Inferring the Demographic History of Inbred Species from Genome-Wide SNP Frequency Data

Paul D. Blischak[1,2], Michael S. Barker[1], Ryan N. Gutenkunst[2]

[1]Ecology and Evolutionary Biology, [2] Molecular and Cellular Biology, University of Arizona; rgutenk@email.arizona.edu; http://gutengroup.mcb.arizona.edu

Summary

To enable demographic history inference in inbred species, we implemented a model for the site frequency spectrum with inbreeding into the inference software dadi. Using simulations, we showed that our approach is unbiased and powerful. We then applied our method to American pumas (not presented) and domesticated cabbage. Our results show that inbreeding can have a strong effect on demographic inference, particularly for parameters involving changes in population size. Given the importance of these estimates for informing practices in conservation, agriculture, and elsewhere, our method provides an important advancement for accurately estimating demographic histories.

Model

The site-frequency-spectrum (SFS) summarizes genetic variation within and between populations using the observed number of SNPs at any given sample frequency. Recent inbreeding increases homozygosity, increasing even entries of the SFS and decreasing odd entries.

Balding and Nichols (1995, 1997) proposed a probability model for inbreeding based on the beta-binomial distribution. Here $g \in 0,1,2$ is the individual genotype, p is the population allele frequency and F is the inbreeding coefficient.

$$Pr(G_i = g|p,F) = \mathcal{BB}\left(g, \alpha = p\left[\frac{1-F}{F}\right], \beta = (1-p)\left[\frac{1-F}{F}\right]\right)$$

The expected number of derived alleles D in a sample of n individuals is then a sum over all possible ways generating genotypes that sum to $D=d$.

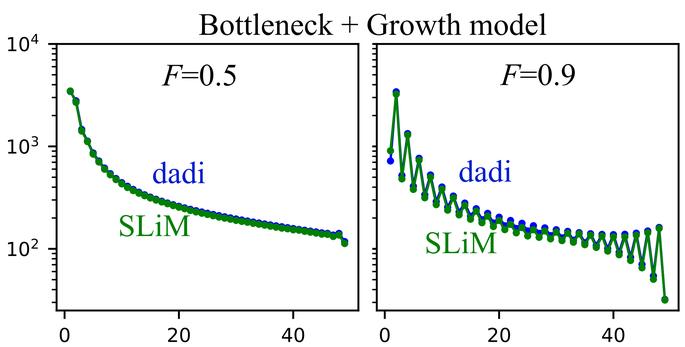
$$P(D = d|p,F) = \sum_{R \in p_n(d)} \frac{n!}{n_0!\, n_1!\, n_2!} \left[\prod_{r \in R} \mathcal{BB}(r,\alpha,\beta)\right]$$

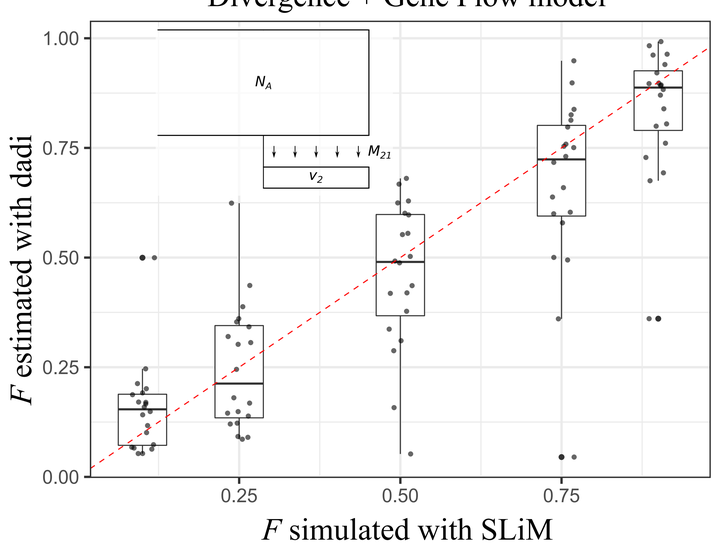
Availability

Published in Molecular Biology and Evolution:
https://doi.org/10.1093/molbev/msaa042

Implemented in dadi:
https://bitbucket.org/gutenkunstlab/dadi

Validation





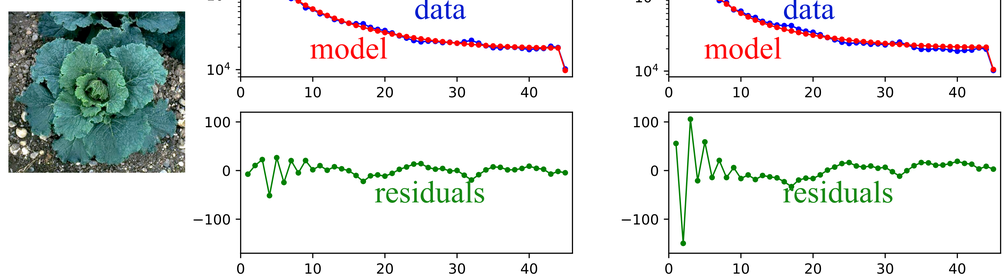
To validate our approach, we compared with forward simulations in SLiM 3 (Haller and Messer 2019). We found good qualitative agreement in the spectra and good quantitative agreement in inferred parameter values.

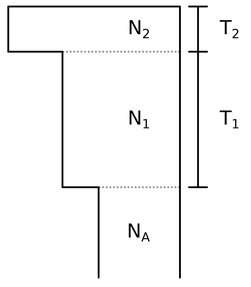
Acknowledgements

This research was supported by an NSF Postdoctoral Research Fellowship (IOS-1811784 to PDB) and by the National Institute of General Medical Sciences of the NIH (R01GM127348 to RNG).

Application to Cabbage

We applied our approach to data from domesticated cabbage (B. oleracea var. capitata, Chen et al. 2016a,b), which is thought to have been domesticated roughly 500 years ago. Models without inbreeding qualitatively failed to the fit the SFS and inferred an implausible domestication time and population crash. Including inbreeding fixed both issues.





Parameter	Estimate With Inbreeding	Estimate Without Inbreeding
N_A	17,500 (16,900–18,100)	19,100 (18,500–19,800)
N_1	31,600 (28,900–34,700)	123,000 (80,400–190,000)
N_2	215,000 (4,910–9,370,000)	592 (547–641)
T_1	16,600 (12,900–21,200)	5,870 (5,200–6,620)
T_2	322 (94.2–1,097)	38.3 (32.5–45.1)*
F	0.578 (0.557–0.599)	–

Discussion

Our model of recent inbreeding performs well in simulation. Application to empirical data shows that neglecting inbreeding can severely bias demographic history inference. Note that our model is for recent inbreeding. Ancient inbreeding can be accounted for by scaling the effective population size based on the inbreeding coefficient (Charlesworth 2003).

References

Balding, DJ and Nichols, RA 1995. Genetica, 96: 3–12.
Balding, DJ and Nichols, RA 1997. Heredity, 108: 583–589.
Charlesworth, D 2003. Philosophical Transactions of the Royal Society B, 358: 1051–1070.
Cheng, F et al. 2016a. Scientific Data, 3: 160119.
Cheng, F et al. 2016b. Nature Genetics, 48: 1218–1224.
Haller, BC and Messer, PW 2019. Molecular Biology and Evolution, 36: 632–637.