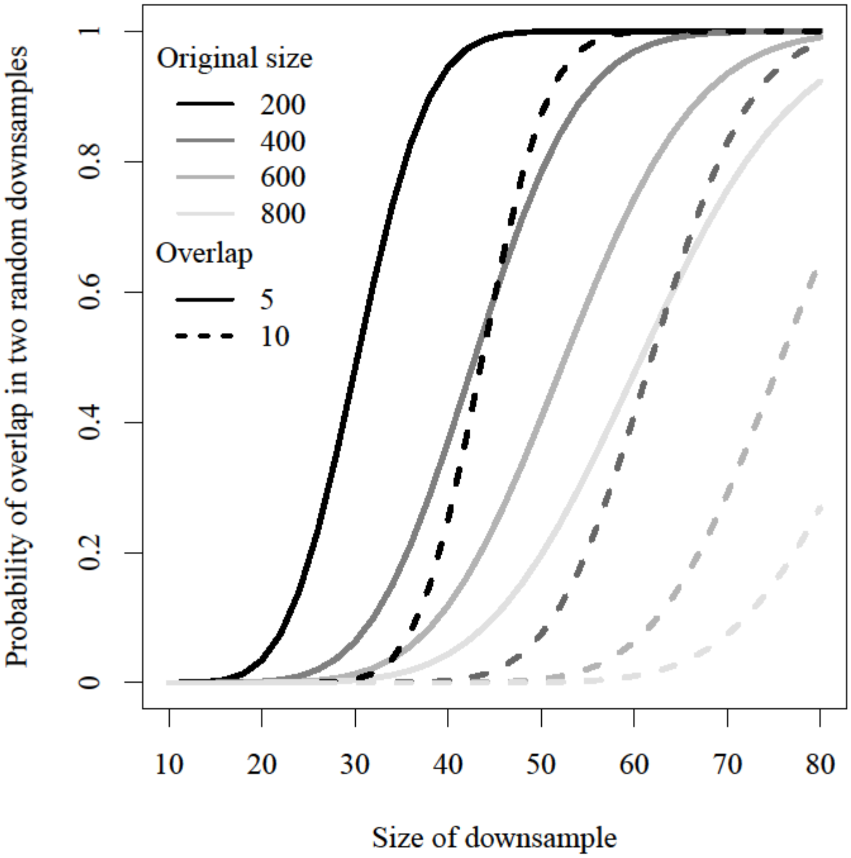
Appendix S4. Overlap between subsamples

Subsampling datasets of limited size could come with a bias: as the proportion of individuals sampled from the original dataset increases, the likelihood of repeatedly selecting the same individuals in consecutive subsamples invariably increases, making the different replicates more and more similar. Realistically, some level of overlap must be accepted in studies based on real data because of their limited sample size (as opposed to pure simulation studies where sample size is/should not be limiting).

Most previous studies based on subsampling of empirical datasets faced this problem (e.g. 10,000 datasets generated from original sample sizes of 100; Pruett & Winker 2008), but only one tried to estimate its extent (Miyamoto *et al.* 2008). In the former case, the chances of drawing extremely similar subsampled datasets were very high, leading the authors to potentially erroneous conclusions, i.e. the underestimation of sample sizes required to achieve reliable parameter estimates.

Here, we quantified the extent of this problem by computing the probability that two subsampled datasets of size *m*, taken in an original dataset of size *n*, present an overlap of at least *r* individuals. This is illustrated below, with four levels of original sample size (200, 400, 600 and 800 samples) and two different overlaps (5 and 10 samples). Clearly, very large original sample sizes are needed to strictly avoid overlap, as even with an original size of 800 samples, there is a 0.2 probability to observe an overlap of 10 samples in a subsampled dataset of 75 samples.

With our larger empirical dataset (726 samples), there is a 0.077 probability to observe an overlap of at least 15 samples for two subsamples of 76 samples (i.e. 20% overlap at maximum). With simulated datasets (500 samples), this probability is 0.15. When interested in sex ratio, the maximum sample sizes are halved. With our larger empirical dataset (363 samples of each sex), there is a 0.01 probability to observe an overlap of at least 6 samples (i.e. 20% overlap approximately) for two subsamples of 28 samples. With simulated datasets (500 samples), this probability is 0.07. Therefore, these probabilities did not, in our opinion, dramatically affect our results and conclusions.



Probability that two independent subsampled datasets present an overlap of at least 5 (solid lines) or 10 individuals (dashed lines) as a function of the size of the subsample. We present here these probabilities for four original sample sizes: 200, 400, 600 and 800 (from black to light gray).

References

Miyamoto, N., Fernández-Manjarrés, J.F., Morand-prieur, M.-E., Bertolino, P. & Frascaria-Lacoste, N. (2008). What sampling is needed for reliable estimations of genetic diversity in *Fraxinus excelsior* L. (Oleaceae)? *Annals of Forest Science*, **65**, 403–403.

Pruett, C. & Winker, K. (2008). The effects of sample size on population genetic diversity estimates in song sparrows *Melospiza melodia*. *Journal of Avian Biology*, **39**, 252–256.