Robust demographic inference from genomic and SNP data

Laurent Excoffier
Isabelle Duperret, Emilia Huerta-Sanchez, Matthieu Foll, Vitor Sousa, Isabel Alves

Computational and Molecular Population Genetics Lab (CMPG)
Institute of Ecology and Evolution
University of Berne
Swiss Institute of Bioinformatics
Past demography affect genetic diversity

Stationary population

Past

Present

Mixture of rare and frequent mutations

Recent expansion

Few and mostly rare mutations

Recent contraction

Very deep lineages separating little differentiated clades
Site Frequency Spectrum (SFS) depends on past demography
Joint SFS (2D-SFS)

Model of Isolation with migration (IM)

$T_{DIV}$

$m_{12}$

$N_A$

$N_1$

$N_2$

$T_{DIV}=100$

$T_{DIV}=300$

$T_{DIV}=1000$

$T_{DIV}=10,000$
Problems with estimation of demographic parameters from SFS

Can one learn history from the allelic spectrum?

Simon Myers\textsuperscript{a}, Charles Fefferman\textsuperscript{b}, Nick Patterson\textsuperscript{a,*}

\textsuperscript{a} Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge MA 02142, United States
\textsuperscript{b} Department of Mathematics, Fine Hall, Washington Road, Princeton, NJ 08544, United States

Received 17 March 2007
Available online 30 January 2008

A demographic history with the same spectrum as a constant size population

\begin{figure}
\centering
\includegraphics[width=\textwidth]{sfs_diagram.png}
\caption{A demographic history with the same spectrum as a constant size population.}
\end{figure}
Estimation of demographic parameters from SFS with dadi

Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data

Ryan N. Gutenkunst¹*, Ryan D. Hernandez², Scott H. Williamson³, Carlos D. Bustamante³

Program dadi : Diffusion Approximation for Demographic Inference  http://code.google.com/p/dadi/

dadi estimates the site frequency spectrum based on a diffusion approximation
Advantages of SFS for parameter inference

• Accuracy of estimates increases with data size, but computing time does not

• Can be used to study complex scenarios (e.g. as complex as ABC)

• Very fast estimations (as compared to ABC, or full likelihoods)
Potential problems

• Maximization of the CL is not trivial (precision of the approximation and convergence problems)

• Ignores (assumes no) LD

• Need to repeat estimations to find maximum CL

• Needs genomic data (several Mb)
  – difficult to have gene-specific estimates

• Next-generation sequencing data must have high coverage to correctly estimate SFS (likely to miss singletons or show errors).

• SFS needs to be estimated from the NGS reads (ML methods: Nielsen et al. 2013, Keightley and Halligan, 2011)
Estimating the SFS with coalescent simulations

The probability of a SFS entry $i$ can be estimated under a specific model $\theta$ from its expected coalescent tree as (Nielsen 2000) a ratio of expected branch lengths

$$p_i = \frac{E(t_i \mid \theta)}{E(T \mid \theta)}$$

$t_i$: total length of all branches directly leading to $i$ terminal nodes

$T$: total tree length.

This probability can then be estimated on the basis of $Z$ simulations as

$$\hat{p}_i = \frac{\sum_j b_{kj}}{\sum_j T_j}$$

where $b_{kj}$ is the length of the $k$-th compatible branch in simulation $j$. 
Likelihood

The (composite) likelihood of a model $\theta$ is obtained as a multinomial sampling of sites (Adams and Hudson, 2004)

$$CL = \Pr(SFS_{obs} \mid \theta) \propto P_0^M (1 - P_0)^S \prod_{i=1}^{n-1} p_i^{m_i}$$

$M$ : number of monomorphic sites
$S$ : number of polymorphic sites
$P_0$ : probability of no mutation on the tree
$p_i$ : probability of the $i$-th SFS entry
$m_i$: number of sites with derived frequency $i$

This can be generalized for the joint SFS of two or more populations
fastsimcoal2 program

• Uses coalescent simulations to estimate the SFS and approximate the likelihood
  – Large number of simulations per point (>50000)
• Uses a **conditional expectation maximization** (CEM) algorithm to find maxCL parameters
• Relatively fast and can explore wide and unbounded parameter ranges
• Can handle an arbitrary number of populations
• For more than 4 populations, we use a composite composite-likelihood
  \[ CL_{1234...} = CL_{12} \times CL_{13} \times CL_{14} \times ... \times CL_{23} \times ... \]
Approximation of the SFS

Chen (2012) TPB
Coalescent approach to infer the expected joint SFS numerically

Divergence model

\[ T_{\text{DIV}} = 10 \]

\[ T_{\text{DIV}} = 100 \]
Bottleneck model

Simulation of 20 Mb data
10 cases, 50 runs/case

\[ \frac{\partial a}{\partial i} \]

\[ \frac{\partial a}{\partial i} \]

\[ 9/10 \]
IM model

\[ \frac{\partial}{\partial t} \]

\[ m_{12} \]

\[ m_{21} \]

\[ T_{DIV} \]
Pseudo human evolution model

\[ \frac{\partial a}{\partial i} \]
Herarchical island model

12 populations in two continent-island models

Migration rates over 3 orders of magnitude are well recovered !!!
Application: Complete genomics data

Four sampled human populations:

4 Luhyah from Kenya (LWK)
9 Europeans (CEU)
9 Yoruba (YRI)
5 African Americans (ASW)

(sequenced at 51-89x per genome)

Data:

Multidimensional SFS estimated from:
239, 120 SNPs in non-coding and non CpG regions
Each SNP more than 5 Kb away from the other
Model of admixture in African Americans

West-African meta-population

Luhya (Kenya)
Ghost (East-African) meta-population

Afr. Am.
Northern Europeans

Yoruba (Nigeria)
Model of admixture in African Americans
Models of African population divergence

Two models with different degrees of realism and complexity

IM model

2 continent-island model

The estimation of each model were performed separately for the San (109,020 SNPs) and the Yoruba (81,383 SNPs) SNP panels.
Models of African population divergence

IM model

Model A - San panel vs. Yoruba panel

Good agreement between panels
Models of African population divergence

2 continent-island model

Akaike’s weights of evidence in favor of model B are close to 1 for both panels.
## Models of African population divergence

### Model B

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Panel 4 (San)</th>
<th>Panel 5 (Yoruba)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point estimation</td>
<td>95% CI&lt;sup&gt;ab&lt;/sup&gt;</td>
</tr>
<tr>
<td>( N_{ANC} )</td>
<td>9612</td>
<td>8977–10424</td>
</tr>
<tr>
<td>( N_{AFR} )</td>
<td>23849</td>
<td>21634–44081</td>
</tr>
<tr>
<td>( N_{S} )</td>
<td>180,771</td>
<td>16598–411442</td>
</tr>
<tr>
<td>( N_{Y} )</td>
<td>96,071</td>
<td>2464–461785</td>
</tr>
<tr>
<td>( N_{Y} )</td>
<td>3,704</td>
<td>412–6996</td>
</tr>
<tr>
<td>( N_{Y} )</td>
<td>10251</td>
<td>2456–461785</td>
</tr>
<tr>
<td>( N_{Y} )</td>
<td>644</td>
<td>85–4553</td>
</tr>
<tr>
<td>( 2N_{M_S} )</td>
<td>5.9</td>
<td>4.6–14</td>
</tr>
<tr>
<td>( 2N_{M_Y} )</td>
<td>37.4</td>
<td>5–77</td>
</tr>
<tr>
<td>( T_d )</td>
<td>1,475 y</td>
<td>10–100</td>
</tr>
<tr>
<td>( \alpha_{YS} )</td>
<td>0.19</td>
<td>0.04–0.28</td>
</tr>
<tr>
<td>( \alpha_{YS} )</td>
<td>0.08</td>
<td>0.04–0.18</td>
</tr>
<tr>
<td>( m_{SY} )</td>
<td>4.45E-05</td>
<td>2.3E-06–9.9E-04</td>
</tr>
<tr>
<td>( m_{YS} )</td>
<td>1.11E-04</td>
<td>1.2E-05–6.3E-04</td>
</tr>
<tr>
<td>( T_{YE} )</td>
<td>4,250 y</td>
<td>101–691</td>
</tr>
<tr>
<td>( T_{DS} )</td>
<td>138,250 y</td>
<td>2482–9710</td>
</tr>
</tbody>
</table>
Inference of archaic admixture in modern humans

Simple model (proof of concept)

Data set:
Non coding DNA and non CpG sites.
Altai Neandertal (Prüfer et al. 2013), unfiltered vcf
271,994 regions of 100 bp in non-coding DNA
Ancestral state deduced by 1000G for 26,466,040 bp (26.5Mb)
All regions are at least 5 Kb apart from each other

Complete genomics
CHB or TSI samples
(4 inds / pop)
Inference of archaic admixture in modern humans

Very preliminary results

Admixture level
CHB: 1.2% (0.94-1.43)
TSI: 1.3% (0.85-1.45)

Recent admixture
TSI: 875 gen (790-1030)
CHB: 950 gen (810-1200)
<25,000 y
(assuming u=2e-8)
Possible extensions

• Multiprocessor version of fsc
• MCMC (Beaumont 2004, Garrigan 2009)
• Multilocus SFS
• Coalescent simulations through pedigrees
Thanks to:
Isabelle Duperret
Emilia Huerta-Sanchez
Isabel Alves
Vitor Sousa
Matthieu Foll
Rasmus Nielsen
CMPG lab

David Reich
Nick Patterson