Using Individual Human Genomes to Illuminate the Mysteries of

Early Human History

Ilan Gronau, postdoc at the Siepel lab, Department of Biology

Statistics and Computational Biology, Cornell University In the decade since the publication of the first draft of the human

In the decade since the publication of the first draft of the human genome sequence, there has been a growing effort to sequence more human individuals. These data provide a rich source of information about human evolution, however, this potential has not yet been fully realized. We recently developed a modeling framework based on coalescent theory, which allows inference of key demographic parameters of ancient human population history from a small number of individual genome sequences. We implemented a Bayesian inference algorithm for infering ancestral population sizes, divergence times, and migration rates from a set of sequence alignments at many neutrally evolving loci along the genome. Our algorithm, called G-PhoCS (Generalized Phylogenetic Coalescent Sampler), draws its inference from the patterns of variation in the genealogies at many neutrally evolving loci. Essentially, it exploits the fact that even small numbers of present-day genomes represent many ancestral genomes, which have been shuffled and assorted by the process of recombination. Because the sequences provide only very weak information about the genealogy at each locus, the method integrates over candidate genealogies using Markov chain Monte Carlo (MCMC) sampling. The implementation of G-Phocs was designed to be efficient enough to facilitate analysis of genome-wide sequence data from multiple individuals.

We used G-PhoCS to examine the published genome sequences of six individuals from six different population groups from East-Asia, Europe, Western and Southern Africa. One of these individuals is a member of the Khoisan-speaking hunter-gatherer population of Southern Africa, known collectively as the San. The San exhibit one of the highest known levels of genetic divergence from other human populations, and are therefore highly informative about ancient human demography. Applying G-PhoCS to these sequences, we were able to provide highly confident estimates of divergence times and ancestral population sizes, taking into account various scenarios of gene flow between populations. Our main focus was on the time of divergence of the San population from other populations, as well as the divergence of Eurasian populations from African populations (also referred to as the "out-of-Africa" date).

In this talk, I will present our Bayesian inference algorithm, G-PhoCS, highlighting the main modeling challenges tackled in its design and implementation. I will also describe the pipeline we developed for preparing the sequence data for analysis, making sure to account for various potential sources of bias. I will summarize the demographic estimates we obtained in our data analysis, and compare them to previously published estimates based on genetic data. To conclude, I will mention some projects we are currently working on in the Siepel lab, involving analysis of a recently published set of 54 individual human genomes from a wide array of populations.

See: http://www.nature.com/ng/journal/v43/n10/full/ng.937.html
Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A.   Bayesian inference of ancient human demography from individual genome sequences. Nature Genetics 43 1031–1034. 2011

G-PhoCS web site: http://compgen.bscb.cornell.edu/GPhoCS/