ChromoPainter and FineSTRUCTURE: Inference of population structure using dense haplotype data

Daniel Lawson
Garrett Hellenthal
Daniel Falush
Simon Myers

Department of mathematics
University of Bristol

dan.lawson@bristol.ac.uk

www.paintmychromosomes.com

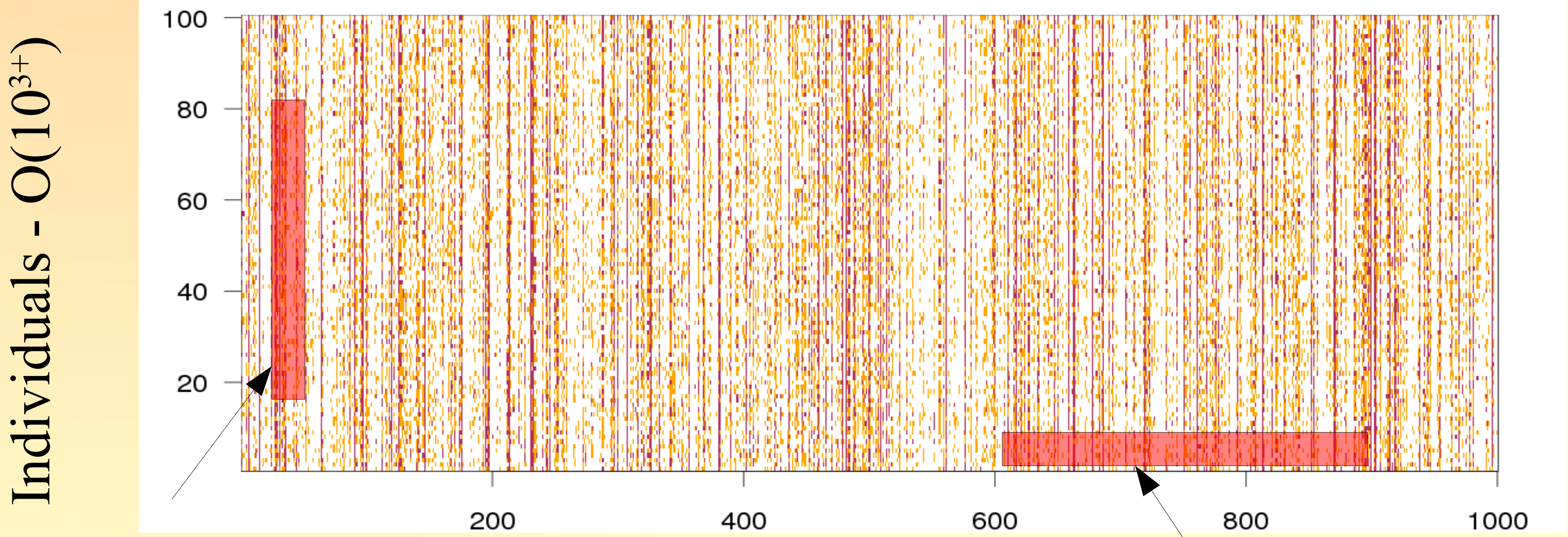
PART 1: A new challenge of modern genetics data

- CHALLENGE:

  Datasets are getting LARGER and MORE COMPLEX

- AIM OF THIS TALK:

  Understand ancestry patterns from such data

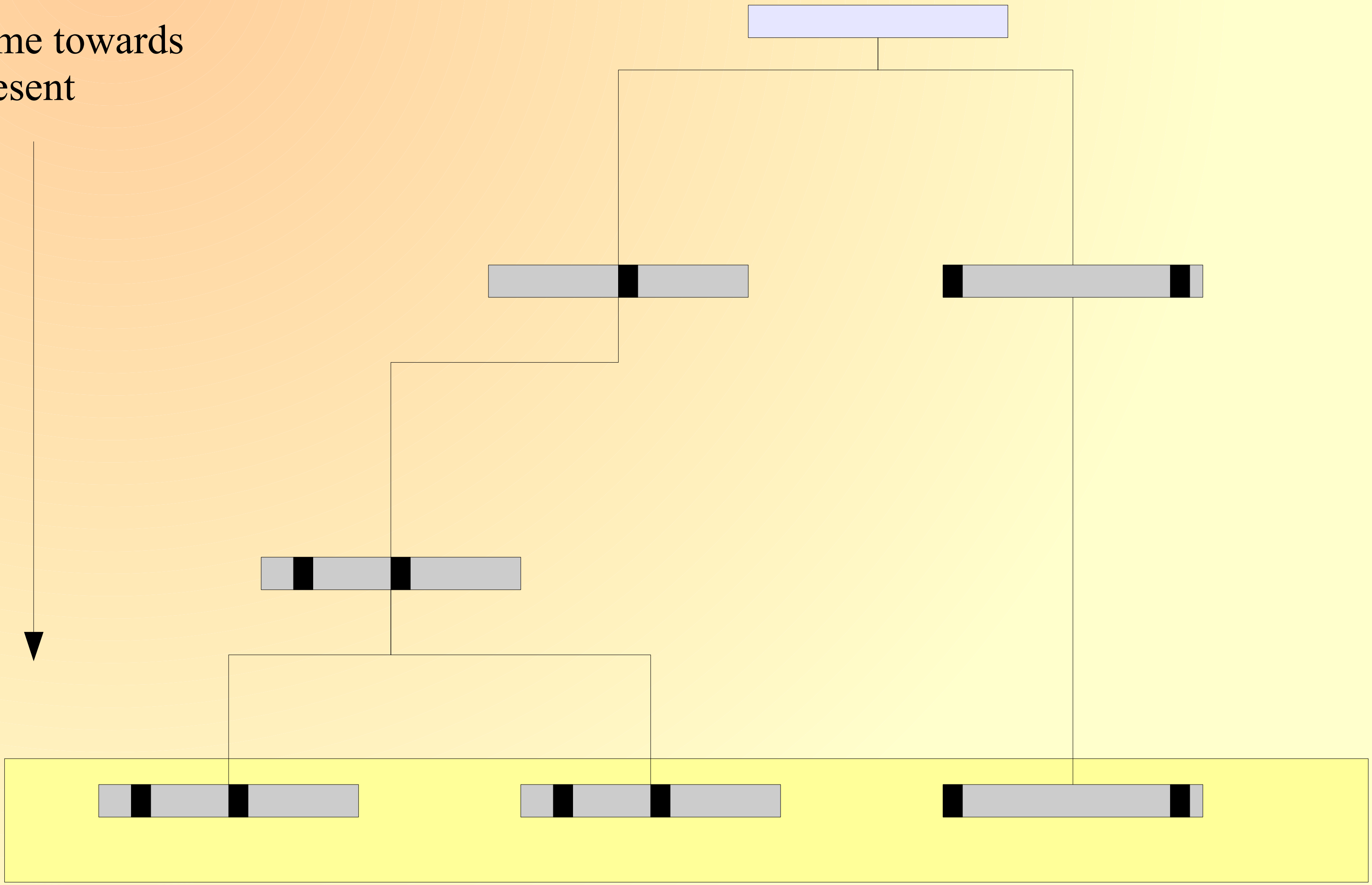


Correlations due to relatedness

Correlations due to linkage

Ancestral Tree

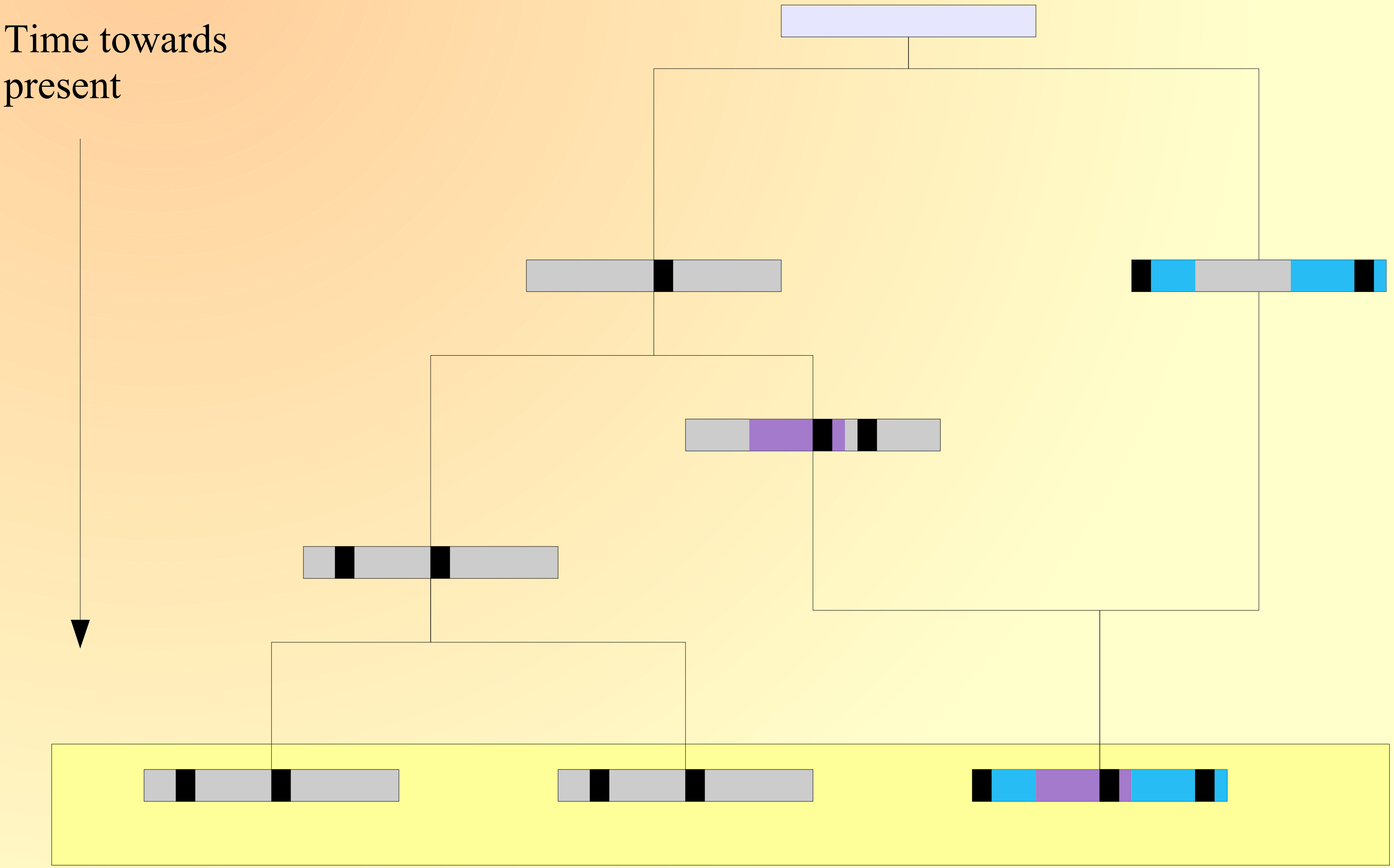
Time towards present



Hein, Schierup and Wiuf 'Gene Genealogies, Variation and Evolution', OUP 2005

Ancestral Recombination Graph

Time towards present



Hein, Schierup and Wiuf 'Gene Genealogies, Variation and Evolution', OUP 2005

Ancestral Recombination Graph - Summary

- Ancestral Recombination Graph (ARG) model
  - backwards in time, ignore unobserved ancestors

is equivalent to the

- Forwards in time model
  - Random mating, within known size populations
  - No selection

- Inference under the ARG is impossible for reasonable datasets

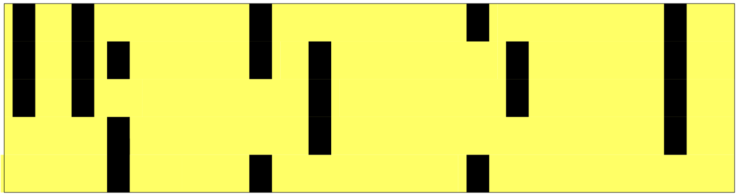
Sex, sample randomisation

- ARG-based inference 'impossible'
- Population model:
  - Assume individuals exchangable within populations
  - Simple distribution (Dirichlet...) model for SNP frequencies in each
- Gives likelihood for frequency of SNPs
  - Assume no linkage (linkage approximations exist)
- Gives popular STRUCTURE* model
  - Still can't cope with large datasets
- Can we do this well on genomic data?

**Pritchard, Stephens and Donnelly, Genetics, 155:945-959, 2000*

Outline: The process

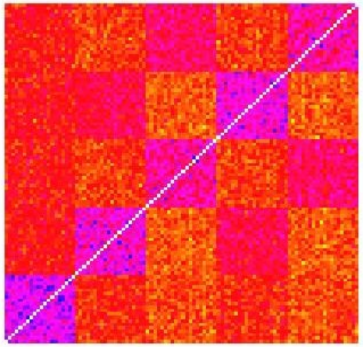
SNPs $O(10^{6-9})$



Step 1: *ChromoPainter*

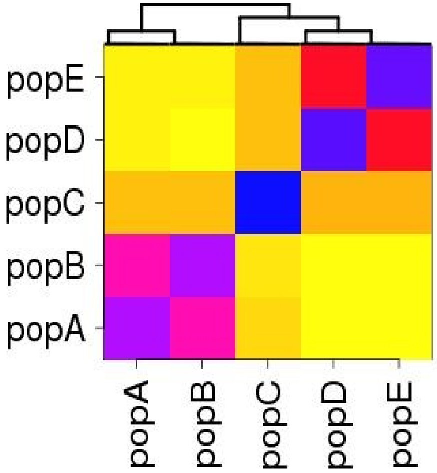
Individuals $O(10^{3-4})$

Coancestry matrix



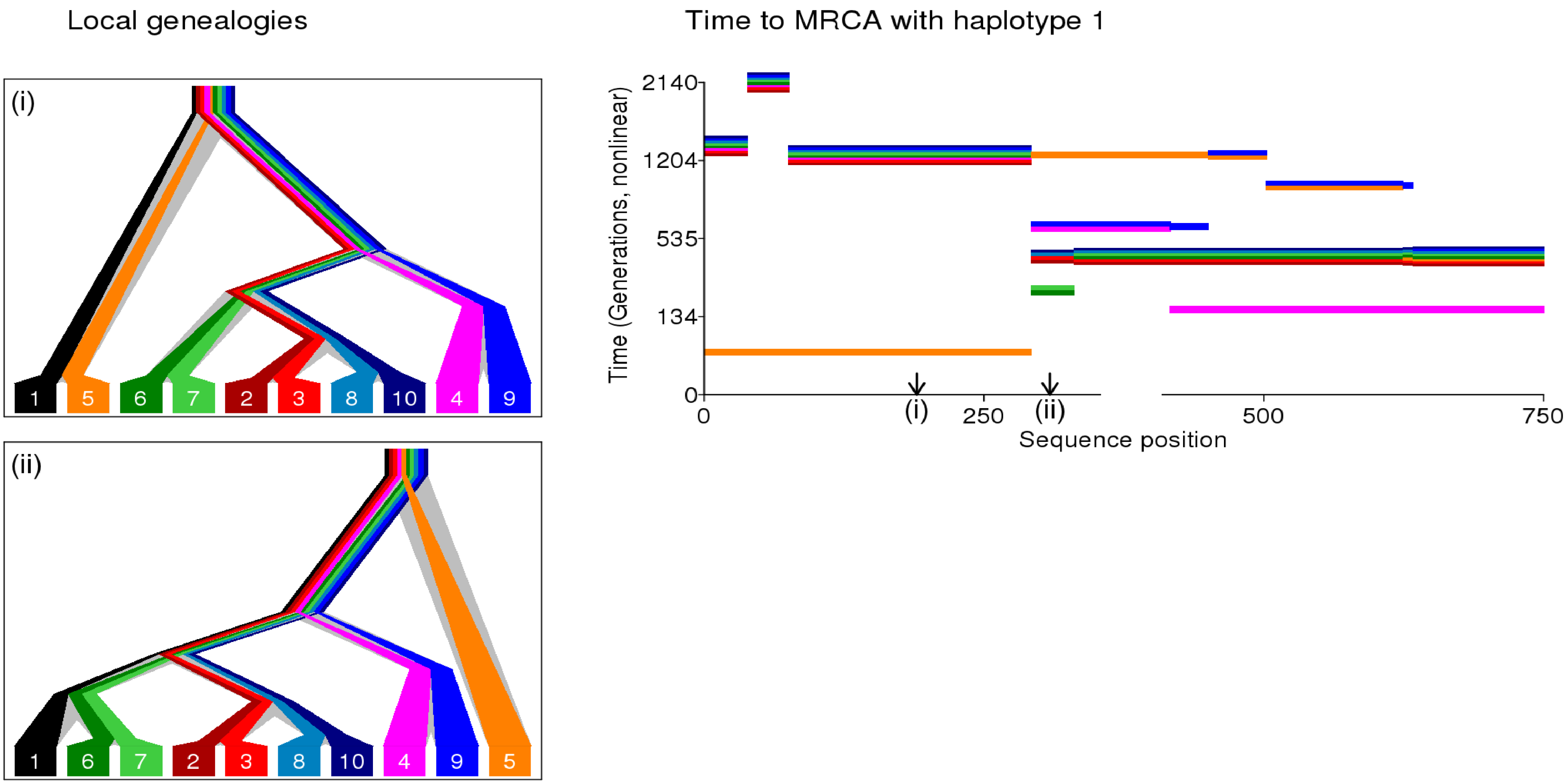
Step 2: *Finestructure*

Populations $O(10^{2-3})$



1) ChromoPainter: SNPs are converted to detailed co-inheritance information

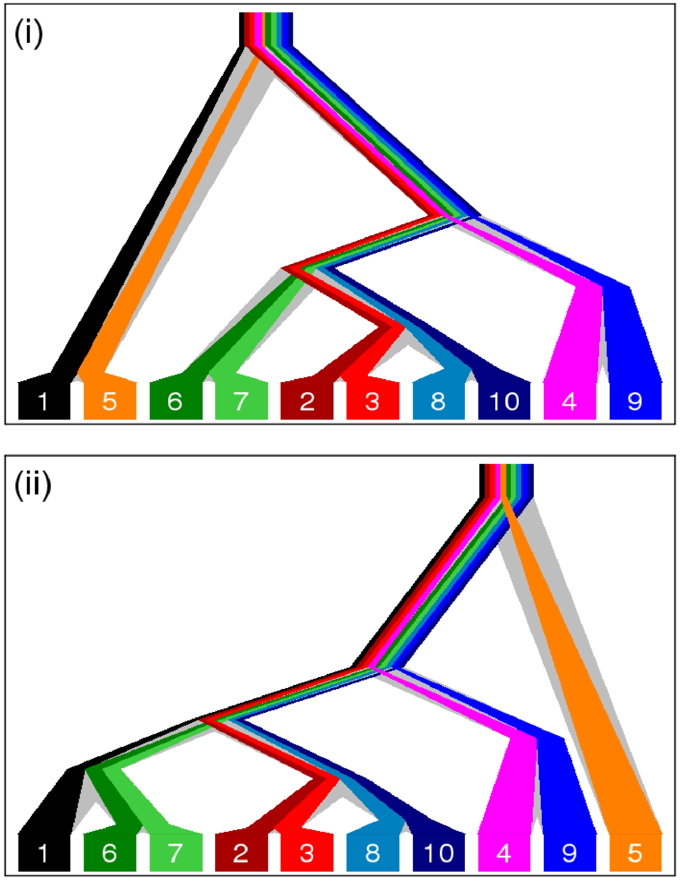
2) Finestructure: analyse the population structure



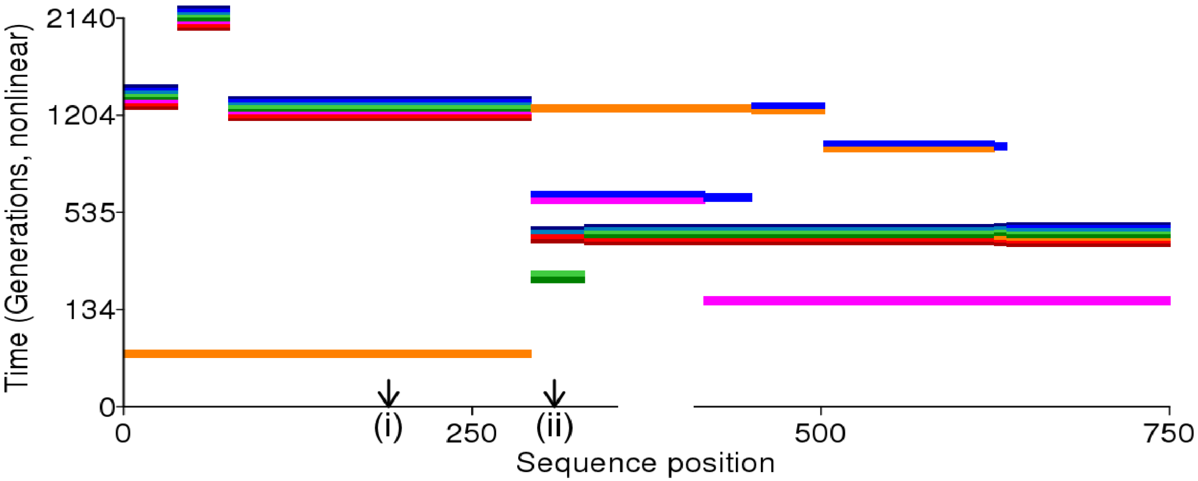
ChromoPainter step

See: Li and Stephens , Genetics 165:2213-2233, 2003

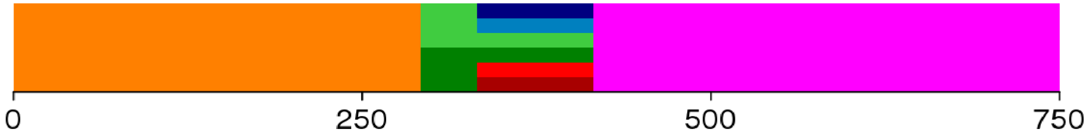
Local genealogies



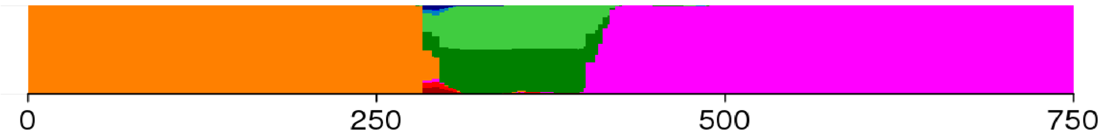
Time to MRCA with haplotype 1



True 'nearest neighbour' distribution of haplotype 1



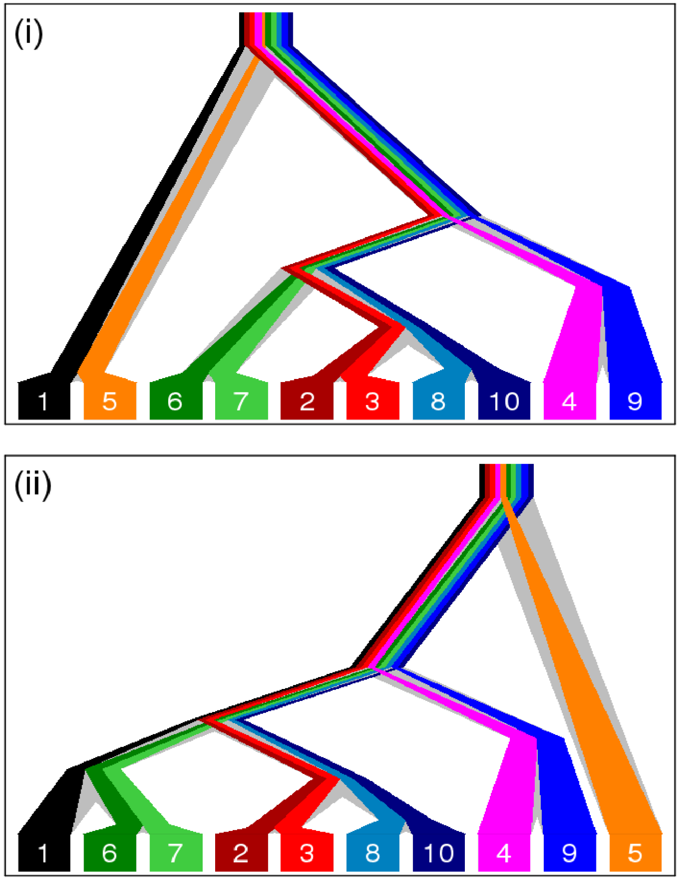
Mean painting of haplotype 1



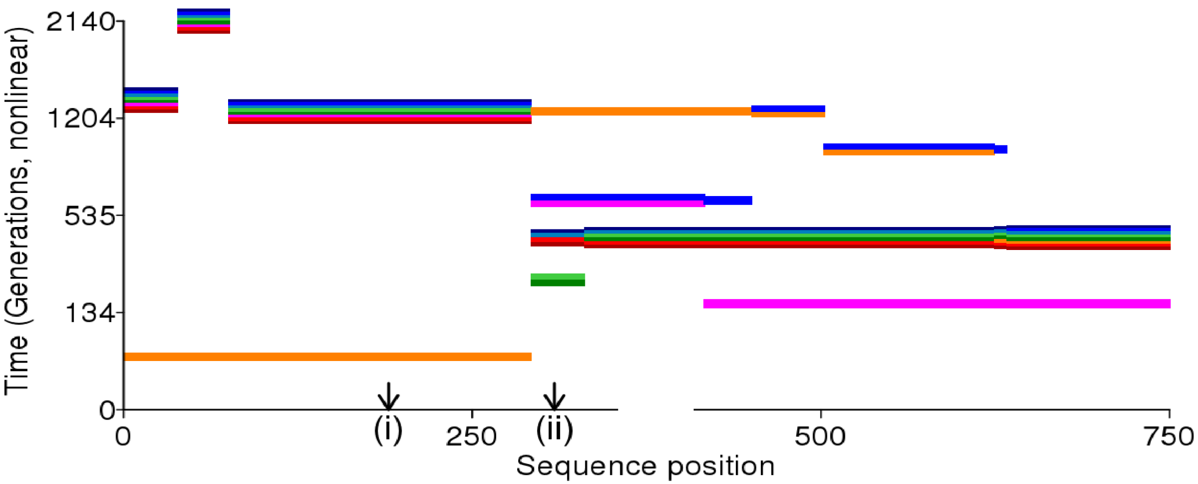
ChromoPainter step

See: Li and Stephens , Genetics 165:2213-2233, 2003

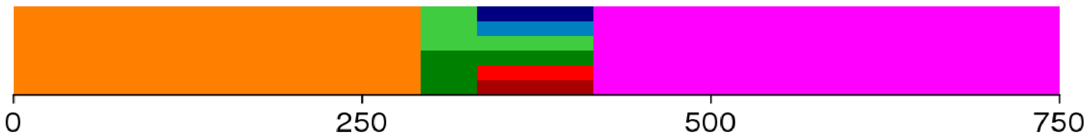
Local genealogies



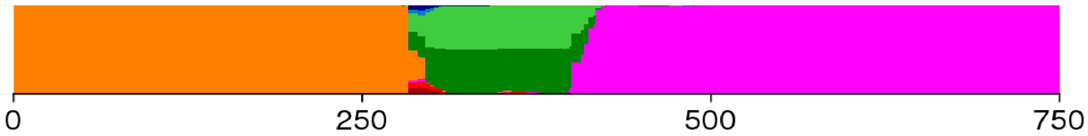
Time to MRCA with haplotype 1



True 'nearest neighbour' distribution of haplotype 1



Mean painting of haplotype 1



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

ChromoPainter step

See: Li and Stephens , Genetics 165:2213-2233, 2003

fineSTRUCTURE: partition model

- Individuals exchangable within populations

$$x_{ab} = \sum_{i \in a, j \in b} x_{ij}$$

- Populations donate chunks independently at a characteristic rate P_{ab}

$$p(X|P) = \prod_{a,b=1}^{K} \left(\frac{P_{ab}}{\hat{n}_b} \right)^{x_{ab}}$$

Coancestry matrix

Donation frequency of population

Population assignment

Number of individuals to donate from

Popu model

- Individua ns

-

ADVERT!

www.paintmychromosomes.com

Paper to appear:

PloS Genetics

Coancestry matrix

n assignment

from

Probability of a partition

- Dirichlet process prior for partition η:

$$\eta \sim \alpha^{K} \prod_{b=1}^{K} \Gamma(\hat{n}_b)$$

$$\{P_1, \cdots, P_K\} | \eta = \prod_{b=1}^{K} G_0$$

- Rows of P_{ab} (i.e. G_0) are Dirichlet *(containing hidden biological parameters)*…
- … so conjugate, and we integrate out P_{ab}

(Idea: add each individual, update Dirichlet posterior, use as prior for the next individual)

- MCMC sampling of partitions

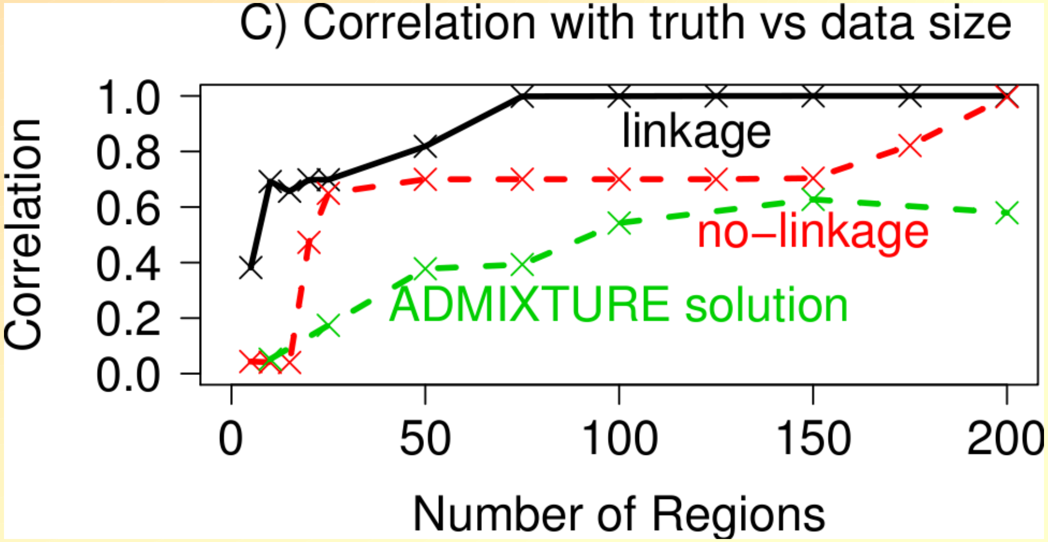
Proven theoretical results

- To O(N), the C*oancestry matrix* is a rotation of the *eigenvector matrix*
  - *If SNPs are uncorrelated*
  - *and the number of individuals is large*
- To O(N), the fineSTRUCTURE likelihood is equivelent to the STRUCTURE* likelihood
  - *if SNPs are uncorrelated,*
  - *drift is weak,*
  - *genotyped SNPs are not very rare*
- With linkage model we do better.

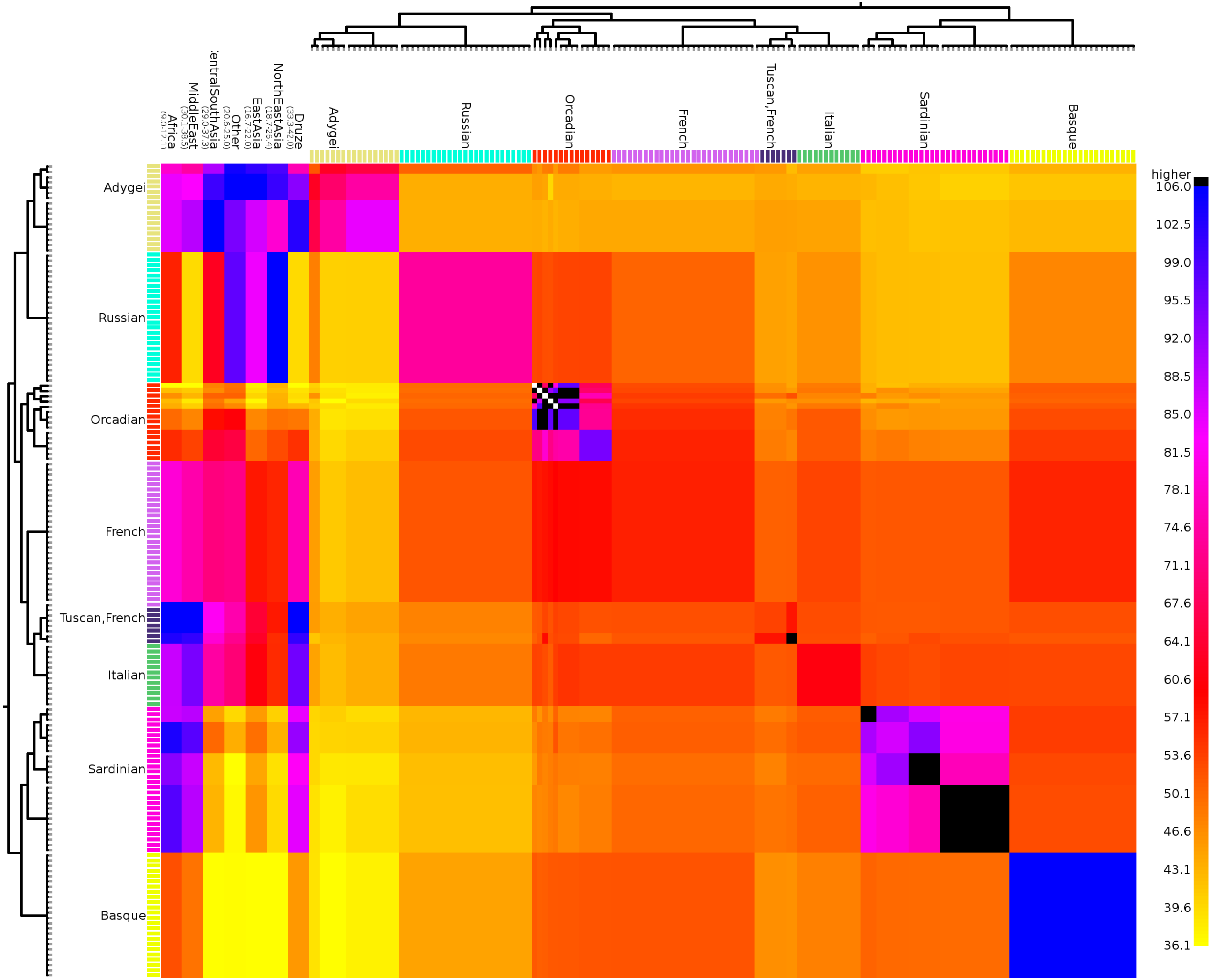
**Pritchard, Stephens and Donnelly, Genetics, 155:945-959, 2000*
Calculations due to Simon Myers

Some checks

- Excellent MCMC Mixing
- Simulated data: complex demographic scenario*
- Confirm theoretical results



*using SFS_CODE: Hernandez, RD, Bioinformatics, 24:2786-2787, 2008



Africa
(9.0-12.1)
MiddleEast
(30.1-38.5)
entralSouthAsia
(29.0-37.3)
Other
(20.6-25.0)
EastAsia
(16.7-22.0)
NorthEastAsia
(18.7-26.4)
Druze
(33.3-42.0)
Adygei
Russian
Orcadian
French
Tuscan,French
Italian
Sardinian
Basque
Adygei
Russian
Orcadian
French
Tuscan,French
Italian
Sardinian
Basque
higher
106.0
102.5
99.0
95.5
92.0
88.5
85.0
81.5
78.1
74.6
71.1
67.6
64.1
60.6
57.1
53.6
50.1
46.6
43.1
39.6
36.1

Acknowledgements:

Garrett Hellenthal

(Oxford)

(painting algorithm)

Simon Myers

(Oxford)

(theory)

Daniel Falush

(Max Planck Institute)

(concept)

Peter Green (Bristol) – Grant, support
Bluecrystal HPC facilities @ Bristol

See Also:

- Bruce Winney: People of the British Isles (POBI). *Saturday 12:20 C15*

- ChromoPainter Code & GUI
- FineSTRUCTURE Code & GUI: www.paintmychromosomes.com

The future – Admixture model

- Pure population structure is not correct – recent mixing leads to admixture
  - Seek conjugate mixture model for individuals
  - Hierarchical Dirichlet Process!
  - Interpretation: Pure populations created by drift, we see mixtures
- Better model:
  - Allow drift and admixture to both occur in real time
  - Requires more sophisticated model, can we keep conjugacy?
  - (Matrix Coalescent* results available)
  - Dirichlet diffusion tree** concept

**Wooding and Rogers, Genetics, 161:1641-1650, 2002*

***Neal, in J. M. Bernardo, et al. (ed.), Bayesian Statistics 7, pp. 619-629, 2003*

Posterior evaluation

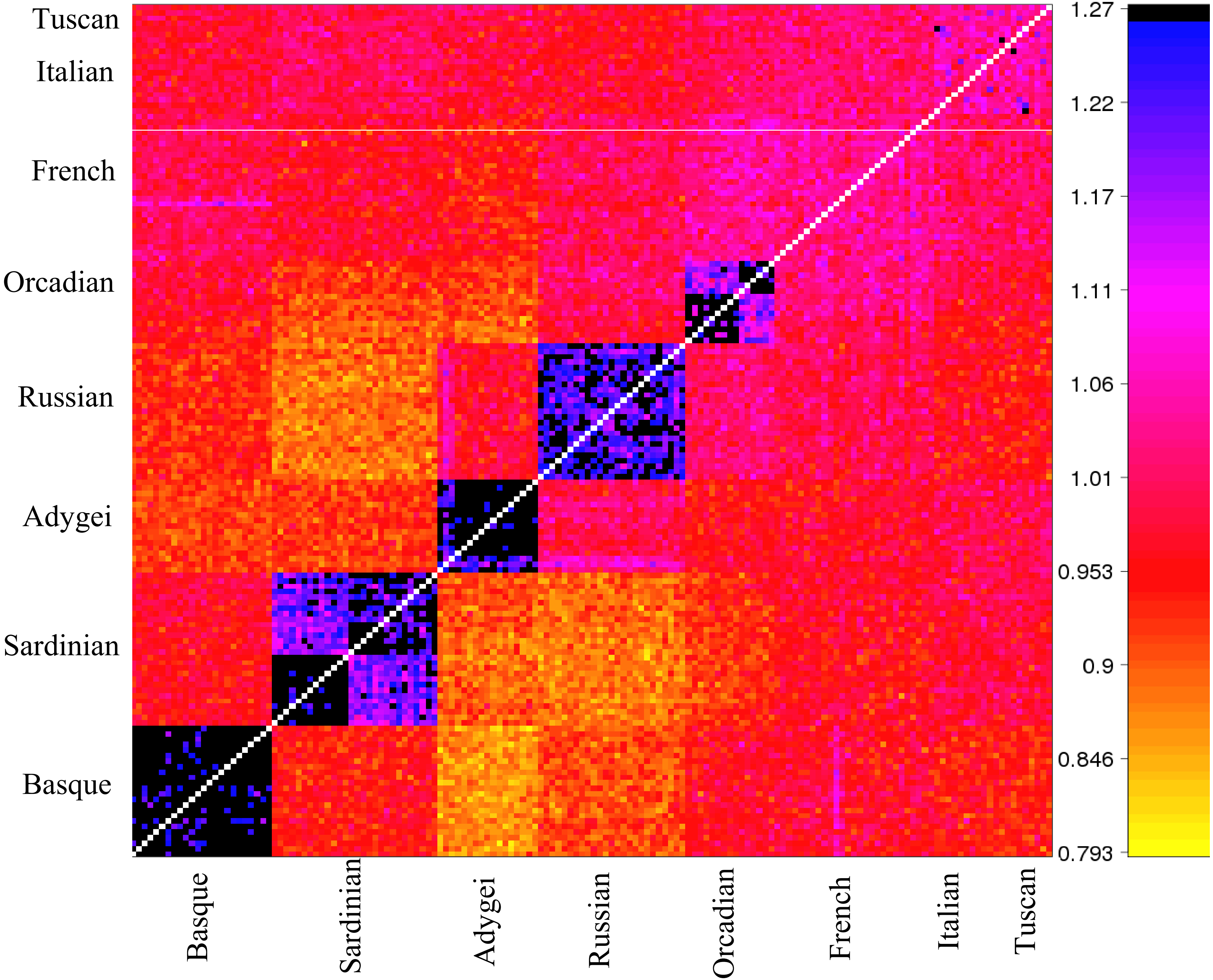
- MCMC update of hyperparameters and partitions
- Partition moves:
  - Move an individual
  - Merge
  - Split
  - Merge and resplit
- Merge/split 'nearly Gibbs' move:

$$p(q_m; a, b) = p(q_1)\, p(q_2|q_1) \cdots p(q_m|q_{1:m-1})$$

$$p(q_m = a) \approx \hat{n}_a \int F(x_m|P_m)\, dH_{<m,S_a}(P_m)$$

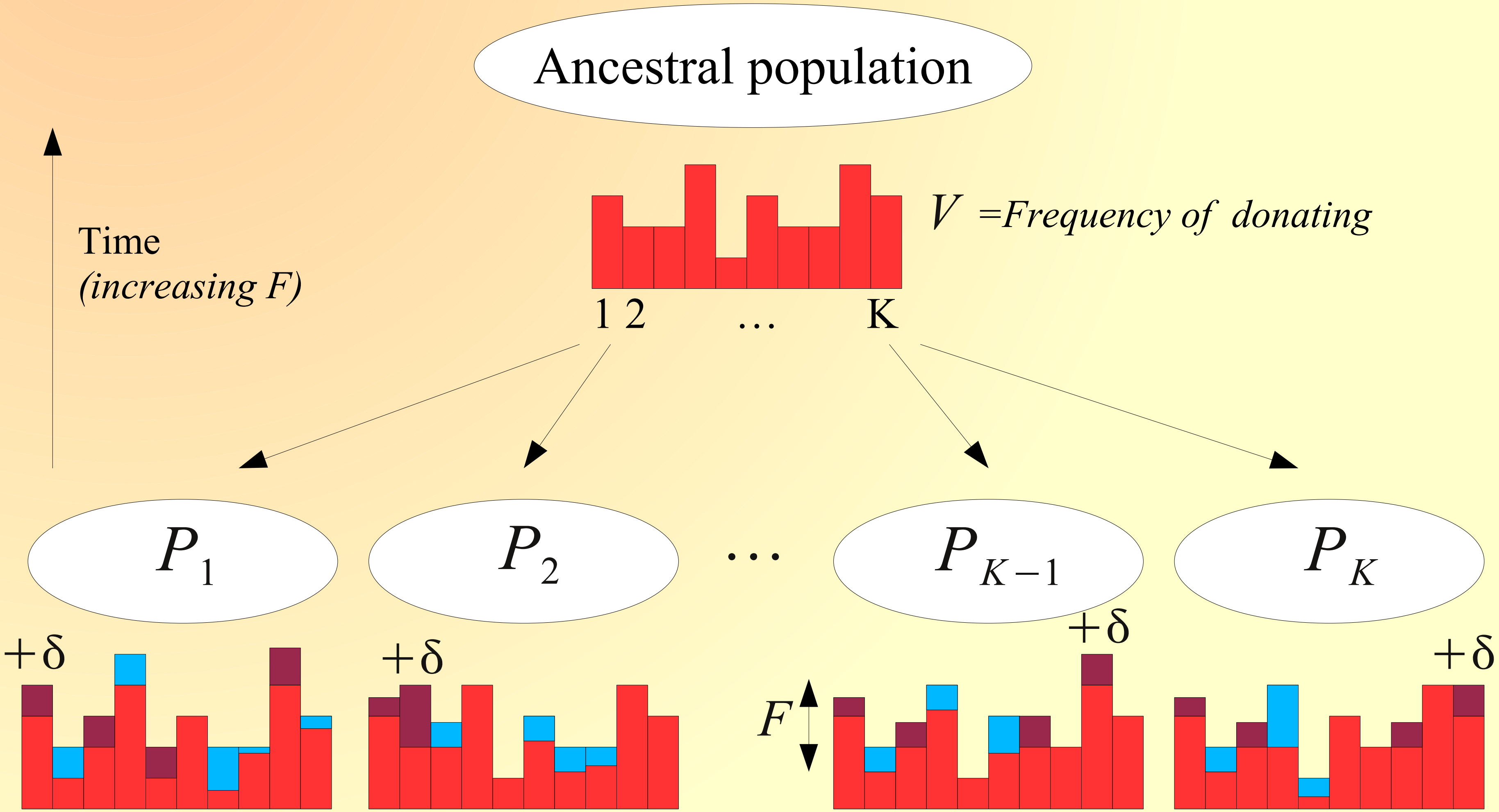
(Not exact as the 'unsplit' population interacts with the remaining dataset)

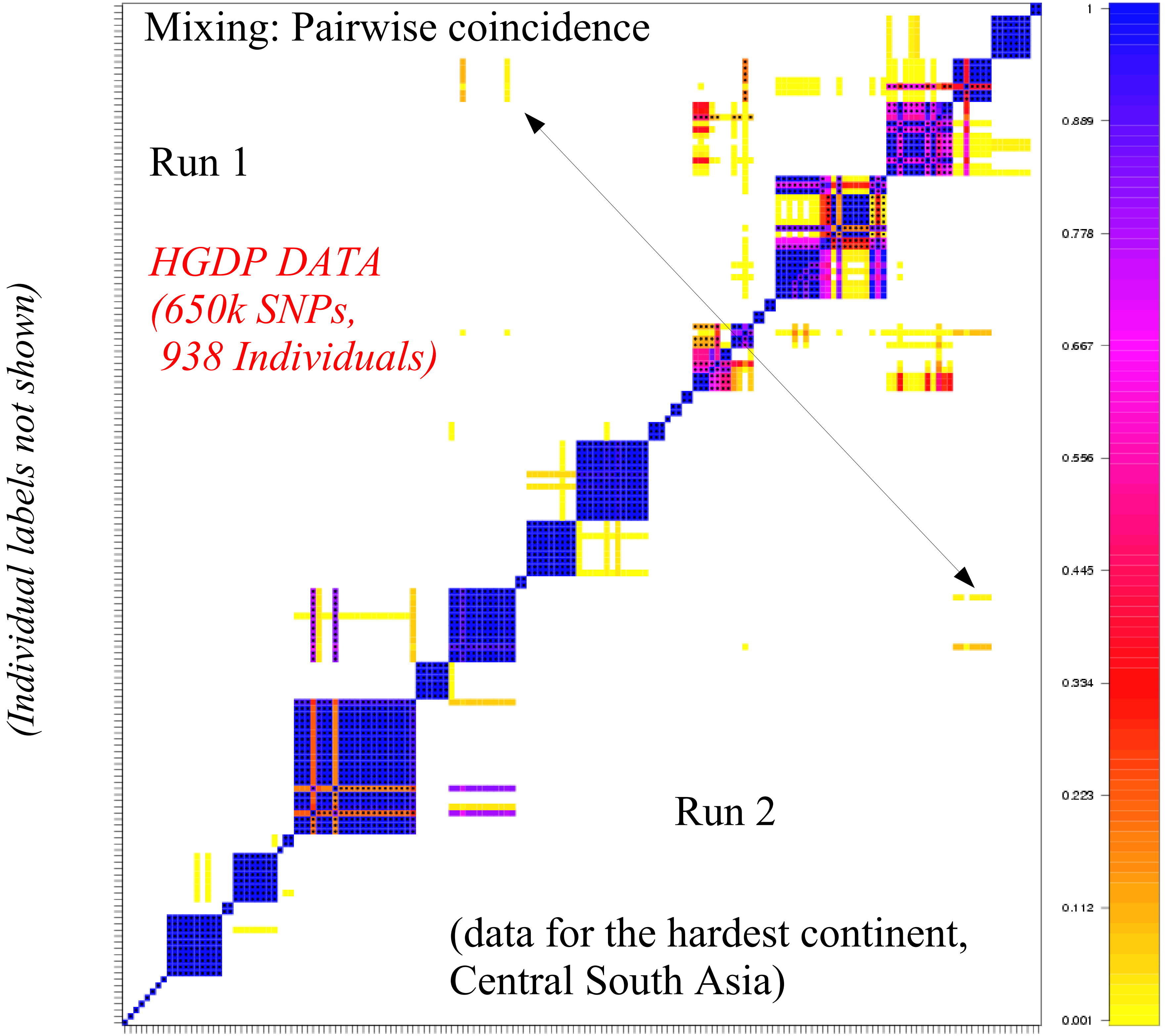
Simple case: Pella and Masuda Canadian J. Fish. Aquatic Science 63:576-596, 2003


Tuscan
Italian
French
Orcadian
Russian
Adygei
Sardinian
Basque
Basque
Sardinian
Adygei
Russian
Orcadian
French
Italian
Tuscan
1.27
1.22
1.17
1.11
1.06
1.01
0.953
0.9
0.846
0.793

Weak Biological Model for prior

'Correct' Ancestral Recombination Graph for the limit of large populations at large time with simple population structure





Mixing: Pairwise coincidence

Run 1

HGDP DATA (650k SNPs, 938 Individuals)

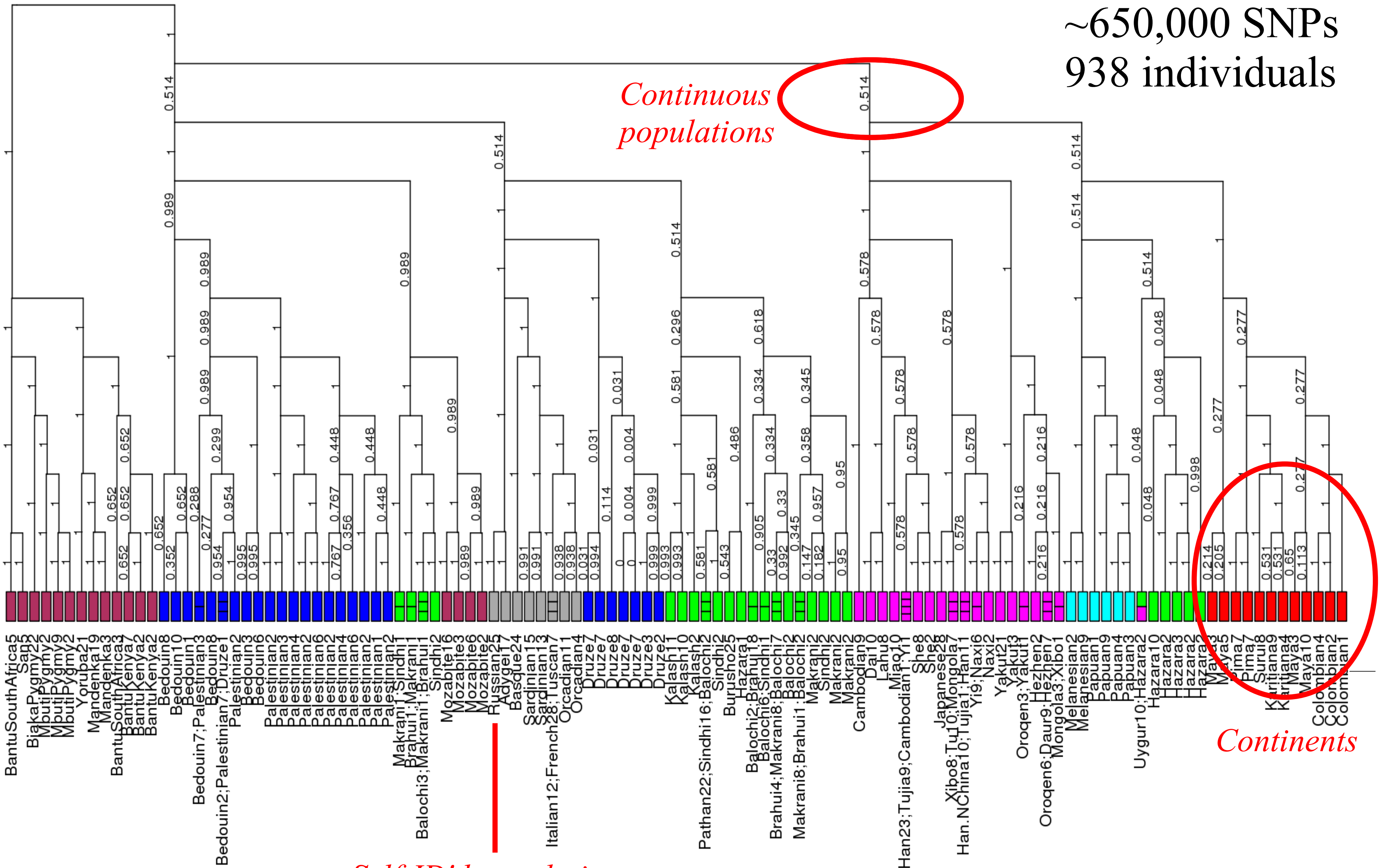
Run 2

(data for the hardest continent, Central South Asia)

(Individual labels not shown)

MAP tree: whole world HGDP data

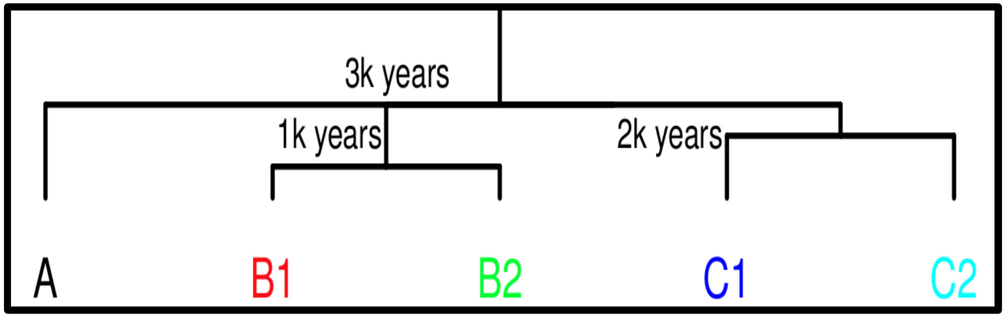
~650,000 SNPs
938 individuals



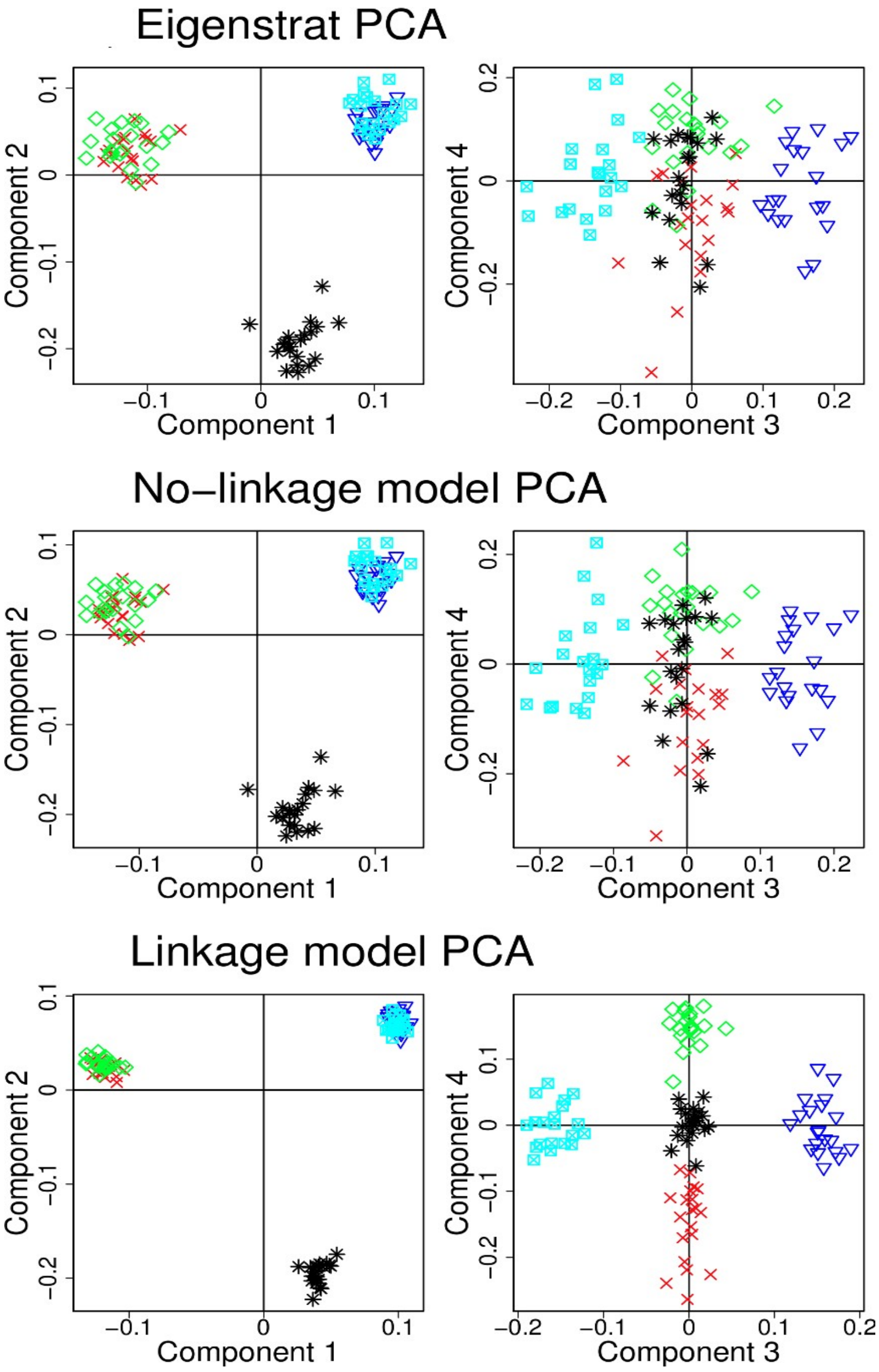
Continuous populations

Continents

Self-ID'd populations



Simulation scenario: 'Europe'


Eigenstrat PCA

No–linkage model PCA

Linkage model PCA

Posterior evaluation: building block

- Sample from posterior

$$p(q_m; a, b) = p(q_1)\, p(q_2 | q_1) \cdots p(q_m | q_{1:m-1})$$

- Metropolis-Hastings proposal for a split:
  - Random individuals creates population *a* and *b* from *c*
  - Move rest from *c* with probability

$$p(m; a) \propto \hat{n}_a \int F(x_m | p_m)\, dH_{<m, S(p_m)}$$

$$\approx n_a \frac{P(S_a, \{i=1,\cdots,m\})\, P(S_c, \{i=1,\cdots,m\})}{P(S_a, \{i=1,\cdots,m-1\})\, P(S_c, \{i=1,\cdots,m-1\})}$$

(Not exact as the 'unsplit' population interacts with the remaining dataset)

Exact case: Pella and Masuda Canadian J. Fish. Aquatic Science 63:576-596, 2003

Probability of a partition

Rows of P_{ab} are Dirichlet

- Conjugate to multinomial, sum to 1
- Weak prior

Compute posterior incrementally due to conjugacy

$$p(x_a|q)=\prod_{m\in a}\int F(x_m|P_a,q)\,dH_{<m,S_a}(P_a)$$

$$dH_{<m,S_a}(P_a)=Dirichlet(P_a;\{\beta_{ab}+x_{<m,b}\}_{b=1,\cdots,K})$$

(Idea: add each individual, update Dirichlet posterior, use as prior for the next individual)

Final model

- Posterior

$$p(\eta|X)\propto\alpha^{K}\prod_{a=1}^{K}\Gamma(\hat{n}_a)\frac{\Gamma(\beta_a)}{\Gamma(x_a+\beta_a)}\prod_{b=1}^{K}\frac{\Gamma(x_{ab}/c+\beta_{ab})}{\Gamma(\beta_{ab})\hat{n}_b^{x_{ab}}}$$

- Prior for hyperparameters

$$\beta_{ab}=\begin{cases}\gamma V_b & if\ a\neq b\\ \gamma(1+\delta)V_b & if\ a=b\end{cases}$$

Drift due to mutation

Ancestral donation frequency

$$\gamma=(1-F)/F$$

Drift in allele frequency