

Mendelian Simulation of STR Genotypes from Existing Sample Data to Create Standard Truth Pedigrees for Benchmarking New Familial Inference Methods

Stephen D. Turner, Christopher A. Hulme-Lowe, Christopher R. Simpson, Carlos A. Acevedo

Background

Current methods used in forensic DNA analysis evaluate the likelihood of a match based on STR length. Recently there has been interest in extending the use of DNA databases beyond exact matching toward identifying potential familial relationships between individuals in a DNA database and/or with an unidentified person of interest whose DNA is found at a crime scene.

Objective

When developing algorithms for identifying relationships, inferring pedigrees, and quantifying uncertainty, it is important to have known ground truth data against which novel algorithms can be benchmarked.

Technical Approach

We developed procedures for sampling individuals from an existing database and simulating offspring having STR genotypes following Mendelian rules of inheritance. By adjusting parameters such as population size, rate of gravity, endogamy, etc., population dynamics can be tailored to mimic those of a population of interest.

Results

The resulting simulated population contains multigenerational pedigrees having individuals with STR genotypes where all individuals' relationships are known (e.g., parent-offspring, full sibling, half-sibling, and potentially more distant second- and third-degree relationships). The I-T-O method was used to assess combined kinship index under the hypothesis of several relationship types, demonstrating the utility of simulated pedigrees in benchmarking kinship inference methods.

Conclusions

This methodology can be used to simulate pedigrees with STR data from any population at any number of loci where existing data are available, not just those where STR population allele frequencies are publicly available. The resulting simulated data can be used as a standard to assist development and conduct rigorous benchmarking of new methods for familial DNA searching, relationship inference, and quantifying uncertainty.

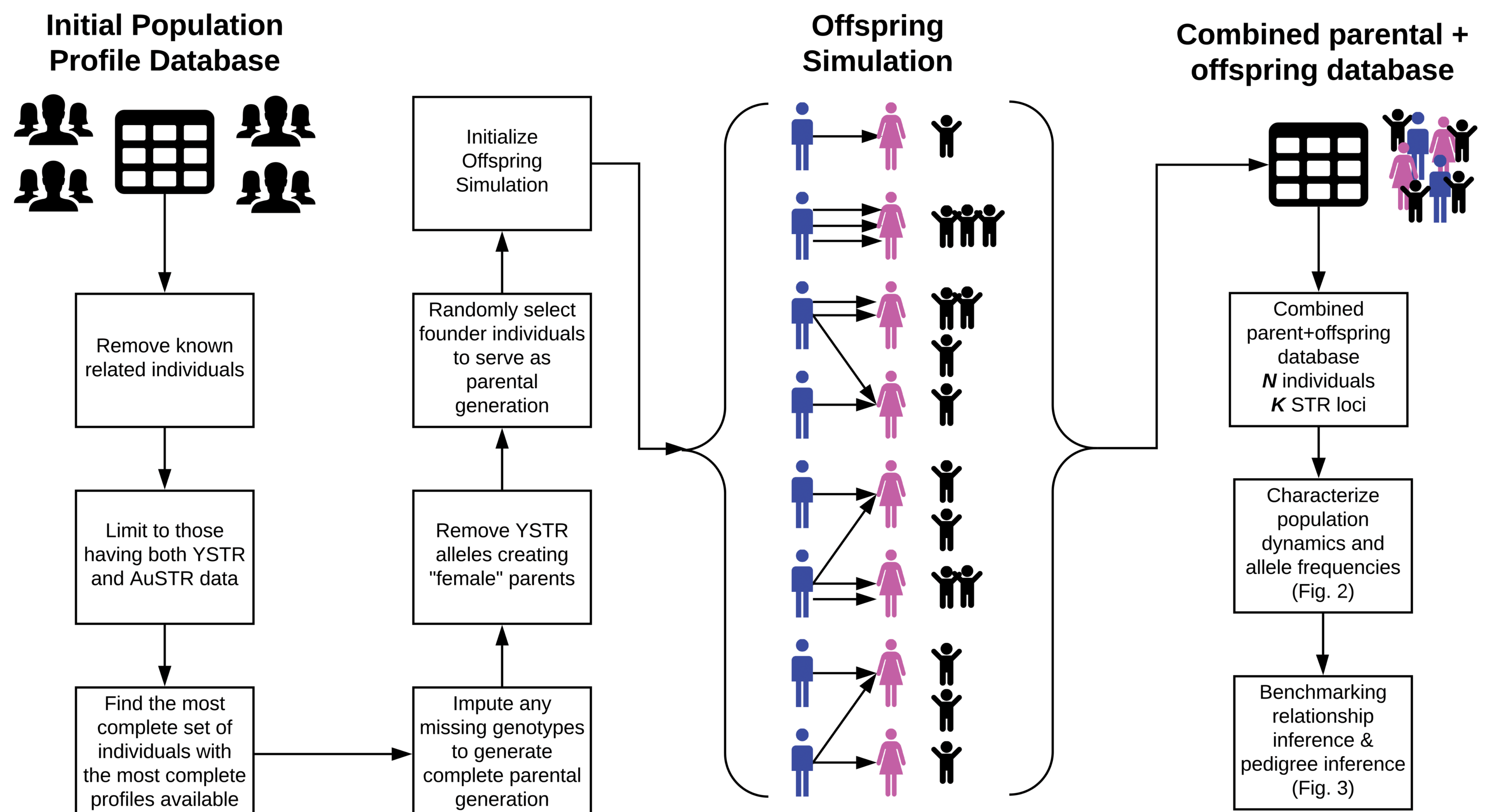


Fig. 1: Database curation and offspring simulation. Left: STR genotypes from samples in an existing database undergo cleaning and filtering to start with the most complete set of individuals with the most complete set of STR genotypes, and "female" profiles are created by removing YSTR alleles. Center: offspring simulation from a selection of founder parents. Mating events may result in 1–3

offspring. Some males mate with multiple females to generate half-sibling relationships. All these rates are customizable. Final output are parents and simulated offspring with STR genotypes and known relationship status. Right: these data are initially checked for allele frequency consistency and population dynamic realism (Fig. 2). The data can then be used for benchmarking (Fig. 3).

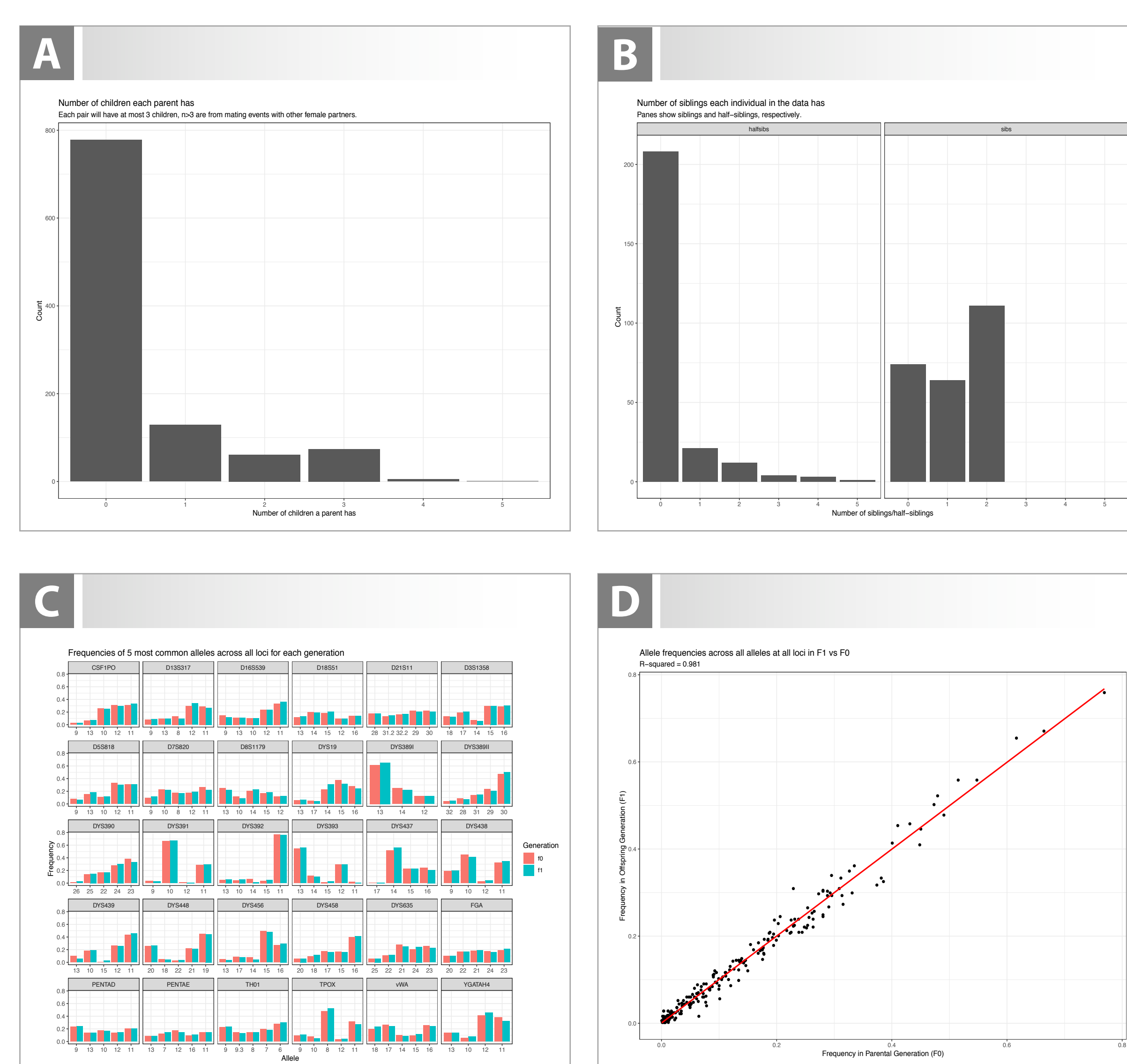


Fig. 2: Population dynamics and allele frequencies. (A) The number of children produced by individuals in the F0 (parental) generation (B) The number of half and full siblings individuals in the F1 each have. (C) Frequencies of the most common alleles at each locus in the F0 and F1 generations. (D) Scatterplot of allele frequencies in the F1 versus the F0 generation. These show that the allele frequencies in the offspring generation recapitulate the original frequencies.

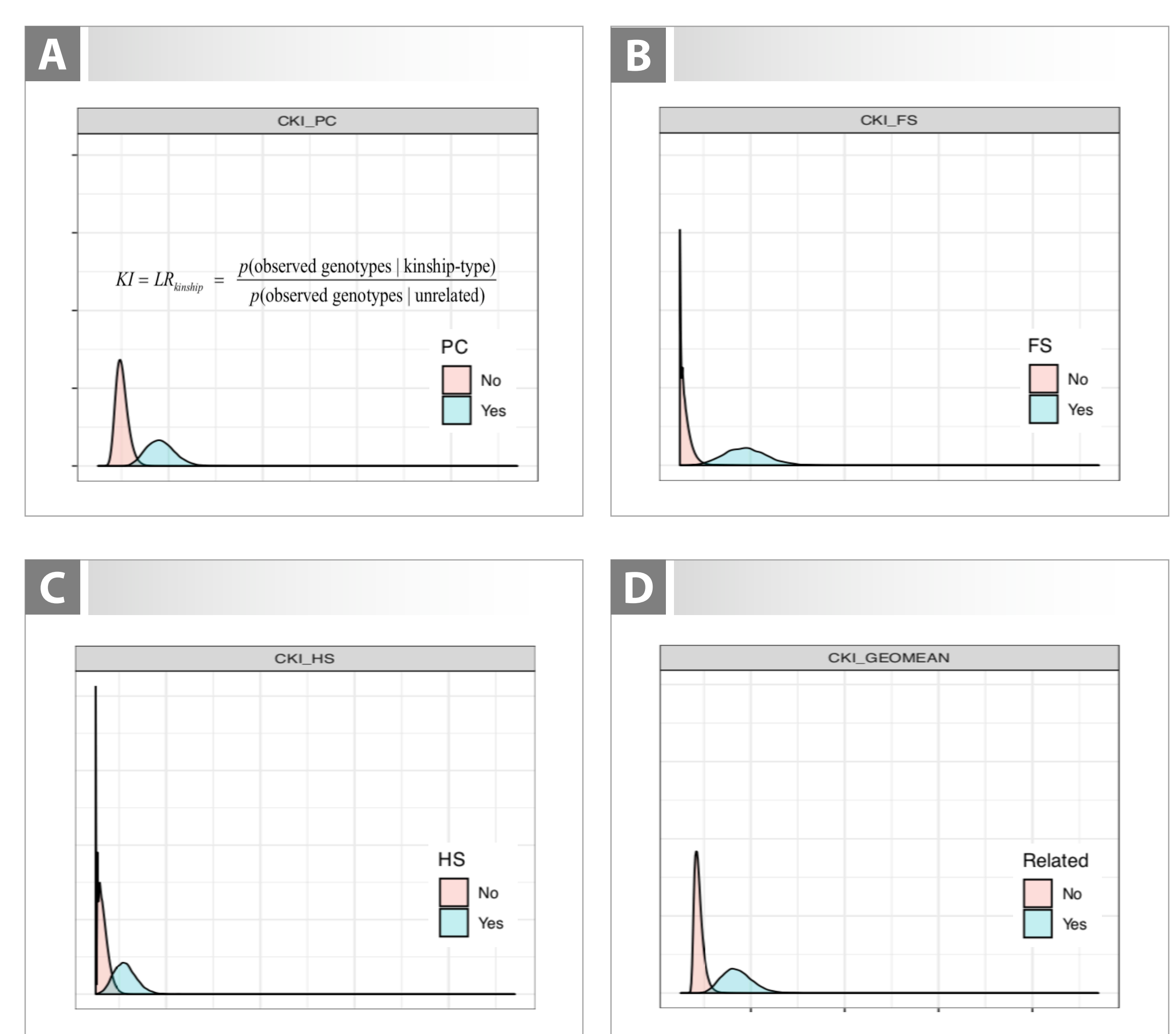
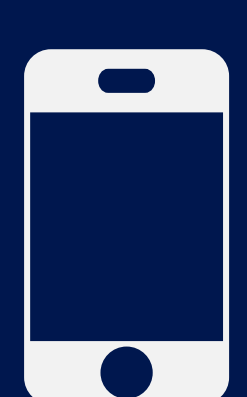


Fig. 3: Combined kinship index for different relationship types. Each panel shows the density distribution of combined kinship index (CKI) calculated using the I-T-O method described in the "CODIS 7.0 SP4 PopStats Calculation Specification" which computes the conditional probability of a relative when the genotype of the other relative of a specified kinship is known. CKI for parent-child (PC), full-sibling (FS) and half-siblings (HS) shown in (A), (B), and (C), respectively. (D) shows the geometric mean of all CKI measures for any relationship.



Scan QR Code to
Download this Poster

signature
science LLC

www.signaturescience.com