# Ancestry Inference

Jonathon Chow

2022-10-04

Section 1

# Introduction

# Ancestor inference

- inheritable diseases, conservation genetics, the ancestry and migration patterns of natural populations
- PSD model
- PCA-based methods
- model-based approaches

## Algorithms

- STRUCTURE (MCMC) (Pritchard, Stephens, and Donnelly 2000)
- FRAPPE (EM) (Tang et al. 2005)
- ADMIXTURE (SQP) (Alexander, Novembre, and Lange 2009)
- sNMF (SNMF) (Frichot et al. 2014)
- fastSTRUCTURE (VI) (Raj, Stephens, and Pritchard 2014)
- TeraStructure (SVI) (Gopalan et al. 2016)

## Data

- TGP (Abecasis et al. 2012)
- HGDP (Cann et al. 2002; Cavalli-Sforza 2005; Li et al. 2008)

Section 2

# Models and Methods

# PSD model

- observed variable: genotype matrix $G$
- latent variable: matrix $Z$ of the true origin of genes
- parameters: population scale matrix $P$, gene scale matrix $F$
- hyperparameter: population number $K$

# EM algorithm

- E-step: compute the expectation $a_{ijk}$ and $b_{ijk}$
- M-step: compute the maximization and update the parameters $p_{ik}$ and $f_{kj}$
- convergence criterion: the log-likelihood of incomplete data $\mathcal{L}(G|P,F)$ converges

# SQP algorithm

- update parameters: update $P$ and $F$ block by block alternately
- convergence criterion: the log-likelihood of incomplete data $\mathcal{L}(G|P,F)$ converges

# VI algorithm

- update parameters: update variational parameters $\tilde{z}_{ij}^{a}$, $\tilde{p}_i$, $\tilde{f}_{kj}^{1}$, $\tilde{f}_{kj}^{2}$
- convergence criterion: the ELBO converges

# SVI algorithm

- sample: sample a SNP
- update parameters: iteratively update local parameter $F_j$ at the SNP until it converges, then update global parameter $P$
- convergence criterion: the log-likelihood at the validation set converges

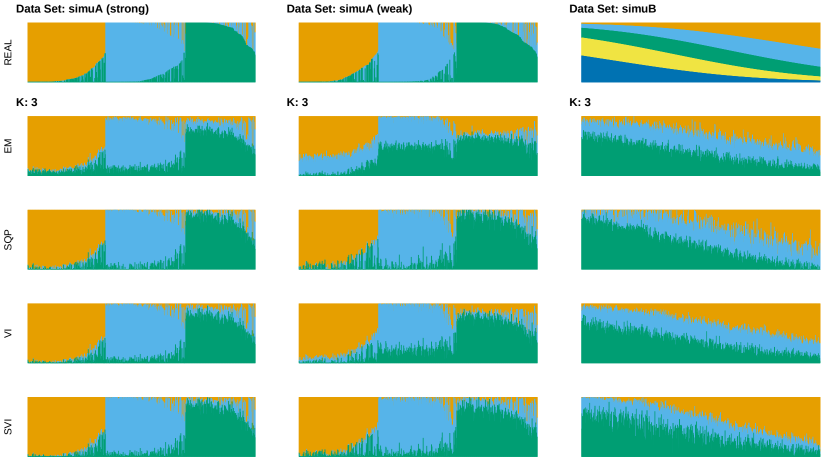
# Relationships with other models

- Poisson NMF model, multinomial topic model (Carbonetto et al. 2021)
- PLSA model (Hofmann 2001)
- LDA model (Blei, Ng, and Jordan 2003)

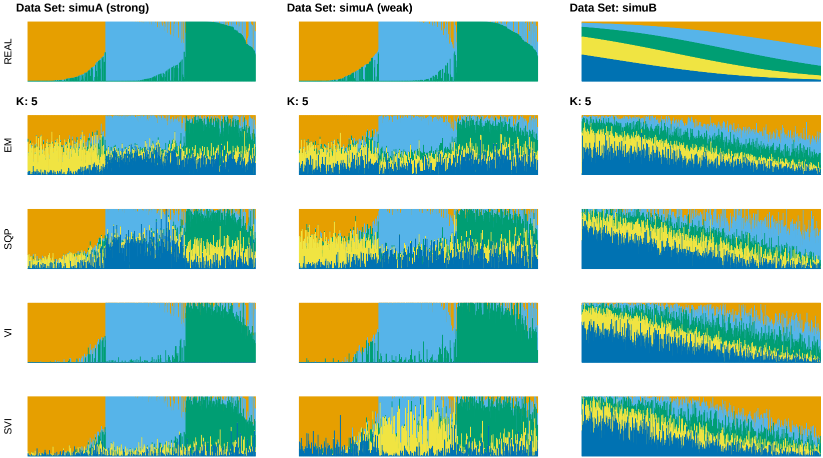
Section 3

# Applications

# Simulated data set

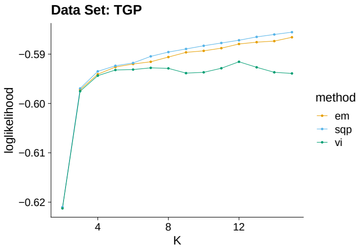


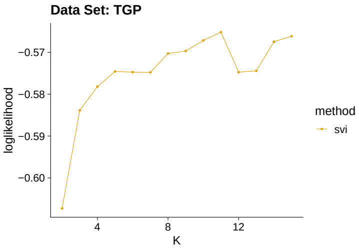
# Simulated data set

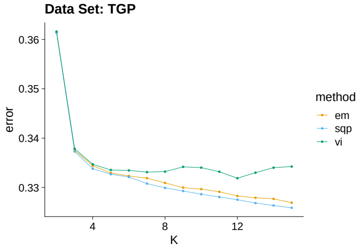


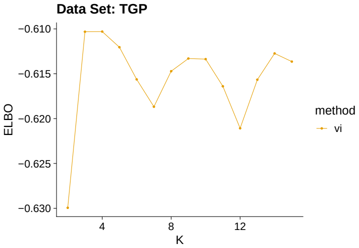
# TGP data set

- choose $K$


Data Set: TGP


Data Set: TGP

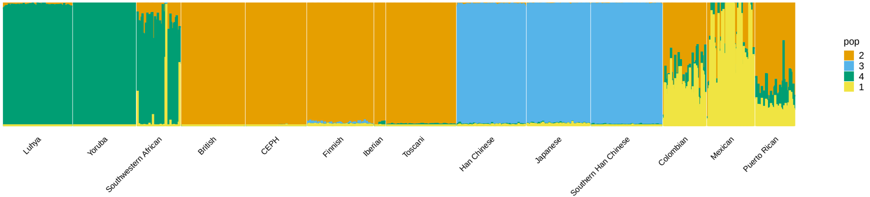

Data Set: TGP


Data Set: TGP

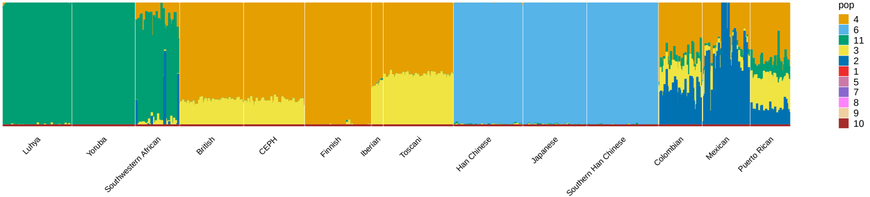
# TGP data set

- structure plot

**Data Set: TGP (full) | Method: SVI (1e+6 interations) | K: 4**

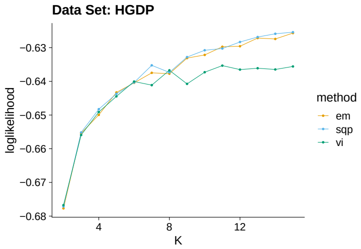


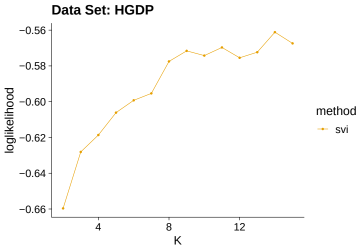
**Data Set: TGP (full) | Method: SVI (1e+6 interations) | K: 11**

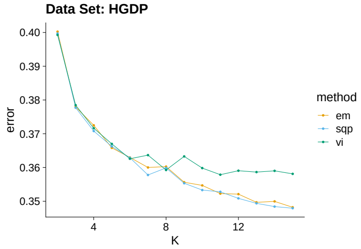


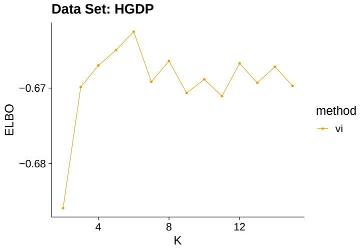
# HGDP data set

- choose $K$





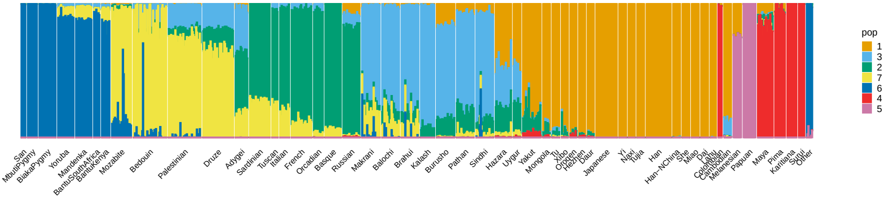




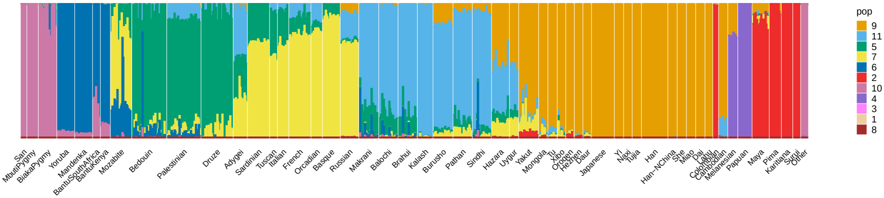
# HGDP data set

- structure plot

Data Set: HGDP (full) | Method: SVI (1e+6 interations) | K: 7



Data Set: HGDP (full) | Method: SVI (1e+6 interations) | K: 11

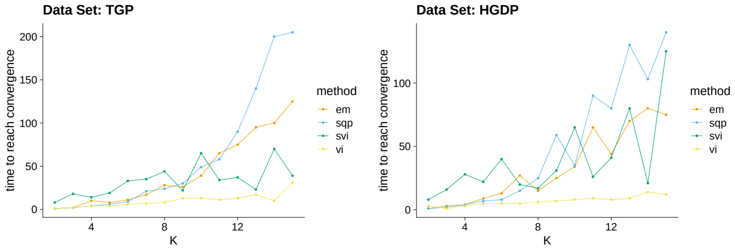


Section 4

# Discussion

# Algorithm evaluation

- convergence accuracy
- convergence efficiency



- algorithm selection criteria

Section 5

# Literature Cited

# Literature Cited I

Abecasis, Goncalo R, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, et al. 2012. "An Integrated Map of Genetic Variation from 1,092 Human Genomes." *Nature* 491 (7422): 56–65.

Alexander, David H, John Novembre, and Kenneth Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research* 19 (9): 1655–64.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Cann, Howard M, Claudia De Toma, Lucien Cazes, Marie-Fernande Legrand, Valerie Morel, Laurence Piouffre, Julia Bodmer, et al. 2002. "A Human Genome Diversity Cell Line Panel." *Science* 296 (5566): 261–62.

Carbonetto, Peter, Abhishek Sarkar, Zihao Wang, and Matthew Stephens. 2021. "Non-Negative Matrix Factorization Algorithms Greatly Improve Topic Model Fits." *arXiv Preprint arXiv:2105.13440*.

# Literature Cited II

Cavalli-Sforza, L Luca. 2005. "The Human Genome Diversity Project: Past, Present and Future." *Nature Reviews Genetics* 6 (4): 333–40.

Frichot, Eric, François Mathieu, Théo Trouillon, Guillaume Bouchard, and Olivier François. 2014. "Fast and Efficient Estimation of Individual Ancestry Coefficients." *Genetics* 196 (4): 973–83.

Gopalan, Prem, Wei Hao, David M Blei, and John D Storey. 2016. "Scaling Probabilistic Models of Genetic Variation to Millions of Humans." *Nature Genetics* 48 (12): 1587–90.

Hofmann, Thomas. 2001. "Unsupervised Learning by Probabilistic Latent Semantic Analysis." *Machine Learning* 42 (1): 177–96.

Li, Jun Z, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, et al. 2008. "Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation." *Science* 319 (5866): 1100–1104.

# Literature Cited III

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly. 2000. “Inference of Population Structure Using Multilocus Genotype Data.” *Genetics* 155 (2): 945–59.

Raj, Anil, Matthew Stephens, and Jonathan K Pritchard. 2014. “fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets.” *Genetics* 197 (2): 573–89.

Tang, Hua, Jie Peng, Pei Wang, and Neil J Risch. 2005. “Estimation of Individual Admixture: Analytical and Study Design Considerations.” *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 28 (4): 289–301.