Identifying fine population structure using genomic scale data

Daniel Lawson

Heilbronn Institute for Mathematical Research
School of mathematics
University of Bristol

dan.lawson@bristol.ac.uk

Work with:
Garrett Hellenthal
Daniel Falush
Simon Myers

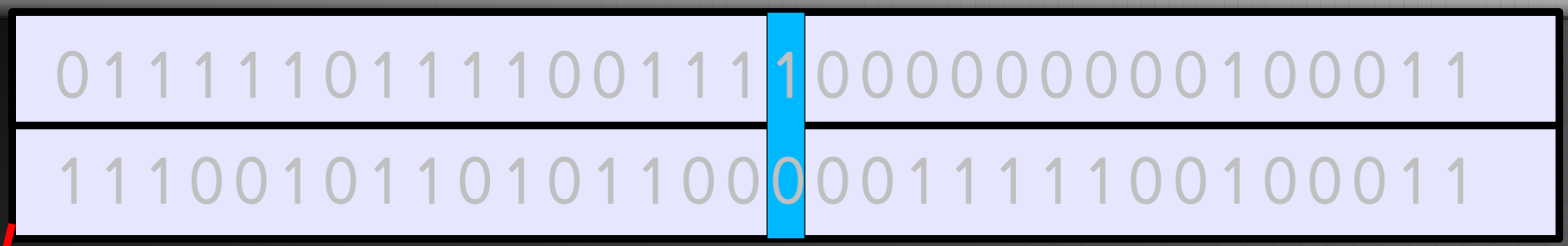
www.paintmychromosomes.com

Talk outline

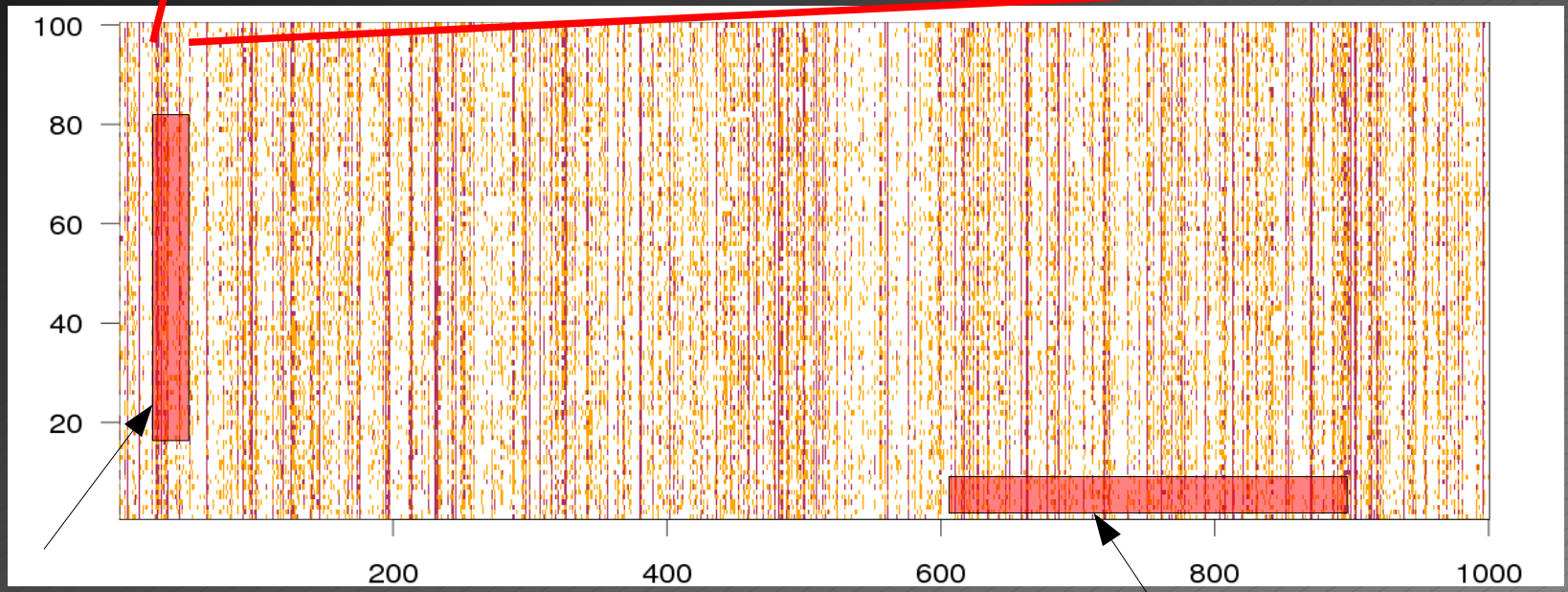
1. Ancestry Overview
2. Previous population model motivation (STRUCTURE)
3. Summary Statistic approaches
4. ChromoPainter ancestry summary model
5. FineSTRUCTURE population model
6. Applications and benefits

Genetics data

Individual *i*

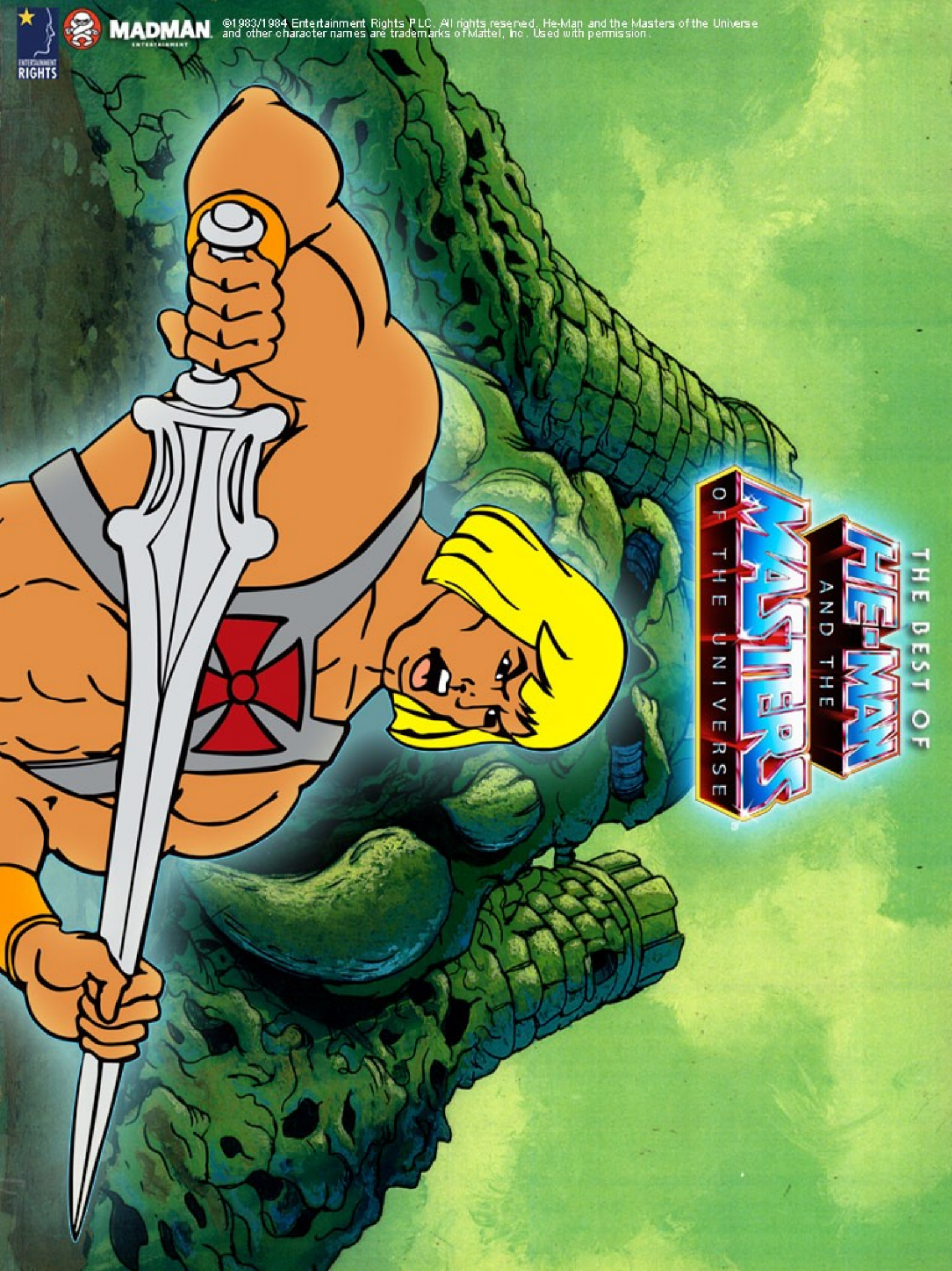


SNP *l*



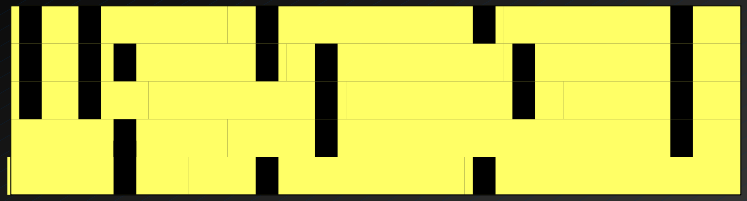
Daniel Lawson, University of Bristol
dan.lawson@bristol.ac.uk


JS3



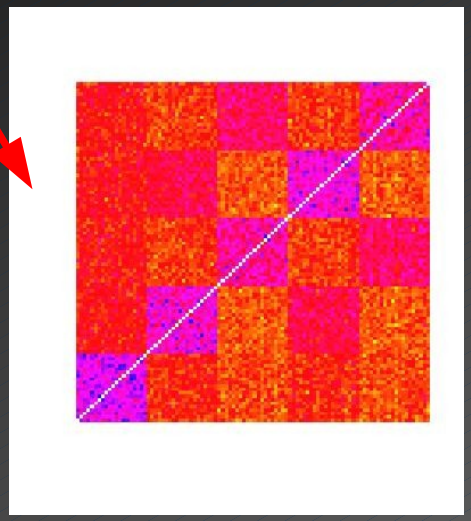
Outline: The process

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



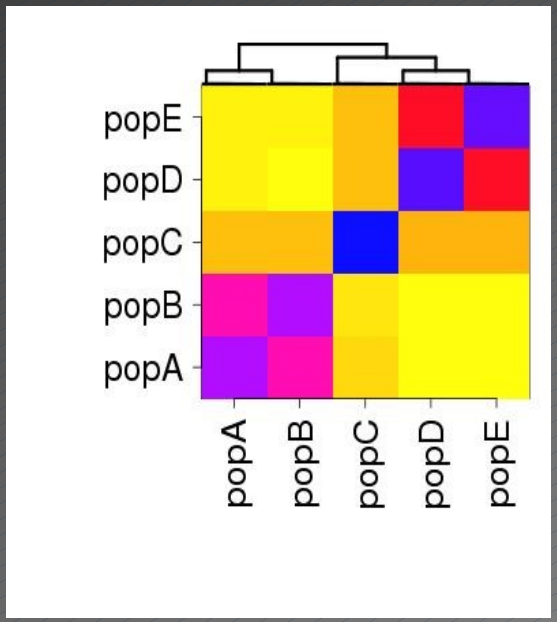
Step 1: Similarity matrix

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



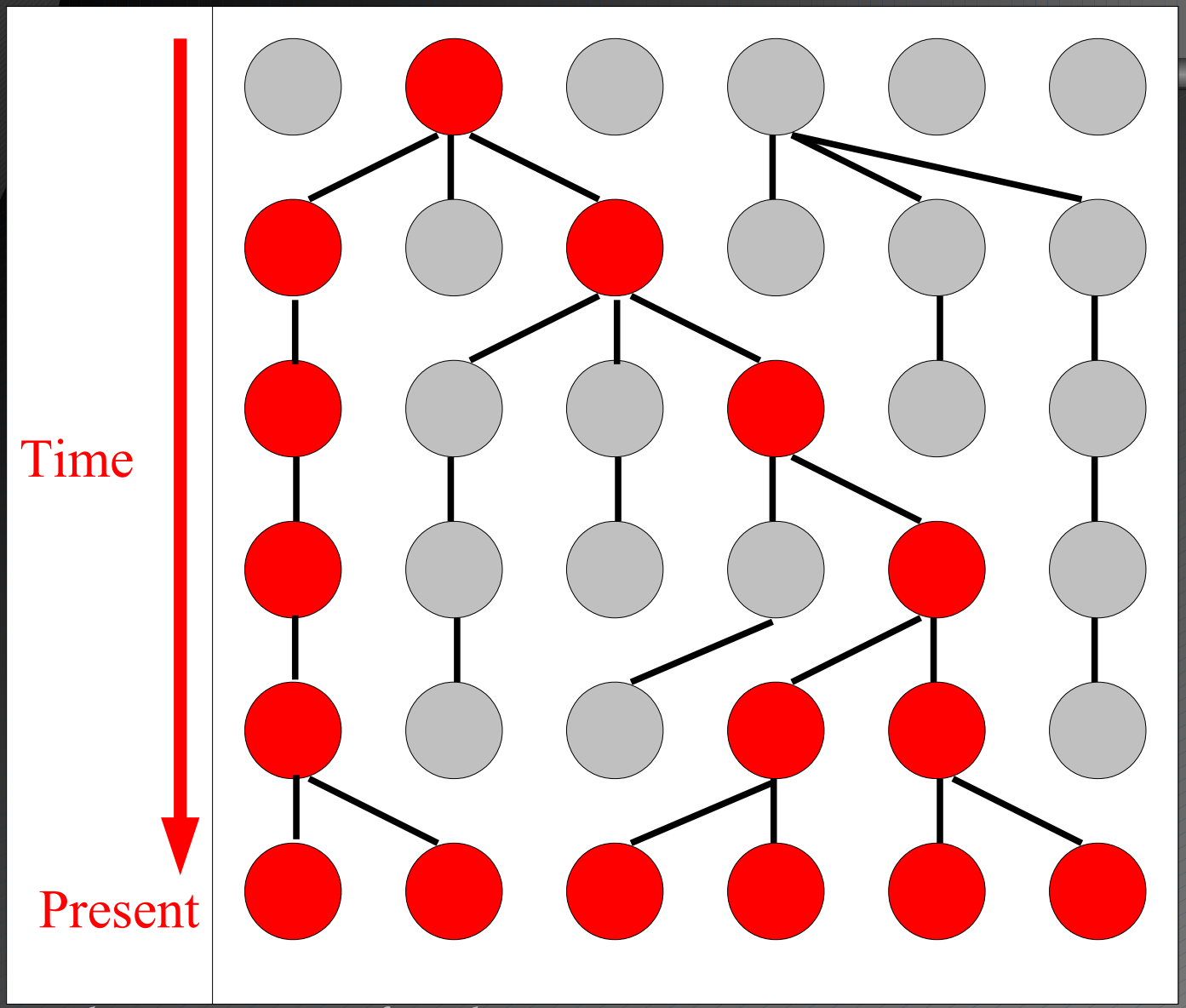
Step 2: Clustering

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



Step 1: SNPs are converted to similarity matrices
Step 2: Analyse the population structure

Ancestry process



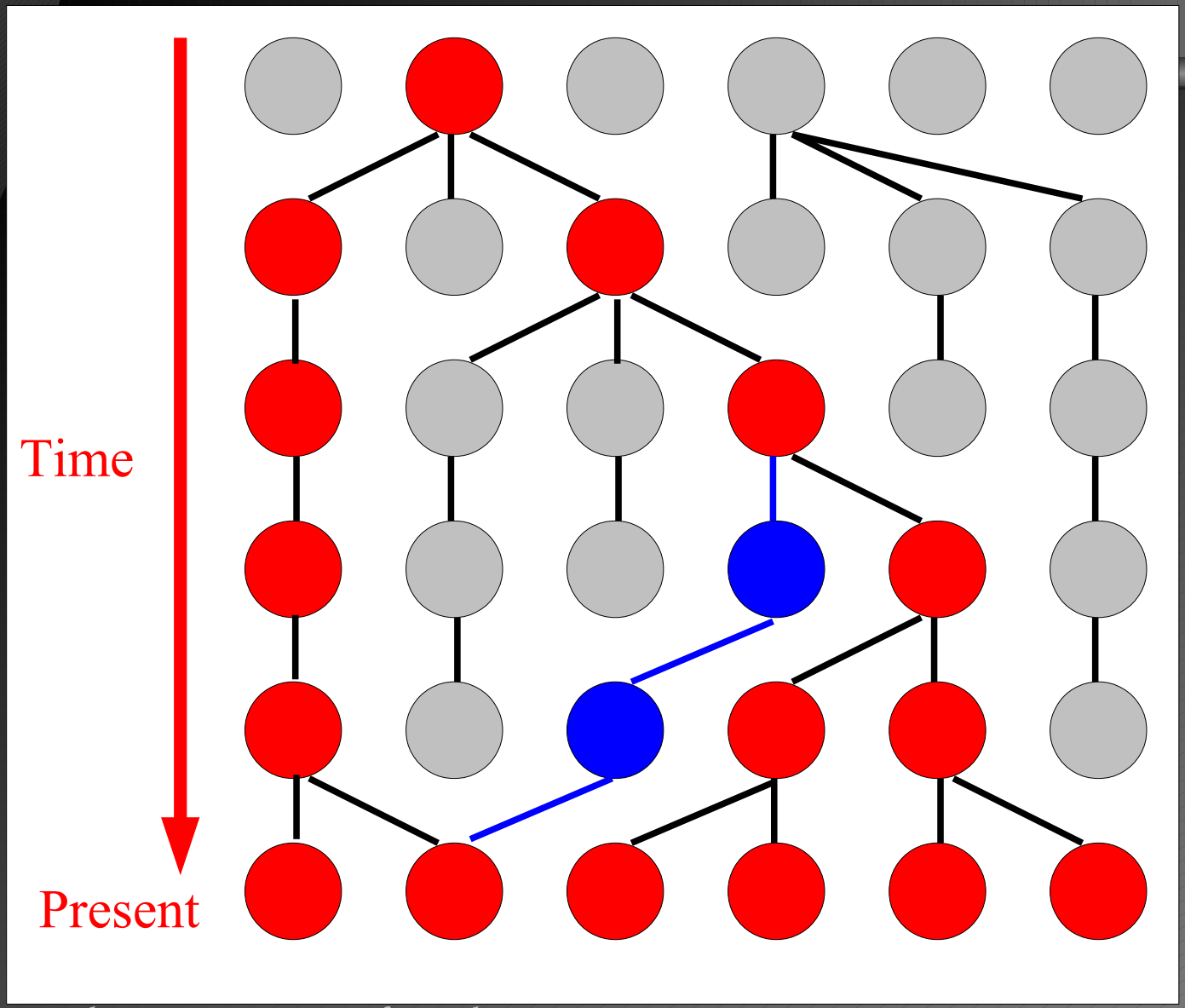
- Each generation randomly chooses parent
- Red individuals ancestors to those sampled
- Take limit

$$N \to \infty$$

keeping N/T constant

- Rate of *coalescence* between pairs $\to 1$
- Lots known about this *Coalescent Tree* distribution.

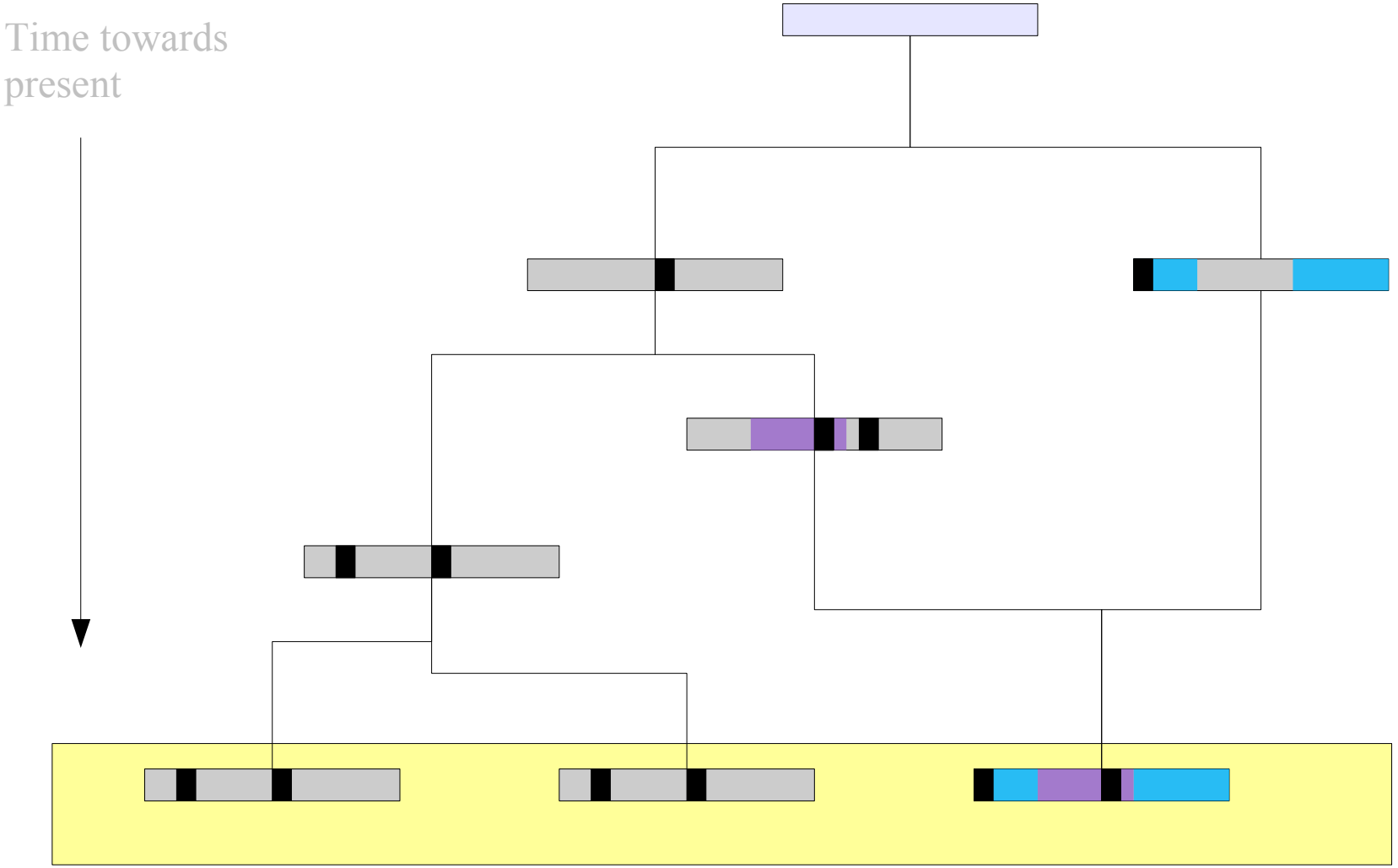
Ancestry process with recombination



- Probability ρ of getting DNA from two parents
- Creates a graph structure
- Coalescence rates unchanged...
- But now a birth/death process
- Easily simulated but few analytical results

Daniel Lawson, University of Bristol
dan.lawson@bristol.ac.uk

Ancestral Recombination Graph



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

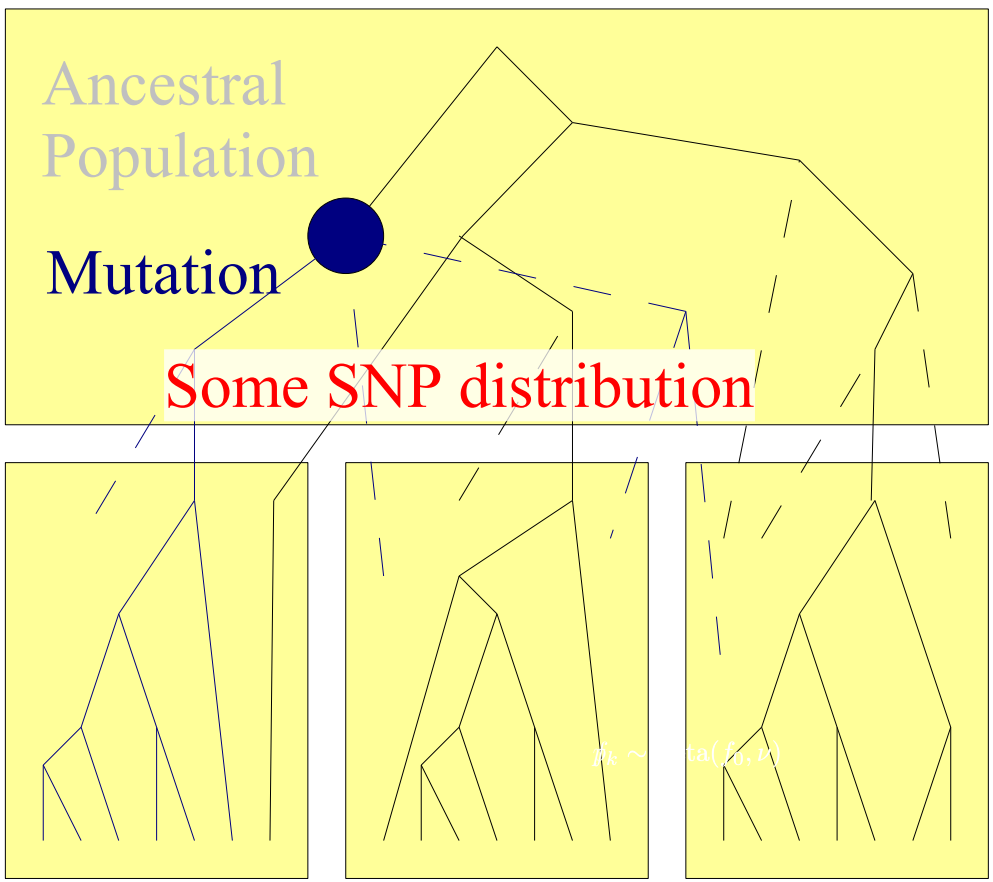
Ancestral Recombination Graph - Summary

- 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
- But when recombination is large, each SNP is independently drawn from a random tree

Genetic drift



'Independent drift' of SNP distribution

- Kimura derived exact distribution for this case... (nasty)
- Wright studied a related model, leads to Beta distribution of SNP frequency

$$p_k \sim \mathrm{Beta}(p_0, \nu)$$

- Normal distribution approximation exists... simpler to handle covariance effects

$$p_k \sim \mathrm{N}\left(p_0, \sigma^2 p_0(1-p_0)\right)$$

STRUCTURE(AMA) Model

$$x_{il}|q_i, \mathbf{p}_l \sim \text{Binom}(p_{q_i l})$$

(SNPs depend on pop SNP frequency)

$$p_{kl} \sim \text{Beta}(\beta_l)$$

(Pop SNP frequency drifts independently)

$$\beta_l \sim P(\beta_l)$$

(Some ancestral frequency model)

$$\mathbf{q}, K \sim \text{DP}(\alpha, G_0)$$

(Some prior on assigning individuals to populations)

- $\mathbf{p}$ is a vector of length K
- The Dirichlet Process prior is for K and the number of individuals in each population
- STRUCTURE* uses a different prior for $\mathbf{q}, K$

- Pella & Masuda 2006
- STRUCTURAMA (Huelsenbeck & Andolfatto 2007)

Daniel Lawson, University of Bristol
dan.lawson@bristol.ac.uk

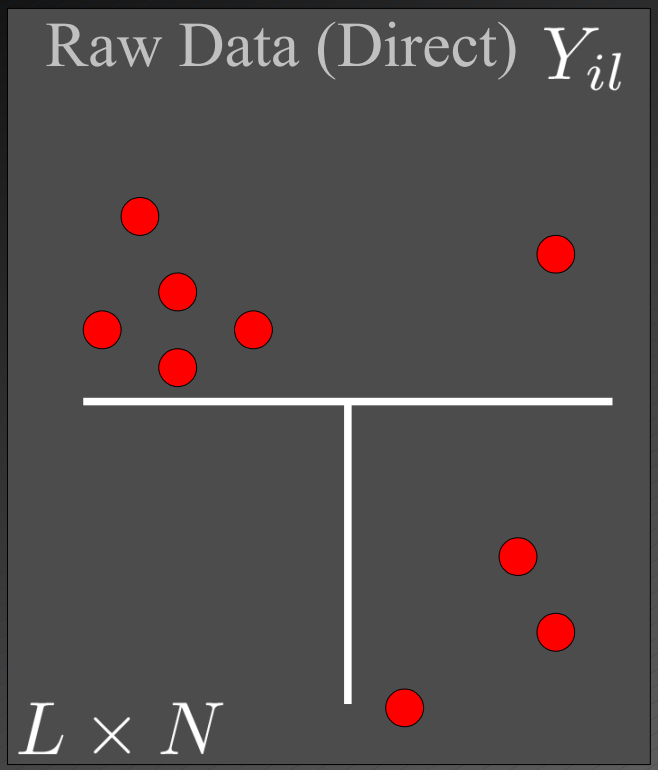
**Pritchard, Stephens and Donnelly, Genetics, 155:945-959, 2000*

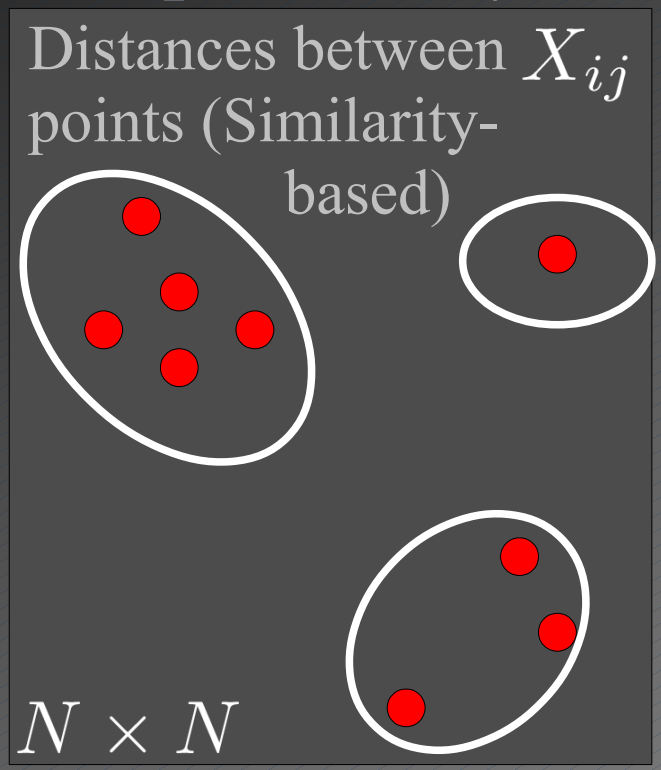
Extension to large datasets

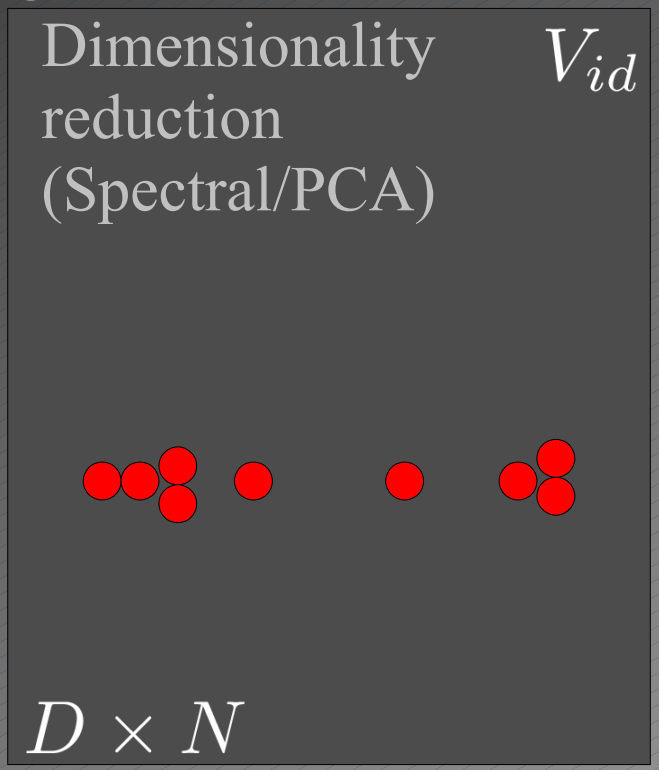
- Lots of latent parameters (SNP frequencies, etc) makes MCMC slow and sticky for large datasets
- Also, deal with correlations between SNPs
- Return to model?
- Start with summary statistic approaches

Clustering

- K clusters (which should be inferred)
- Find 'best' assignment of individuals into populations
- 'Best' is some score, potentially using:







$L \gg N \gg D$

Daniel Lawson, University of Bristol
dan.lawson@bristol.ac.uk

Pairwise summary statistics

Allele Sharing distance/Identity by state
Edit distance/'Norm' distance e.g. L1, L2

$$\sum_l |Y_{il} - Y_{jl}|$$

Count the number of shared SNPs

Covariance

$$\sum_l (Y_{il} - f_l)(Y_{jl} - f_l)$$

PCA uses the covariance

Normalised covariance

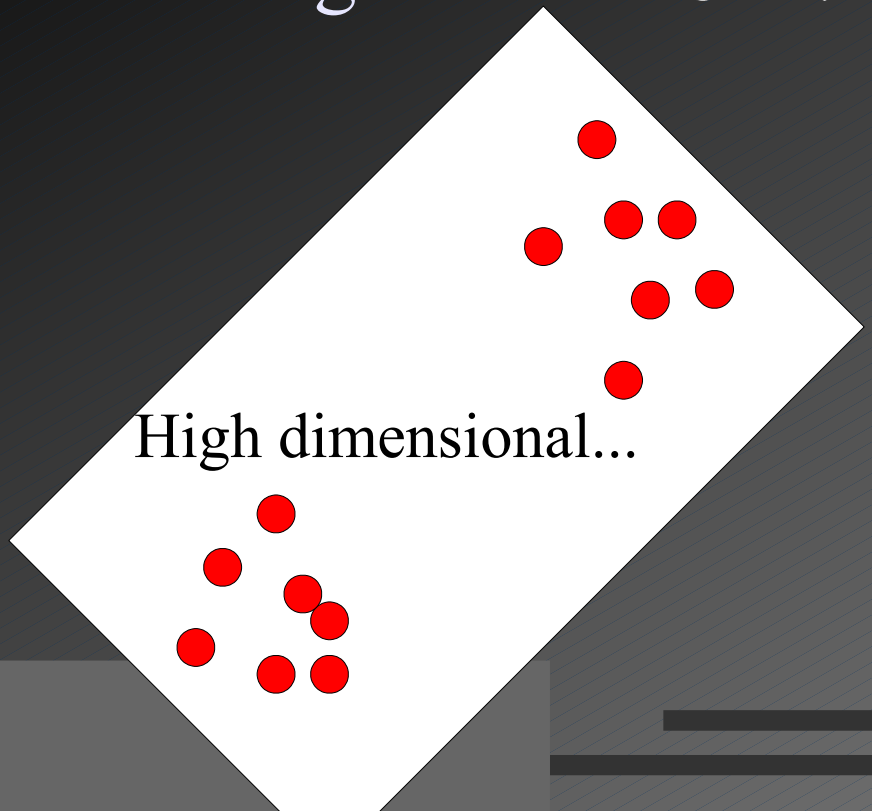
$$\sum_l \frac{(Y_{il} - f_l)(Y_{jl} - f_l)}{f_l(1 - f_l)}$$

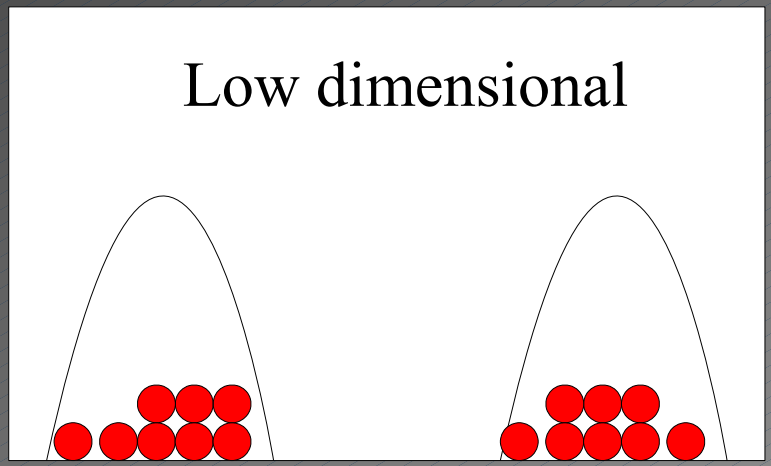
EIGENSTRAT use this normalisation to do PCA

Dimensionality reduction

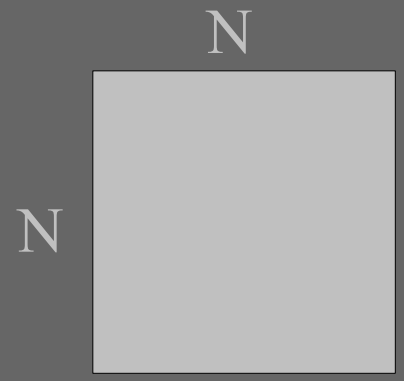
SVD of Y_{il}
= PCA of $\mathrm{cov}(Y_{il})$ (suitably normalised)
= MDS on Y_{il} (suitably normalised)
… all give UDV^T

Eigenvalues $D = \mathrm{Diag}(\lambda)$
Eigenvectors V
(Left Eigenvectors U do differ)





How many dimensions?



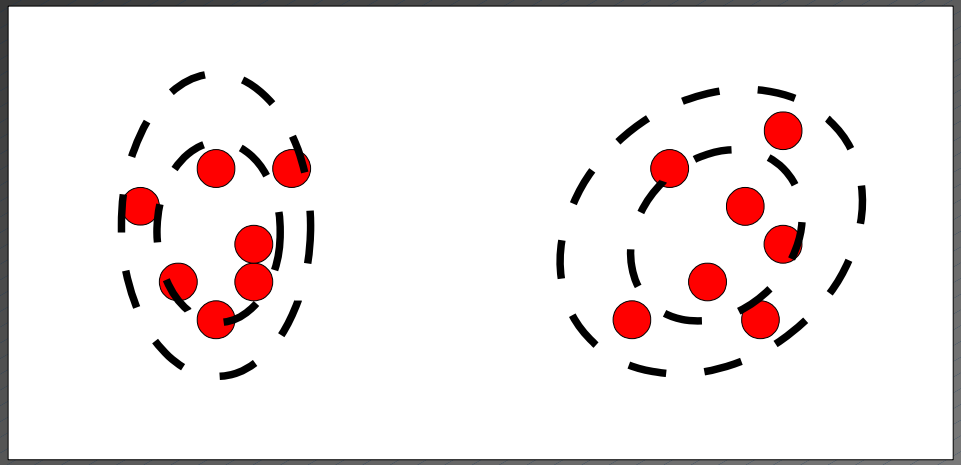
- MAP (Minimum Average Partial Correlation), Velicer 1976
  - Remove largest eigenvalue
  - Compute (partial) correlation between remaining eigenvectors and the data (accounting for previous eigenvectors)
  - Repeat

- Parallel Analysis (Horn 1965)
  - Simulate many matrices the same shape, with the same mean and variance as the data
  - Keep all components bigger than some quantile of the simulated values

- Tracy-Widom Distribution (TW 1994, Patterson et al 2006)
  - Theoretical distribution of the EVs of an *L* by *N* matrix (*L* is the (effective) number of SNPs)
  - Remove biggest EV if bigger than some quantile
  - Repeat

Clustering: MVN

- Multivariate Normal ("Soft K-Means", implemented by MCLUST in R)
- Infer mean and variance for each cluster
- BIC model selection for *K*
- *(Alternatives: K-means, UPGMA, etc)*



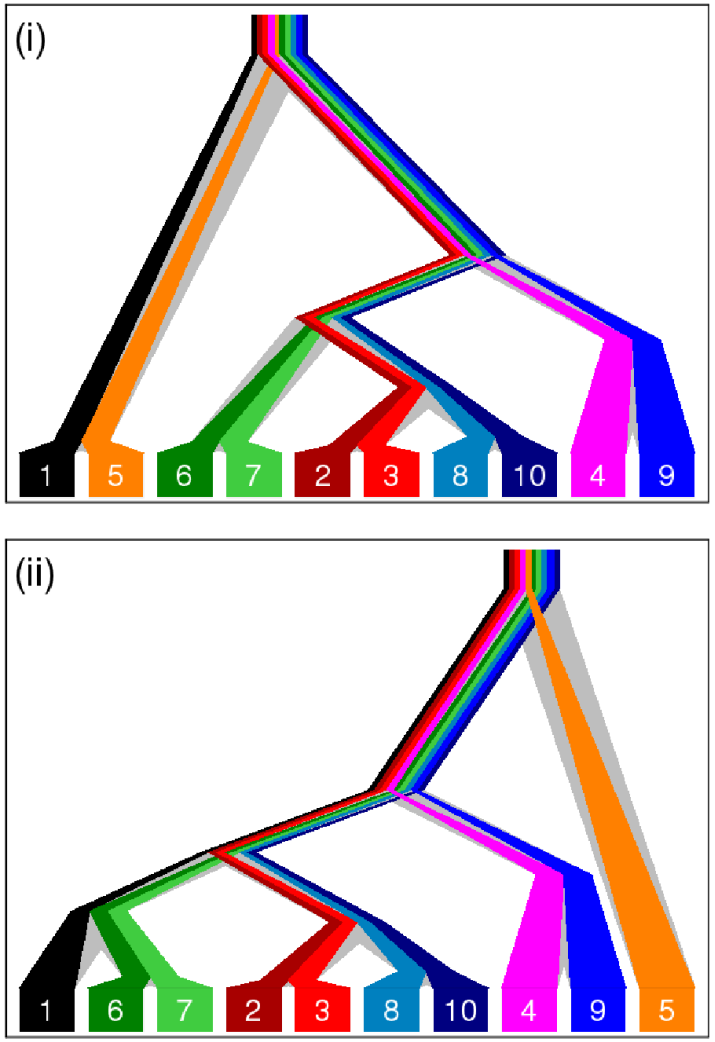
The coancestry matrix

- (unlinked) coancestry matrix:

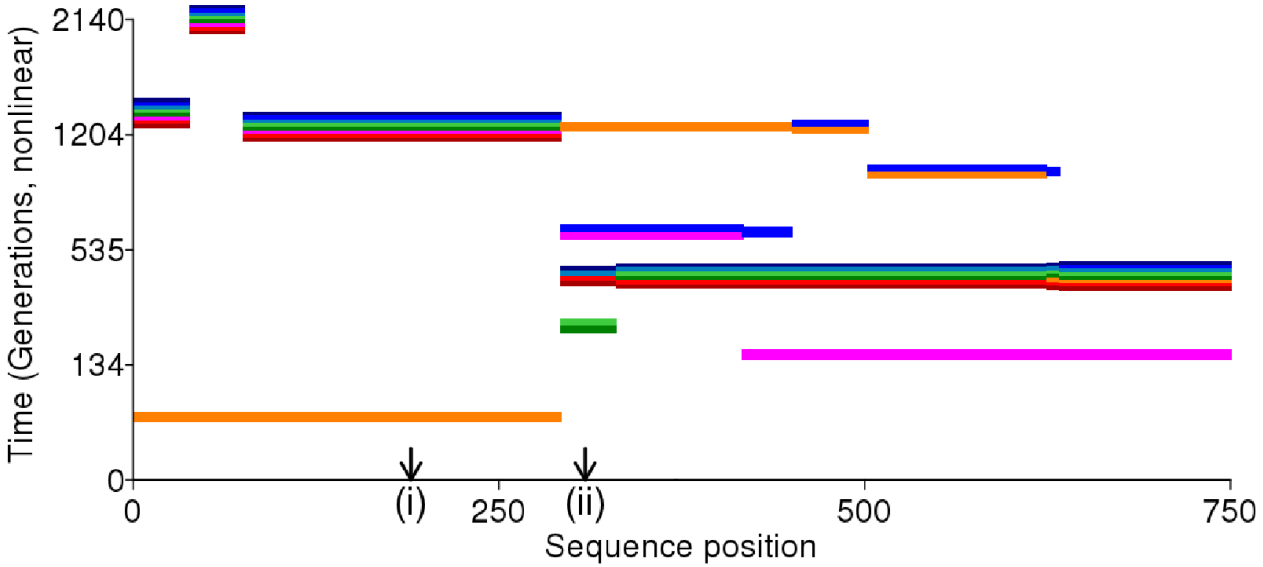
$$Y_{ij} = \sum_{l=1}^{L} \frac{x_{il}x_{jl}}{\sum_{j \neq i} x_{jl}} + \frac{(1-x_{il})(1-x_{jl})}{\sum_{j \neq i}(1-x_{jl})}$$

- To O(N), Coancestry matrix is a rotation of the (EIGENSTRAT PCA) eigenvector matrix
  - *If SNPs are uncorrelated*
  - *and the number of individuals is large*
- To O(N), Coancestry matrix is a *sufficient statistic* for the STRUCTURE likelihood
  - *If additionally, drift is small*

Local genealogies



Time to MRCA with haplotype 1



True 'nearest neighbour' distribution of haplotype 1

Mean painting of haplotype 1

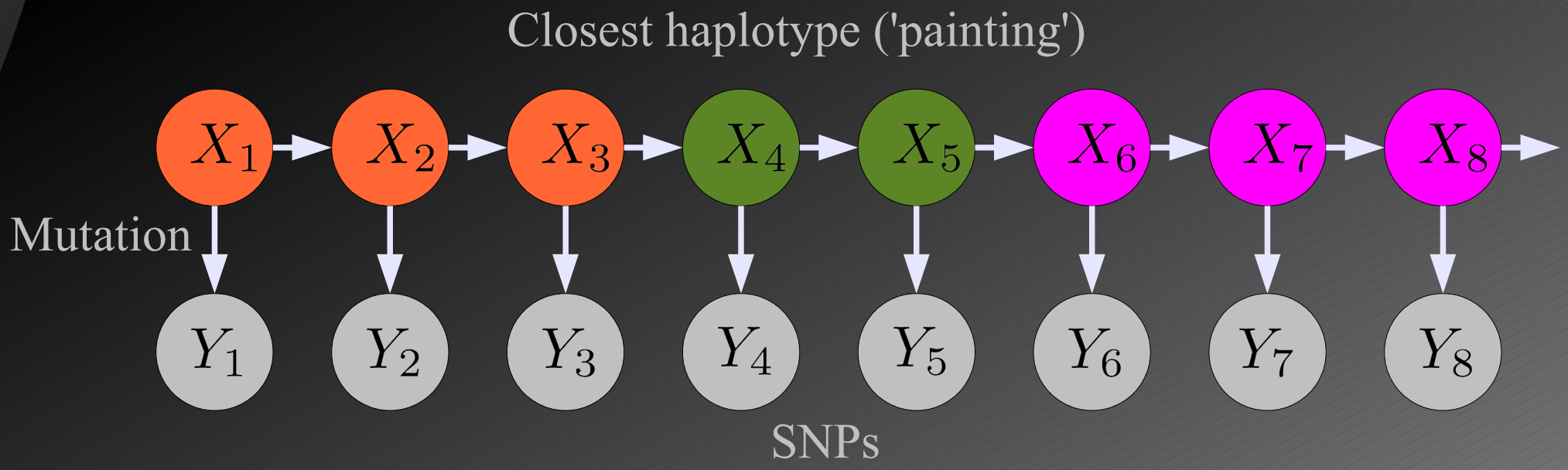
- ChromoPainter 'Coancestry' similarity matrix
- Unlinked limit: normalised allele sharing

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

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

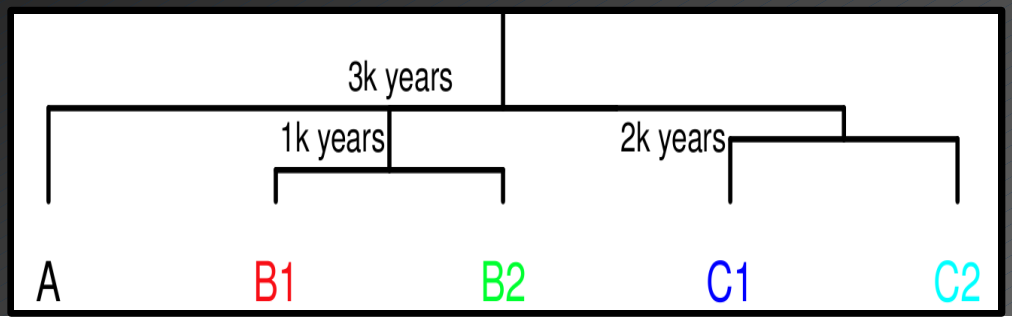
Li and Stephens Hidden Markov Model

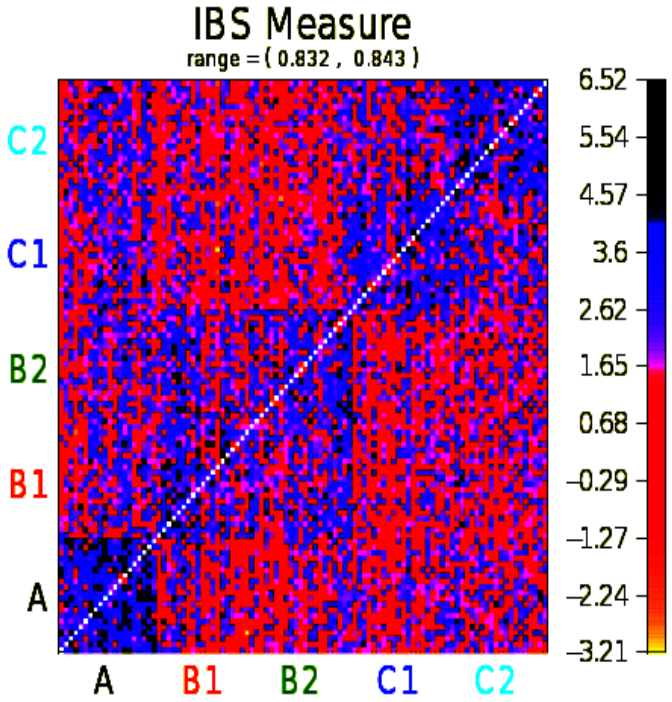


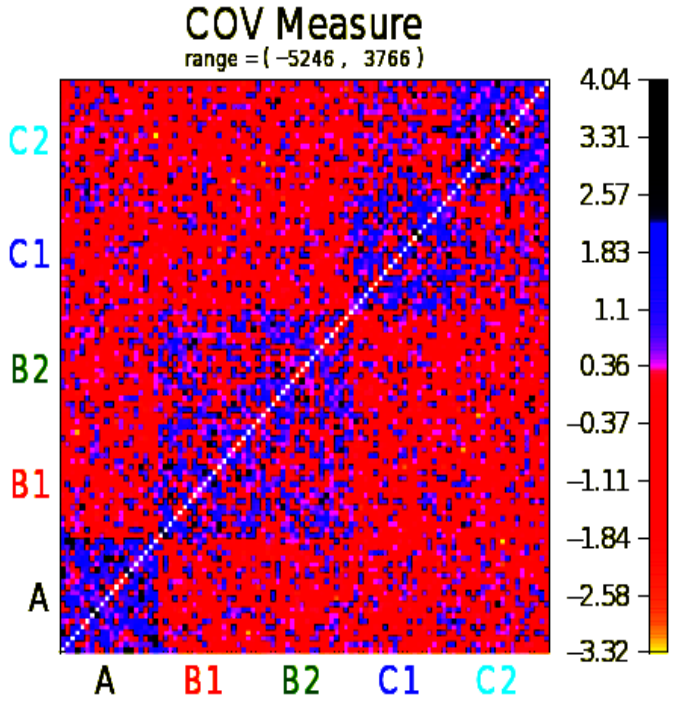
- Switches occur at constant rate ρg_l *(Scaled Genetic distance)*
- Mutations occur at constant rate θ
- Can efficiently infer $X_{1\ldots L}$ (painting)
- And $X_l | X_{l-1}$ (switches, the number of 'chunks')

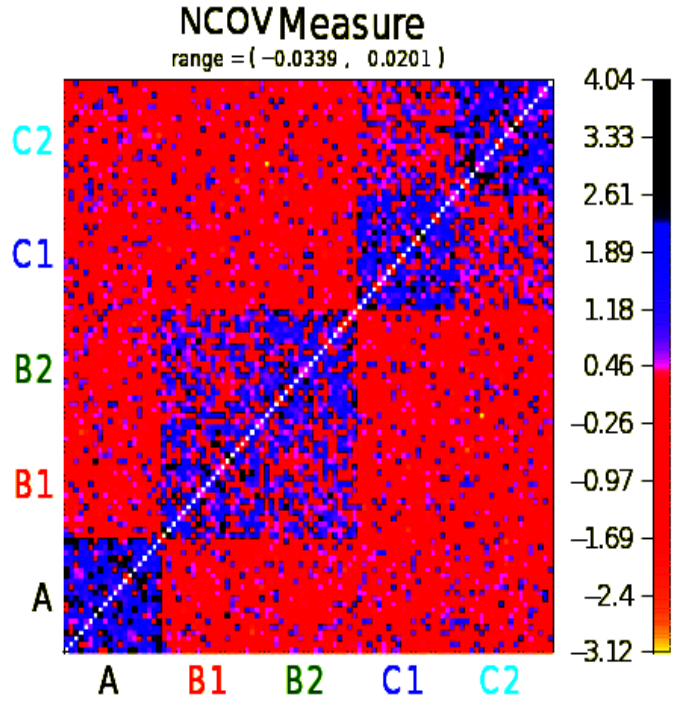
Similarity measures of simulated data

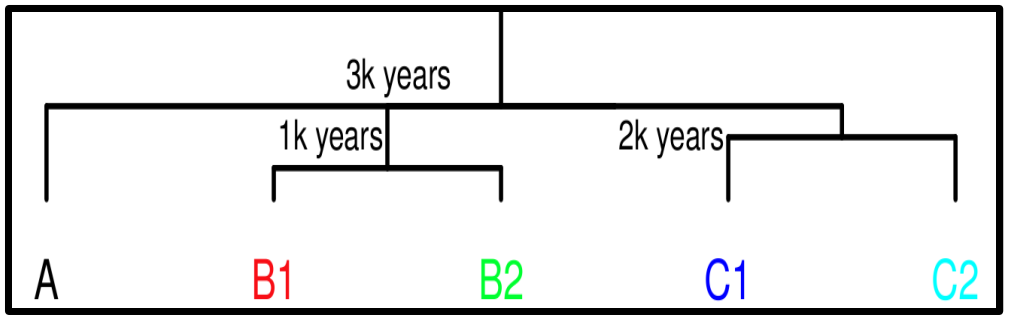
Simulation scenario: 'Europe'





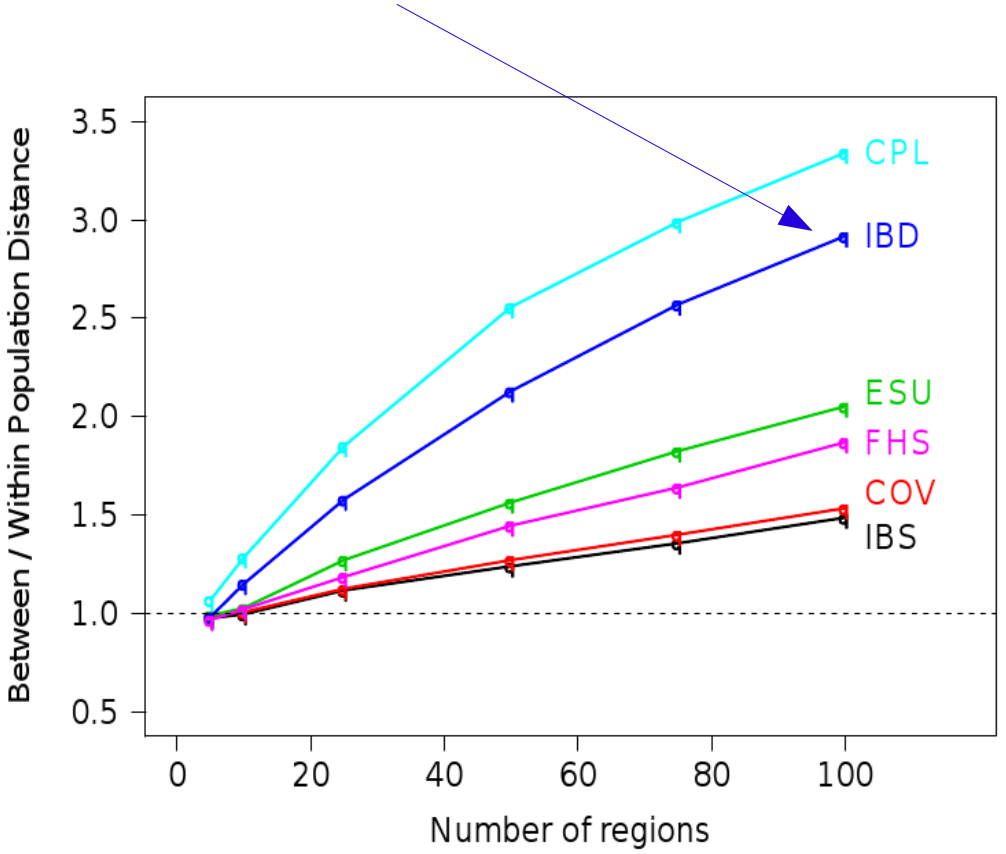


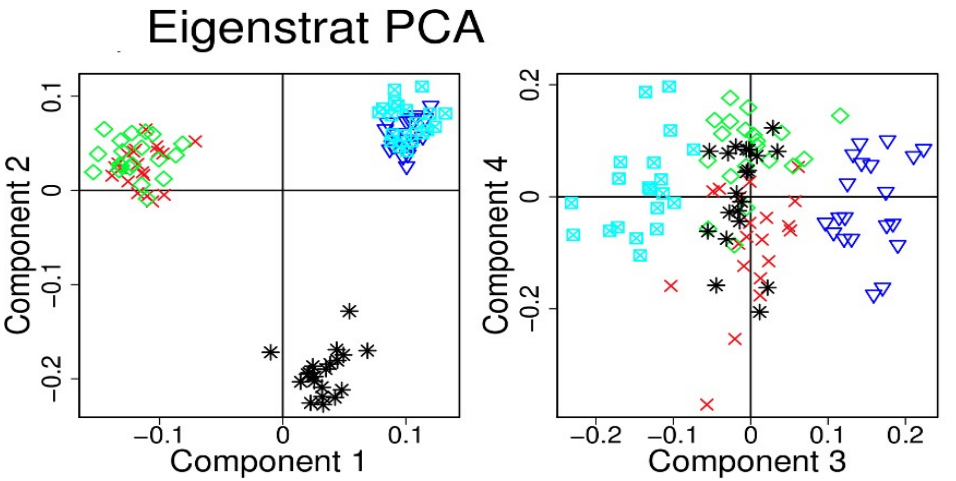


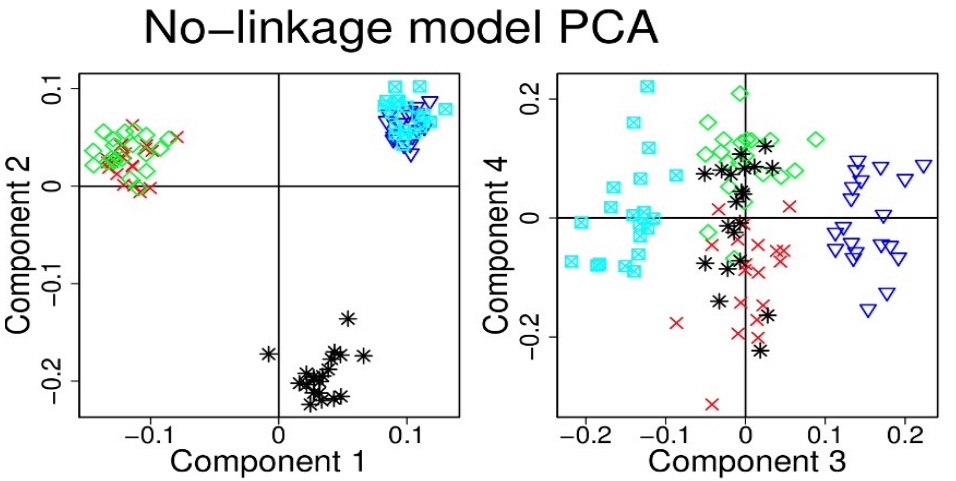


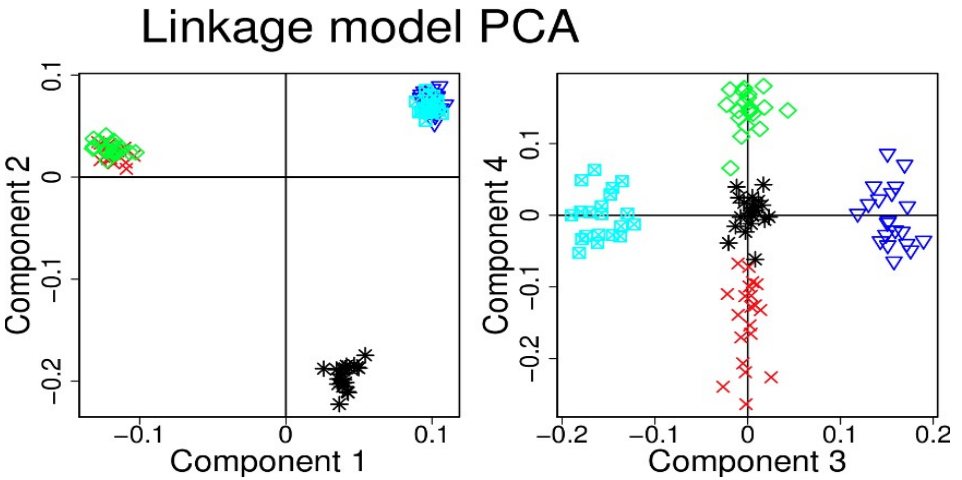
Simulation scenario: 'Europe'

FastIBD (Browning & Browning 2011)









FineSTRUCTURE
Population structure model

- Individuals exchangeable 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

Number of individuals to donate from

Population assignment

Probability of a partition

- Dirichlet Process prior for partition η :

$$\eta \sim \alpha^K \prod_{k=1}^{K} \Gamma(\hat{n}_k) \qquad \{P_1, \cdots, P_K\}|\eta = \prod_{b=1}^{K} G_0$$

- Rows of P_{ab} (i.e. G_0) are Dirichlet *(containing hidden biological parameters)*…
- … so conjugate, and we integrate out P_{ab}
- No individual/population level parameters

Additional results

- To O(N), FineSTRUCTURE likelihood is equivalent to the STRUCTURE* likelihood
  - *if SNPs are uncorrelated,*
  - *drift is weak,*
  - *genotyped SNPs are not very rare*
- Empirically, with linkage model we do better.

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

More details

- MCMC based Bayesian inference
- Post-processing step for a tree relating populations
  - Identify & merge close relatives first
- Infer *K*: Structure bar plots don't make sense
- New ways to visualise results needed...

HGDP dataset

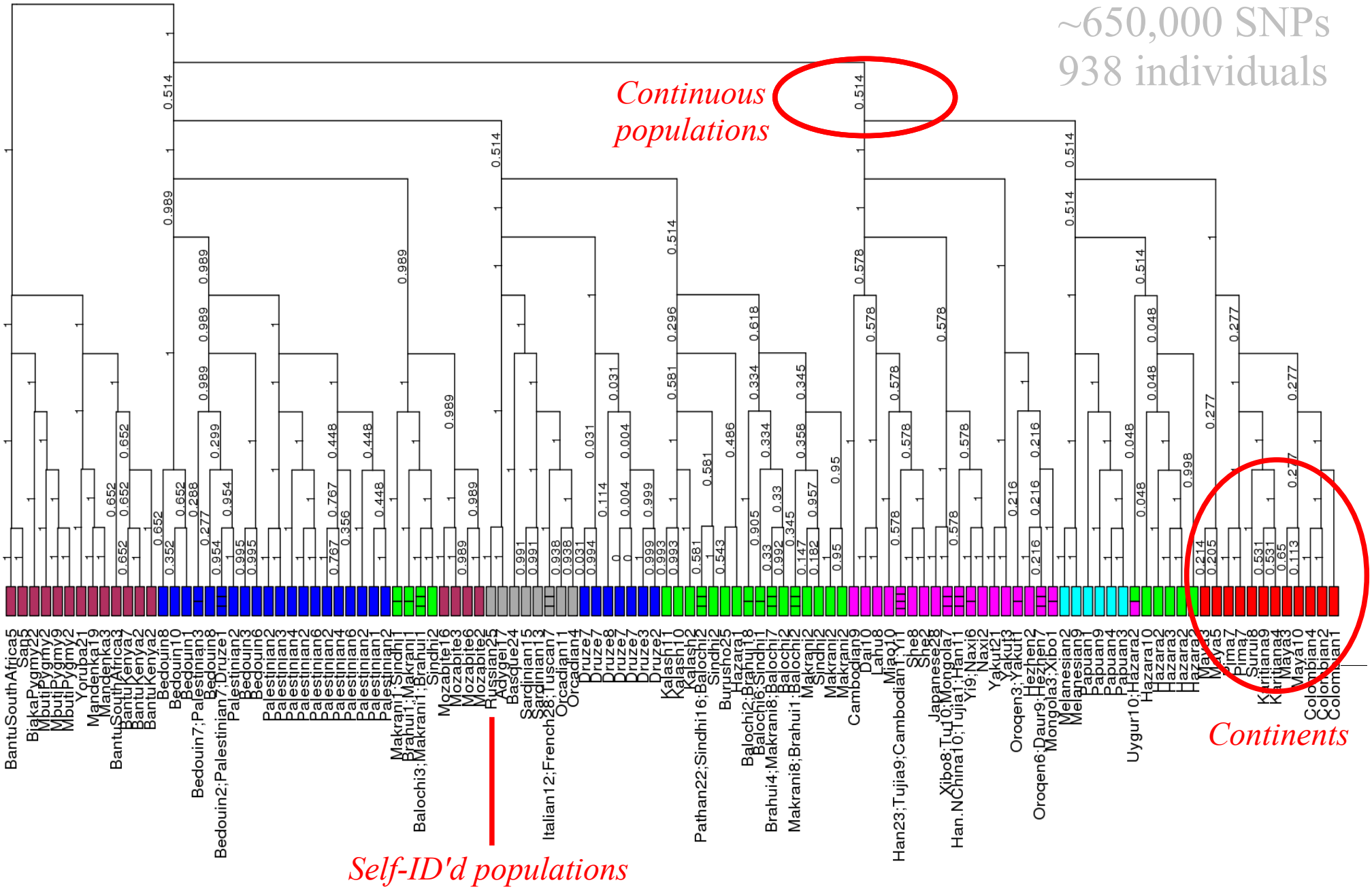
650K SNPs on 938 individuals from 53 pops (5-45 inds/pop)

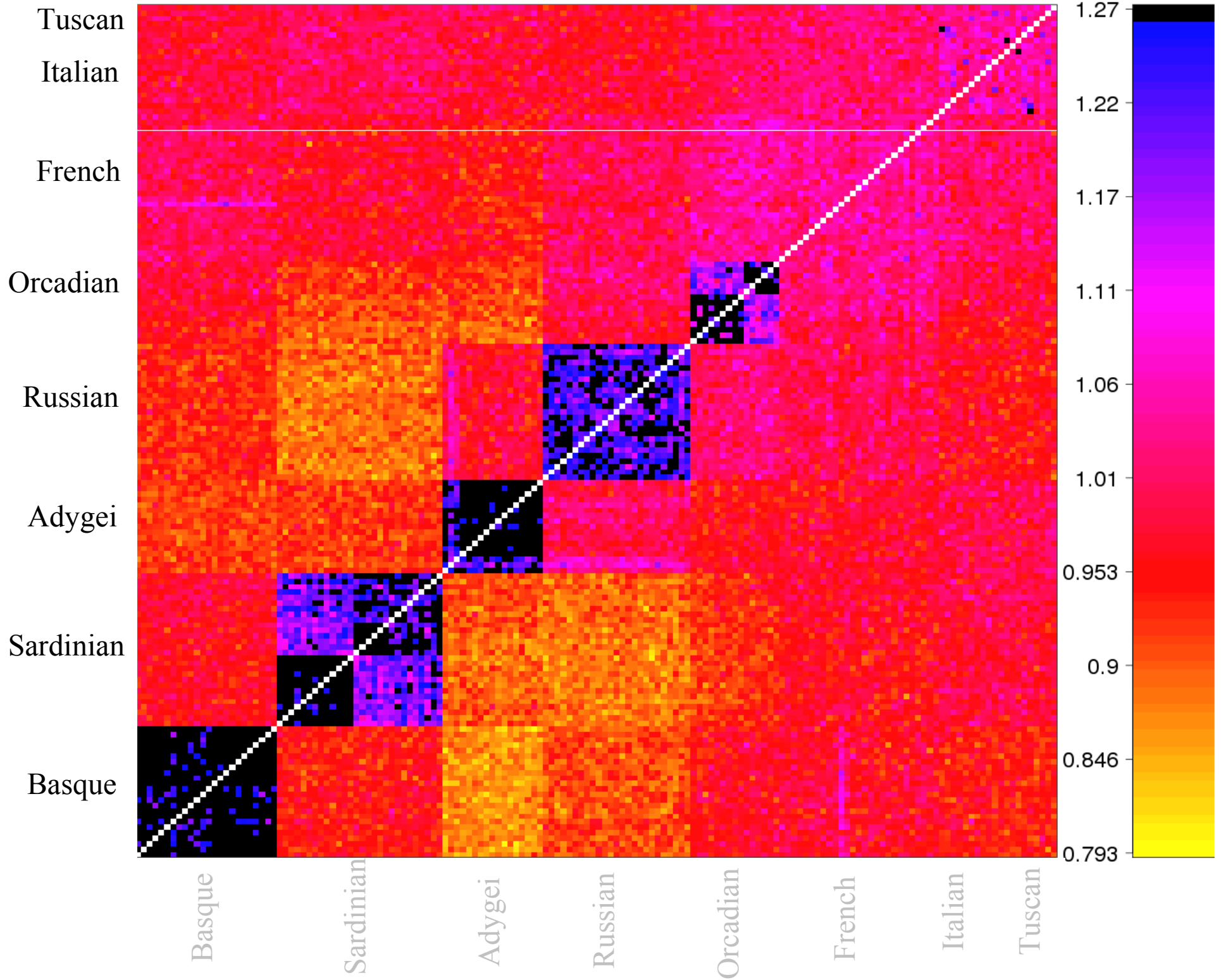
website: *http://hgdp.uchicago.edu/*

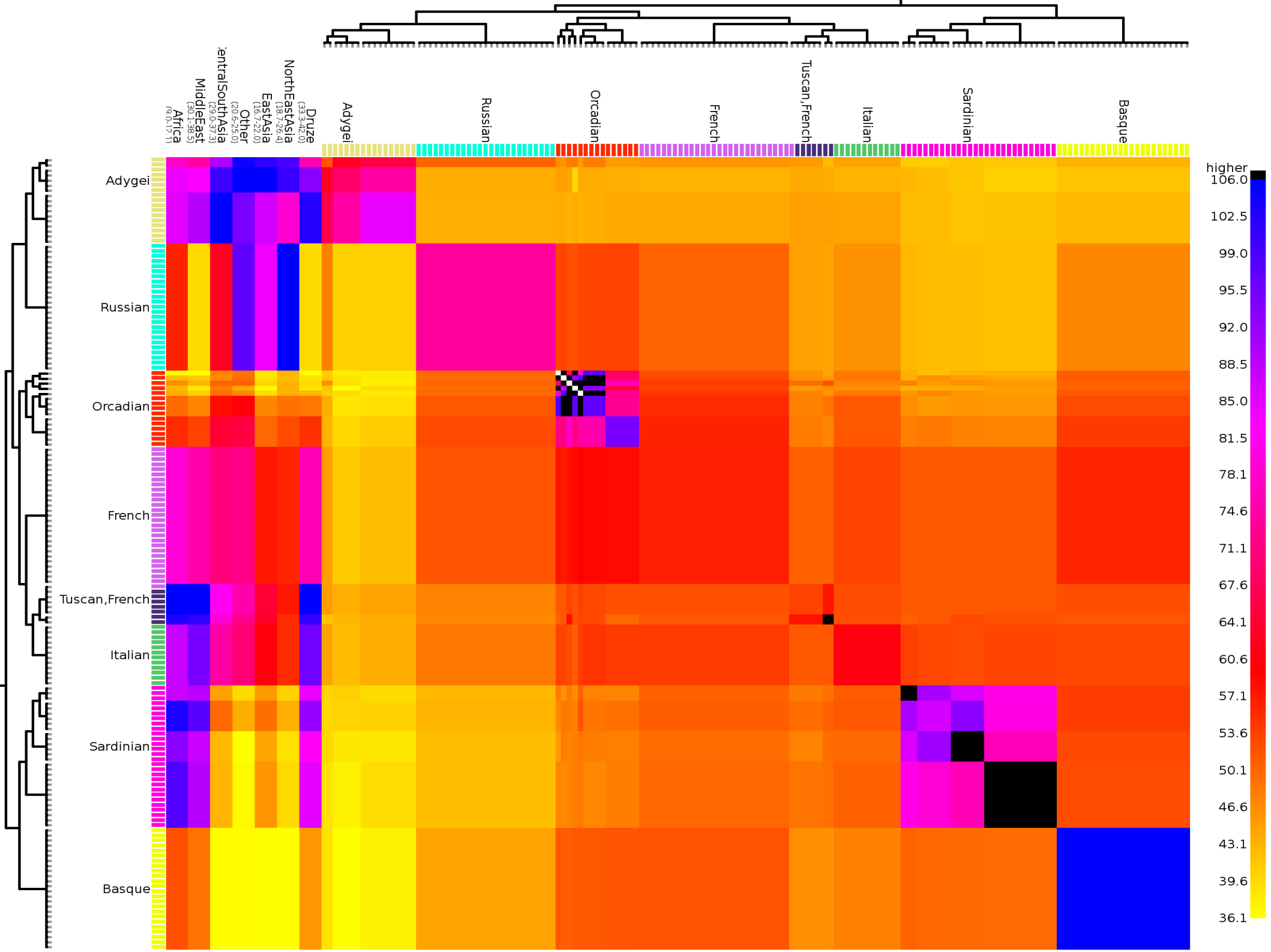
(picture from Cvalli-Sforza (2005) *Nat Rev Genet*)

MAP tree: whole world HGDP data

~650,000 SNPs
938 individuals




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

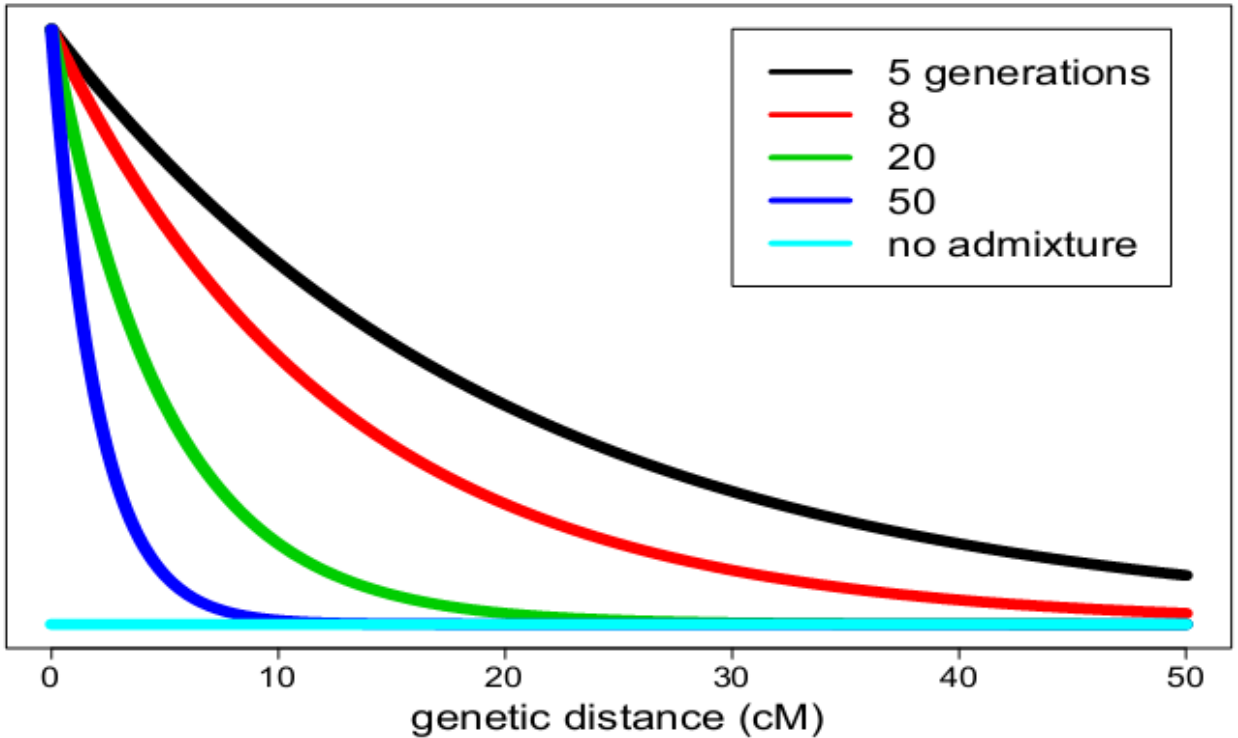

Africa (6.0-17.1)
MiddleEast (30.1-38.5)
CentralSouthAsia (29.0-37.3)
Other (20.6-25.0)
EastAsia (16.7-22.0)
NorthEastAsia (18.7-26.4)
Druze (33.3-42.0)
Adygei
Russian
Orcadian
French
Tuscan,French
Italian
Sardinian
Basque
higher
106.0
102.5
99.0
95.5
92.0
88.5
85.0
81.5
78.1
74.6
71.1
67.6
64.1
60.6
57.1
53.6
50.1
46.6
43.1
39.6
36.1

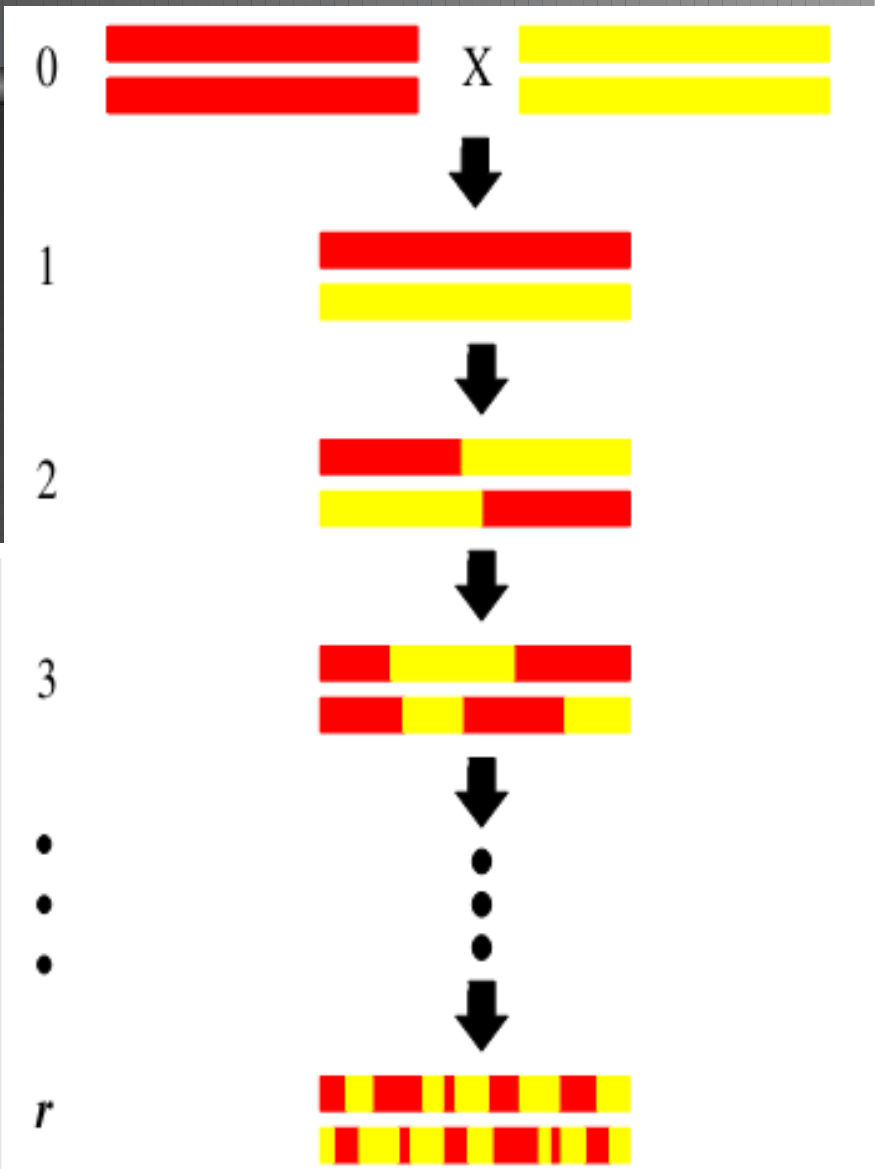
Other uses

- Other quantities:
  - Matrix L_{ij} of the length of haplotype chunks
  - Matrix M_{ij} of mutations on haplotype chunks
  - $P(X_{l+d} = b | X_l = a)$ Correlation between 'origin' of SNPs as a function of genetic distance
- Applications
  - Population inference
  - Identifying groupable individuals (e.g for GWAS, demographic inference, other modelling)
  - … Haplotype-based association studies?
  - Admixture dating

Admixture dating

- Expected length of haplotypes between populations halves each generation
- Exponential tract length distribution

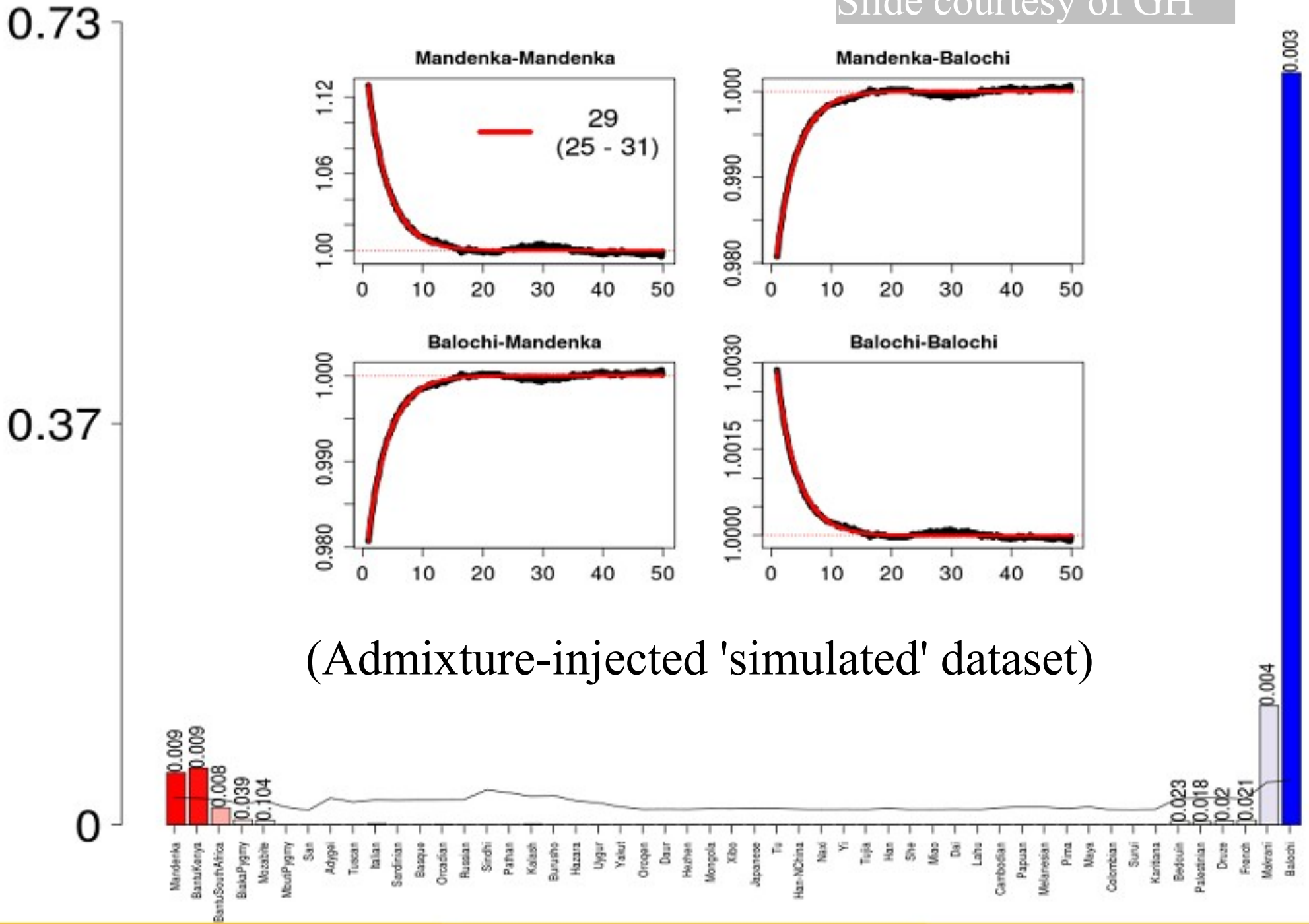




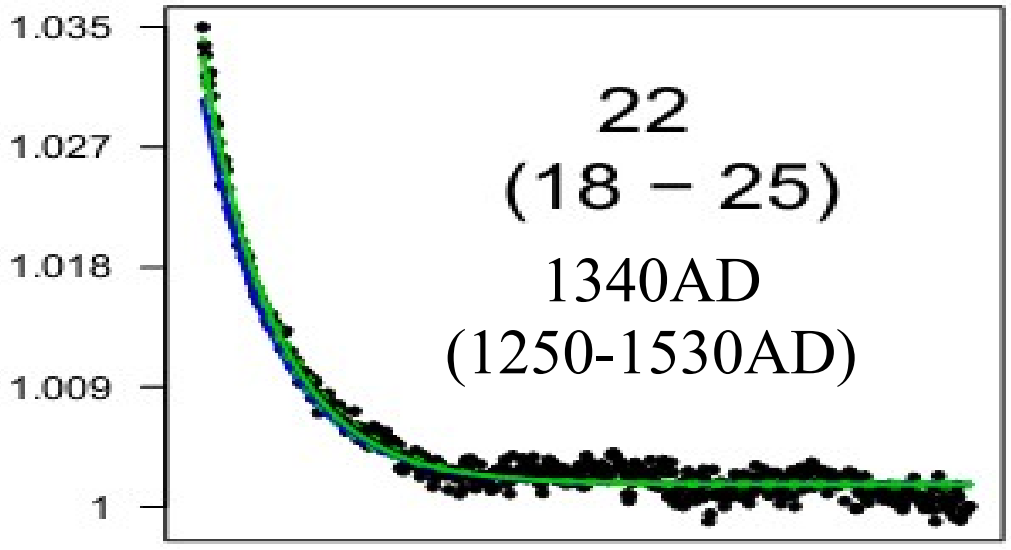
Slide courtesy of GH

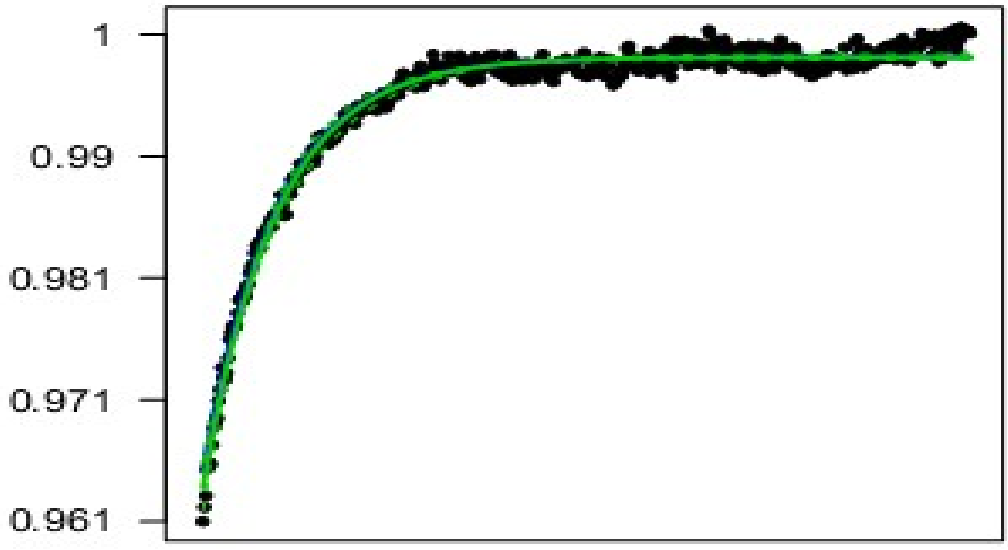
Brahui-Yoruba 95/5% admixture (30 gens)

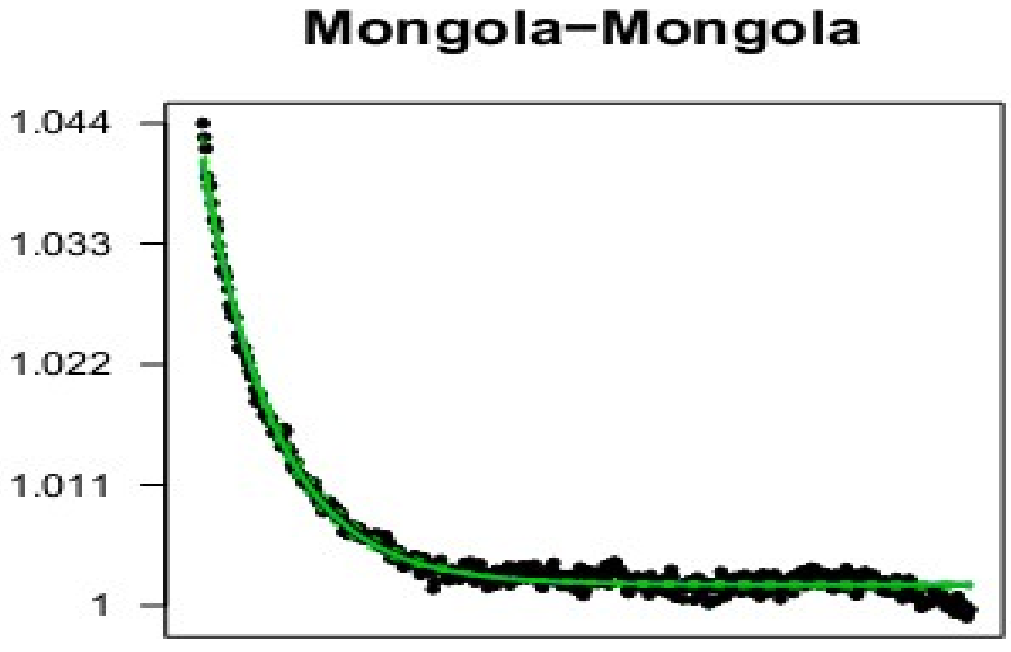
Slide courtesy of GH

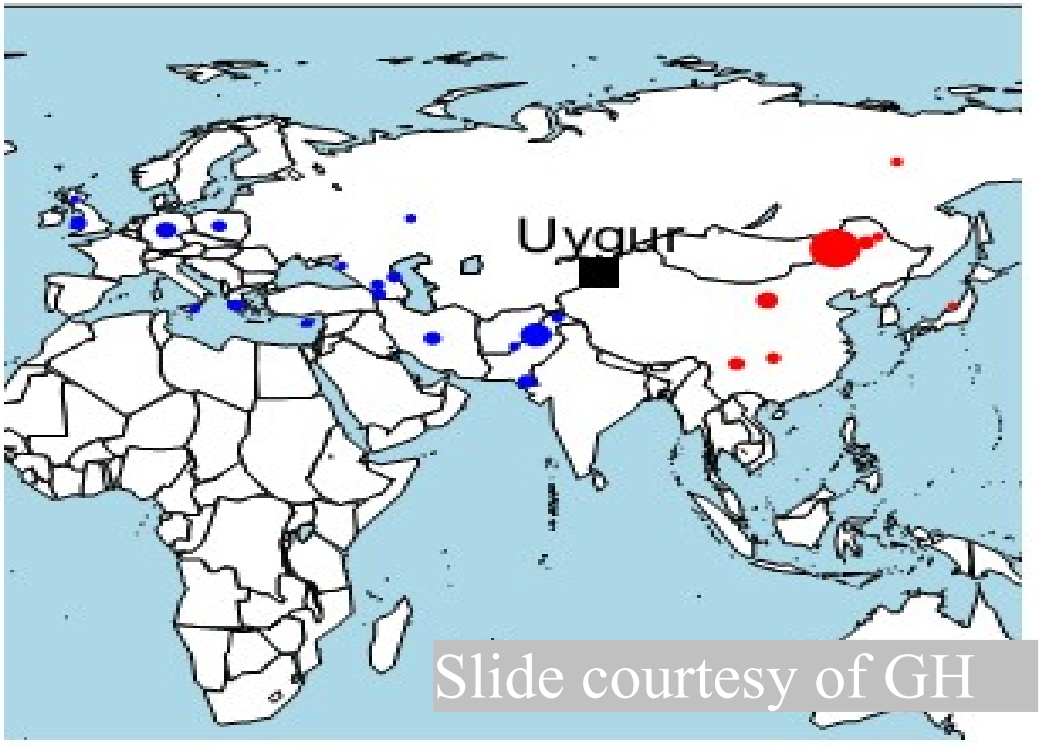


(Admixture-injected 'simulated' dataset)

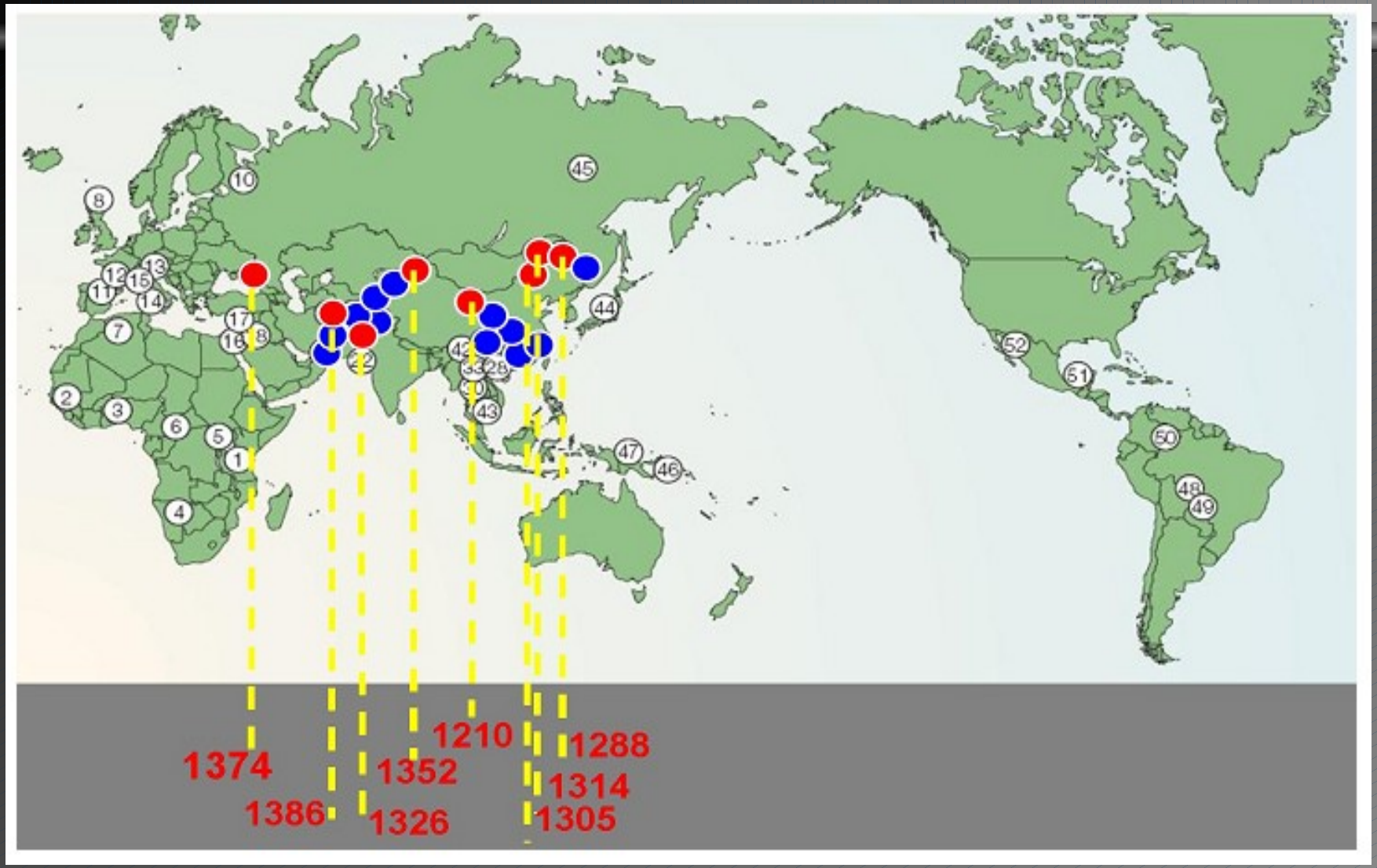

Pathan–Pathan
1.035
1.027
1.018
1.009
1
22
(18 – 25)
1340AD
(1250-1530AD)


Pathan–Mongola
1
0.99
0.981
0.971
0.961


Mongola–Mongola
1.044
1.033
1.022
1.011
1


Uygur
Slide courtesy of GH

Former Mongolian Empire



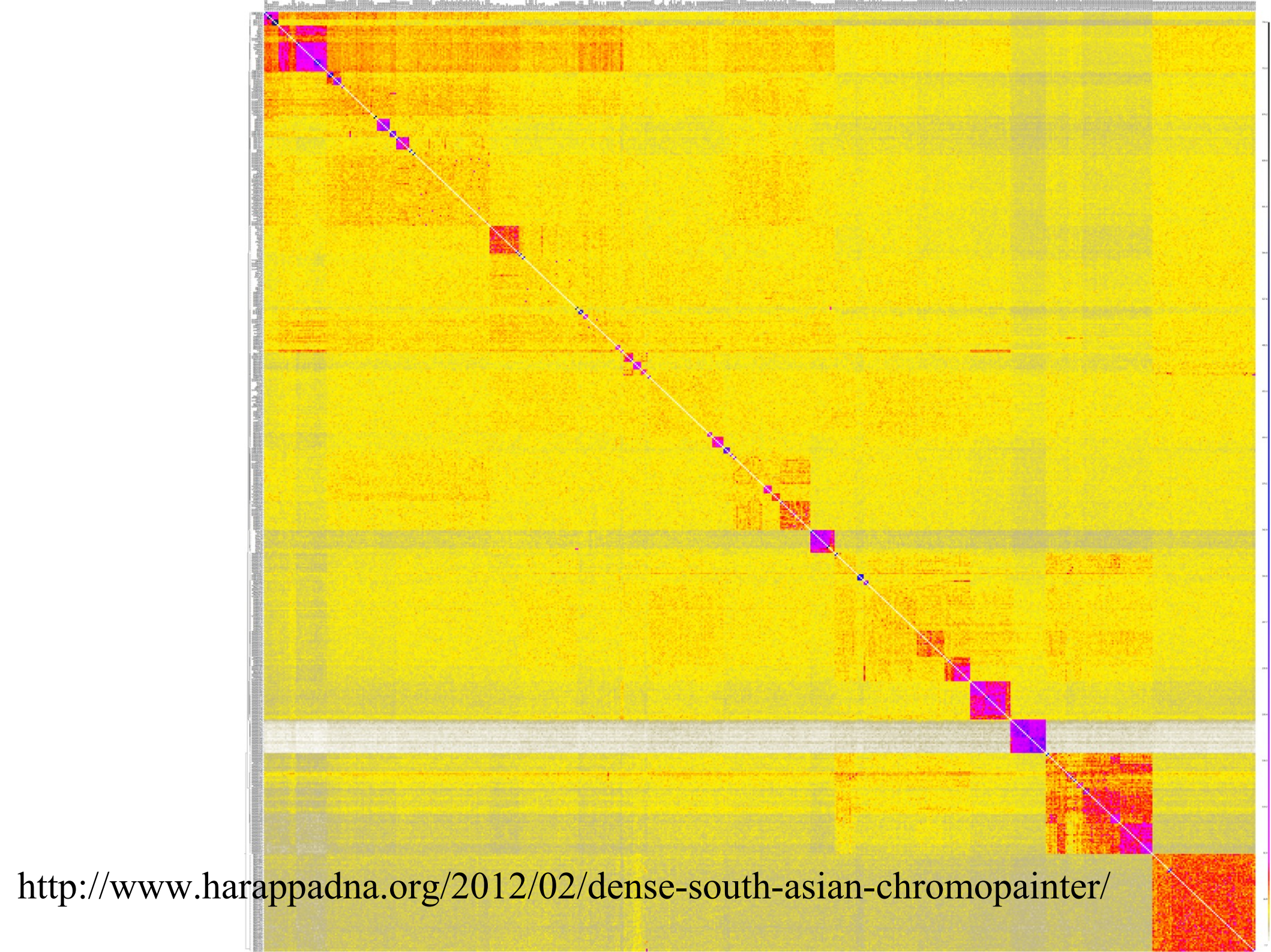
Daniel Lawson, University of Bristol
dan.lawson@bristol.ac.uk

Slide courtesy of GH

Personal genomics (e.g. 23andMe)

(Sharing approx 250K SNPs)

- http://fennoscandia.blogspot.co.uk (~200 Scandanavian individuals)
- http://www.harappadna.org (~700 South Asian individuals)
- http://dodecad.blogspot.co.uk (~400 Balkans/West Asia individuals)
- http://eurogenes.blogspot.co.uk (~500 European individuals)
- http://magnusducatus.blogspot.co.uk (~100 Lithuanian individuals)



http://www.harappadna.org/2012/02/dense-south-asian-chromopainter/

Summary statistics (requiem)

- FastIBD (Browning & Browning 2011) is another useful linked summary statistic
  - Different features to ChromoPainter
- If you know the correct number of PCA components to retain, you can do very well with PCA and simple models
  - But it goes very wrong in some (real) circumstances
- New linked approaches will be developed
- ChromoPainter/FineSTRUCTURE pipeline is a lot more robust than model-free alternatives

Further info

- ChromoPainter/FineSTRUCTURE software/code/info:
  www.paintmychromosomes.com
- Includes GUI: Windows/Linux/Mac (cmd line only)
- ChromoPainter/FineSTRUCTURE publication: *Lawson et al 2012 PLoS Genetics*
- Summary statistics review: *Lawson & Falush 2012 An. Rev. Hum. Genet. (to appear)*
- Admixture dating paper in preparation
- POBI paper in preparation
- 'Admixture' model in the works

Acknowledgements:

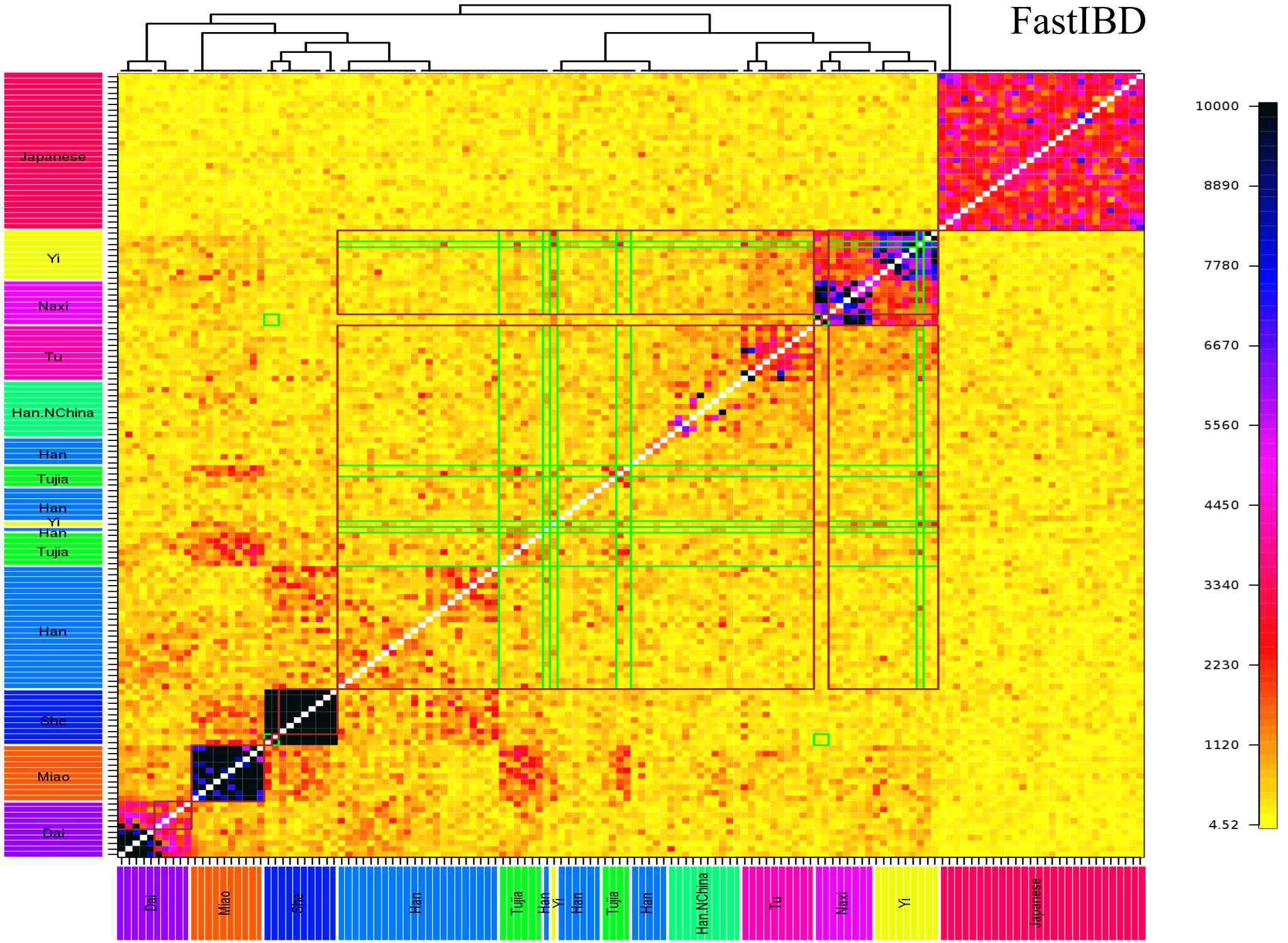
fineSTRUCTURE

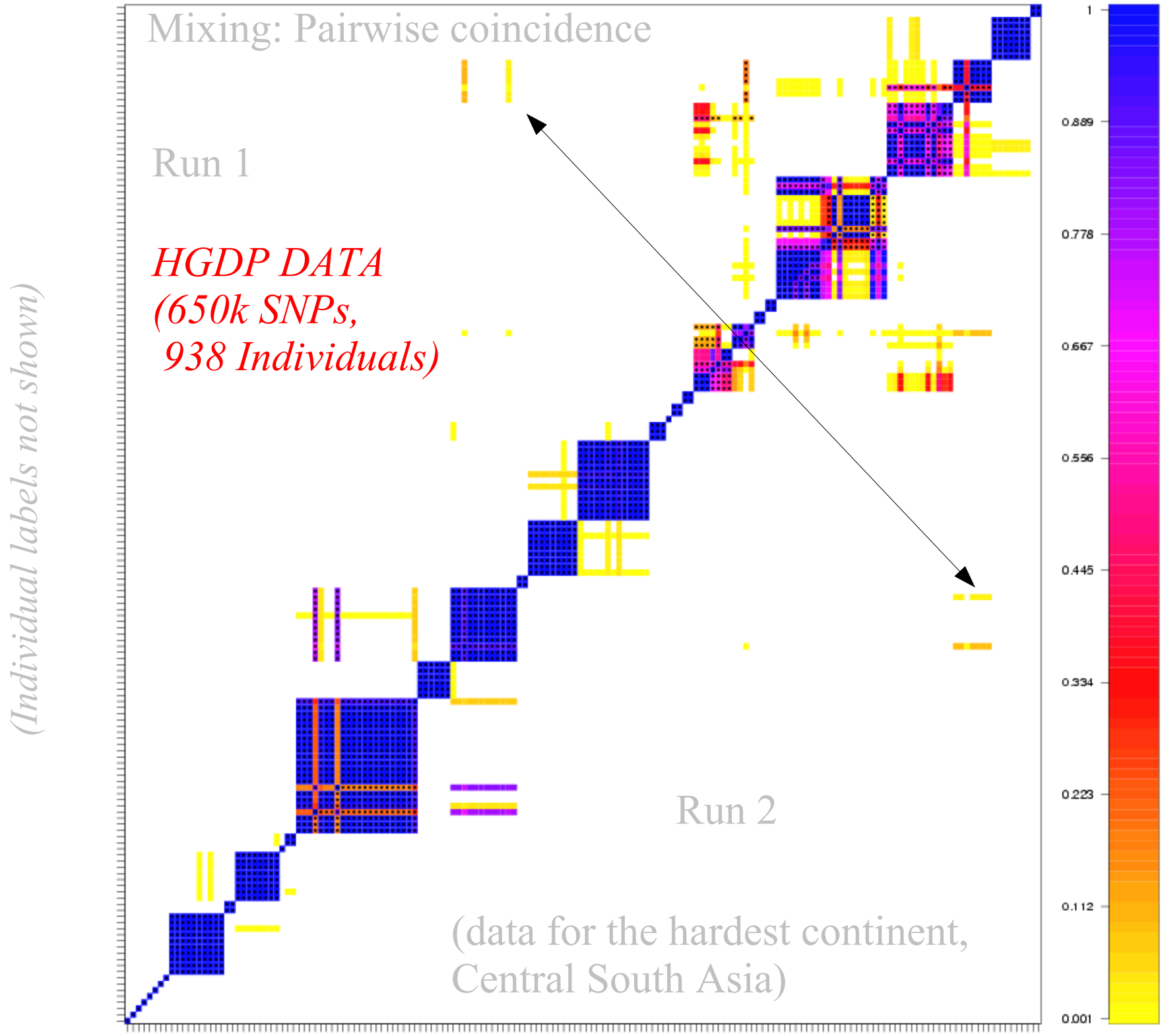
Garrett Hellenthal
(Oxford)
(CP algorithm, admixture dating)

Simon Myers
(Oxford)
(theory, admixture dating)

Daniel Falush
(Max Planck Institute)
(CP/FS concept)

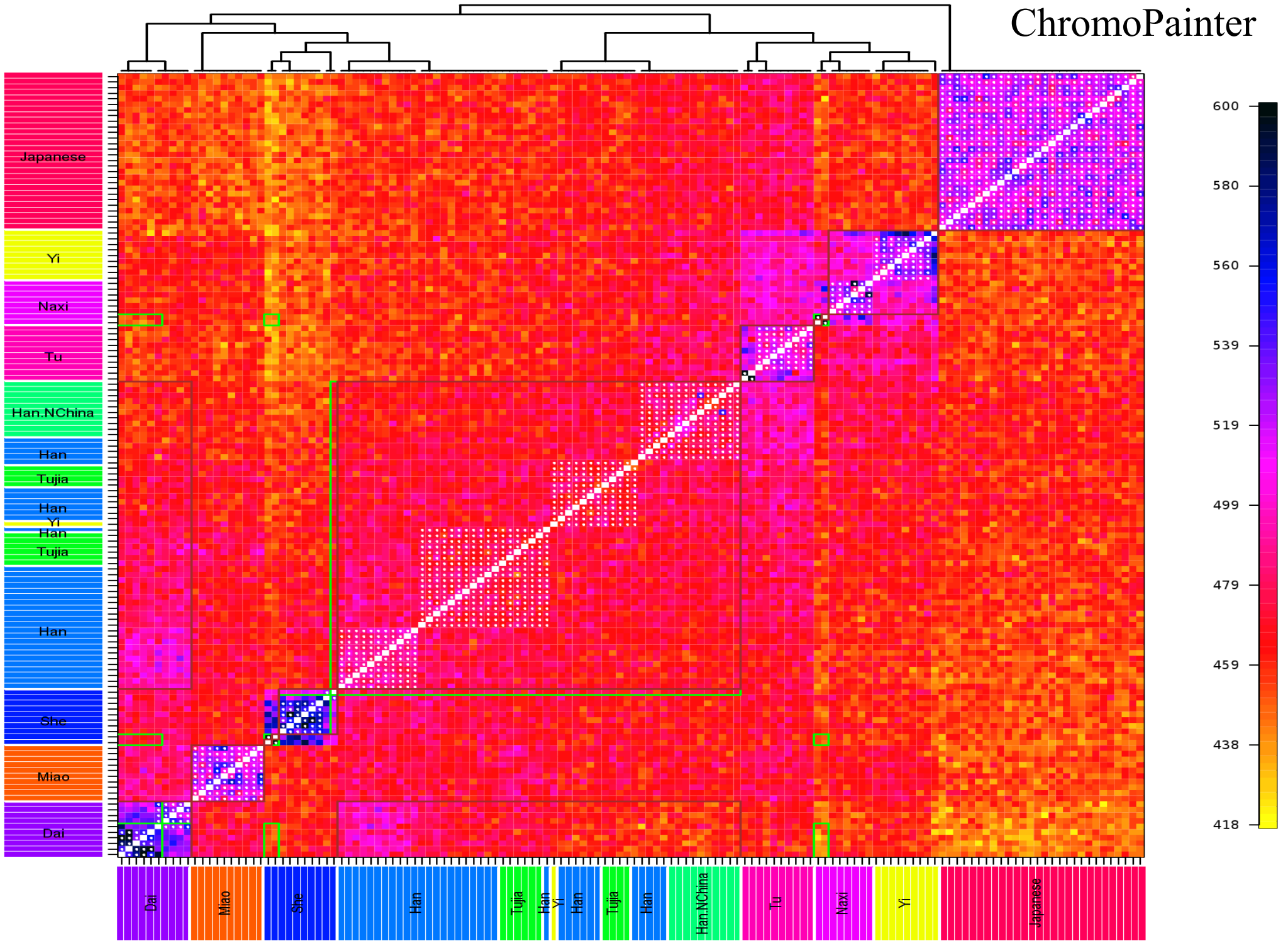
- Peter Green (Bristol) – Grant, support
- Bluecrystal HPC facilities @ Bristol
- www.paintmychromosomes.com


FastIBD
Japanese
Yi
Naxi
Tu
Han.NChina
Han
Tujia
Han
Yi
Han
Tujia
Han
She
Miao
Dai
Dai
Miao
She
Han
Tujia
Han
Yi
Han
Tujia
Han
Han.NChina
Tu
Naxi
Yi
Japanese
10000
8890
7780
6670
5560
4450
3340
2230
1120
4.52

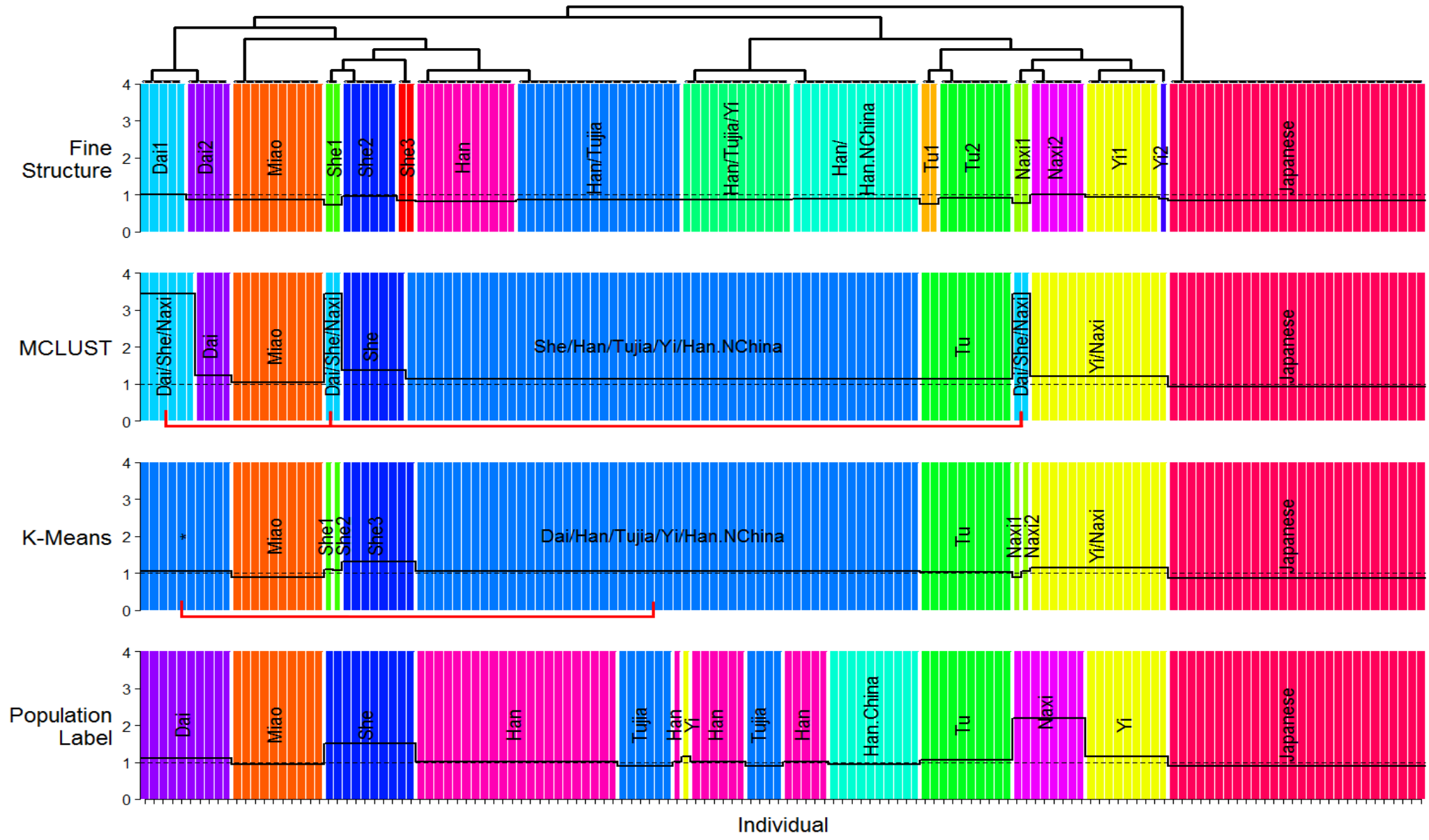

Mixing: Pairwise coincidence
Run 1
HGDP DATA
(650k SNPs,
938 Individuals)
Run 2
(data for the hardest continent,
Central South Asia)
(Individual labels not shown)
1
0.889
0.778
0.667
0.556
0.445
0.334
0.223
0.112
0.001

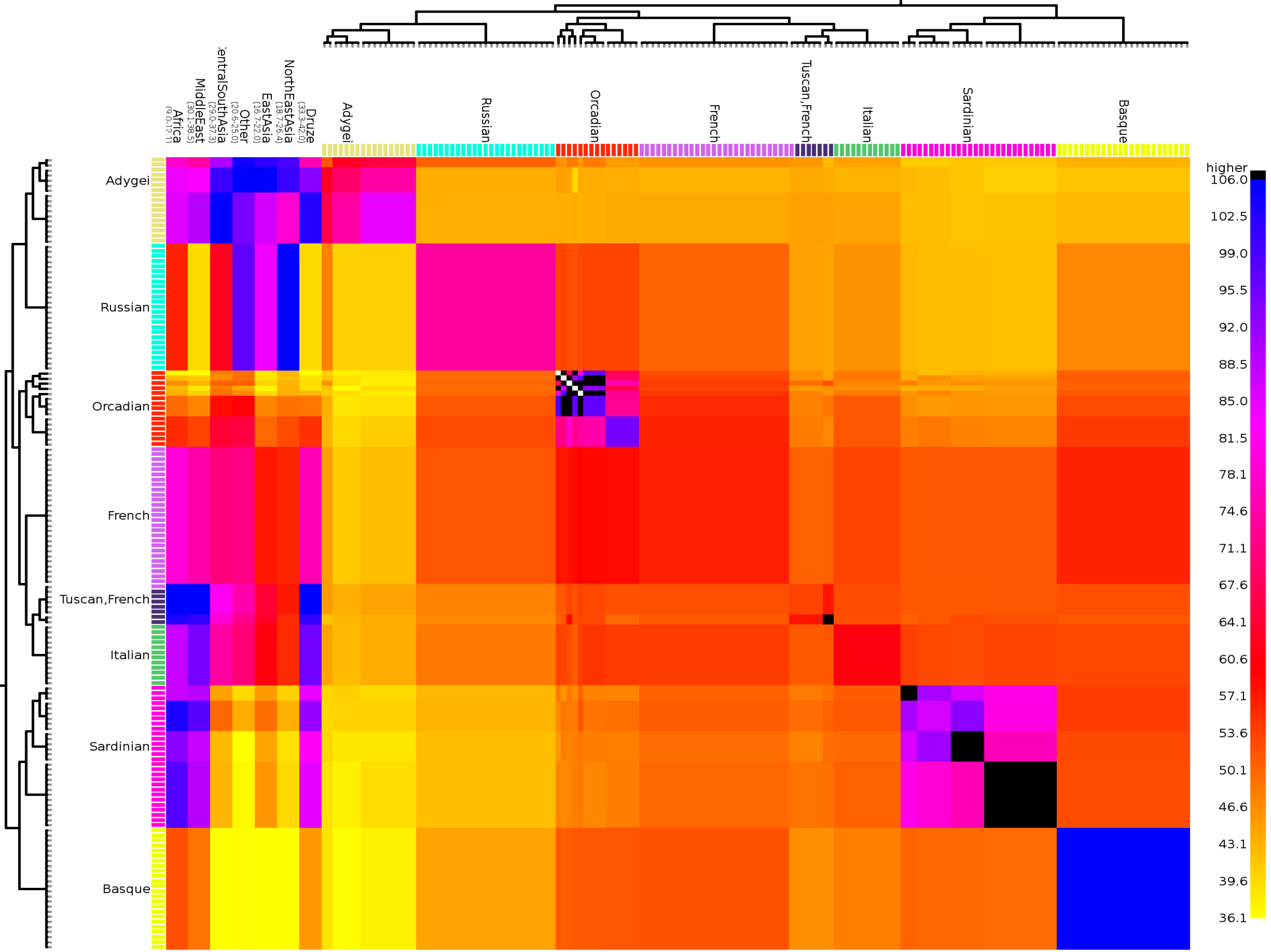
FastIBD (Browning and Browning 2011)

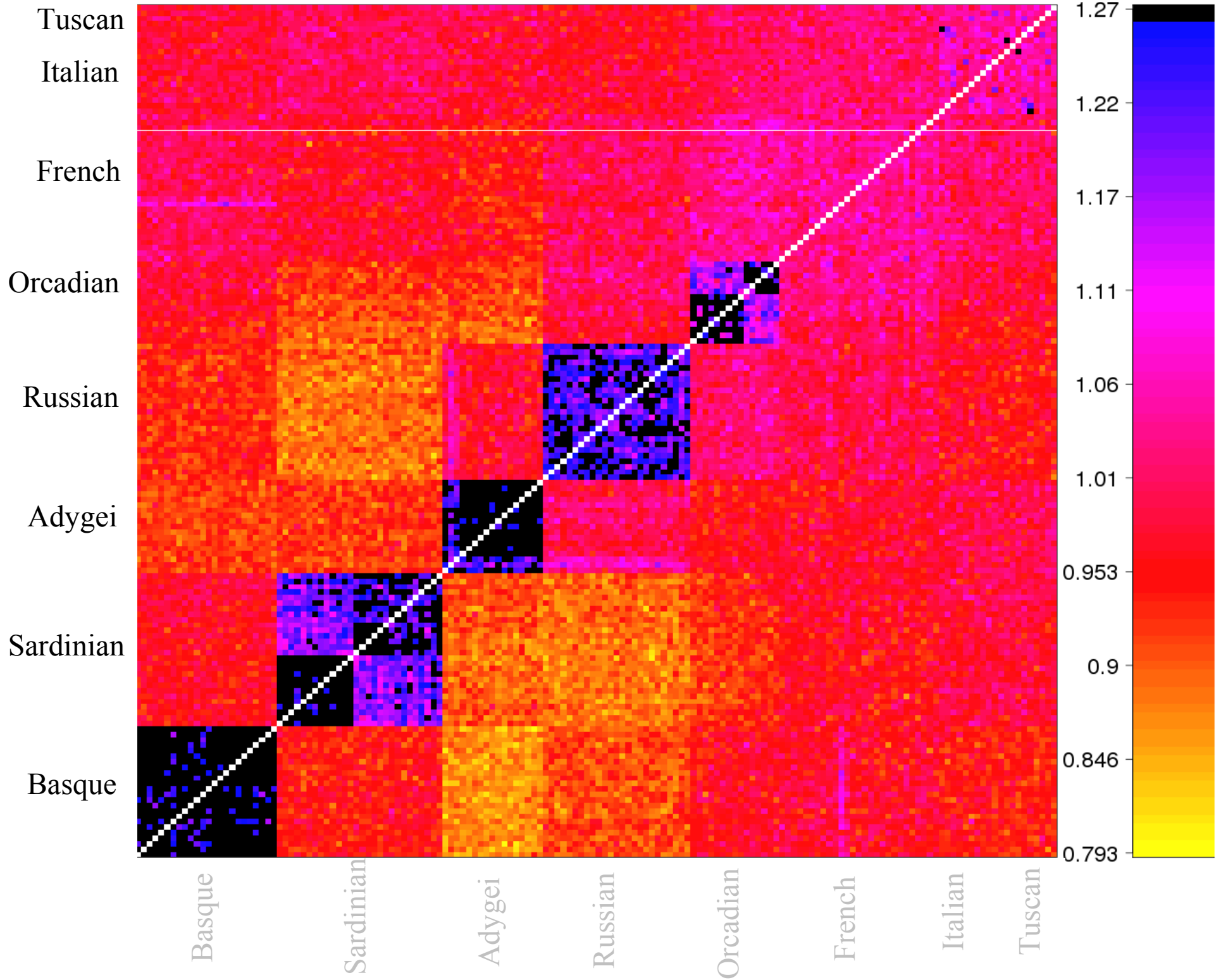
- Alternative linked model: Identify *r* closest segments of DNA for each pair of individuals
- Genetic lengths of each are related to time since common ancestor
- Similarity measure: sum of the genetic lengths found for each pair
- Somewhat heuristic, has some tuning parameters, but empirically works well


ChromoPainter
Japanese
Yi
Naxi
Tu
Han.NChina
Han
Tujia
Han
Yi
Han
Tujia
Han
She
Miao
Dai
600
580
560
539
519
499
479
459
438
418
Dai
Miao
She
Han
Tujia
Han
Yi
Han
Tujia
Han
Han.NChina
Tu
Naxi
Yi
Japanese

Comparison of clusterings

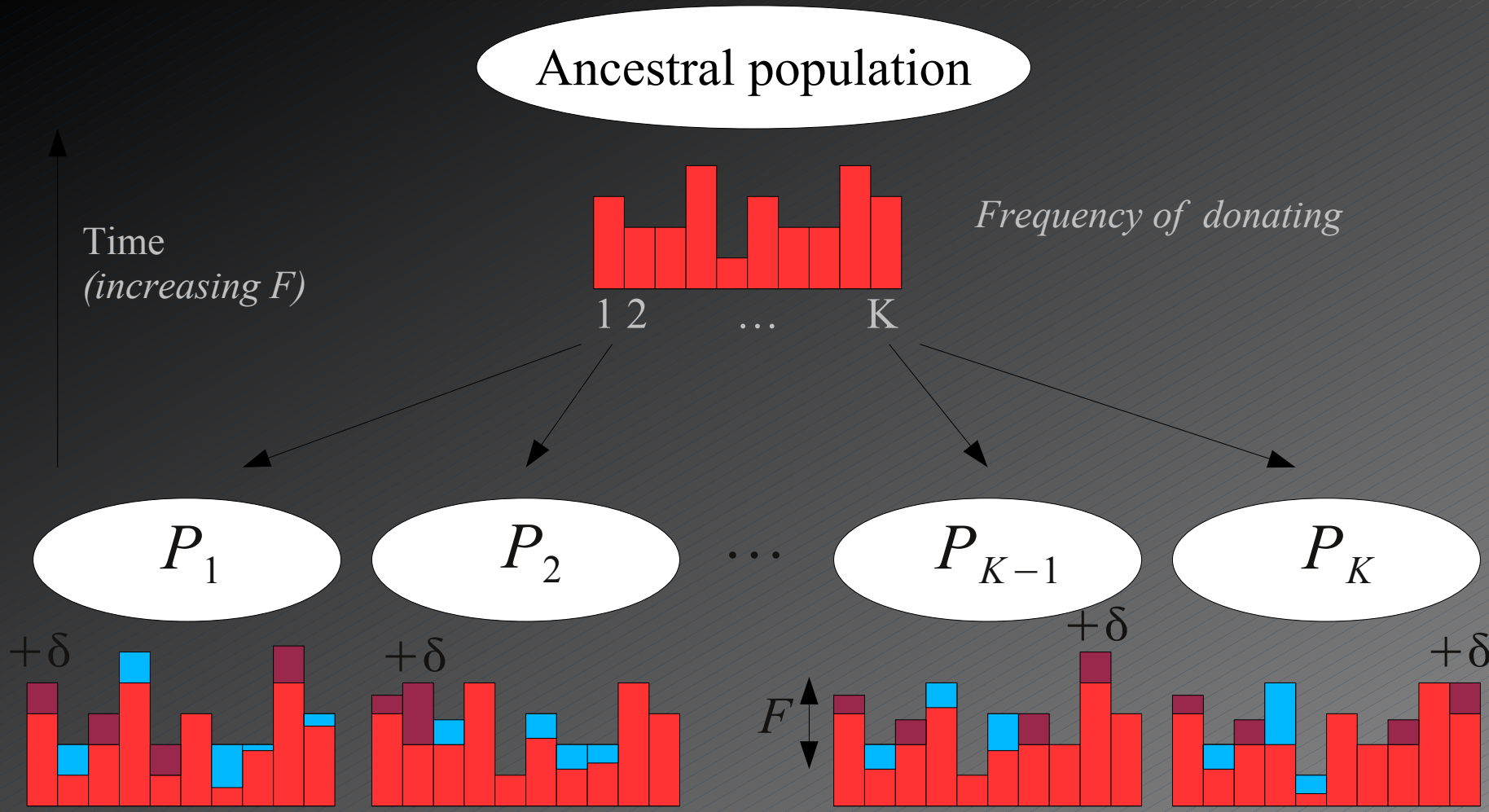



Africa (6.0-17.1)
MiddleEast (30.1-38.5)
CentralSouthAsia (29.0-37.3)
Other (20.6-25.0)
EastAsia (16.7-22.0)
NorthEastAsia (18.7-26.4)
Druze (33.3-42.0)
Adygei
Russian
Orcadian
French
Tuscan,French
Italian
Sardinian
Basque
higher
106.0
102.5
99.0
95.5
92.0
88.5
85.0
81.5
78.1
74.6
71.1
67.6
64.1
60.6
57.1
53.6
50.1
46.6
43.1
39.6
36.1


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

Weak Biological Model for prior

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



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

Clustering into k clusters

- Find "similar" individuals
- Three main approaches:
  - Cluster on raw data Y_{il}
  - Cluster on similarity matrix X_{ij}
  - Cluster on dimensionality reduced version of data, e.g. MDS/PCA/SVD V_{id}

$$l \in [1, L]$$

$$i, j \in [1, N]$$

$$d \in [1, D]$$

Recall: $L \gg N \gg D$

$O(k) = O(d)$?

- Lowest dimension description usually best...
- Raw data approach terrible here (without good model)

The future – Admixture model

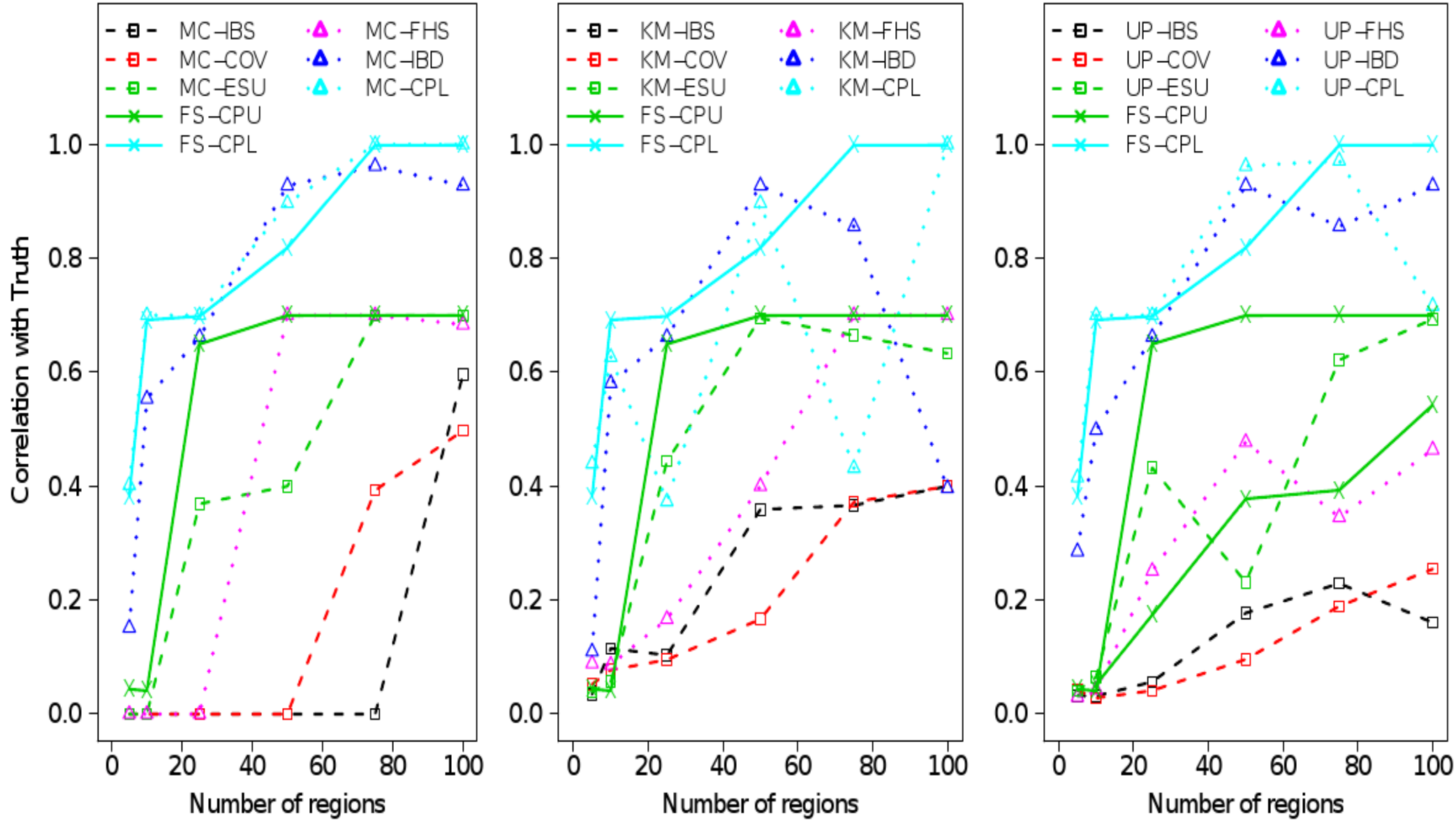
- Pure population structure is not correct – recent mixing leads to admixture
  - Seek conjugate mixture model for individuals
  - Hierarchical Dirichlet Process!
  - Interpretation: Pure populations created by drift, we see mixtures
- Better model:
  - Allow drift and admixture to both occur in real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

Daniel Lawson, University of Bristol
dan.lawson@bristol.ac.uk

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

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

Comparison of linked methods



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)

Probability of a partition

Rows of P_{ab} are Dirichlet

– Conjugate to multinomial, sum to 1

– Weak prior

Compute posterior incrementally due to conjugacy

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

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

(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

Posterior visualisation

- Too many populations!
- Pairwise coincidence matrix
- Create MAP (maximum a posteriori) tree from MAP partition
  - Show partition split posterior support
- (Population summary of data matrix X)

Comparison of linked methods