

“Inference of Population Structure using Dense Haplotype Data”

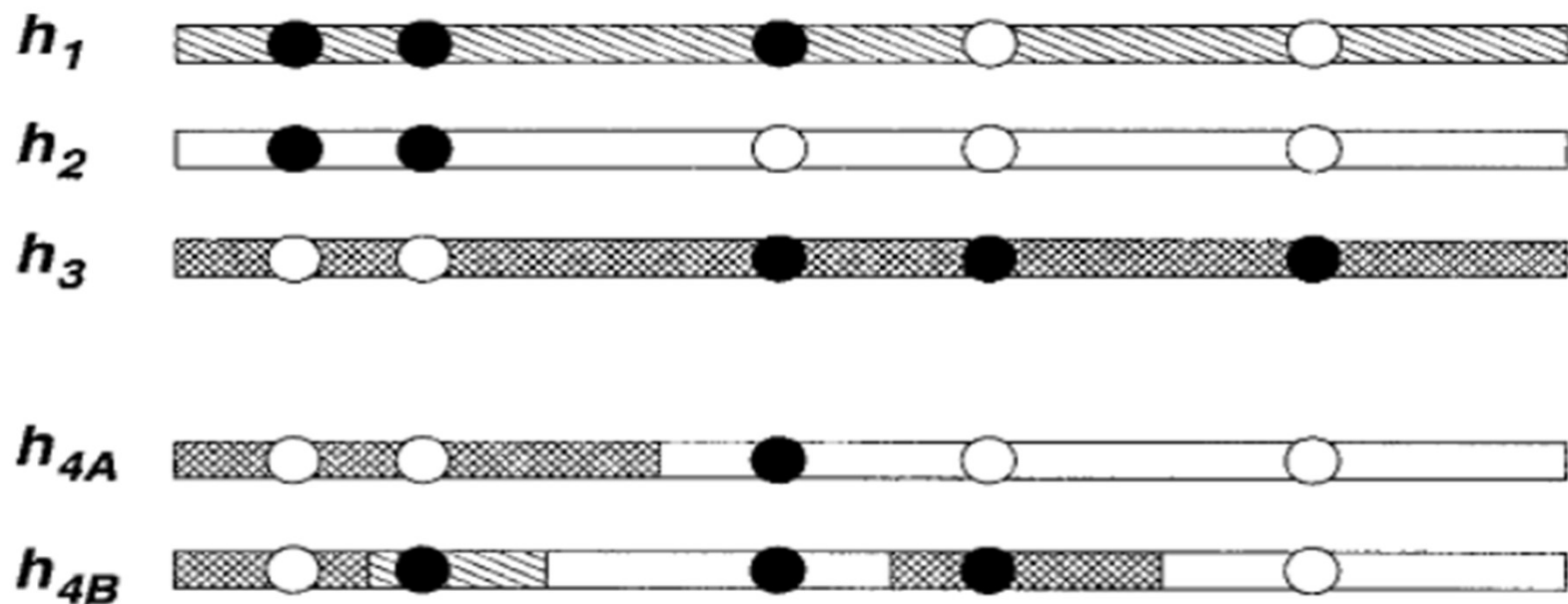
Written by Lawson, Hellenthal, Myers, and Falush (2012)
Presented by Seth Temple

Process

1. Phase data
2. “Paint chromosomes” to calculate coancestry matrix
3. Use coancestry matrix to infer populations
 - a. Principal Components Analysis (PCA)
 - i. Eigenstrat, Price et al (2006)
 - b. fineSTRUCTURE
 - i. structure
 - ii. ADMIXTURE (2009)
4. (Optional) Visualize with a tree diagram

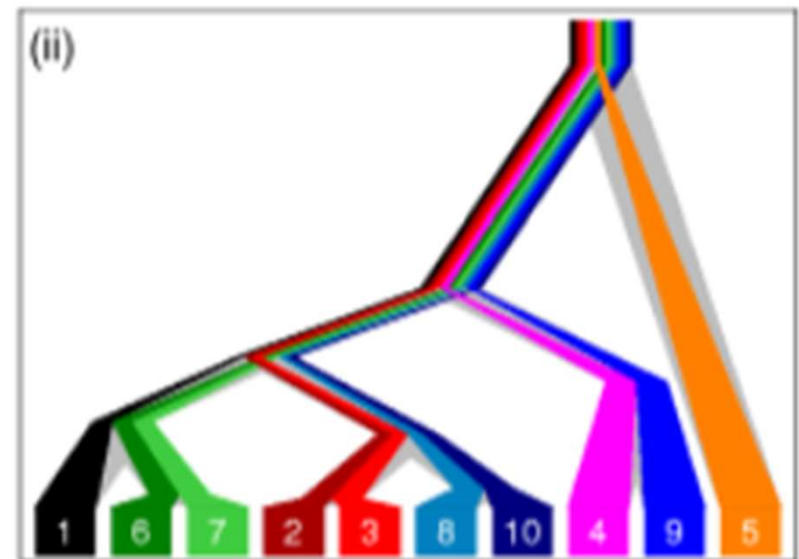
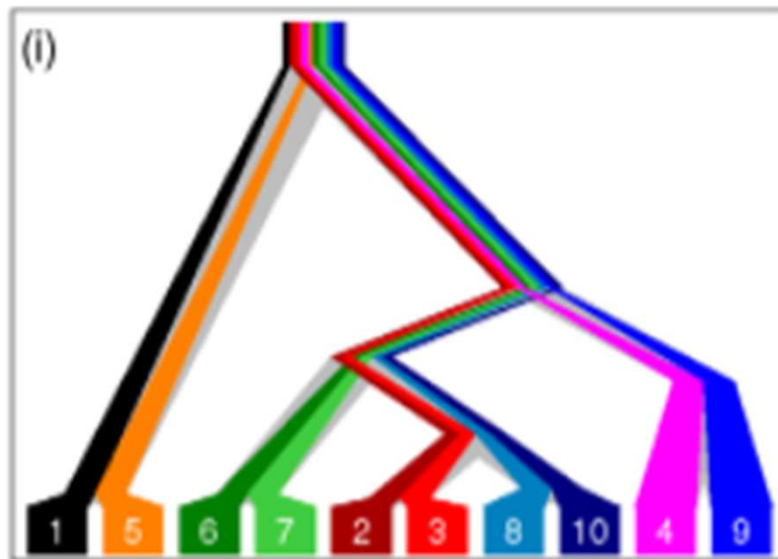
Li and Stephens (2003)

Hidden Markov Model (HMM) to describe genealogical process



ChromoPainter

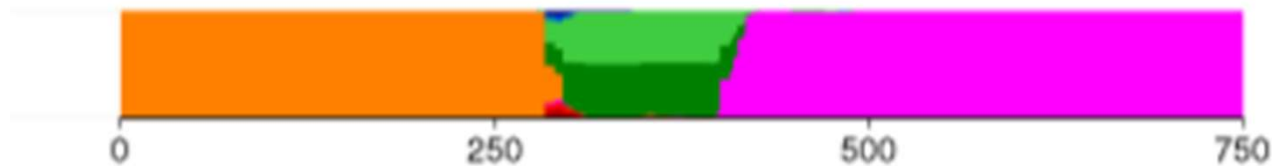
HMM like Li and Stephens (2003), but without ordering haplotypes



Coancestry Matrix

- x_{ij} is the expected number of chunks inferred from donor j to haplotype i
- Parameter ρ used to consider linked vs unlinked loci

E) Mean painting of haplotype 1

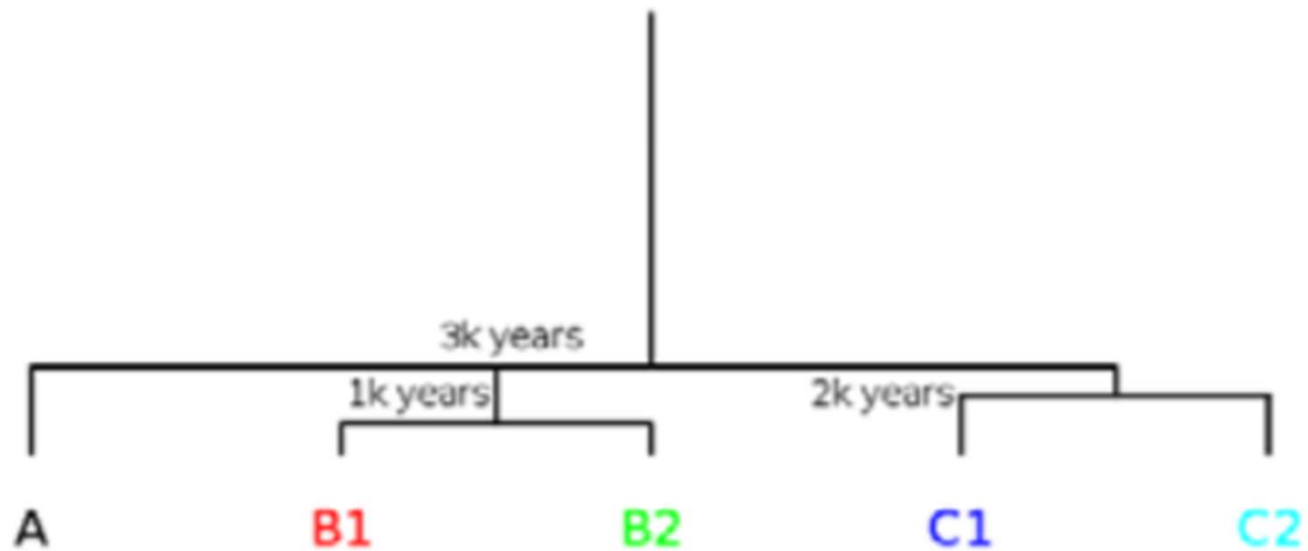


F) Coancestry matrix row for haplotype 1

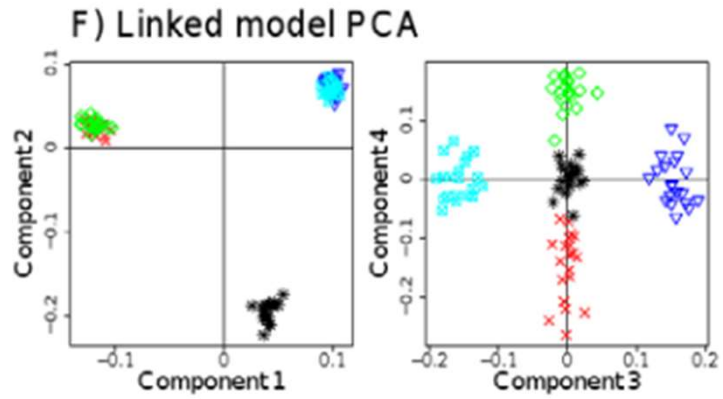
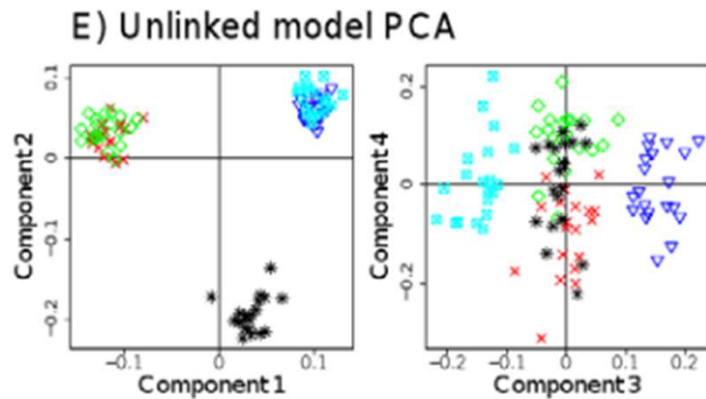
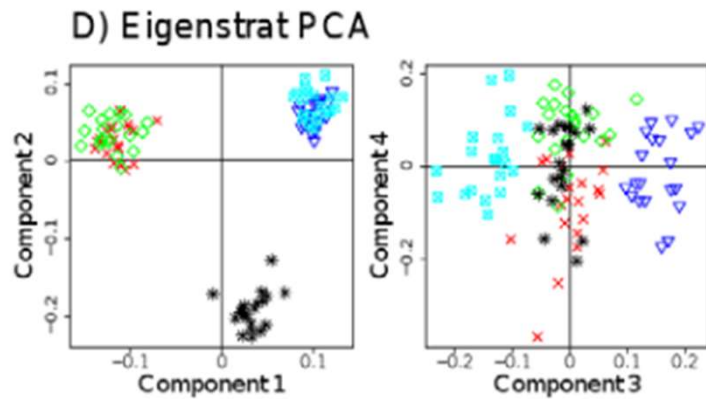
	Donor haplotype									
	1	2	3	4	5	6	7	8	9	10
Haplotype 1	0	0.08	0.09	1.1	1.24	0.52	0.52	0.06	0.01	0.06

Simulation Study

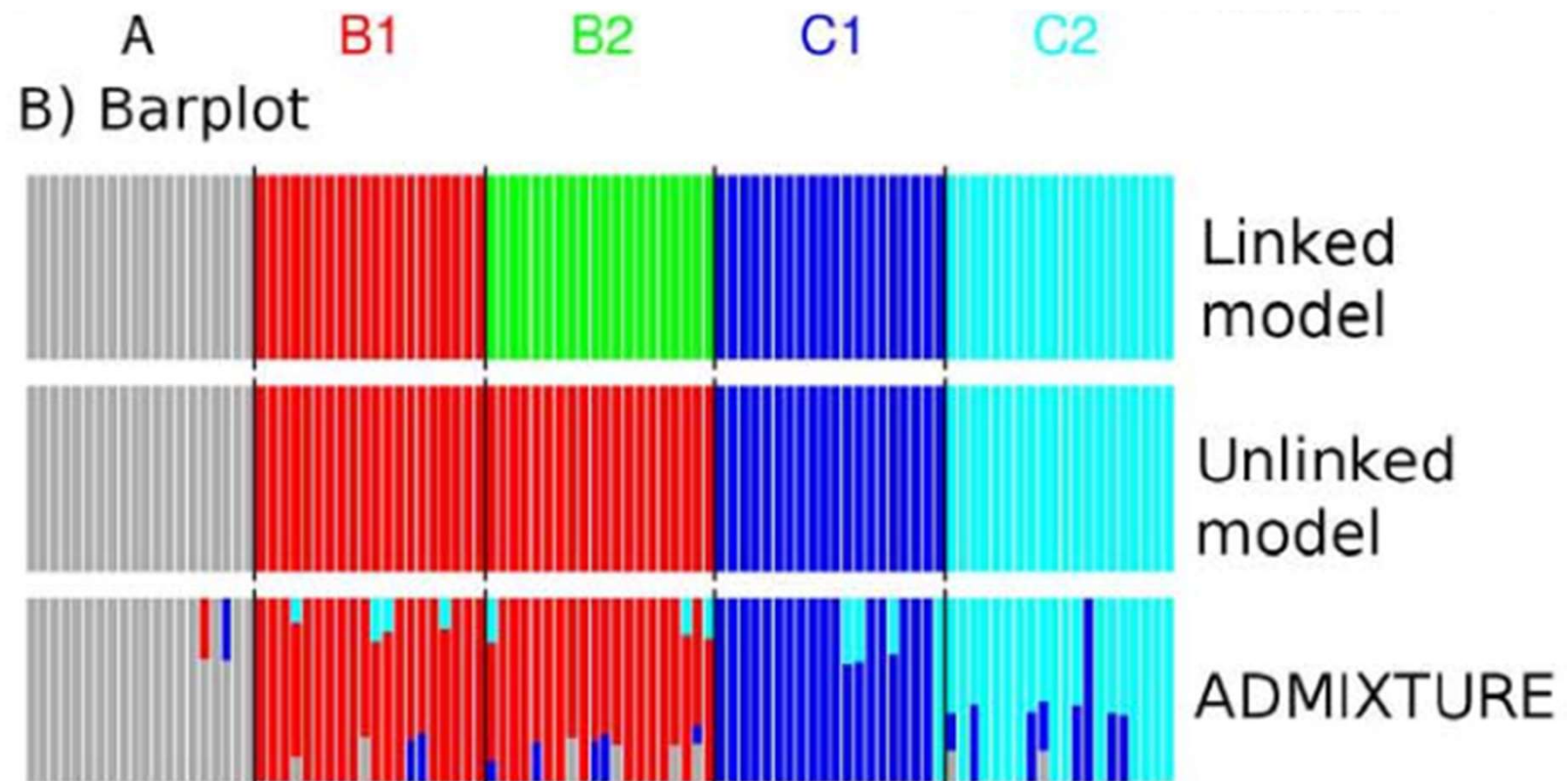
B) Population Tree



Simulation Study: PCA



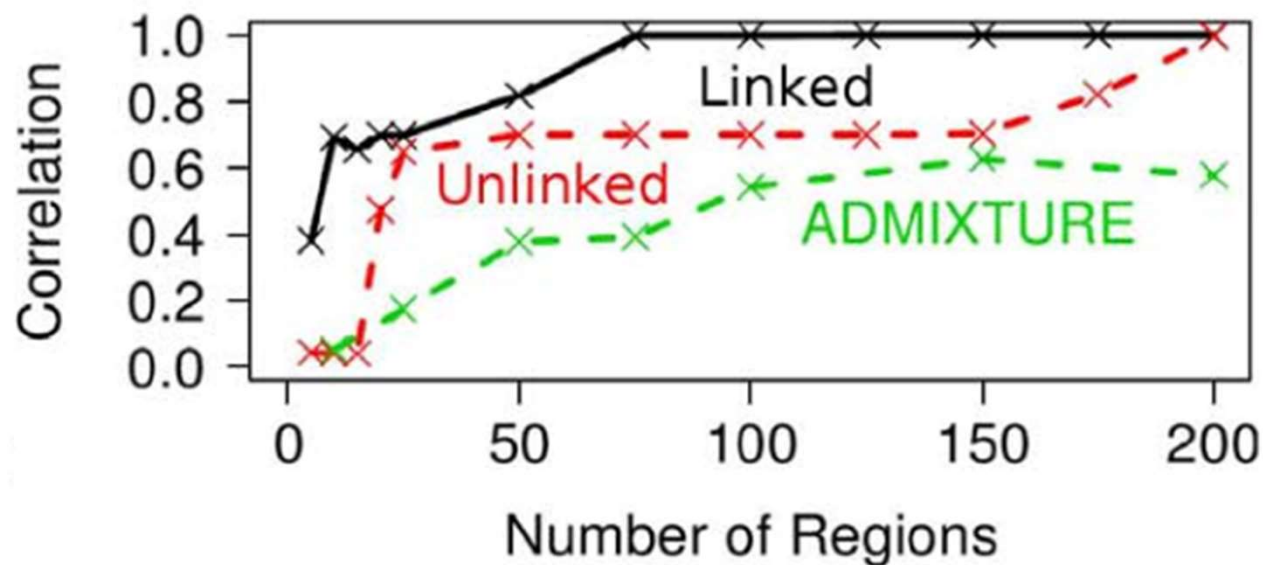
Simulation Study: Model-based Approaches



Simulation Study: Model-based Approaches

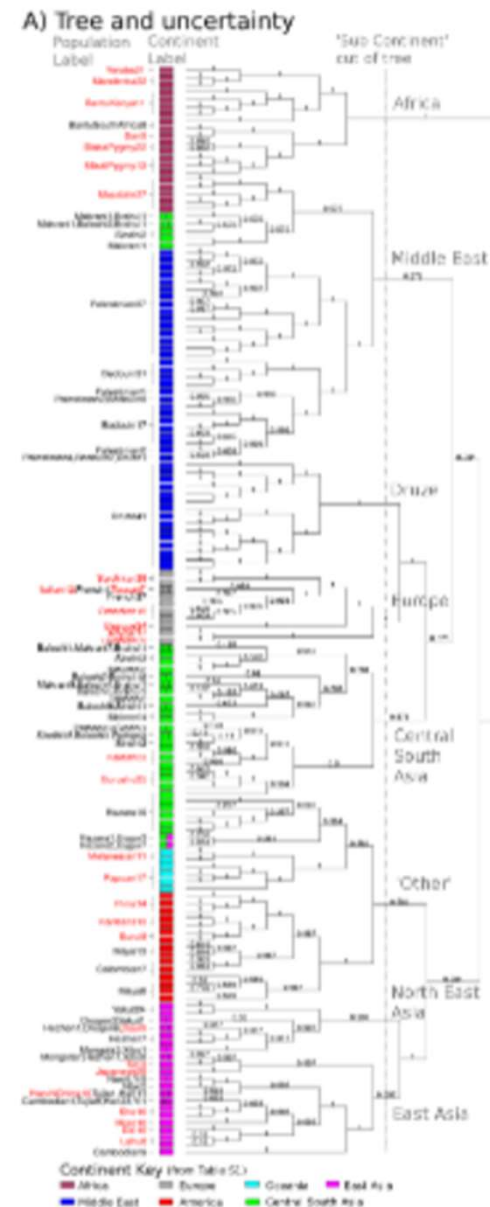
“using linkage information will therefore be expected to allow clear identification of more subtle structure than is detectable otherwise”

D) Correlation with truth versus Data Size



K > 10 Populations

- structure-like methods struggle for large K
- Dirichlet Process prior and reversible-jump MCMC
- Ad hoc tree building on the left
- 226 populations inferred from 938 individuals



Takeaways

- LD baked into coancestry matrix enables fine grained population inference using dense data
 - Previous approaches could only handle $K < 10$
- Adapted existing PCA and model-based approaches for coancestry matrix input
- Admixture not addressed yet in the software