Size matters: how sample size and molecular marker choice affect mixed stock analyses

Gustavo Stahelin[1,2], Katherine Mansfield[1]

[1]University of Central Florida, Department of Biology, Orlando, FL, USA. [2]Fundação Projeto Tamar, Salvador, BA, Brazil.

Introduction

Small sample sizes can generate misleading research findings.

Mixed stock analysis is a powerful tool to assess genetic composition and source populations in mixed aggregations.

A variety of molecular markers exist for population differentiation.

There is variation in the capacity to differentiate source population depending upon molecular marker used.

Yet, there is no direct guidance on the effects of sample size or molecular marker on model accuracy.

Objective

Simulate the effect of sample size and molecular marker type on MSA model accuracy.

Methods

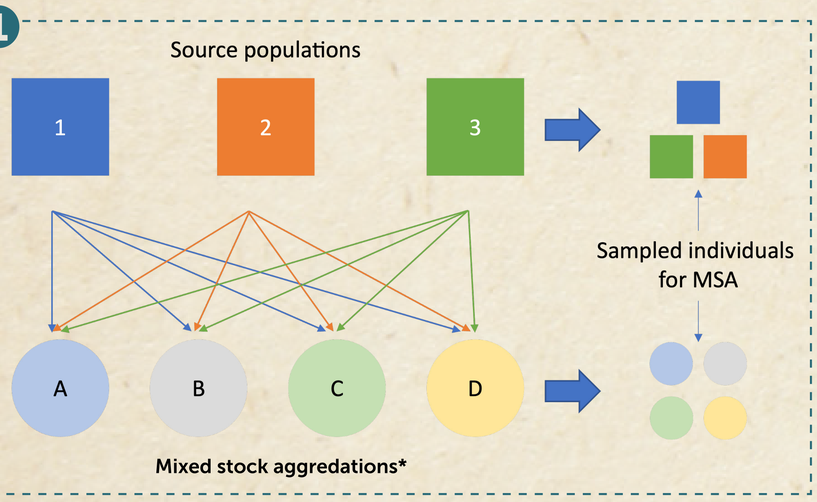
Simulated three source populations and four mixed stock aggregations (Fig. 1).

Randomly sampled in rookeries and mixed stock aggregations to simulate field data collection.

Ran mixed stock analysis (mixstock R package).

We considered two scenarios, one with greater haplotype frequency overlap between source populations and one with lower overlap (Fig. 2).

1



*Known true contribution from each source population.

Figure 1: Diagram indicating simulations. Source populations and source population sampling (square), random migration of individuals between areas, and mixed stock aggregations and mixed stock aggregation sampling (circle) were generated for each simulation.

2

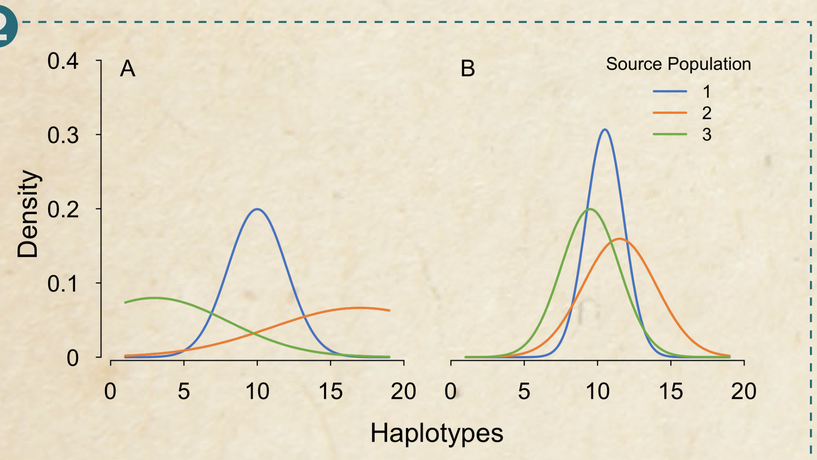


Figure 2: Histogram of simulated haplotype frequencies (n = 20). (A) low-resolution genetic marker scenario (B) high-resolution genetic marker scenario. Colors represent the different rookeries.

We used 30,000 simulations with the following variables in each scenario:

- Source population size: 500, 1,000, 10,000, and 50,000;
- Mixed stock aggregation size: 500, 1,000, and 5,000;
- Number individuals sampled: 25 to 200 samples

We compared model accuracy for each scenario by calculating the absolute difference between true contribution to mixed stock aggregations and mean model estimate, comparing credible interval width, and calculating the probability of true contribution to be contained inside credible intervals.

Results

Haplotype frequencies of rookeries can be accurately characterized in studies using MSA with 30-50 samples.

Model accuracy is maximized with **higher-resolution markers and near 100 samples** from mixed stock aggregations (Fig. 3).

Using **lower-resolution markers** (e.g., 400 bp mtDNA fragments), model accuracy is only maximized by **using >150 samples** from mixed stock aggregations (Fig. 3).

3

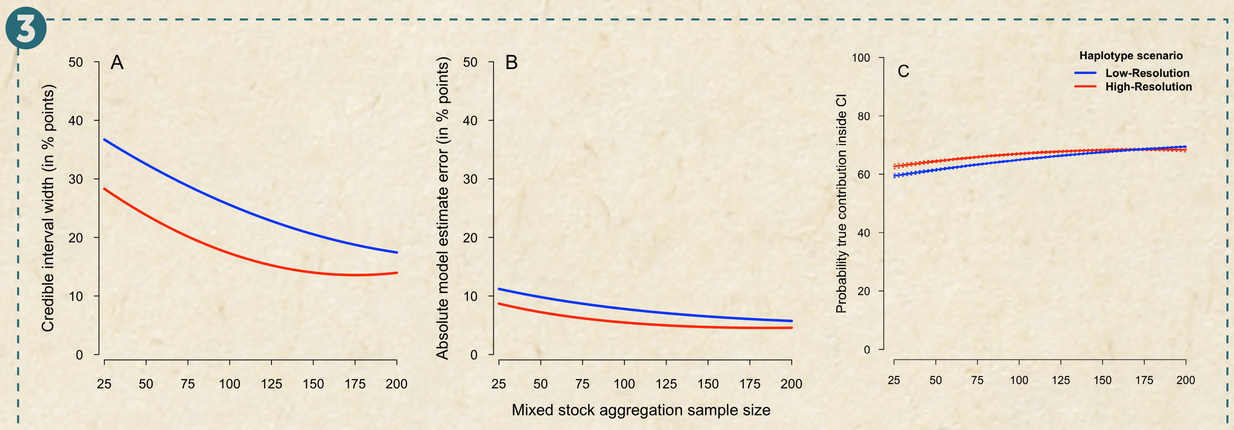


Figure 3: Comparing model accuracy using different metrics as a function of mixed stock aggregation sample size between the two tested haplotype scenarios: low-resolution (blue) and high-resolution marker (red). (A) Absolute credible interval width, (B) absolute difference between true contribution and mean model estimate, and (C) probability of true contribution to be contained inside the credible interval. Dashed lines represent 95% credible interval.

Discussion

Large sample sizes may be hard to achieve and expensive to process.

We strongly recommend the use of higher resolution markers to better distinguish rookeries and reduce sample sizes.

Need for rookeries reassessment using higher-resolution markers.

More accurate MSA estimates can improve our understanding of connectivity and sea turtle **dispersal patterns** across habitats and, ultimately, **conservation efforts** for populations.

ACKNOWLEDGMENTS

We thank Dr. D. Jenkins for his guidance while designing the simulations and statistical testing. The analyses presented here were made possible by computational resources available at Stokes High Performing Computing Cluster at University of Central Florida Advanced Research Computing Center. G. Stahelin is a grantee of the '*Ciência Sem Fronteiras*' scholarship program.