesting phenomena.

In this talk I will discuss recent statistical and computational innovations which enable us to infer demography using modern data sets consisting of hundreds of whole-genome sequences obtained from populations all over the world. These include momi, a new software package for stable and rapid computation of the expected sample frequency spectrum under complex demographic scenarios involving numerous diverged populations, as well as SMC++, a new probabilistic framework which couples the genealogical process for a given individual with allele frequency information for a large panel of related samples. I will demonstrate how we are using these tools to learn about how human populations expanded in the last 12,000 years, understand the mysterious origins of ancient DNA samples, and estimate when Europeans acquired lighter skin and the ability to digest lactose. Finally, I will discuss some of the statistical aspects of these estimators, in particular an information-theoretic lower bound on the error attainable by any possible demographic inference procedure.

All relevant theory will be introduced during the talk; no prior knowledge of population genetics is assumed. Portions of this work are joint with Jack Kamm, Pier Palamara, and/or Yun Song.