Supplementary Material for manuscript: On the importance of being structured: instantaneous coalescence rates and human evolution - Lessons for ancestral population size inference

Mazet Olivier, Rodríguez Willy, Grusea Simona, Boitard Simon, Chikhi Lounès

November 4, 2015

The Supplementary Figures were obtained with data simulated using the *ms* software (Hudson, 2002), as is the main manuscript. For each scenario, we simulated independent values of T_2 and used them to estimate the IICR at various time points t_i, using the following equation:

$$\widehat{IICR(t_i)} = \frac{1 - \widehat{F_{T_2}(t_i)}}{\widehat{f_{T_2}(t_i)}} \tag{1}$$

where $\widehat{F_{T_2}(t_i)}$ is the estimated or empirical cumulative distribution function of T_2 and $\widehat{f_{T_2}(t_i)}$ is an estimated approximation of its density around t_i.

1 Supplementary Figure 1

For Figure S1 the *ms* commands were

- *ms 2 1000000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 1 -eN 1 0.1* for panel (a)
- *ms 2 1000000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 1 -eN 1 0.02* for panel (b).

[Figure 1 about here.]

2 Supplementary Figure 2

For Figure S2 the ms commands were

- *ms 2 1000000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 1 -eN 1 0.1 -eM 1 0.1* for panel (a)
- *ms 2 1000000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 1 -eN 1 0.02 -eM 1 0.02* for panel (b)

[Figure 2 about here.]

3 Supplementary Figure 3

For Figure S3 the ms commands used were:

- *ms 2 1000000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 0.5 -eN 1 0.5 -eN 0.01 0.1 -eN 0.06 1 -eN 0.2 0.5 -eN 1 1 -eN 2 2*
- *ms 2 1000000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 1 -eN 1 0.5 -eN 0.01 0.1 -eN 0.06 1 -eN 0.2 0.5 -eN 1 1 -eN 2 2*
- *ms 2 1000000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 10 -eN 1 0.5 -eN 0.01 0.1 -eN 0.06 1 -eN 0.2 0.5 -eN 1 1 -eN 2 2*
- *ms 2 1000000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 20 -eN 1 0.5 -eN 0.01 0.1 -eN 0.06 1 -eN 0.2 0.5 -eN 1 1 -eN 2 2*
- *ms 2 1000000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 50 -eN 1 0.5 -eN 0.01 0.1 -eN 0.06 1 -eN 0.2 0.5 -eN 1 1 -eN 2 2*
- *ms 2 1000000 -T -L -I 10 2 0 0 0 0 0 0 0 0 0 100 -eN 1 0.5 -eN 0.01 0.1 -eN 0.06 1 -eN 0.2 0.5 -eN 1 1 -eN 2 2.*

[Figure 3 about here.]

4 Supplementary Figure 4

This figure is identical to the figure in the main manuscript except that no recent human expansion was simulated since the PSMC is not reliable for the corresponding time window.

[Figure 4 about here.]

References

Hudson RR (2002). Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337–338. **URL:** *http://bioinformatics.oxfordjournals.org/content/18/2/337.abstract*

Li H, Durbin R (2011). Inference of human population history from individual whole-genome sequences. *Nature* **475**: 493–496.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, *et al.* (2008). The diploid genome sequence of an Asian individual. *Nature* **456**: 60–65.

List of Figures

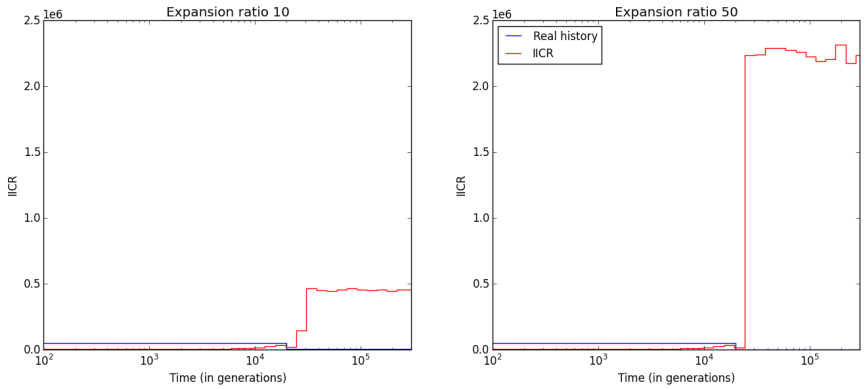

Figure S1: Inferred population size changes for n-island models with a recent increase in population size and constant M. This figure shows the predicted population size changes that will be inferred for an n-island model under the assumption that populations are not structured. For both panels the x-axis represents time in generations, whereas the y-axis represents real or inferred population size in units of $2 \times N \times n$, where N is the haploid deme size, and n the number of islands. We simulated an n-island model with $n = 10$ and computed the IICR using the simulated T_2 values (see main manuscript). Data where simulated under an n-island model with $n = 10$ and $M = 1$. In both panels all the demes increased in size at $T = 2,000$ generations. The IICR is plotted in red whereas the actual size of the metapopulation is plotted in blue. In panel (a) the population increased by a factor 10, whereas in panel (b) it increased by a factor 50. The inferred history of population size changes does not reflect the actual changes. Instead of a single population increase, the IICR suggests that the population has been decreasing over a very long time scale with a bottleneck and a quick recovery followed by a final bottleneck.

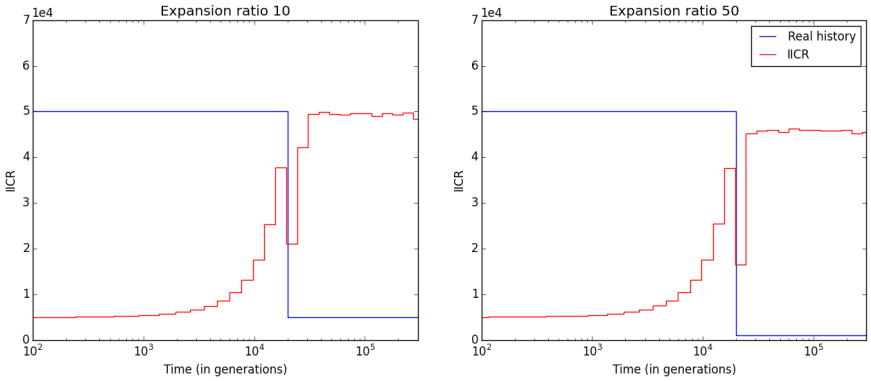

Figure S2: Inferred population size changes for n-island models with a recent increase in population size and constant migration rate, m. This figure shows the predicted population size changes that will be inferred for an n-island model under the assumption that populations are not structured. For both panels the x-axis represents time in generations, whereas the y-axis represents real or inferred population size in units of $2 \times N \times n$, where N is the haploid deme size, and n the number of islands. We simulated an n-island model with $n = 10$ and computed the IICR using the simulated T_2 values and equation 1. Data where simulated under an n-island model with $n = 10$. In both panels all the demes increased in size at $T = 2,000$ generations and migration rate $m = M/2N$ was constant (i.e. we changed the value of M in the same way we changed the size of the metapopulation). The IICR is plotted in red whereas the actual size of the metapopulation is plotted in blue. In panel (a) the population increased by a factor 10, whereas in panel (b) it increased by a factor 50.

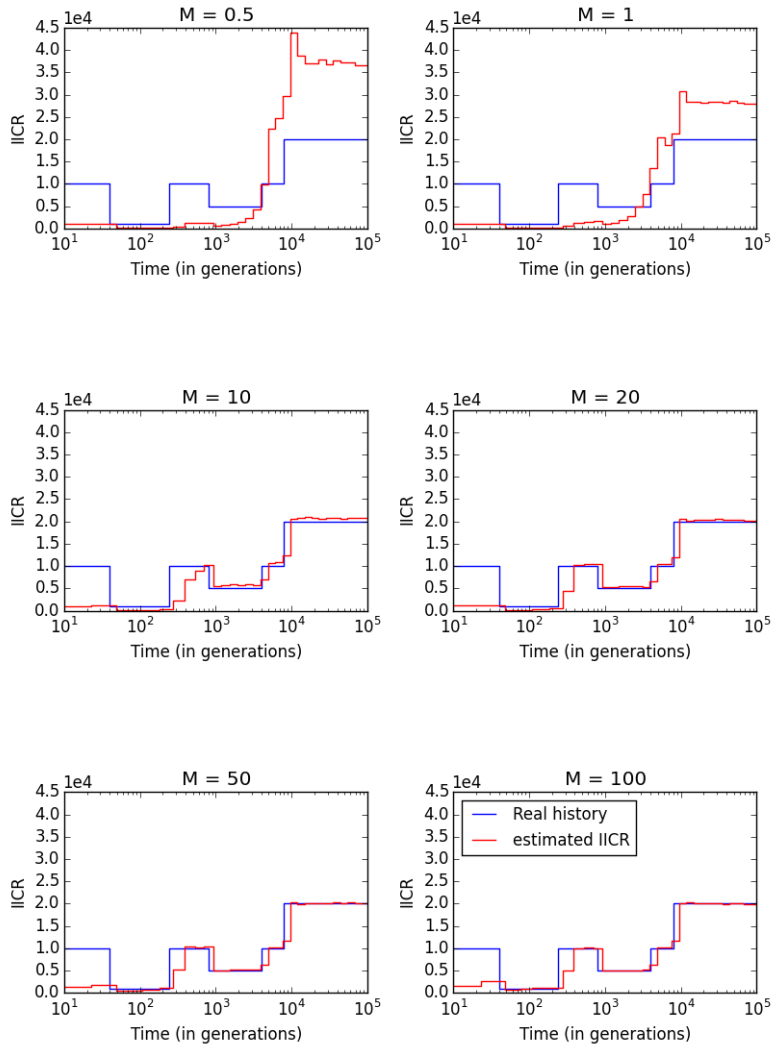

Figure S3: Inferred population size changes for an n-island model with constant migration rates and changes in metapopulation size. For both panels the x-axis represents time in generations, whereas the y-axis represents the IICR or actual population size in units of $2 \times N \times n$, where N is the haploid deme size, and n the number of islands. In all cases the metapopulation size varied following a series of stepwise changes represented by the blue line. The various panels correspond to various values for M the number of haploid genomes exchanged, namely $M = 0.5, 1, 10, 20, 50, 100$. For values of $M \geq 10$ (*i.e.* $F_{ST} \leq 0.025$ the changes in metapopulation sizes are either well or relatively well estimated when they are ancient but even for $M = 100$ the very recent history is poorly inferred. We also note that even for $M = 1$ the ancient history is poorly correlated to the IICR. Altogether this suggests that the PSMC (and related methods) will perform worse since it should infer the IICR from DNA sequences which is more hard due to the fact that the T_2 distribution must be inferred first.

.

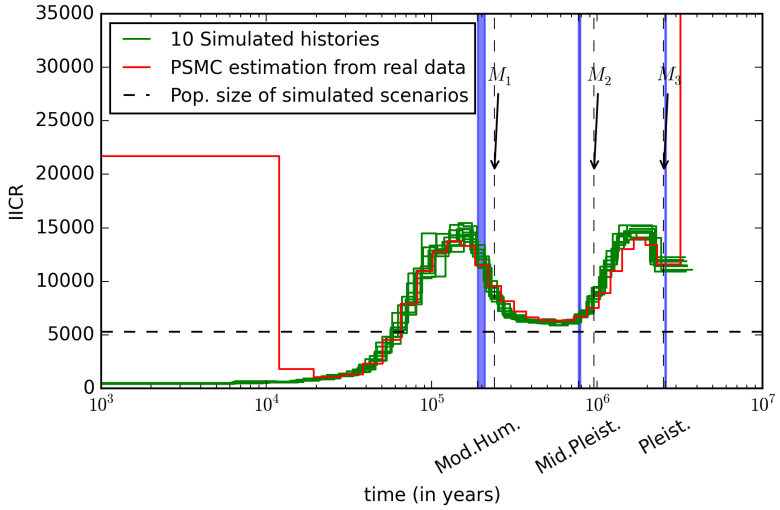

Figure S4: Human history with changes in migration rates. This figure shows, in red, the history of population size changes inferred by Li and Durbin from the complete diploid genome sequences of a Chinese male (YH) (Wang *et al.*, 2008). The 10 green curves correspond to the IICR of ten independent replicates of the same demographic history involving three changes in migration rates and no population size change. The x-axis represents time in years in a log scale, whereas the y-axis represents real or inferred population size in units of diploid genomes. The times at which these changes occur are represented by the vertical arrows at 2.52 MY ago, 0.95 MY ago and 0.24 MY ago. The blue shaded areas correspond to (i) the beginning of the Pleistocene (Pleist.) at 2.57-2.60 MY ago, (ii) the beginning of the Middle Pleistocene (Mid. Pleist.) at 0.77-0.79 MY ago, and (iii) the oldest known fossils of anatomically modern humans (AMH), at 195-198 KY ago. Following Li and Durbin (2011) we assumed that the mutation rate was $\mu = 2.5 \times 10^{-8}$ and that generation time was 25 years. We also kept their ratio between mutation and recombination rates. Each deme had a size of 530 diploids and the total number of haploid genomes was thus constant and equal to $10,600$.