Inferring the joint demographic history of multiple populations

Ryan N. Gutenkunst
T-6 & CNLS, Los Alamos National Lab
with Ryan D. Hernandez, Scott H. Williamson, Carlos D. Bustamante
September 7, 2009


LA-UR: 09-03343
Genetic demographic models

1. Complement archeological and linguistic data.
2. Constrain the neutral background in genome scans for natural selection.
3. Assist in the design of further genetic studies.

figure from http://gmed.bu.edu/about/
Frequency spectrum

Pop 1: 2 individuals sampled

Pop 2: 1 individual sampled

A complete summary of the data in the absence of linkage. Encompasses many common statistics, e.g. Tajima’s D, $F_{ST}$. 
Frequency spectrum gallery

\[ \tau = 2N_\text{At} \]

\[ N_1 = \nu_1 \cdot N_\text{A} \]

\[ M = 2N_\text{At} \]

\[ N_2 = \nu_2 \cdot N_\text{A} \]

\[ \theta = 4N_\text{At} \]

\[ \nu_1 = \nu_2 = 0.5 \]

\[ \nu_1 = 0.9, \nu_2 = 0.1 \]

SNPs per bin

\[ 1000, 100, 10, 2 \]

\[ F_{ST} = 0.38 \]

\[ F_{ST} = 0.20 \]

\[ F_{ST} = 0.08 \]

\[ \tau = 0.5 \]

\[ \tau = 0.5 \]

\[ \tau = 0.5 \]

pop 1

pop 2

pop 2

pop 2
Forward simulation via diffusion

$\phi(x,y,t)$: density of SNPs at freq $x$ in pop 1 and $y$ in pop 2.

$$\frac{\partial \phi}{\partial \tau} = \frac{1}{2} \frac{\partial^2}{\partial^2 x} \left[ \frac{x(1-x)}{v_1} \phi \right] - \frac{\partial}{\partial x} \left[ \left( M_{1\leftarrow 2}(y-x) + \gamma_1 x(1-x) \right) \phi \right]$$

$$+ \frac{1}{2} \frac{\partial^2}{\partial^2 y} \left[ \frac{y(1-y)}{v_2} \phi \right] - \frac{\partial}{\partial y} \left[ \left( M_{2\leftarrow 1}(x-y) + \gamma_2 y(1-y) \right) \phi \right]$$

Splittings & Admixture

Pop 3 has fraction $f$ ancestry from pop 1 and $1-f$ from pop 2.

$$\phi(x,y,z) = \phi(x,y) \delta(z - [fx + (1-f)y])$$
$\phi$ to spectrum to likelihood

$$FS[i, j] = \int_0^1 dx \int_0^1 dy \binom{n_1}{i} x^i (1-x)^{n_1-i} \binom{n_2}{j} y^j (1-y)^{n_2-j} \phi(x, y)$$

... can also model (some) ascertainment

$\phi$

log-likelihood $= \log \left[ \prod_{i} \prod_{j} \text{Poisson}\left(\text{drawing Data}[i, j] | FS[i, j]\right) \right]$

... assuming no linkage
Our implementation: dadi
(Diffusion Approximations for Demographic Inference)

- Up to three interacting populations
- Very flexible model specification, arbitrary parameter time courses
- 1 pop, ~3 params: ~1 minute to fit
  2 pops, ~6 params: ~10 minutes to fit
  3 pops, ~12 params: ~3 hours to fit
- Available at http://dadi.googlecode.com
Human data

219 autosomal genic regions resequenced by the NIEHS Environmental Genome Project (5 Mb non-coding → 28k SNPs)

YRI: Yoruba; 12 ind.
CEU: European descent; 22 ind.
CHB: Han Chinese; 12 ind.

Polarize w/ chimp + correction
Assumed 25 yr generation time
Out-of-Africa spectrum
Out-of-Africa model
Maximum likelihood parameters

Data

Model

Residuals

Units
migration: per chromosome per generation
growth: % per generation

Data

Model

Residuals

7,300
12,300
2,100
1,000
510
0.40%
0.55%
220 kya
140 kya
21 kya

YRI (Africa)

CHB (East Asia)

CEU (Europe)

(m_{AF-AS} = 1.9 \times 10^{-5})
Accounting for linkage

- Conventional bootstrap of data (over genes)
  Parameter uncertainties: ~20% for most params, ~50% for ancient times

- Parametric bootstrap using simulated data

LD: real vs. simulated data

Goodness-of-fit

p-value ~ 0.02
Selection on nonsynonymous SNPs

Prediction for segregating SNPs

Previously inferred distribution of selection coefficients

\[-s \sim \Gamma(0.184, 8200)\]

Boyko et al.

Comparison with EGP data

\[
\text{proportion}
\]

\[
\text{singleton, freq > 10\%, YRI only, CEU only, CHB only, YRI/CEU only, YRI/CHB only, CEU/CHB only, YRI/CEU/CHB}
\]

Noncoding data
Noncoding model
Nonsynonymous data
Nonsynonymous model
Summary

• The frequency spectrum is a powerful tool for demographic inference.
• An efficient diffusion approach allows for statistical analysis of complex models, including selection.

Future Questions

• Applications: humans within Africa, 1000 Genomes, other species...
• GPU *dadi*
• Two-locus diffusion theory

Thanks

Adam Auton
Jim Booth
Michael Hammer
Amit Indap
Debbie Nickerson
Rasmus Nielsen
Nick Patterson
Molly Przeworski
Monty Slatkin
Jeff Wall
Carsten Wiuf

NIEHS EGP
Kavli ITP

http://dadi.googlecode.com