Motivation

- In population genetics, unsupervised clustering is widely used to infer ancestry proportions from genetic data.
- Aligning and interpreting clustering results produced by different algorithms, or even multiple runs of the same algorithm, is hard.

Challenges:

- label-switching
- arbitrariness in number of clusters
- genuine multimodality

Goals:

- a unified measure to quantify the cost of alignment
- a formalized pipeline to perform efficient alignment.

Background: Population Structure

- Membership coefficient: the proportion of the individual's genome originating in each of the \( K \) clusters (inferred populations).

Data 1: Membership coefficients of 978 human individuals inferred based on 791 loci (Fortier et al. 2020)

Data 2: Membership coefficients of 600 chicken individuals inferred based on 27 dinucleotide microsatellite loci (Rosenberg et al. 2001)

Cost of Misalignment

- Individual membership under Dirichlet model
  - In replicate 1: \( p \sim \text{Dir}(a) \Rightarrow a = (a_1, ..., a_K) \)
  - In replicate 2: \( q \sim \text{Dir}(\Phi(a)) \), with permutation \( \Phi(a) \).
  - Contribution of an individual to the distance between two replicates, 1 and 2:
    \[
    X = \| p - q \|_2^2
    \]
  - Mean of \( X \) when two replicates are aligned, i.e., \( a = \Phi(a) \):
    \[
    A_{aa} = \frac{4 \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} a_i a_j}{(a_i + 1) a_j^2}
    \]
  - Mean of \( X \) when two replicates are possibly permuted: \( A_{a,\Phi(a)} \)
  - Mean contribution of an individual to the cost of misalignment:
    \[
    C_{a,\Phi(a)} = \frac{A_{a,\Phi(a)} - A_{aa}}{2} = \frac{1}{a_i} \sum_{i=1}^{K} a_i (a_i - a_{\Phi(i)})
    \]

Demonstration on Real Data

- Relative difference between empirical and theoretical costs demonstrated on human population structure replicates.

Results and Conclusions

- For fixed Dirichlet parameters, the misalignment cost increases with the Hamming distance between permutations (Fig 3).
- The misalignment cost accurately reflects the empirical cost on low variance data (Fig 2), thus can contribute to improving cluster alignment algorithms seeking to find optimal permutations of replicates.
- Clustering alignment can be performed with high efficiency and accuracy through numerical optimization and community algorithms.
- Visualization of clustering modes in multipartite graph helps interpret the genuine multimodality of the inferred population structure.

Next steps:

- Incorporate leader clustering as an optional heuristic method to achieve higher efficiency when number of replicates is large.
- Compile the alignment pipeline into a distributable package easily accessible to researchers in the population genetics community.