Duke Center for Statistical Genetics and Genomics

ASHG
VIRTUAL MEETING 2020
OCTOBER 27-30

Estimating ancestry from a population kinship matrix under arbitrary ancestral subpopulation structure

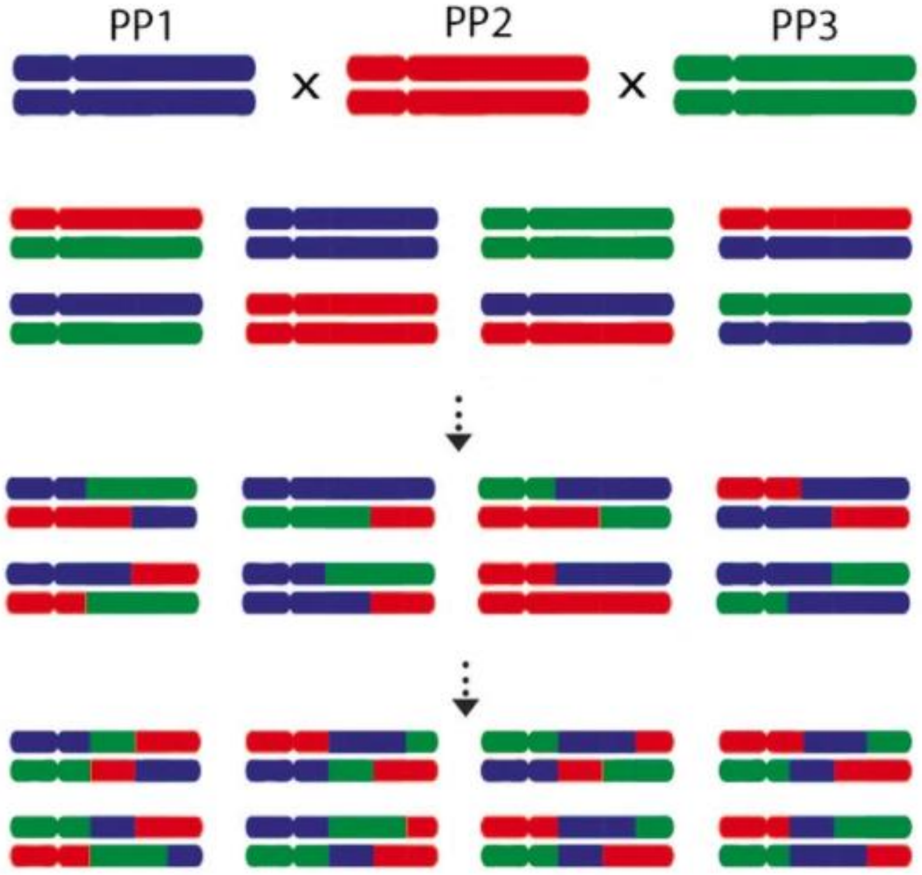
Amika Sood, Alejandro Ochoa
Duke University

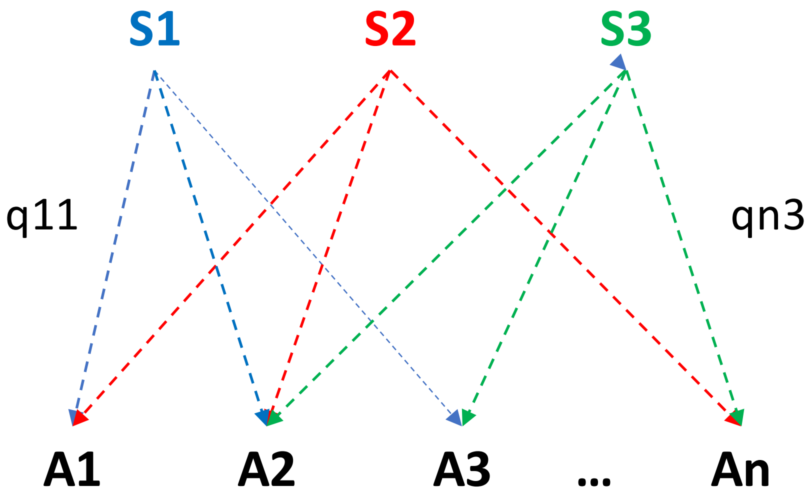
Disclosure Slide

Financial Disclosure for:
Amika Sood
Post-Doctoral Research Associate
Duke University

I have nothing to disclose

Motivation

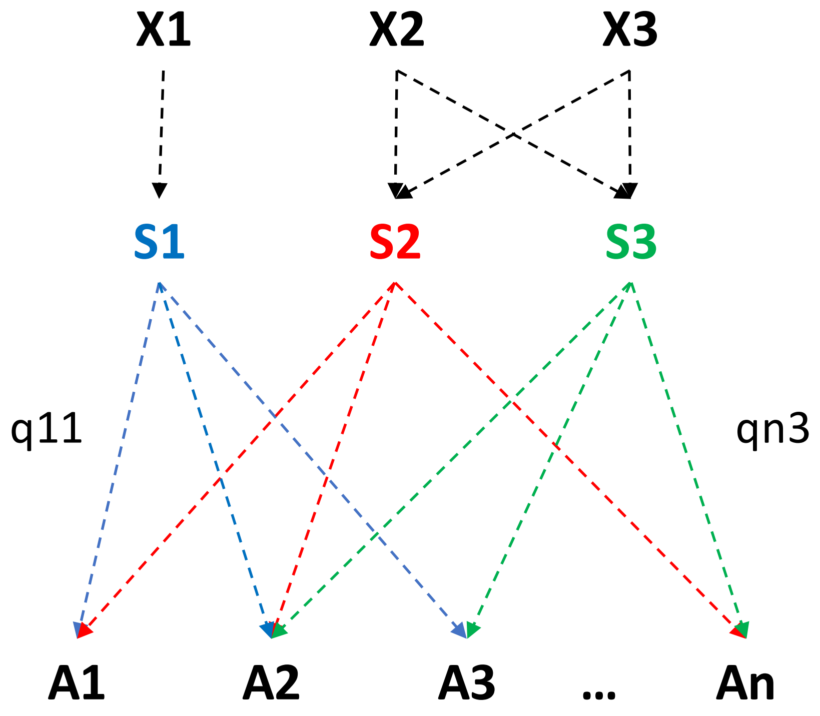




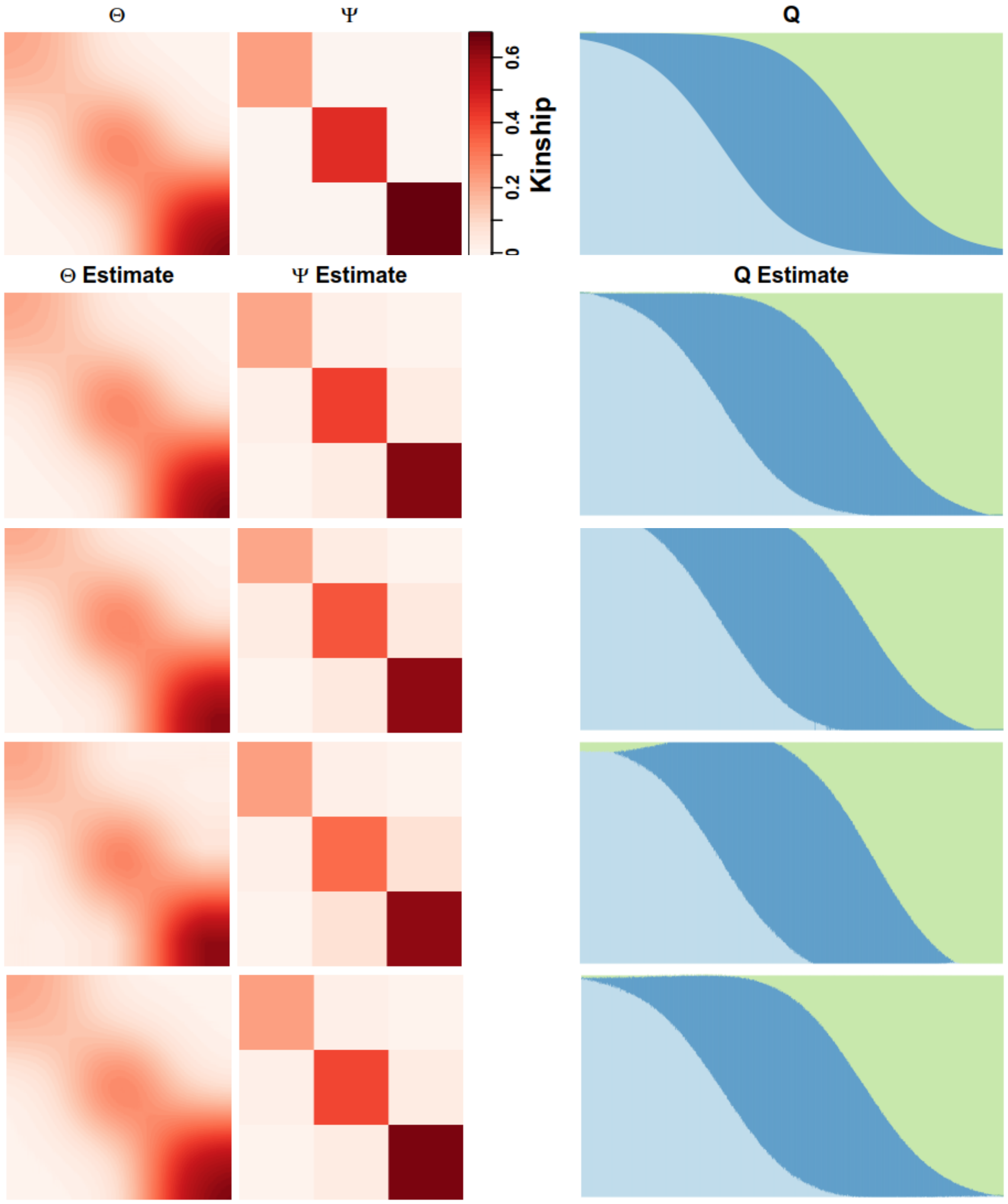
Soares-Souza, G., Borda, V., Kehdy, F. et al. Admixture, Genetics and Complex Diseases in Latin Americans and US Hispanics. Curr Genet Med Rep ***6,*** *208–223 (2018)*

Model and Data Generation

- $P \in R^{m \times k}$, $p_{ij} \in [0,1]$
- $Q \in R^{n \times k}$, $q_{ij} \geq 0, \sum_j q_{ij} = 1$
- $\pi = PQ^T$
- The coancestry matrix of P should equal the coancestry of the ancestral/intermediate subpopulations, Psi.



Performance of existing approaches



Independent ancestral populations

Admixture:
Model-based ancestry estimation

fastStructure :
Variational Bayesian framework

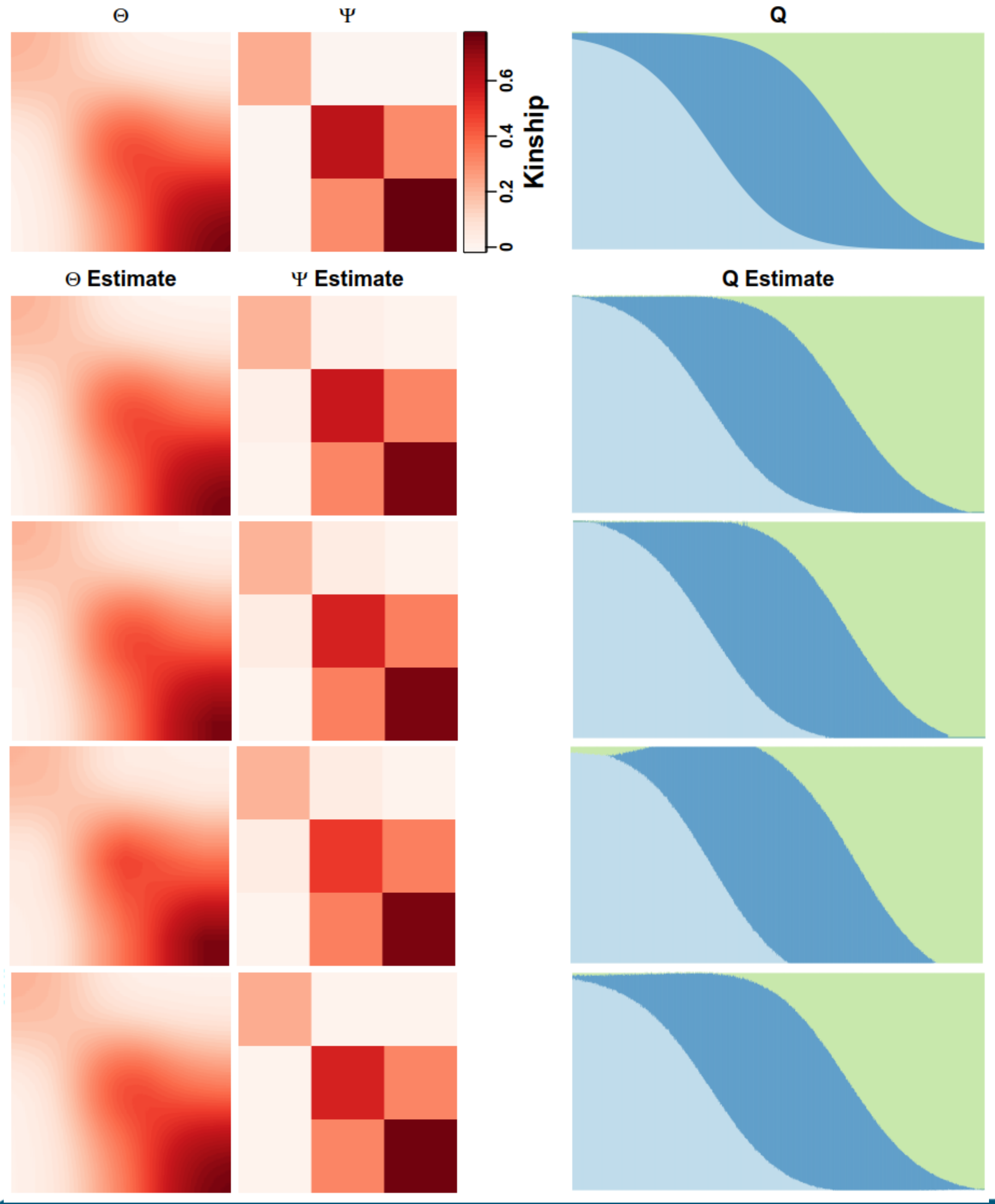
sNMF:
NMF and least-squares optimization

ALStructure:
PCA based method

- 1000 individuals
- 100, 000 loci
- 3 populations
- FST = 0.3

Performance of existing approaches

Correlated ancestral populations



Admixture:
Model-based ancestry estimation

fastStructure :
Variational Bayesian framework

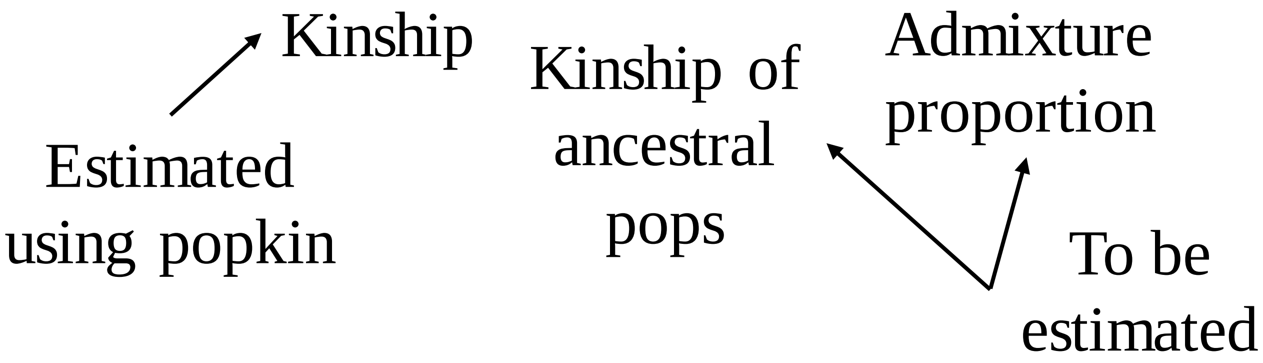
sNMF:
NMF and least-squares optimization

ALStructure:
PCA based method

- 1000 individuals
- 100, 000 loci
- 3 populations
- FST = 0.3

Optimization

$$\Theta = Q\Psi Q^T$$



Kinship

Estimated using popkin

Kinship of ancestral pops

Admixture proportion

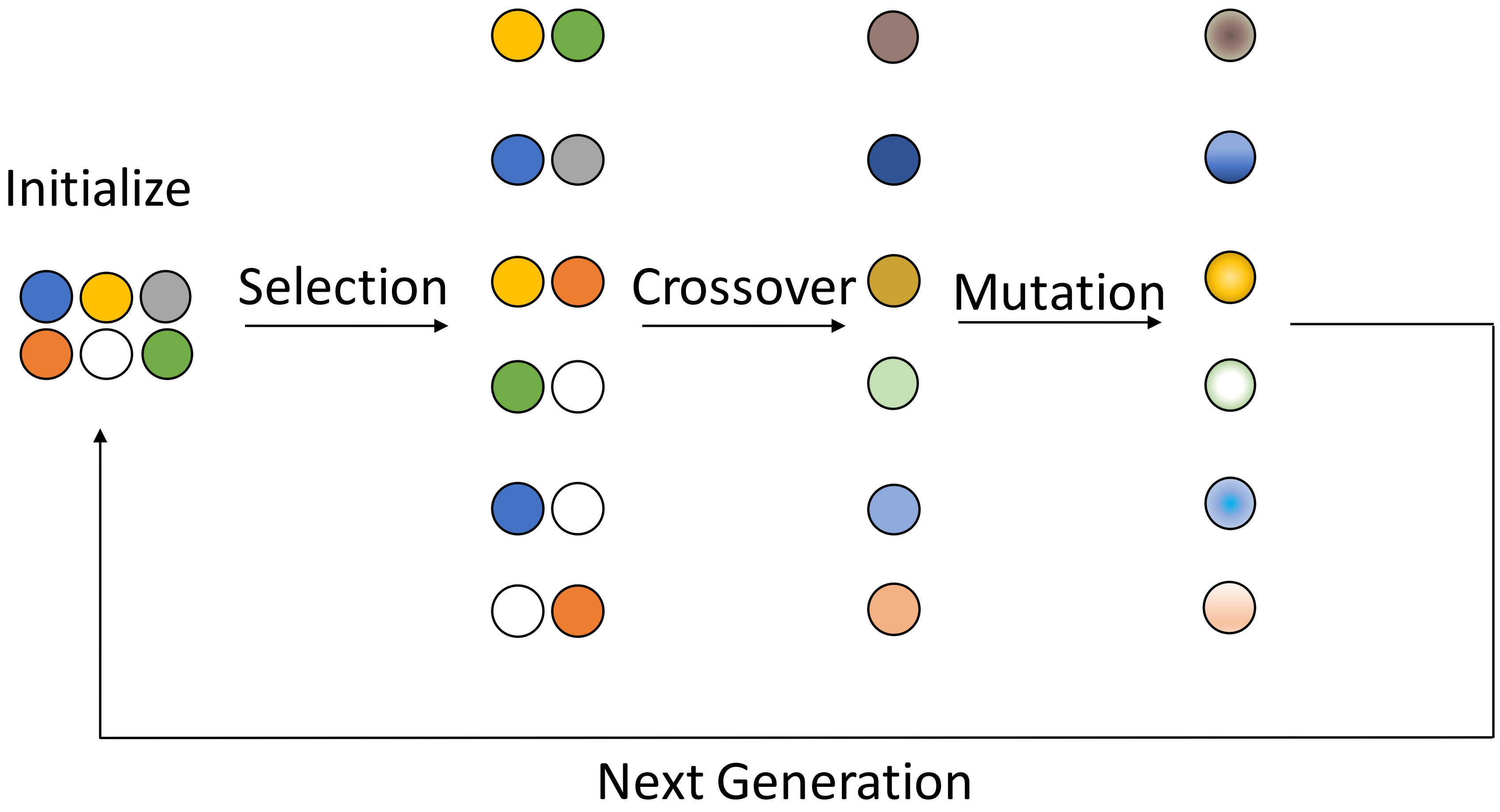
To be estimated

$$F = \|Q\Psi Q^T - \Theta\|^2$$

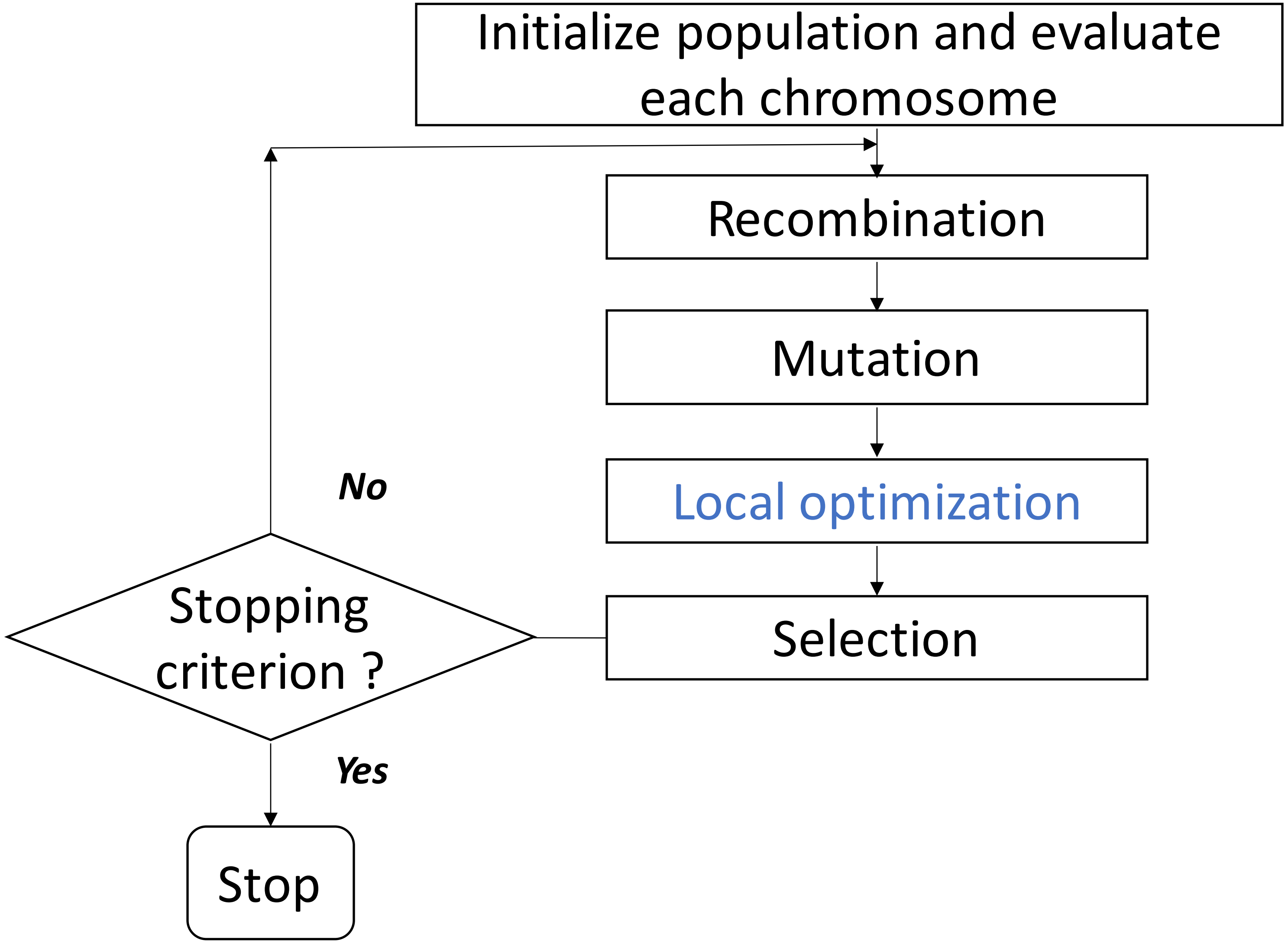
- Multiple minima problem,
- Factorization is not unique.

$$Q' = QR$$
$$\Psi' = R^{-1}\Psi(R^{-1})^T$$

Genetic Algorithm



Memetic Algorithm



Initialize population and evaluate each chromosome

Recombination

Mutation

Local optimization

Selection

Stopping criterion ?

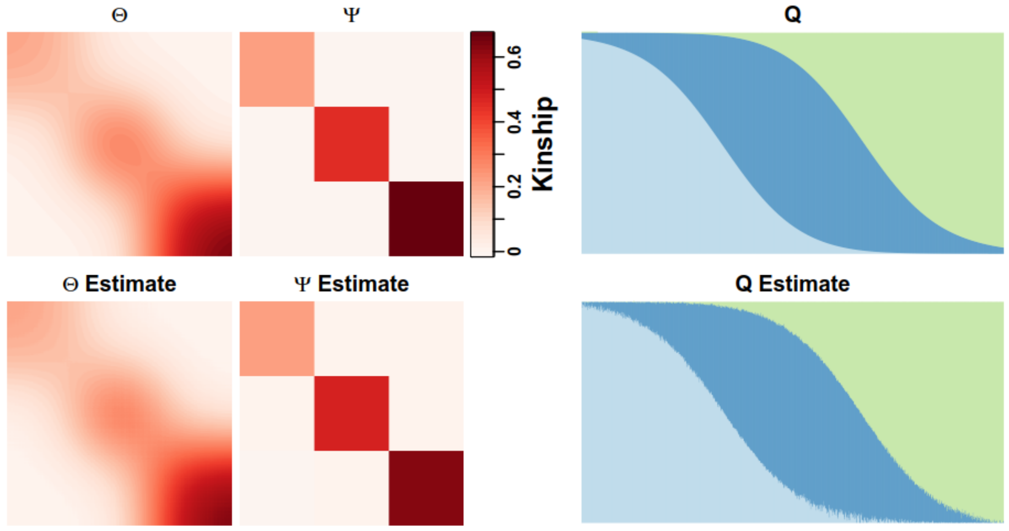
No

Yes

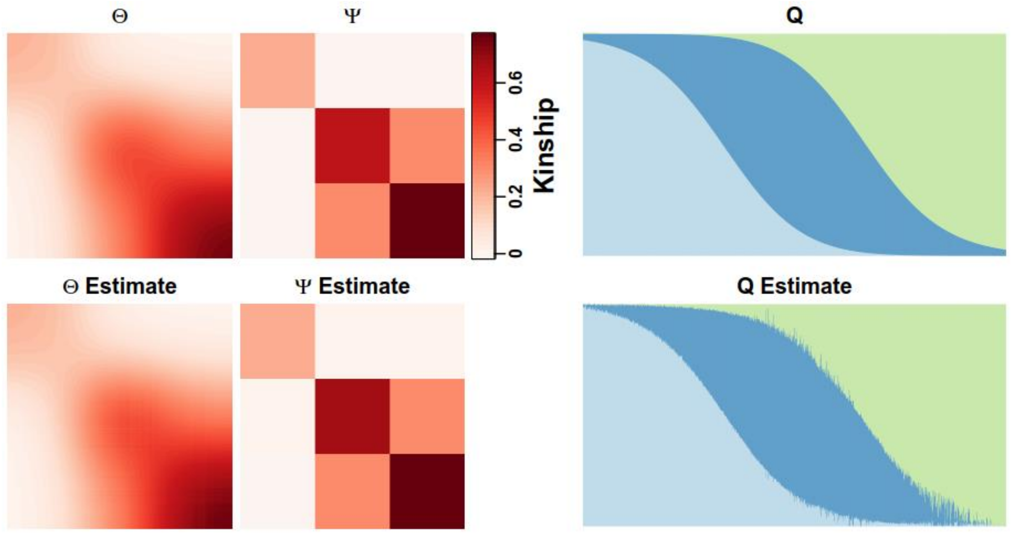
Stop

Memetic Algorithm

Independent ancestral populations

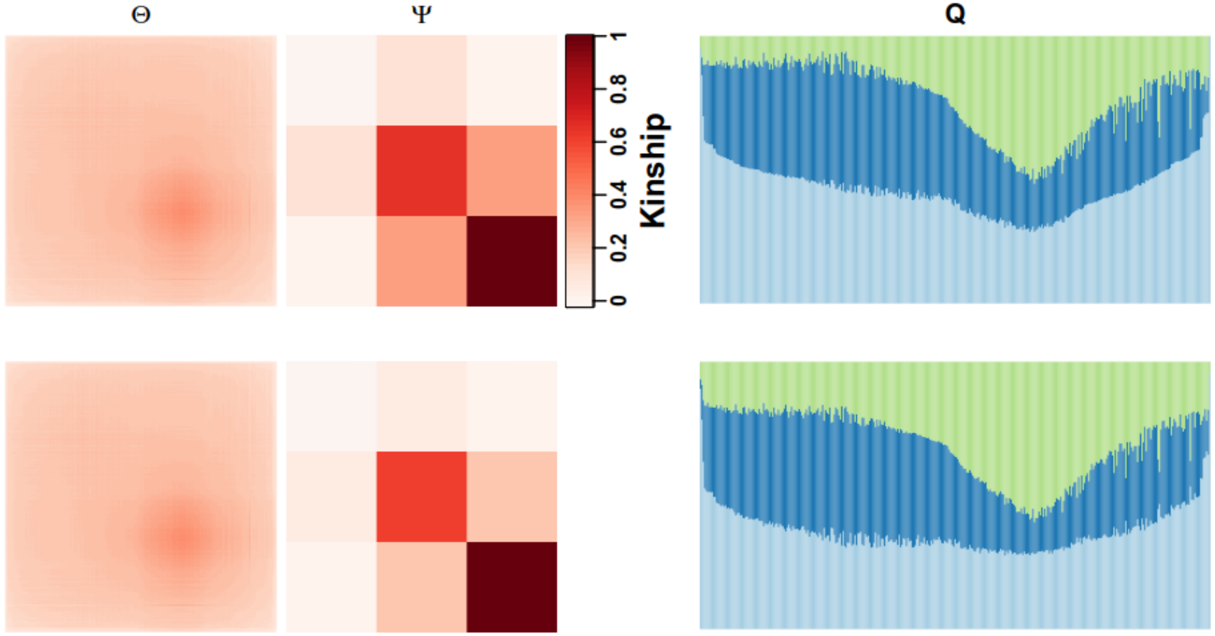


Correlated ancestral populations



- 1000 individuals,
- 100, 000 loci, and
- 3 subpopulations

Hispanic individuals in the 1000 Genomes Project



Admixture

Memetic Algorithm

6.7 million high-frequency SNPs

347 individuals

Conclusion and Next steps

- First likelihood free approach to jointly estimate admixture and kinship between ancestral subpopulations without bias.
- Estimated from a kinship matrix only.
- Not dependent on the number of loci, therefore run time is only dependent on the number of individuals and populations.

Acknowledgements

Goldberg lab

Amy Goldberg, PhD
Katharine Korunes, PhD
Marina Voinson, PhD
Iman Hamid

Annette Jackson, PhD
Rasheed Gbadegesin, M.D.
Eileen Chambers, M.D.
Cliburn Chan, PhD
Brian Shaw, M.D.

Thank you