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

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

2018 WSPG Cesky Krumlov
22 Jan 2020

vmsousa@fc.ul.pt
Outline part II

Example of Applications:

- Human dispersal out of Africa (high quality whole-genome) – lessons on choice of models
- Human colonization of Siberia and America (ancient whole-genome data) - lessons on dealing with sequencing errors
- 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


Nature(2016)

Ewaninga Rock Carvings Conservation Reserve, NT, Australia
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.
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)
We always get results…

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

Alves et al. (2012) Plos Genetics;
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
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).
Model captures the higher derived allele sharing between Eurasians and Yoruba.

D-statistics suggest that Yoruba and Eurasians share more derived alleles than Yoruba and Australians.
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
The population history of northeastern Siberia since the Pleistocene

Martin Sikora1,4,3*, Vladimir V. Pitulko2,4,3*, Vitor C. Sousa3,4,5,43, Morten E. Allentoft1,43, Lasse Vinner1, Simon Rasmussen6,41, Ashot Margaryan1, Peter de Barros Damgaard1, Constanza de la Fuente1,42, Gabriel Renaud1, Melinda A. Yang7, Qiaomei Fu7, Isabelle Dupanloup8, Konstantinos Giampoudakis9, David Nogués-Bravo9, Carsten Rahbek9, Guus Kroonen10,11, Michaël Peyrot11, Hugh McColl1, Sergey V. Vasilyev12, Elizaveta Veselovskaya12,13, Margarita Gerasimova12, Elena Y. Pavlova2,14, Vyacheslav G. Chasnyk15, Pavel A. Nikolskiy2,16, Andrei V. Gromov17, Valeri I. Khartanovich17, Vyacheslav Moiseyev17, Pavel S. Grebenyuk18,19, Alexander Yu. Fedorchenko20, Alexander I. Lebedintsev18, Sergey B. Slobodin18, Boris A. Malyarchuk21, Rui Martiniano22, Morten Meldgaard1,23, Laura Arppe24, Jukka U. Palo25,26, Tarja Sundell27,28, Kristiina Mannermaa27, Mikko Putkonen25, Verner Alexandersen29, Charlotte Primeau29, Nurbol Baimukhanov30, Ripan S. Malhi31,32, Karl-Göran Sjögren33, Kristian Kristiansen33, Anna Wessman27,34, Antti Sajantila25, Marta Mirazon Lahr1,35, Richard Durbin22,36, Rasmus Nielsen1,37, David J. Meltzer1,38, Laurent Excoffier4,5# & Eske Willerslev1,36,39,40#.

Nature (2019)
Colonization of Siberia

Yana RHS (31,600 years ago)
Whole-genome depth of coverage 25x

Kolyma (9,800 years ago)
Whole-genome depth of coverage 14x
Hypothesis: Continuity vs Replacement of populations

**Data:** Ancient and present-day samples; 625 blocks of 1Mb (~1.5 Million SNP), far from genic regions and CpG islands

**Method:** Composite likelihood - *fastsimcoal2* (Excoffier et al, 2013 Plos Genetics)
Hypothesis: Continuity vs Replacement of populations

For instance:

\( \beta = 1 \) indicates continuity: Kolyma descends from Yana

\( \beta = 0 \) indicates replacement of Yana by Kolyma
Site frequency spectrum is affected by damage patterns in ancient DNA

- High proportion of singletons in Kolyma probably reflect errors
- Thus, all analyses were performed discarding the singletons

![Graph showing relative number of SNPs for different populations.]

Proportion of singletons in Kolyma is reduced to 1/3 of original!

Derived allele frequency in:
- Sardinian
- Yana
- Karitiana
- Kolyma
- Han

#SNPs original dataset: 1,518,818
#SNPs after discarding transitions G->A,C->T: 938,911
Model comparison and likelihood profiles consistent with replacement with gene flow
Model comparison and likelihood profiles consistent with replacement with gene flow.
Estimates of best nested model indicate replacement with gene flow
Siberia and colonization of the Americas

Yana RHS (31,600 years ago)  
Whole-genome depth of coverage 25x

USR1 (11,500 years ago) Alaska  
Whole-genome depth of coverage 14x

Kolyma (9,800 years ago)
Estimates consistent with replacement with gene flow

- Kolyma is the closest population to Native Americans (USR1 and Karitiana)
- Native Americans with a contribution of up to 20% from Yana
Summary: 3 migration waves

• Ancient North Siberians (Yana) reached Siberia before 30 ka (thousand-years ago)
Summary: 3 migration waves

• Ancient North Siberians (Yana) reached Siberia before 30 kya
• Paleo-Siberians (Kolyma) migrated after Last Glacial Maximum (26.5 ka)
• Native-Americans are closer to Kolyma, with 20% of Yana contribution
Summary: 3 migration waves

- Ancient North Siberians (Yana) reached Siberia before 30 ka
- Paleo-Siberians (Kolyma) likely migrated after Last Glacial Maxima
- Native-Americans are closer to Kolyma, with 20% of Yana contribution
- Paleo-Siberians (Kolyma) were replaced by Neo-Siberians, likely associated with the cooler period “Younger Dryas” (12.8-11.5 ka)
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

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
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
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

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
Effect of HW filtering on demographic estimates
Removing sites with HWE excess and deficit leads to different estimates

- High migration between all groups of populations (2Nm~20)
- No evidence of a strong bottleneck signal associated with colonization of SH

REFERENCE \( N_{\text{ANC}} = 100,000 \)

\[ T_3 = 1.58 \ (\sim 127 \text{ kya}) \]
\[ N_{\text{south}} \sim 902,000 \]
\[ N_{\text{OFF N}} = 325,000 \]
\[ N_{\text{OFF S}} = 27.8 \]
\[ N_{\text{ON N}} = 287,000 \]
\[ N_{\text{ON S}} = 292,000 \]
\[ N_{\text{ANC ON}} \sim 400,000 \]
\[ N_{\text{BOT}} = 1582 \]
\[ T_2 = 0.28 \ (\sim 23 \text{ kya}) \]
\[ T_1 = 0.19 \ (\sim 16 \text{ kya}) \]
Sawflies and RAD data

MOLECULAR ECOLOGY

History, geography and host use shape genomewide patterns of genetic variation in the redheaded 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*

Needle width
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.
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
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>$\log_{10}(\text{Lhood})$ ALL SNPs</th>
<th>$\log_{10}(\text{Lhood})$ 1 SNP</th>
<th># Parameters</th>
<th>AIC</th>
<th>$\Delta$AIC</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><strong>-46462.24</strong></td>
<td><strong>-7361.5</strong></td>
<td><strong>15</strong></td>
<td><strong>33931.01</strong></td>
<td><strong>0.00</strong></td>
<td><strong>0.851</strong></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>
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.
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

- The sample size can vary across SNPs due to missing 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>

![Graph showing frequency of derived alleles](#)
Estimating SFS from observed data

- The sample size can vary across SNPs due to missing 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>
Estimating SFS from observed data

- The sample size can vary across SNPs due to missing 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. (2011) PNAS
Acknowledgements

Martin Sikora
Laurent Excoffier
Isabelle Dupanloup
Stephan Peischl
Eske Willerslev

Thank you!

Catherine R. Linnen
Stefan Laurent
Jeffrey D. Jensen
Susanne Pfeifer
Hopi E. Hoekstra
Laurent Excoffier

MCSA 2018-2020: MAPgenome (N.799729)