∂a∂i Report

Alex Landy

10/5/2015

INTRODUCTION

∂a∂i (Diffusion Approximations for Demographic Inference) is a python module that uses the allele frequencies of a set of loci from individuals from up to three populations to estimate complex demographic histories. ∂a∂i uses allele frequency data which is arranged in an allele frequency spectrum (AFS). The AFS is a series of statistics which can be used to describe genetic polymorphisms within a population, and is commonly used in the field of population genetics (Chen 2012). This module models important demographic parameters such as population size, migration rates, and divergence times among these populations. The use of AFS in ∂a∂i can be quite powerful when it contains many markers from a large contiguous genomic dataset.

To estimate demographic parameters, ∂a∂*i* uses diffusion approximation of the joint frequency spectrum (JFS) between populations (Gutenkunst et al. 2009). The JFS combines the AFS from two or more populations and can be used to infer complex demographic parameters (Chen 2012). This diffusion approximation efficiently simulates the JFS. ∂a∂i sets up a partial derivate equation (PDE) and solves this numerically and then allows for calculation of the maximum composite likelihood for a variety of demographic parameters. ∂a∂*i* incorporates the python module *MS* to generate bootstrap samples to generate a likelihood ratio test of the estimated parameters.

Aside from optimizing demographic parameters, ∂a∂i can also be used to compare competing demographic models. These models are defined *a priori* and can be used to determine divergence time, population differentiation (*Fst*), and the role of migration in population differentiation.

Here, we use ∂a∂i to test two demographic models and to estimate demographic parameters describing population growth, divergence, and subsequent migration between the Yoruba (YRI) people of Africa and CEPH Europeans founders (CEU).

METHODS

A joint frequency spectrum was previously created from 5 MB of genomic data from 12 YRE and 22 CEU individuals (Livingston et al. 2004). This dataset includes over 15,000 segregating sites. Using these data we will estimate the YRE size after growth (nu1F), CEU population size at the time of divergence/bottleneck (nu2B), population size of CEU after growth (nu2F), migration rate between YRE and CEU (m), growth rate of YRE (Tp), and time of population split (t) between two demographic models.

Two models are being compared in this analysis. These models are identical except one includes a migration rate parameter and the other does not. This will allow us to assess the importance of migration in the history of these populations.

Initially ∂a∂i requires somewhat arbitrary estimates of the demographic parameters to begin with as well as their upper and lower bounds (shown in Table 1). Defaults of these initial values are provided however, users can define these as well. Ultimately ∂a∂i will optimize these parameters. We excluded very small population sizes and high migration rates as they take considerable time to evaluate.

Table 1

Parameters	(nu1F, nu2B, nu2F, m, Tp, T)
p0	[2, 0.1, 2, 1, 0.2, 0.2]
Upper Bound	[100, 100, 100, 10, 3, 3]
Lower Bound	[1e-2, 1e-2, 1e-2, 0, 0, 0]

The primary function of ∂a∂i is to optimize these parameters. Optimization is done by using a diffusion approximation and calculating composite likelihoods. The diffusion approximation calculates the expected JFS, which can then be compared to the observed JFS allowing composite likelihoods of the parameters to be calculated. Composite likelihoods are easier and faster to calculate than correct likelihoods but still provide a similar maximum for the function to optimize.

One drawback of using composite likelihoods is that they lack confidence intervals. ∂a∂i incorporates the *MS* simulation module within python to perform a parameter bootstrap test to provide confidence intervals and to test the models.

RESULTS

∂a∂i was able to estimate the demographic parameters of the YRI and CEU populations and to estimate their demographic history (Table 2; Figure 1). The model including migration was best supported by ∂a∂i (p-value for rejecting no-migration model: 1.000). This p-value was estimated by a bootstrap performed by the module *MS* within ∂a∂i. The estimated *Fst* between these two populations was 0.158.

∂a∂i optimized the demographic parameters of our model. The estimated population size of YRI after growth (AncSize after growth) was 2.58072 (Table 2). The estimated population size of CEU at divergence (bottleneck Popsize2) was 0.08473, and this population grew to a size of 1.34003 (Popsize2) after divergence. Migration rate (migration m) between YRI and CEU was estimated to be 1.24147. YRI grew to its full size in t=0.13019 (time for ancestor to grow) and the populations split soon after reaching maximum population size, t=0.13998 (Population split). Units for each of these parameters are relative to estimated effective population size (see the ∂a∂i manual for further details).

Table 2

Parameter	Estimate
AncSize after growth	2.58072
bottleneck Popsize2	0.08473
Popsize2	1.34003
migration m	1.24147
time for ancestor to grow	0.13019
Population split	0.13998

Maximum log composite likelihood: -1153.96

Optimal value of theta: 2757.36

Figure 1

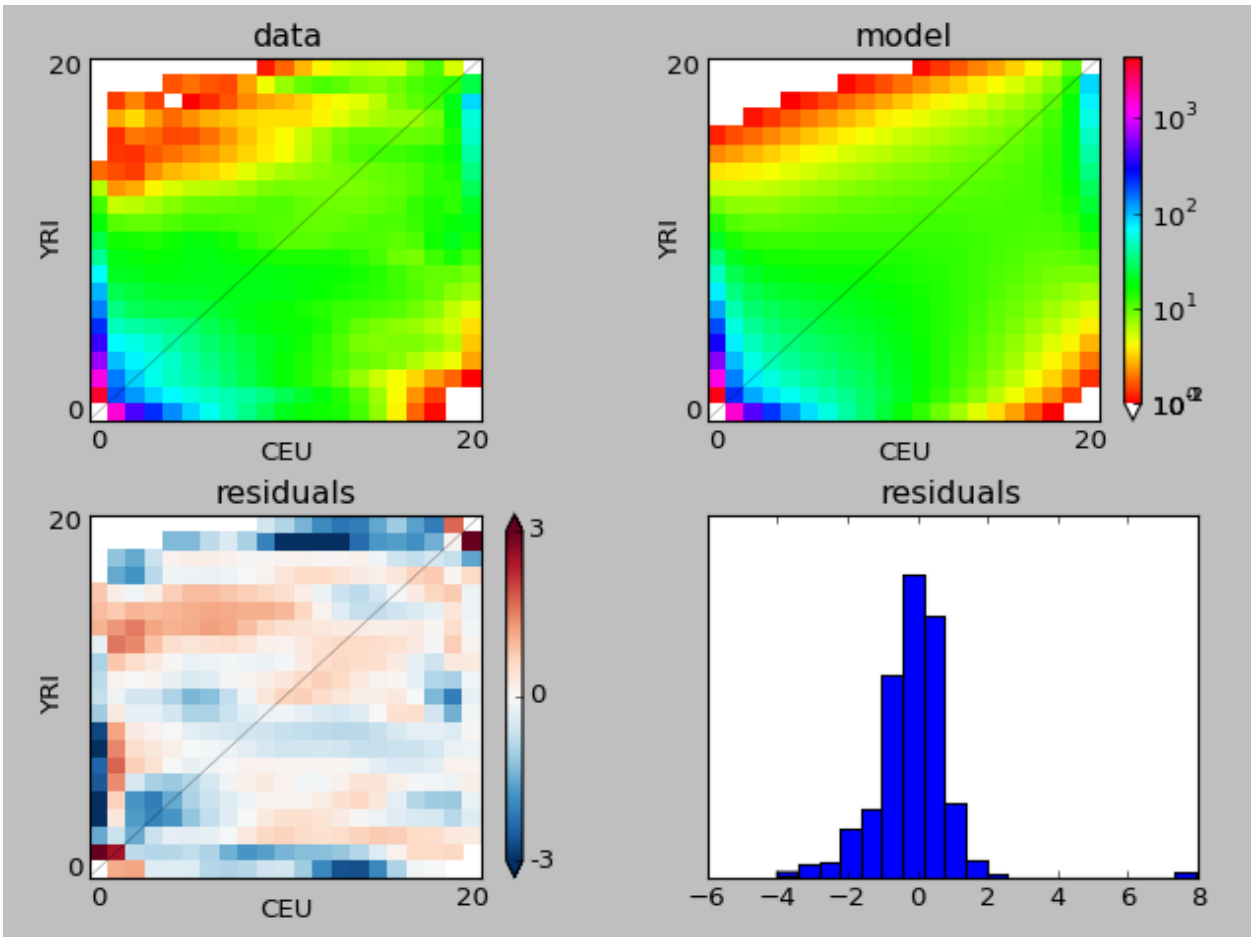

Figure 1. Plot of YRI_CEU data compared to the best-fit model data. Upper-left shows plot of the YRI_CEU joint frequency spectrum (JFS). Upper-right shows the 2D JFS of the best-fit model data (with migration). Each entry in the FS is colored according to the logarithm of the number of variants within it. Lower-left shows the residuals of the data, and lower-right represents the residuals as a histogram.

Figure 2

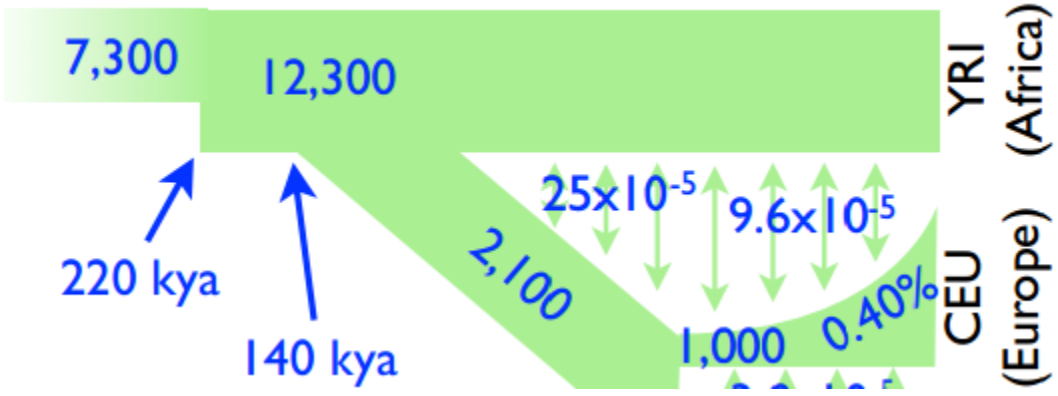

Figure 2. Modified image from Gutenkunst et al. 2009, visualizing optimization of demographic parameters in ∂a∂i and how they can be used to determine the evolutionary history of two populations.

DISCUSSION

∂a∂i is a population genetics python module that uses the JFS to estimate historical demographic parameters between two or three populations. In this report we estimated the historical parameters of the human populations of YRI and CEU. In this analysis we also tested two competing models (migration vs. no migration) and showed that the migration model best predicts the patterns in genetic divergence observed in the data.

Overall ∂a∂i was an efficient and relatively fast method for estimation of these population parameters and the comparison of competing models. In order to use ∂a∂i effectively, however, a thorough understanding of the program is required. Specifically, an understanding of the format of the model files, the program file, and the use of the module *MS* is needed. ∂a∂i also requires the user to have a basic to intermediate understanding of python.

LITERACTURE CITED

Chen, H. 2012. The joint allele frequency spectrum of multiple populations: A coalescent theory approach. Volume 81 (2): 179–195

RN Gutenkunst, RD Hernandez, SH Williamson, CD Bustamante "Inferring the joint demographic history of multiple populations from multidimensional SNP data" PLoS Genetics 5:e1000695 (2009).

Livingston RJ, et al. 2004. Pattern of sequence variation across 213 environmental response genes. Genome Research. 14:1821