Although coalescent models have come to play a central role in population genetics, there are some situations where genealogies may not lead to efficient inference.

- Unphased SNP data: with short reads, it may not be possible to accurately infer haplotypes.
- Mixtures: with short reads, it may not be possible to assemble sequences corresponding to different individuals.
- Selection: although the coalescent can be extended to include the effects of selection, in general the resulting processes are difficult to simulate.
- Complex models: likelihood surfaces may be difficult to reconstruct in models with many parameters using inherently noisy stochastic simulations of the coalescent.
An alternative approach is to use the distribution of allele frequencies in sampled populations to make inferences about the evolutionary forces operating in those populations.

- The **allele frequency spectrum (AFS)** is the distribution of the frequencies or counts of derived alleles in a sample of individuals calculated over all segregating sites.

- The **folded allele frequency spectrum (folded AFS)** is the distribution of the frequencies or counts of minor alleles in a sample calculated over all segregating sites.

- The AFS and folded AFS will be similar when the derived alleles are also the minor alleles, but this will not always be the case.

- However, the folded AFS can always be calculated, whereas the AFS can only be calculated if we can reliably distinguish derived from ancestral alleles.

- When it can be calculated, the AFS will be more informative about the population history than the folded AFS.
Allele frequency spectra of autosomal and X-linked genes in M. musculus

The allele frequency spectrum for a sample of $n$ chromosomes can be represented by a vector

$$ a = (a_0, \cdots, a_n) $$

where $a_i$ denotes the number of segregating sites at which the derived allele is carried by exactly $i$ chromosomes.

For example, $a_1$ is the number of singletons in the sample.

Athanosios et al. (2014)
Folded and Unfolded AFS in D. melanogaster

Cooper et al. (2015)
To use the allele frequency spectrum for inference, we need to be able to calculate its likelihood under models that account for the evolutionary processes thought to be operating on the population.

- The probability density of the frequency $x$ of the derived allele will depend on the choice of the model and its parameter through some function $f(x|\Theta)$.

- If $n$ chromosomes are sampled at random from a population in which the frequency of the derived allele is $x$, the probability that it will be sampled $k$ times is given by the binomial distribution:

$$P(k|x) = \binom{n}{k} x^k (1-x)^{n-k}.$$  

- The unconditional probability of sampling the derived allele $k$ times can be calculated by integrating the preceding conditional probability with respect to the density of $x$:

$$P(k|\Theta) = \int_0^1 f(x|\Theta) \cdot \binom{n}{k} x^k (1-x)^{n-k} dx.$$  

If we assume that the segregating sites are independent of one another, then the total probability of the allele frequency spectrum $A = (a_1, a_2, \cdots, a_{n-1})$ can be calculated as follows:

$$
\mathbb{P}(A|\Theta) = \prod_{k=1}^{n-1} e^{-\mathbb{P}(k|\Theta)} \frac{\mathbb{P}(k|\Theta)^{a_k}}{a_k!}.
$$

- This assumes an infinite sites model, with new mutations produced by a Poisson process.
- It also assumes that the segregating sites in the sample are unlinked (and hence statistically uncorrelated conditional on $\Theta$).
- With linked sites, we can treat this as a **composite likelihood**, which is known to give consistent estimates of parameter values under many neutral models, even though it underestimates the variance in these estimates.
- Bootstraps (conventional and parametric) can be used to obtain confidence intervals when working with composite likelihoods.
Joint Allele Frequency Spectrum

If individuals are sampled from different populations, the distribution of allele frequencies across populations can be characterized using the joint allele frequency spectrum, which specifies the joint distribution of the counts of derived alleles in different populations.

- For example, with two populations, we need to keep track of the number of segregating sites at which the derived allele was sampled $i$ times in the first population and $j$ times in the second population.
- To evaluate the likelihood of the joint AFS, we then need a model that will allow us to calculate the joint probability distribution $f(x_1, x_2)$ of the derived allele frequencies in the two populations.
- Dependence between the allele frequencies in the different populations will arise because of shared ancestry and migration.
Joint Allele Frequency Spectra of Two Populations (Simulated)

C

D

$\nu_1, \nu_2 = 0.5, 0.5$

$\nu_1, \nu_2 = 0.9, 0.1$

$\tau = 0.1$

$\tau = 0.3$

$\tau = 1.0$

$M = 0.1$

$M = 2$

$M = 10$
In principal, the distribution of the derived allele frequencies in a population could be calculated under either the Wright-Fisher model or some other suitable Markov chain (e.g., the Moran model).

- In practice, this is difficult because usually analytical expressions are not available for these models, which instead must be solved using calculations involving large matrices.
- In addition, changing the effective population size changes the state space of the Wright-Fisher model, which will increase computational costs.

Fortunately, these complications can be sidestepped to a degree by working with an approximation to the Wright-Fisher model which is accurate when the effective population size is sufficiently large, say $N_e > 100$. 
The following figure shows a series of simulations of the Wright-Fisher model for 100 generations for $N = 10$ (blue), 100 (red), 1000 (orange), and 10,000 (green).

Notice that both the size of the fluctuations and the total change in $p$ over 100 generations decrease as the population size is increased.
A different picture emerges if we plot each sample path against time measured in units of $N$ generations. Here the total rescaled time is $t = 1$.

In this case, the jumps become smaller, but the total change in $p$ over $N$ generations does not tend to 0.
Diffusion Processes in Population Genetics

In fact, it can be shown that as $N \to \infty$, the rescaled processes $(p^{(N)}(Nt) : t \geq 0)$ converge to a limiting process known as a **diffusion approximation**.

- Diffusion processes are Markov processes with continuous sample paths, i.e., there are no jumps.
- Diffusion approximations can be derived for the Wright-Fisher model with mutation and selection provided that these other processes are sufficiently weak.
- The transition densities $f(p; t)$ satisfy a partial differential equation known as the **Kolmogorov forward equation**:

$$\frac{\partial}{\partial t} \phi(x; t) = \frac{\partial^2}{\partial x^2} \left( \frac{x(1 - x)}{4N_e} \phi(x; t) \right) - \frac{\partial}{\partial x} \left( sx(1 - x)\phi(x; t) \right)$$

- This equation can be solved numerically using sophisticated techniques available for PDE's that do not require stochastic simulations.
Diffusion approximations can also be derived for models with more than one population. In this case, the joint density of the allele frequencies in the different populations can be found by solving the following PDE:

\[
\frac{\partial}{\partial t} \phi = \frac{1}{2} \sum_{i=1}^{P} \frac{\partial^2}{\partial x_i^2} \left\{ \frac{x_i(1-x_i)}{4\nu_i} \phi \right\} - \sum_{i=1}^{P} \frac{\partial}{\partial x_i} \left\{ \gamma_i x_i(1-x_i) + \sum_{j=1}^{P} M_{i\to j}(x_j-x_i) \phi \right\}
\]

where

- \( \nu_i \) is the relative effective population size of population \( i \);
- \( \gamma_i = 2Ns_i \) is the scaled selective coefficient of the derived allele in population \( i \);
- \( M_{i\to j} = 2Nm_{i\to j} \) is the scaled migration rate from \( j \) to \( i \).

This must be solved numerically, but again this can be done without stochastic simulations.
Example: Out of Africa analysis
Example: Out of Africa analysis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Maximum likelihood</th>
<th>Conventional bootstrap 95% confidence interval</th>
<th>Parametric bootstrap bias-corrected 95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_A$</td>
<td>7,300</td>
<td>4,400–10,100</td>
<td>6,300–9,200</td>
</tr>
<tr>
<td>$N_{AF}$</td>
<td>12,300</td>
<td>11,500–13,900</td>
<td>11,100–13,100</td>
</tr>
<tr>
<td>$N_B$</td>
<td>2,100</td>
<td>1,400–2,900</td>
<td>1,700–2,600</td>
</tr>
<tr>
<td>$N_{EU0}$</td>
<td>1,000</td>
<td>500–1,900</td>
<td>500–1,500</td>
</tr>
<tr>
<td>$r_{EU}$  (%)</td>
<td>0.40</td>
<td>0.15–0.66</td>
<td>0.26–0.57</td>
</tr>
<tr>
<td>$r_{AS}$  (%)</td>
<td>0.55</td>
<td>0.23–0.88</td>
<td>0.32–0.79</td>
</tr>
<tr>
<td>$m_{AF-B}$ ($\times 10^{-5}$)</td>
<td>25</td>
<td>15–34</td>
<td>19–36</td>
</tr>
<tr>
<td>$m_{AF-EU}$ ($\times 10^{-5}$)</td>
<td>3.0</td>
<td>2.0–6.0</td>
<td>1.6–7.6</td>
</tr>
<tr>
<td>$m_{AF-AS}$ ($\times 10^{-5}$)</td>
<td>1.9</td>
<td>0.3–10.4</td>
<td>0.7–6.9&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>$m_{EU-AS}$ ($\times 10^{-5}$)</td>
<td>9.6</td>
<td>2.3–17.4&lt;sup&gt;b&lt;/sup&gt;</td>
<td>5.7–20.2</td>
</tr>
<tr>
<td>$T_{AF}$ (kya)</td>
<td>220</td>
<td>100–510</td>
<td>90–410</td>
</tr>
<tr>
<td>$T_{B}$ (kya)</td>
<td>140</td>
<td>40–270</td>
<td>60–310</td>
</tr>
<tr>
<td>$T_{EU-AS}$ (kya)</td>
<td>21.2</td>
<td>17.2–26.5</td>
<td>17.6–23.9</td>
</tr>
</tbody>
</table>

<sup>a</sup>See Figure 2B for model schematic. Growth rates $r$ and migration rates $m$ are per generation.

<sup>b</sup>One low-migration outlier was removed for each of these estimations.

doi:10.1371/journal.pgen.1000695.t001
Example: Settlement of the New World analysis
Example: Settlement of the New World analysis

<table>
<thead>
<tr>
<th>parameter</th>
<th>maximum likelihood</th>
<th>conventional bootstrap 95% confidence interval</th>
<th>parametric bootstrap bias-corrected 95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{EU0}$</td>
<td>1,500</td>
<td>700–2,100</td>
<td>900–2,200</td>
</tr>
<tr>
<td>$r_{EU}$ (%)</td>
<td>0.23</td>
<td>0.08–0.45</td>
<td>0.16–0.34</td>
</tr>
<tr>
<td>$N_{AS0}$</td>
<td>590</td>
<td>320–800</td>
<td>410–790</td>
</tr>
<tr>
<td>$r_{AS}$ (%)</td>
<td>0.37</td>
<td>0.16–0.60</td>
<td>0.24–0.51</td>
</tr>
<tr>
<td>$N_{MX0}$</td>
<td>800</td>
<td>160–1,800</td>
<td>140–1,600</td>
</tr>
<tr>
<td>$r_{MX}$ (%)</td>
<td>0.50</td>
<td>0.14–1.17</td>
<td>0.41–0.98</td>
</tr>
<tr>
<td>$m_{EU-AS} \times 10^{-5}$</td>
<td>13.5</td>
<td>7.5–32.2</td>
<td>9.9–20.8</td>
</tr>
<tr>
<td>$T_{EU-AS}$ (kya)</td>
<td>26.4</td>
<td>18.1–43.1</td>
<td>21.7–30.7</td>
</tr>
<tr>
<td>$T_{MX}$ (kya)</td>
<td>21.6</td>
<td>16.3–26.9</td>
<td>18.6–24.7</td>
</tr>
<tr>
<td>$f_{MX}$ (%)</td>
<td>48</td>
<td>42–60</td>
<td>41–55</td>
</tr>
</tbody>
</table>

*See Figure 3B for model schematic. Growth rates $r$ and migration rates $m$ are per generation. $f_{MX}$ is the average European admixture proportion of the Mexican-Americans sampled.

doi:10.1371/journal.pgen.1000695.t002