Inferring the demographic history of multiple populations from genomic polymorphism data

Ryan N. Gutenkunst[1], Ryan D. Hernandezr[2], Scott H. Williamson[3], and Carlos D. Bustamante[4]

***Short Abstract* — Models of population history inferred from genetic data complement archeology and serve as null models in genome scans for natural selection. Most current inference methods are computationally limited to considering simple models or non-recombining data. We introduce a method [1] based on a diffusion approximation to the joint frequency spectrum (FS) of genetic variation between populations. We apply our method to human data from Africa, Europe, and East Asia, building to date the most complex statistically well-characterized model of human migration out of Africa. This model makes testable predictions and forms a basis for future studies of natural selection in and between human populations.**

I. Methods

Genomic scale inference requires an efficient data summary; we employ the frequency spectrum (FS). Given sequence from individuals in P populations, the FS is a P-dimensional matrix. Each entry records the number of single nucleotide polymorphisms (SNPs) in which the derived allele was found in the corresponding number of samples from each population. For example, the [2,0] entry records the number of derived alleles seen twice in population 1 and zero times in population 2. In the absence of linkage, the FS is a complete summary of the data, and it is known that linkage does not bias demographic inference [2].

For any given model, we simulate the expected FS by numerically solving a diffusion equation. We then calculate the composite likelihood of the data and optimize to find the maximum-likelihood model. Because we use a composite likelihood, statistical tests must use bootstrapping, but the efficiency of our approach makes this feasible. Our implementation, ∂a∂i, can model up to three populations and is available at http://dadi.googlecode.com.

II. Application

We model human migration out of Africa, using 5 Mb of noncoding sequence generated by the Environmental Genome Project [3] from each of 12 Yoruba (YRI), 22 CEPH European (CEU) and 12 Han Chinese (CHB) individuals. Figure 1 shows the data, along with the maximum-likelihood parameters for our demographic model. These parameters offer insight into human history. For example, our inferred divergence time between Europeans and Asians of 21 thousand years ago suggests that one of these populations must not be descended from the original humans to settle the region. Recent comparisons of modern and ancient DNA in Europe have confirmed that modern Europeans are indeed descended from a later wave of settlers [4,5].

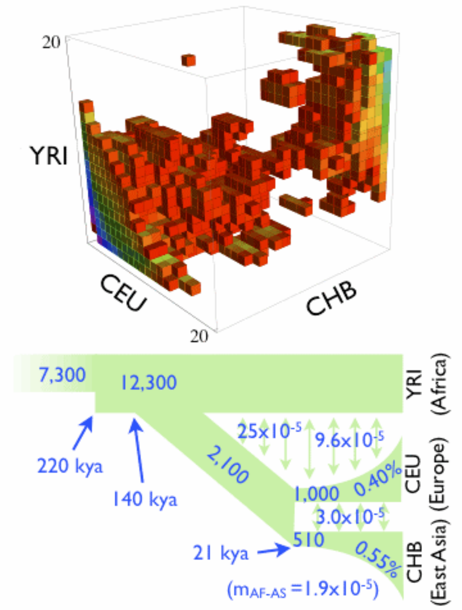

Figure 1: Genetic data and model for human expansion out of Africa [1]. Using ∂a∂i, the 14 model parameters were fit to 5 Mb of noncoding sequence. Parameter uncertainties are typically about 20%. (Parameters are drift effective sizes, divergence times in thousands of years ago, migration rates per chromosome per generation, and growth rates per generation.)

Our methodology is general and widely applicable. To our knowledge, ∂a∂i has been applied to humans [1,6,7,8], cattle [9], rice, and orangutans.

References

[1] Gutenkunst RN, et al. (2009) PLoS Genet 5:e1000695
[2] Wiuf C (2006) J Math Biol 53:821
[3] Livingston RJ, et al. (2004) Genome Res 14:1821
[4] Bramanti B, et al. (2009) Science 326:137
[5] Balaresque P, et al. (2010) PloS Biol 8:e1000285
[6] Andrés AM, et al. (2009) Mol Biol Evol 26:2755
[7] Nielsen R, et al. (2009) Genome Res 19:838
[8] Yi X, et al. (2010) Science 329:75
[9] Murray C, et al. (2010) Phil Trans R Soc B 365:2531

[1] Molecular and Cellular Biology, University of Arizona
E-mail: ryan.gutenkunst@arizona.edu
[2] Human Genetics. University of California, San Francisco
[3] Biological Statistics and Computational Biology, Cornell University
[4] Genetics, Stanford University