Population Genomic Jigsaw Puzzle: Unraveling Inferences from Convolutional Neural Networks with Data Scrambling

Linh N. Tran[1,2], David Castellano[2], Ryan N. Gutenkunst[2]

[1] Genetics Graduate Interdisciplinary Program, [2] Department of Molecular and Cellular Biology, University of Arizona; lnt@arizona.edu

THE UNIVERSITY OF ARIZONA

This research was supported by the NIH-NIGMS (R01GM127348 & R35GM149235 to RNG).

Significance

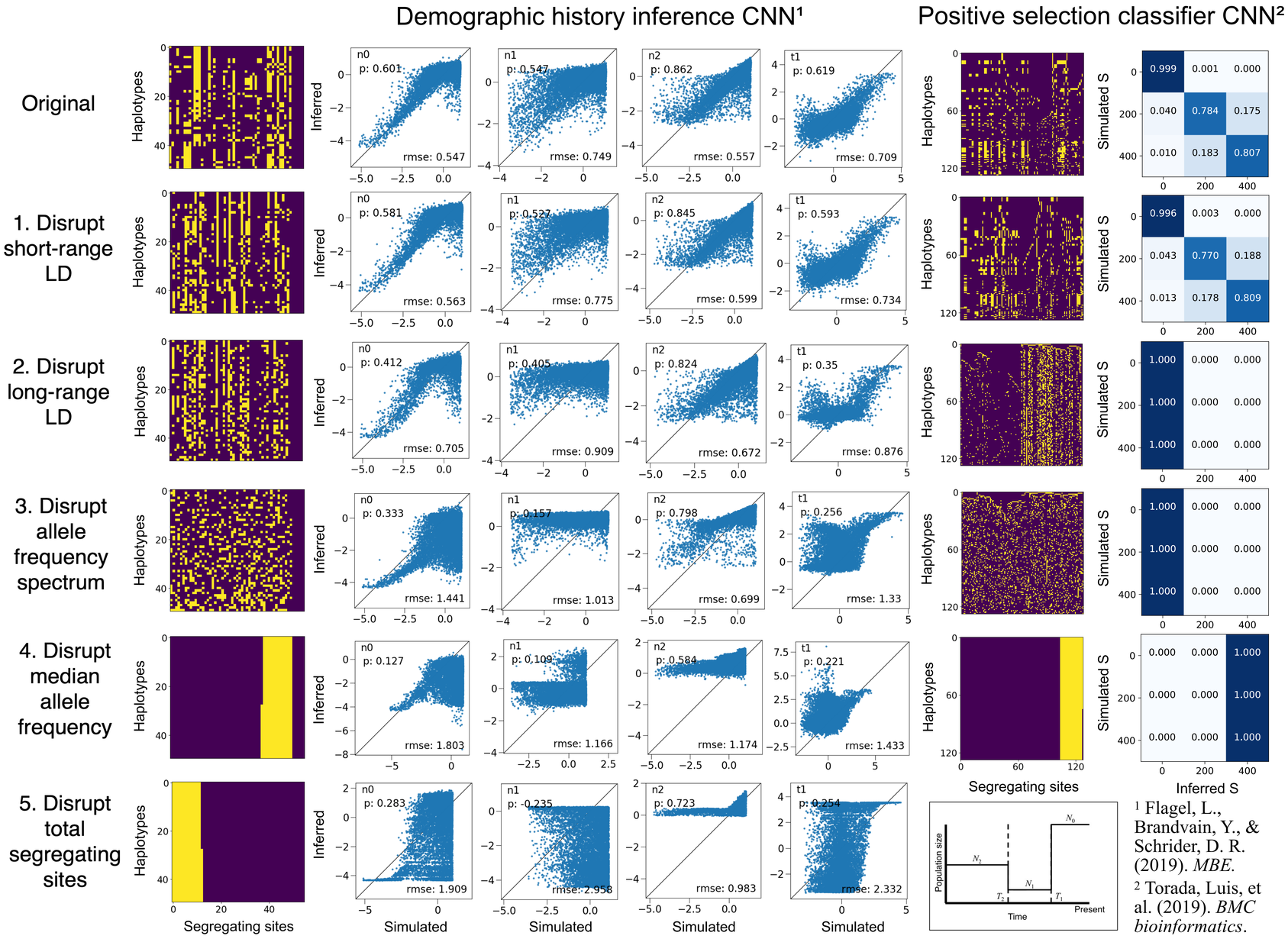
- Convolutional neural networks (CNNs) have been applied to many inference tasks in population genetics but it is often unclear which data features contribute to the reported performance.
- We designed a suite of scramble tests motivated by population genetic understanding of genomic data to provide biological interpretations for and to better evaluate the efficacy of CNN learning.

Method

- We investigated a CNN that infers 5 demographic history parameters (population sizes N_0, N_1, N_2 and times T_1, T_2) of a 3-epoch model, and a CNN that classifies genomic regions into 3 classes: under neutrality (S=0), under weak to moderate selection (S=200), and under strong selection (S=400).
- We reproduced all simulations, CNN training, and testing per the original published work. We then performed various levels of scrambling on the original simulated test data, where each level disrupts known genomic data features, before challenging the original CNNs with each level of scrambled test data.

Results

- With the original test data, we achieved performance similar to the reported results in the published studies.
- To disrupt short-range linkage disequilibrium (LD), we randomly shuffled all columns of segregating sites in the genomic data image. This perturbation has little effect on the performance of the two CNNs, suggesting a minimal contribution of short-range LD patterns to CNN learning.
- To disrupt long-range LD, we randomly shuffled entries within each column after shuffling entire columns as in scramble test 1. The selection CNN is very sensitive to this disruption, as performance completely breaks down with all-neutral classifications. In contrast, the demographic inference CNN remains relatively robust even though performance decreases more noticeably than in test 1.



- To disrupt the allele frequency spectrum (AFS) pattern, we randomly shuffled all entries. The demographic inference CNN performs significantly worse, indicating that AFS signal has a high impact on CNN learning for this task.
- In tests 4 & 5, grouping distinct allele entries into one block alters the average allele frequency across sites compared to test 3, since most sites (columns) here will have an extreme allele frequency (0 or 1). Test 4 further breaks down the demographic inference CNN's performance and causes the selection CNN to exhibit a bias toward strong selection as opposed to neutrality as in test 3.
- For CNNs that have zero-padding as a data preprocessing step (e.g. the demographic inference CNN), the position of the distinct allele block with respect to the padding dictates whether the total segregating sites boundary is preserved in the genomic data image. The demographic inference CNN performs much worse on test 5 compared to test 4, demonstrating the importance of this boundary to the CNN.

[1] Flagel, L., Brandvain, Y., & Schrider, D. R. (2019). *MBE*.

[2] Torada, Luis, et al. (2019). *BMC bioinformatics*.