Demographic inference based on Site frequency spectrum (SFS) – Part II

Vitor Sousa
CE3C – center for ecology, evolution and environmental changes

2018 WSPG Cesky Krumlov
26 Jan 2018

vmsousa@fc.ul.pt
Example of Applications:

- Human dispersal out of Africa (high quality whole-genome) – lessons on choice of models
- Deer mice colonization of Nebraska Sand Hills (targeted re-capture data) – lessons on effects of filtering
- Inferring divergence times and gene flow in sawflies (ddRAD-seq data) – lessons from comparing models
A genomic history of Aboriginal Australia

Anna-Sapfo Malaspina1,2,3, Michael C. Westaway4*, Craig Muller1*, Vitor C. Sousa2,3,*, Oscar Lao5,6,*, Isabel Alves2,3,7,*, Anders Bergström8,*, Georgios Athanasiadis9, Jade Y. Cheng9,10, Jacob E. Crawford10,11, Tim H. Heupink4, Enrico Macholdt12, Stephan Pelsch13, Simon Rasmussen14, Stephan Schiffler13, Sankar Subramanian4, Joanne L. Wright4, Anders Albrechtsen16, Chiara Barbieri12,17, Isabelle Dupanloup1,2,3, Anders Eriksson18,19, Ashot Margaryan1, Ida Moltke16, Irina Pugach12, Thorfinn S. Korneliussen1, Ivan P. Levkivskyi20, J. Victor Moreno-Mayar1, Shengyu Ni12, Fernando Racimo10, Martin Sikora1, Yali Xue8, Farhang A. Aghakhianan21, Nicolas Brucato22, Søren Brunak23, Paula F. Campos1,24, Warren Clark25, Sturla Ellingvåg26, Gudjugudju Fourmile27, Pascale Gerbault28,29, Darren Injie30, George Kok31, Matthew Leavesley32, Betty Logan33, Aubrey Lynch34, Elizabeth A. Matisoo-Smith35, Peter J. McAllister36, Alexander J. Mentzer37, Mait Metspalu38, Andrea B. Migliano29, Les Murgha39, Maude E. Phipps21, William Pomat31, Doc Reynolds40, Francois-Xavier Ricaut22, Peter Siba31, Mark G. Thomas28, Thomas Wales41, Colleen Ma’run Wall42, Stephen J. Oppenheimer43, Chris Tyler-Smith8, Richard Durbin8, Joe Dortch44, Andrea Manica18, Mikkel H. Schierup9, Robert A. Foley1,45, Marta Mirazón Lahr1,45, Claire Bowne46, Jeffrey D. Wall47, Thomas Mailund8, Mark Stoneking12, Rasmus Nielsen1,48, Manjinder S. Sandhu8, Laurent Excoffier2,3, David M. Lambert4 & Eske Willerslev1,8,18

Nature(2016)
Australia harbors some of the oldest modern human remains outside Africa

Many sites and remains dated to be older than 40 kya, suggesting a human settlement 47.5-55 kya
One wave out of Africa vs Two waves out of Africa
83 high-coverage Aboriginal Australians genomes

Average depth of coverage: 65x
Very good quality of genotype calls
Effect of depth of coverage on SFS

- Compared 2D SFS based on depth of coverage of observed data (mean larger than >20x), with a distribution 8 times smaller.

Malaspinas et al. (2016) Nature
A note on recovering the SFS from genomic data

- Simulation study
- Low depth of coverage and missing data lead to biased SFS towards rare variants
83 high-coverage Aboriginal Australians genomes

Average depth of coverage: 65x

Western Central Desert (WCD)
Since we want to infer demography we tried to minimize the number of sites affected by selection:

- 985 1Mb blocks outside genic regions and CpG islands (~4.3 Million SNPs)
- 5 dimensional SFS (16,875 entries)
- Confidence intervals obtained using block-bootstrap

Archaic human genomes:
- 1 Neanderthal (~66 kya)
- 1 Denisovan (~52 kya)

Mutation rate assumed
1.25 x 10^{-8} /site/gen

Generation time
29 years/gen
Towards a model to test the hypotheses: One vs Two waves Out of Africa

- **Data (SFS)**
- (Re-)Define model (hypotheses to test)
- Run fastsimcoal2
- Estimates!
  - Assess the fit to the data

---

**Do you have an outgroup?**
- **Yes** – use the derived (unfolded) SFS
- **No** – use the minor allele frequency spectrum (folded)

**Do you have monomorphic sites?**
- **Yes** - then, given a mutation rate you can infer the absolute times and effective sizes
- **No** – then all your estimates need to be relative to a fixed parameter (fixed Ne or fixed time)
Evidence of two waves Out of Africa:

- Old split leading to colonization of Australia (81 kya)
- More recent split leading to colonization of Eurasia (67 kya)
Towards a model incorporating Neanderthal and Denisovan admixture

- Non-African populations: 1-4% estimated Neanderthal admixture
- Aboriginal Australians and New Guineans: 3-6% estimated Denisovan admixture
- Archaic admixture can affect times of split estimates

Evidence of archaic introgression

Total length (Mb) of:
- Putative Denisovan haplotype (PDH)
- Putative Neanderthal haplotypes (PNH)
Accounting for shared ancestry of Neanderthal and Denisovan

Admixture occurs between modern humans and:
- Denisovan-related (D.R.) population
- Neanderthal-related (N.R.) population

Prüfer et al. (2014) Nature
Two-waves out of Africa

- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events
Two-waves out of Africa

- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events
Two-waves out of Africa

- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events
One wave out of Africa

- Similar divergence times ($\Delta t$ close to zero)
- One single bottlenecks associated with the Out of Africa events
- A major admixture pulse with Neanderthal
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture.

- Similar divergence time ($\Delta t$ close to zero)
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time ($\Delta t$ close to zero)
- Bottleneck associated with the Out of Africa event
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time ($\Delta t$ close to zero)
- Bottleneck associated with the Out of Africa event
- A major admixture pulse with Neanderthal in ancestors of all non-Africans
A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time ($\Delta t$ close to zero)
- Bottleneck associated with the Out of Africa event
- A major admixture pulse with Neanderthal in ancestors of all non-Africans
Model captures aspects about the observed data

Good fit to the marginal 1D site frequency spectrum
What entries are not well fitted?

The model does not fit very well the rare variants (singletons, doubletons) private to a single population.

Pagani et al (2016) suggests two waves: Papuan genomes with signature of admixture with humans from first wave (at least 2% of their genome).
Summary
Aboriginal Australians genomes support a single major wave out of Africa

- Accounting for archaic admixture with Neanderthal and Denisovan was crucial to understand population divergence
- Genomic data consistent with a single major dispersal event out of Africa (60-104 kya)
- Two major dispersal waves into Asia: Aboriginal Australians diverged 51-72 kya from Eurasians
Deer mice from Nebraska Sand Hills

S. Pfeifer, S. Laurent, V. Sousa, C. Linnen, H. Hoekstra, L. Excoffier, J. Jensen
Coat color adaptation in deer mice

*Peromyscus maniculatus*

- Habitat (soil color) correlated with coat phenotype
- Field experiments suggest that light color confers selective advantage against visually hunting predators
- Nebraska Sand Hills were formed 8000 to 15,000 years ago

![Image of deer mice on and off Sand Hills](https://example.com/image.png)

Linnen et al (2013) *Science*

Pfeifer*, Laurent*, Sousa* et al (in press) *MBE*
A transect across the Sand Hills (ON and OFF)

Sample locations “off” and “on” the Sand Hills
- 11 populations
- 330 individuals

- Genomic data (NGS) data
  - Target 10,000 random 1.5kb regions
  - 185kbp region comprising the *Agouti* gene

- Phenotypic data for each individual
Evidence for isolation by distance but three groups

Geographically closer samples are genetically more similar

TESS3 analysis (ancestry estimation accounting for spatial information, Caye et al 2016)
Model-based inference

Is there evidence of gene flow between Off and On the Sand Hills?

Colonization from North

Serial colonization from South

Colonization from South

Serial colonization from North

Legend:
- Bottlenecks associated with founder events

Estimates based on the joint 3D site frequency spectrum (SFS):
- Folded SFS with 140,358 SNPs

Pooled individuals from three groups: north OFF, south OFF and ON the Sand Hills
Deer mice: Pairwise marginal 2D SFS
Since we did not have an outgroup we used the folded SFS
Estimates support south colonization and high gene flow levels

- Recent time of colonization of Sand Hills ~3-5 kya, younger than formation of Sand Hills 8-15 kya
- High migration rates across all populations, inferred for all models

Migration rates above/below arrows in units of 2Nm, i.e. average number of immigrants per generation.
Deer mice: Model fit to marginal SFS
Some lessons I learned working with the deer mice data

- Be careful when applying Hardy-Weinberg filters to your data
- Be careful when filtering on depth of coverage applying the same thresholds for all individuals
The depth of coverage varied considerably across individuals

- Applying the same threshold for all individuals can lead to biases
- Apply a filter on DP for each individual
Effect of DP filters on the SFS

Simulation study

- **DP > 10**: SFS based on called genotypes.

- **DP > 15**: SFS accounting for genotype uncertainty (ANGSD).

Simulated 2 pops SFS sampling 4 diploids from each pop, 200000 SNPs, mean coverage=10x, error rate=0.01. Simulated with correlated allele frequencies model ($F_{ST}=(0.275, 0.01)$)

With DP>15 we have a very good approximation to the correct SFS, even when using the called genotypes.
- High migration between all groups of populations (2Nm~20)
- No evidence of a strong bottleneck signal associated with colonization of SH
Sawflies and RAD data

MOLECULAR ECOLOGY

Molecular Ecology (2016)

History, geography and host use shape genomewide patterns of genetic variation in the reddheaded pine sawfly (Neodiprion lecontei)

ROBIN K. BAGLEY,* VITOR C. SOUSA,† MATTHEW L. NIEMILLER‡ and CATHERINE R. LINNEN*

*Department of Biology, University of Kentucky, Lexington, KY 40506, USA, †cE3c - Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal, ‡Illinois Natural History Survey, Prairie Research Institute, University of Illinois Urbana-Champaign, Champaign, IL 61820, USA
Sawflies *Neodiprion lecontei*

- Hymenoptera
- Plant-feeding insects
- Pine tree specialists

Ovipositor (saw)

Same geographic area

- *N. pinetum*
- *N. lecontei*
ddRAD seq data

- 80 individuals from 77 localities and 13 host species
- 100 bp paired-end reads, mapped to reference genome of *N. lencontei*
- Depth of coverage filter DP>10
Given the detected three groups (North, Central, South):

- What is the population tree topology?
- What are the split times?
- What are the migration levels among groups?
Comparing models with composite likelihoods

- Fastsimcoal2 likelihood is “correct” if all SNPs are independent.
- We can then compare the model likelihoods using Akaike Information Criterion (AIC).

Composite likelihood provide unbiased maximum likelihood parameter estimates, but the likelihoods are inflated.

- Composite likelihood (assuming linked sites are independent)
- “Correct” likelihood (all sites are actually independent)
A strategy to compare models

1. Divide the dataset into LD blocks.
2. Create a dataset with all SNPs (including linked SNPs)
3. For each model, obtain the parameters that maximize the likelihood (this is ok even with linked sites!) and the corresponding expected SFS
4. Create a dataset with “independent” SNPs (1 SNP per RAD tag)
5. Given the expected SFS of each model, compute the “correct” likelihood for each model with the dataset with independent SNPs
6. Compare models with AIC

Observed SFS with ALL SNPs
Run fastsimcoal2
Expected SFS for each model
“Correct” likelihood for each model
Comparing alternative models

Table 2 Summary of the likelihoods for the sixteen demographic models tested. Lhood (ALL SNPs) and Lhood (1 SNP) correspond to the mean likelihood computed with the data sets containing ‘all SNPs’ (including monomorphic sites) and a ‘single SNP’ (without monomorphic sites) per RAD locus, respectively. Mean likelihoods were computed based on 100 expected site frequency spectra simulated according to the parameters that maximized the likelihood of each model. Topology names for each model are as indicated in Fig. S1 (Supporting information). AIC scores and relative likelihoods (Akaike’s weight of evidence) were calculated based on the ‘single SNP’ data set following Excoffier et al. 2013.

<table>
<thead>
<tr>
<th>Topology</th>
<th>Migration allowed?</th>
<th>Exponential growth?</th>
<th>North bottleneck?</th>
<th>log₁₀(Lhood) ALL SNPs</th>
<th>log₁₀(Lhood) 1 SNP</th>
<th># Parameters</th>
<th>AIC</th>
<th>ΔAIC</th>
<th>Relative likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>North–South</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>-46502.02</td>
<td>-7381.4</td>
<td>7</td>
<td>34006.70</td>
<td>75.69</td>
<td>0.000</td>
</tr>
<tr>
<td>North–Central</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>-46475.82</td>
<td>-7369.0</td>
<td>7</td>
<td>33949.44</td>
<td>18.43</td>
<td>0.000</td>
</tr>
<tr>
<td>South–Central</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>-46502.18</td>
<td>-7381.6</td>
<td>7</td>
<td>34007.60</td>
<td>76.59</td>
<td>0.000</td>
</tr>
<tr>
<td>Trifurcation</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>-46501.54</td>
<td>-7380.4</td>
<td>5</td>
<td>33998.07</td>
<td>67.06</td>
<td>0.000</td>
</tr>
<tr>
<td>North–South</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>-46470.49</td>
<td>-7365.0</td>
<td>15</td>
<td>33947.25</td>
<td>16.24</td>
<td>-0.000</td>
</tr>
<tr>
<td>North–Central</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>-46462.24</td>
<td>-7361.5</td>
<td>15</td>
<td>33931.01</td>
<td>0.00</td>
<td>0.851</td>
</tr>
<tr>
<td>South–Central</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>-46467.69</td>
<td>-7363.8</td>
<td>15</td>
<td>33941.57</td>
<td>10.56</td>
<td>0.004</td>
</tr>
<tr>
<td>Trifurcation</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>-46470.28</td>
<td>-7364.7</td>
<td>11</td>
<td>33937.93</td>
<td>6.91</td>
<td>0.027</td>
</tr>
<tr>
<td>North–South</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>-46469.48</td>
<td>-7362.8</td>
<td>18</td>
<td>33942.91</td>
<td>11.90</td>
<td>0.002</td>
</tr>
<tr>
<td>North–Central</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>-46461.17</td>
<td>-7361.7</td>
<td>18</td>
<td>33937.82</td>
<td>6.80</td>
<td>0.028</td>
</tr>
<tr>
<td>South–Central</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>-46463.73</td>
<td>-7363.9</td>
<td>18</td>
<td>33948.15</td>
<td>17.13</td>
<td>-0.000</td>
</tr>
<tr>
<td>Trifurcation</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>-46467.72</td>
<td>-7363.3</td>
<td>14</td>
<td>33937.39</td>
<td>6.37</td>
<td>0.035</td>
</tr>
<tr>
<td>North–South</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-46467.45</td>
<td>-7361.5</td>
<td>20</td>
<td>33940.86</td>
<td>9.85</td>
<td>0.006</td>
</tr>
<tr>
<td>North–Central</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-46461.25</td>
<td>-7362.1</td>
<td>20</td>
<td>33943.82</td>
<td>12.81</td>
<td>0.001</td>
</tr>
<tr>
<td>South–Central</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-46463.58</td>
<td>-7364.1</td>
<td>20</td>
<td>33953.08</td>
<td>22.07</td>
<td>0.000</td>
</tr>
<tr>
<td>Trifurcation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-46466.06</td>
<td>-7362.4</td>
<td>16</td>
<td>33936.93</td>
<td>5.92</td>
<td>0.044</td>
</tr>
</tbody>
</table>

Joint 3D minor allele frequency SFS (11,617 SNPs – ALL SNPs; 4,478 SNPs – 1 SNP per RAD tag)
Estimates favors a scenario where North and Central diverged more recently with asymmetric gene flow.

The inferred population tree topology and divergence times are consistent with divergence and range expansion from different refugia after LGM.

3 pairwise 2D minor allele frequency SFS (15,230 SNPs)
Summary

- Fastsimcoal2 can be applied to RAD seq data
- We used a strategy to obtain (as close as possible) the “correct” likelihood by dividing the data into blocks, inferring the expected SFS for each model with ALL SNPs, and then re-computing the “true” likelihood with independent SNPs (1 SNP per block)
- Despite the reduced number of SNPs we were able to discriminate models based on their likelihoods
Protocol for model comparison based on AIC when we have independent SNPs

• Get the observed SFS
• Define the alternative models
• Perform 50-100 runs under each model
• Select the runs with maximum likelihood under each model
• Compute the AIC (Akaike information criteria) for each model
• Select the model with minimum AIC
Estimating SFS from observed data

- How to deal with missing data?

<table>
<thead>
<tr>
<th>SNP</th>
<th>Freq. derived</th>
<th>Sample size</th>
<th>Rel. freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP1</td>
<td>1</td>
<td>16</td>
<td>1/16</td>
</tr>
<tr>
<td>SNP2</td>
<td>6</td>
<td>12</td>
<td>1/2</td>
</tr>
<tr>
<td>SNP3</td>
<td>1</td>
<td>12</td>
<td>1/12</td>
</tr>
<tr>
<td>SNP4</td>
<td>6</td>
<td>16</td>
<td>3/8</td>
</tr>
</tbody>
</table>

![Bar chart showing frequency of derived alleles](chart.png)
Estimating SFS from observed data

• How to deal with missing data?

<table>
<thead>
<tr>
<th>SNP</th>
<th>Freq. derived</th>
<th>Sample size</th>
<th>Rel. freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP1</td>
<td>1</td>
<td>16</td>
<td>1/16</td>
</tr>
<tr>
<td>SNP2</td>
<td>6</td>
<td>12</td>
<td>1/2</td>
</tr>
<tr>
<td>SNP3</td>
<td>1</td>
<td>8</td>
<td>1/12</td>
</tr>
<tr>
<td>SNP4</td>
<td>6</td>
<td>16</td>
<td>3/8</td>
</tr>
</tbody>
</table>

# derived alleles

frequency

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Estimating SFS from observed data

- How to deal with missing data?
  - Solution:
    - Find minimum sample size
    - Resample without replacement

<table>
<thead>
<tr>
<th>SNP</th>
<th>Freq. derived</th>
<th>Sample size</th>
<th>Rel. freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNP1</td>
<td>1</td>
<td>16</td>
<td>1/16</td>
</tr>
<tr>
<td>SNP2</td>
<td>6</td>
<td>12</td>
<td>1/2</td>
</tr>
<tr>
<td>SNP3</td>
<td>1</td>
<td>8</td>
<td>1/12</td>
</tr>
<tr>
<td>SNP4</td>
<td>6</td>
<td>16</td>
<td>3/8</td>
</tr>
</tbody>
</table>

Gavel et al. (2014) PNAS
FASTSIMCOAL2 INPUT FILES

Vitor Sousa
vmsousa@fc.ul.pt
Cesky Krumlov 2018
## Examples of observed SFS

### 1PopExpInst20Mb_DAFpop0.obs

<table>
<thead>
<tr>
<th>d0_0</th>
<th>d0_1</th>
<th>d0_2</th>
<th>d0_3</th>
<th>d0_4</th>
<th>d0_5</th>
<th>d0_6</th>
<th>d0_7</th>
<th>d0_8</th>
<th>d0_9</th>
<th>d0_10</th>
</tr>
</thead>
<tbody>
<tr>
<td>19973842</td>
<td>24630</td>
<td>810</td>
<td>173</td>
<td>145</td>
<td>111</td>
<td>88</td>
<td>84</td>
<td>61</td>
<td>56</td>
<td>0</td>
</tr>
</tbody>
</table>

### 2PopDivMigr20Mb_jointDAFpop1_0.obs

<table>
<thead>
<tr>
<th>d0_0</th>
<th>d0_1</th>
<th>d0_2</th>
<th>d0_3</th>
<th>d0_4</th>
<th>d0_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1_0</td>
<td>19985747</td>
<td>8350</td>
<td>1628</td>
<td>360</td>
<td>62</td>
</tr>
<tr>
<td>d1_1</td>
<td>9660</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d1_2</td>
<td>4790</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d1_3</td>
<td>3280</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d1_4</td>
<td>2490</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d1_5</td>
<td>1760</td>
<td>13</td>
<td>18</td>
<td>13</td>
<td>19</td>
</tr>
</tbody>
</table>

### 2PopDiv20Mb_jointDAFpop1_0.obs

<table>
<thead>
<tr>
<th>d0_0</th>
<th>d0_1</th>
<th>d0_2</th>
<th>d0_3</th>
<th>d0_4</th>
<th>d0_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1_0</td>
<td>19985547</td>
<td>8211</td>
<td>1415</td>
<td>316</td>
<td>55</td>
</tr>
<tr>
<td>d1_1</td>
<td>1266</td>
<td>101</td>
<td>37</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>d1_2</td>
<td>61142</td>
<td>20</td>
<td>8</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>d1_3</td>
<td>48631</td>
<td>12</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d1_4</td>
<td>47915</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>d1_5</td>
<td>1189</td>
<td>46</td>
<td>22</td>
<td>19</td>
<td>18</td>
</tr>
</tbody>
</table>
Parameter estimation settings files

Additional files necessary to estimate parameters

Estimation file

```
1PopExpInst20Mb/1PopExpInst20Mb.est

// Search ranges and rules file
// ******************************

[PARAMETERS]
//isInt? #name   #dist.#min #max
//all Ns are in number of haploid individuals
1 NPOP logunif 1000   1e7   output
1 NANC logunif  10    1e5   output
1 TEXP unif     10    1e5   output

[RULES]

[COMPLEX PARAMETERS]

0 RESIZE = NANC/NPOP  hide
```

Template file

```
1PopExpInst20Mb/1PopExpInst20Mb.tpl

//Parameters for the coalescence simulation program: fsimcoal2.exe
1 samples to simulate:
//Population effective sizes (number of genes)
NPOP
//Samples sizes and samples age
10
//Growth rates: negative growth implies population expansion
0
//Number of migration matrices: 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix index
1 historical event
TEXP 0 0 0 RESIZE 0 0
//Number of independent loci [chromosome]
1 0
//Per chromosome: Number of contiguous linkage Block: a block is a set of contiguous loci
1
//per Block: data type, number of loci, per generation recombination and mutation rates and optional parameters
FREQ 1 0  2.5e-8 OUTEXP
```
INPUT files for fastsimcoal2:
Defining an evolutionary model with PAR files

Here we simulate 10 recombining segments of 1000 bp DNA, in two populations of sizes 20000 and 1000 having diverged 5000 generations ago from a small population of size 100
TPL files

TPL are like PAR files, but the actual parameter values are replaced by parameter tags. These files are very important! Check carefully all the definitions. Errors in the TPL file are difficult to detect and imply the model specification is incorrect! This means that all inferences will be wrong, and also that all parameter estimates will be incorrect!

Defining population sizes and sample sizes

![Parameter tags](image)

2PopDivMigr10Loci.par

//Parameters for the coalescence simulation program: fsimcoal2.exe
2 samples to simulate:
//Population effective sizes (number of genes)
NPOP1
NPOP2
//Samples sizes and samples age
6
6
//Growth rates: negative growth implies population expansion
0
0

Population effective sizes are given in number of gene copies. For a diploid species with N=500 individuals, this corresponds to a 2N=1000 gene copies, as each individual carries two gene copies at any given site.

The sample size is also given in gene copies. The value of 6 means that we sampled 3 diploid individuals.
The migration matrix can be asymmetric, and in the case the entry $m_{ij}$ list the migration rates backward in time from population $i$ to population $j$. The above-mentioned matrix states that, for each generation backward in time, any gene from population 0 has probability $MIG_{01}$ to be sent to population 1, and that a gene from population 1 has a probability $MIG_{10}$ to move to population 0.

If no migration matrix is defined, no migration is assumed between populations.
A note on looking backward in time

Assuming that we look forward in time and that the size of the arrows are proportion to the migration rate, to what model does the following migration matrix corresponds to?

```plaintext
//Number of migration matrices: 0 implies no migration between demes
1
//migration matrix
0.000 0.005
0.001 0.000
```
A note on looking backward in time

Assuming that we look forward in time and that the size of the arrows are proportion to the migration rate, to what model does the following migration matrix corresponds to?

```
//Number of migration matrices : 0 implies no migration between demes
1
//migration matrix
0.000 0.005
0.001 0.000
```

Note that in the PAR and TPL files everything is backward in time!!

This is the correct model forward in time, meaning there are more migrants moving from pop0 to pop1 each generation.

Backward in time this is the model. Lineages are more likely to move from pop1 to pop0.
Historical events in fastsimcoal2

Historical events can be used to:

- Change the size of a given population
- Change the growth rate of a given population
- Change the migration matrix to be used between populations
- Move a fraction of the genes of a given population to another population. This amounts to implementing a (stochastic) admixture or introgression event.
- Move all genes from a population to another population. This amounts to fusing two populations into one looking backward in time.
- One or more of these events at the same time

Defining the historical events is crucial to have a correct model!
Historical events (backward in time)

Each historical event is coded with a line with the following arguments:
- time
- source
- sink
- migrants
- new deme size
- new growth rate
- migration matrix index

500 0 1 1 1 0 1

500 2 1 1 1 0 1

500 generations ago, 100% (migrants=1.0) of lineages in pop0 (source =0) migrated to pop1 (sink=1). The size of the sink (pop1) remained the same (new deme size=1.0, i.e. N2=2000). The new growth rate is zero. The migration rate that is active after the event is given in the migration matrix 1.
Historical events (backward in time)

Each historical event is coded with a line with the following arguments:

- time
- source
- sink
- migrants
- new deme size
- new growth rate
- migration matrix index

<table>
<thead>
<tr>
<th>time</th>
<th>source</th>
<th>sink</th>
<th>migrants</th>
<th>new deme size</th>
<th>new growth rate</th>
<th>migration matrix index</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>500</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

500 generations ago, 100% of lineages (migrants=1.0) in pop2 (source=2) migrated to pop1 (sink=1). The size of the sink (pop1) remained the same (new deme size=1.0, i.e. N2=2000). The new growth rate is zero. The migration rate that is active after the event is given in the migration matrix 1.
Historical events in fastsimcoal2

Change the size of a given population

1PopContrInst10Loci.par

//Parameters for the coalescence simulation program : fsimcoal2.exe
1 samples to simulate :
//Population effective sizes (number of genes)
1000
//Samples sizes and samples age
10
//Growth rates: negative growth implies population expansion
0
//Number of migration matrices : 0 implies no migration between demes
0

//historical events: time, source, sink, migrants, new deme size, new growth rate, migration matrix index
1 historical event
1000 0 0 0 1000 0 0

• 1000 generations ago, 0% (migrants=0) of lineages in pop0 (source) migrated to pop1 (sink). This means that 100% of lineages remained in pop0.

• The sink population (pop0) has a size 1000 larger after the event (new size=1000). Given that N0=500 diploids at time zero, it implies that NA=500000 diploids.

• The migration matrix valid after the event is the migration rate 0. Since it is not defined it implies no migration.
Historical events in fastsimcoal2

Change the migration matrix to be used between populations

- At generation 1000 in the past, 0% (migrants=0) of lineages migrated from pop0 (source=0) to pop1 (sink=0).
- After the historical event, the deme size of the sink population (pop1) remained the same (new deme size=1).
- After the historical event the growth rate was set to zero.
- After the historical event the migration rate matrix was set to matrix 1, i.e. no migration between populations.
Historical events in fastsimcoal2

Population split (merge populations going backwards in time)

- At generation 5000 in the past, 100% (migrants=1) of lineages migrated from pop1 (source=1) to pop0 (sink=0).
- After the population split, the deme size of the sink population (pop0) is 1500 (new deme size=1500/20000=0.075).
- After the historical event the growth rate of the sink population pop0 is zero.
- After the historical event the migration rate matrix was set to matrix 1, i.e. no migration between populations.
Launching parameter estimations

Command line to estimate parameters

```
fsc22.exe -t 1PopExpInst20Mb.tpl -e 1PopExpInst20Mb.est
-M0.001 -d -n100000 -N100000 -l5 -L20 -q -c4
```

Template file name

Estimation file name

No. of threads

Do ML estimation
And stop criterion

Use derived SFS

Minimum (-n) and maximum (-N) number of simulations to estimate the expected SFS

Minimum (-l) and maximum (-L) number of iterations to perform

Observed SFS file must have the same name as template file and extension _DAFpop0.obs. e.g. 1PopExpInst20Mb_DAFpop0.obs