Evolutionary Genomics - Fastsimcoal2 practical

Thibault Schowing[*]

March 2020

Application to real data from human population

The folder ”/Henn_et_al_data” contains exome data from two human populations (Namibian San, SAN and Mexican Mayan, MAYA) published in *Henn et al.* (2015) PNAS. This data was kindly provided by Stephan Peischl, and cannot be used outside the scope of this course.

The folder /Henn_et_al_data contains the following files:

- 2D-SFS for the two populations, assuming that we sequenced a total of 44Mb
- Original genotypic matrix
- TPL and EST files for a model without gene flow

Infer the time of divergence between two human populations, **San** from Africa and **Maya** from America.

- Start by considering a model without gene flow.
- Then, consider a model with gene flow.

With the given R script, we generate the following command[1]:

```
fsc26.exe  -t NoMig_San_Maya.tpl -e NoMig_San_Maya.est -L 20 -n 100000 -d -M -q -C1 -c2 -B2;
```

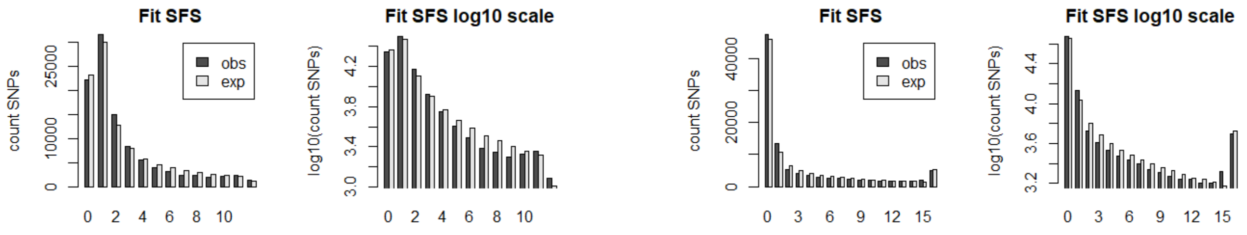
Once we have executed this command, we take the model with the maximum likelihood. As a reminder, **MaxObsLhood** is the maximum possible value for the likelihood if there was a perfect fit of the expected to the observed SFS, i.e. if the expected SFS was the relative observed SFS and **MaxEstLhood** is the maximum likelihood estimated according to the model parameters. From this model we get these estimated values: a first population with 6694 individuals, a second with 32513 individuals that come from an ancestral population of 18581 individuals that diverged 3699 years ago.

```
NPOP1   NPOP2   NANC    TDIV   MaxEstLhood  MaxObsLhood
6994    32513   18581   3699   -476858.5    -475819.9
```

[*] Teacher: Dr. Vitor Sousa, UNIBE, Switzerland

[1] Slightly modified for Windows: copy the fsc26.exe from the Fastsimcoal folder in the folder containing the data and replace ”./fsc26” with ”fsc26.exe”.

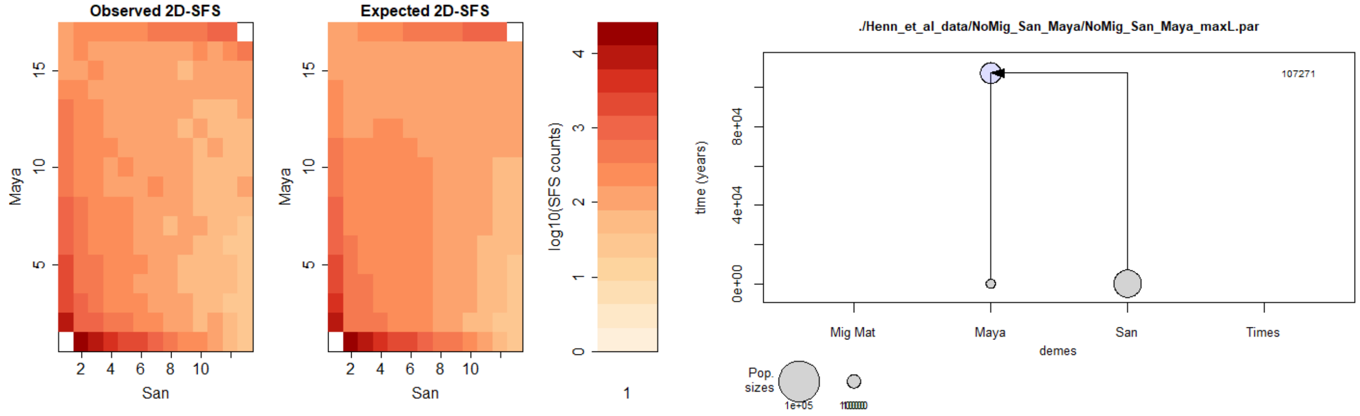
SFS for the San and Maya population without migration



(a) San

(b) Maya

Figure 1: SFS of the San and Maya populations without migration since divergence time. There is a rather good fit between the observed (Real data) and expected (Model) SFS.



(a) 2D SFS - Observed vs Expected

(b) Scenario.

Figure 2: Scenario assuming generations of 29 years and no migration. There is a rather good fit between the two 2D-SFS, the pattern between the two is clearly similar.

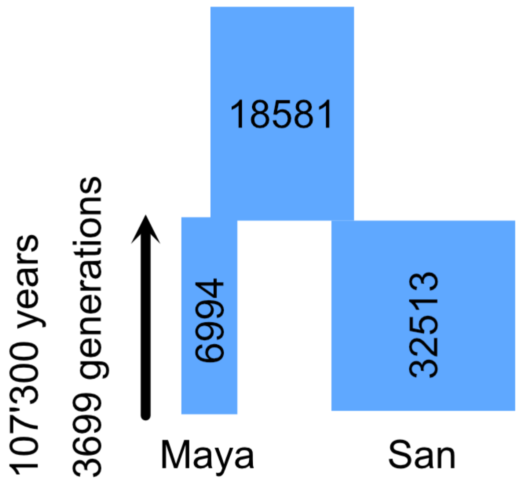


Figure 3: Generated model of the populations without migration.

Same execution but with migration

We execute the following simulation:

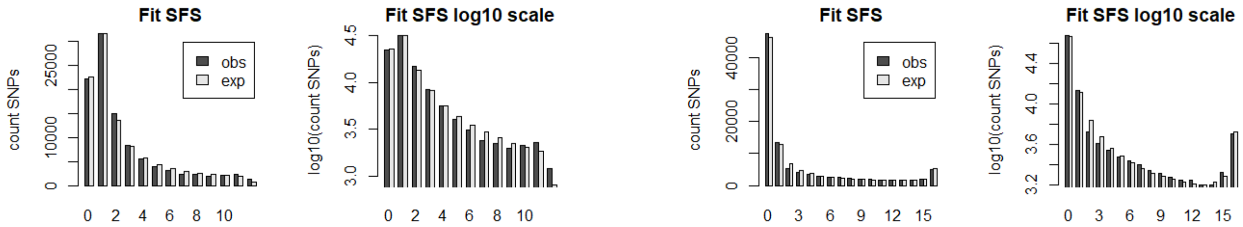
```
./fsc26  -t Mig_San_Maya.tpl -e Mig_San_Maya.est -L 20 -n 100000 -d -M -q -C1 -c2 -B2;
```

The maximum likelihood result gives:

```
NPOP1   NPOP2   NANC    TDIV    N1M12       N2M21
6539    26955   16356   7220    0.3210159   1.085235

RESIZE0  MIG12       MIG21        MaxEstLhood   MaxObsLhood
2.5013   4.90925e-05 4.0261e-05   -476313.5    -475819.9
```

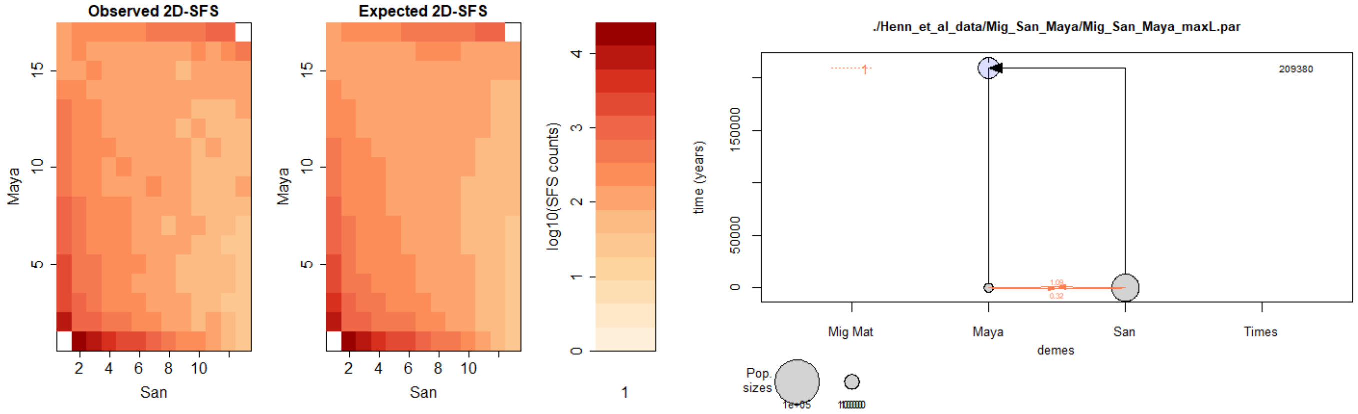
SFS for the San and Maya population with migration



(a) San

(b) Maya

Figure 4: SFS of the San and Maya populations with migration since divergence time.



(a) 2D SFS - Observed vs Expected

(b) Scenario.

Figure 5: Scenario assuming generations of 29 years.

What is the effect of adding gene flow into the model? What is the model that better fits the data?

Adding gene flow allows to transfer genetic variation between the two population. If the migration rate (gene flow) between the two population is high, the two populations will have equivalent allele frequencies and this be close to identical. On the other hand, if the migration rate is null, the two populations can experience speciation due to drift. It has been shown that one migrant per generation can prevent two populations to diverge (Frankham, Richard; Briscoe, David A.; Ballou, Jonathan D. (2002-03-14). *Introduction to Conservation Genetics.* Cambridge University Press.).

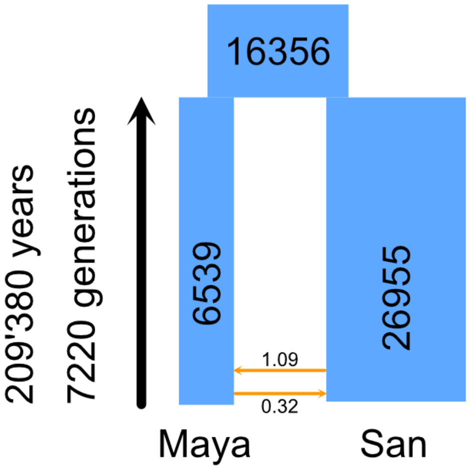


Figure 6: Generated model of the populations with migration. There is on average 1.09 migrant per generation from San to Maya population and 0.32 from Maya to San.

The model including migration fits the data even better. The 2D-SFS's are really alike. We can observe that the San population is estimated to be much smaller since the split when we include migration. It is also shown that the populations split is much closer in time without migration (100'000 years) as compared with the model with migration (200'000 years).

Are these models sufficient to explain the data? Here we analyzed data from exome sequencing. Can we trust these demographic estimates?

The whole exome sequencing has been tested to be an efficient method for population genetics analysis[2] but to explain the data it is important to also rely on historical and geographical data and to compare to other related populations. According to *Henn et al.*, the Out-Of-Africa (OOA) dispersal, that caused a lot of founder events around the world, happened approximately 50'000 years ago, which does not match neither the model with migration, nor the one without. Other OOA dispersals happened[3] at various periods up to until 120k years ago. Parameters of the model, statistical methods, data or exactitude of migration events might need adjustment to support each other.

[2] *Maróti, Z., Boldogkői, Z., Tombácz, D. et al. Evaluation of whole exome sequencing as an alternative to BeadChip and whole genome sequencing in human population genetic analysis. BMC Genomics 19, 778 (2018).* https://doi.org/10.1186/s12864-018-5168-x

[3] *López et al. Human Dispersal Out of Africa: A Lasting Debate. Evolutionary Bioinformatics 2015:11(s2) 57–68 doi: 10.4137/EBo.s33489.*