# Similarity Measures and Clustering In Genetics

Daniel Lawson

Heilbronn Institute for Mathematical Research
School of mathematics
University of Bristol

dan.lawson@bristol.ac.uk

www.paintmychromosomes.com

# Talk outline

Introduction
- Genetic data and overview

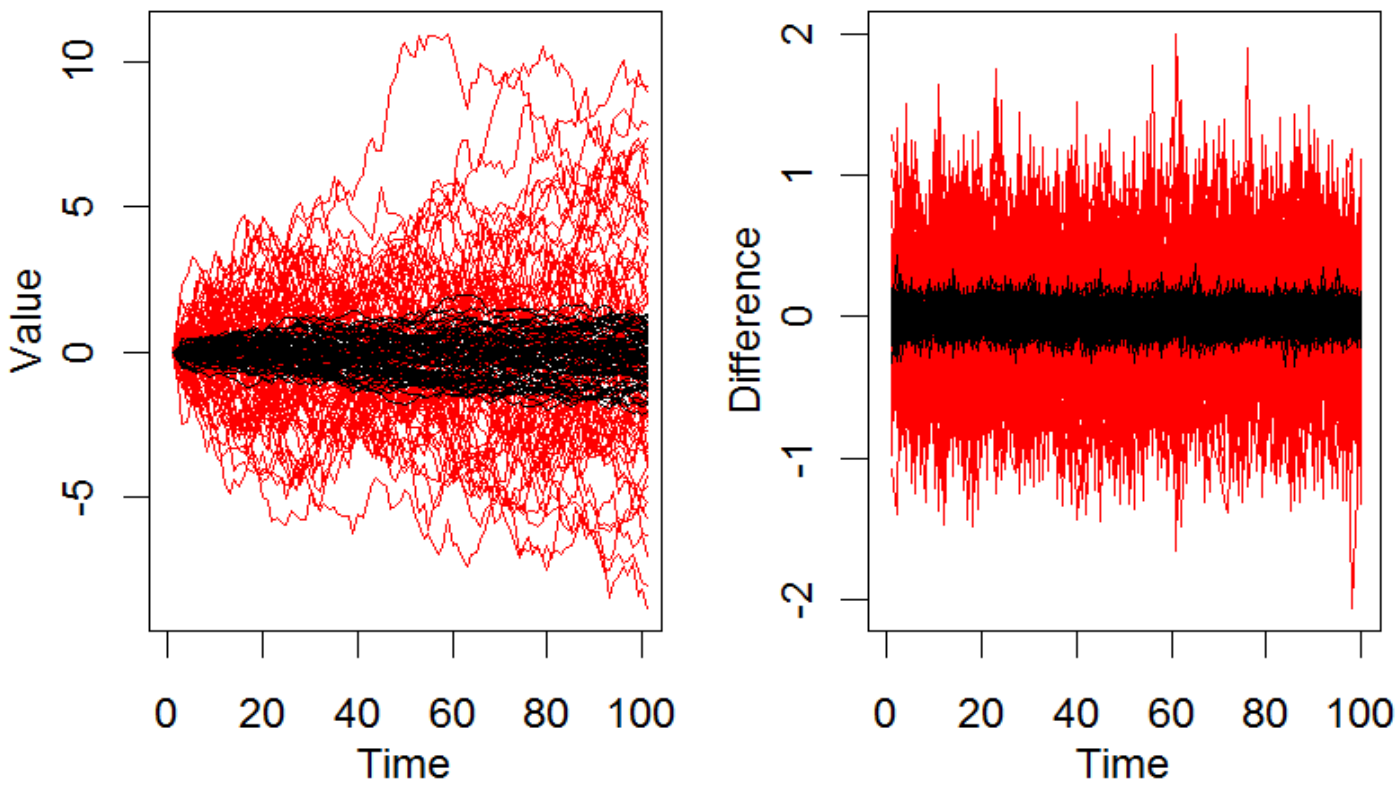
1 Generic approaches
- Similarity measures
- Similarity based clustering
- Spectral methods

2 Genetics models
- Direct model-based clustering
- Model-based similarity measures
- ChromoPainter/FineSTRUCTURE clustering

3 Results for real data

# Motivation: Models and Clustering

- Models for clustering essential
- Choice of measure strongly influences clustering
- Example: Random walk



# Genetics data

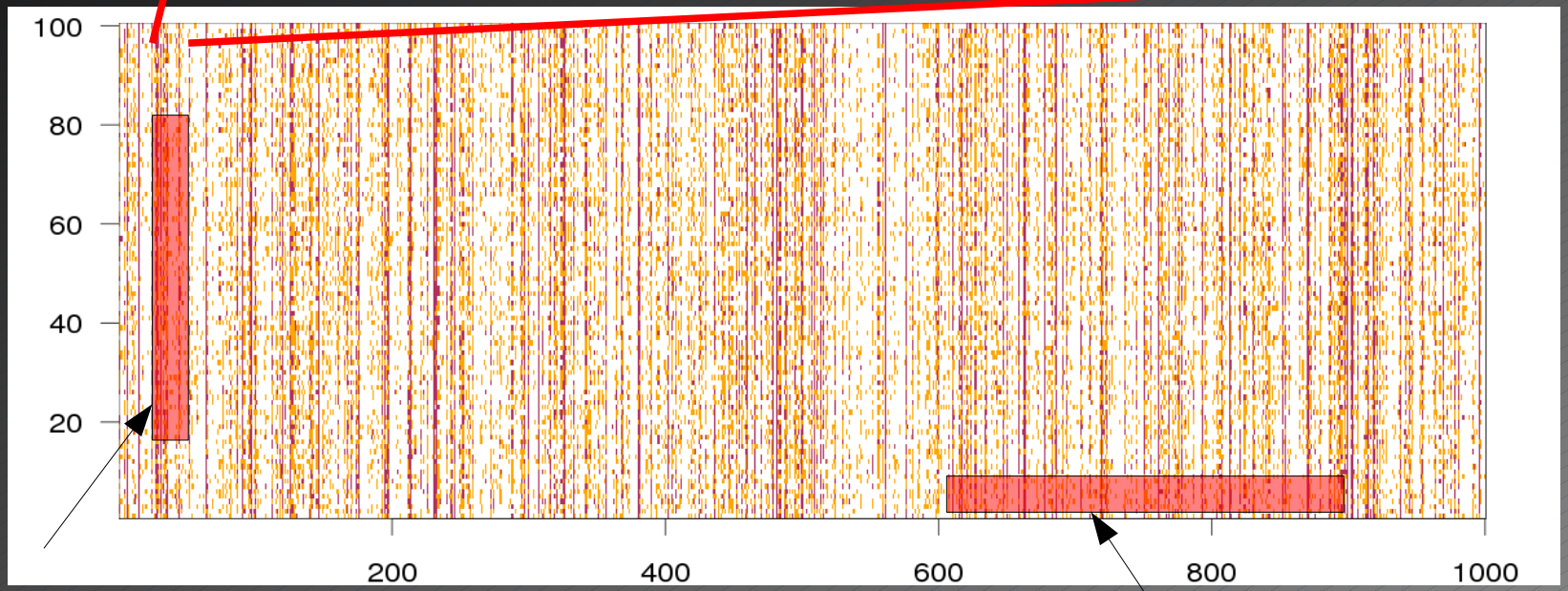
Individual $i$

0 1 1 1 1 1 0 1 1 1 1 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1

1 1 1 0 0 1 0 1 1 0 1 0 1 1 0 0 0 0 0 1 1 1 1 1 0 0 1 0 0 0 1 1

SNP $l$

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




*Correlations due to relatedness*

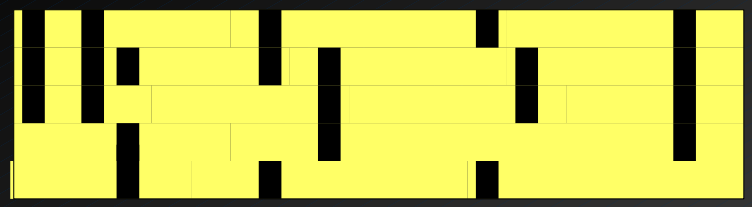
*Correlations due to linkage*

Genome position – $O(10^{6-9})$

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

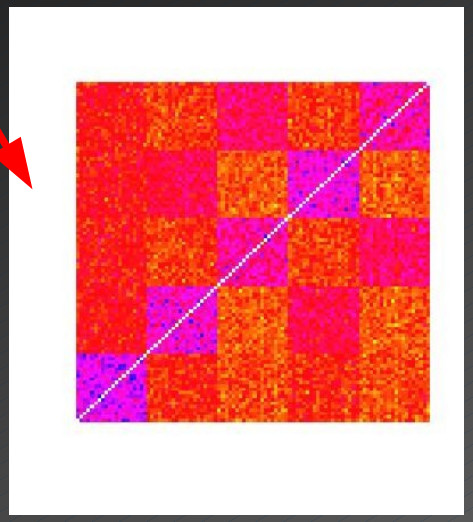
# Outline: The process

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



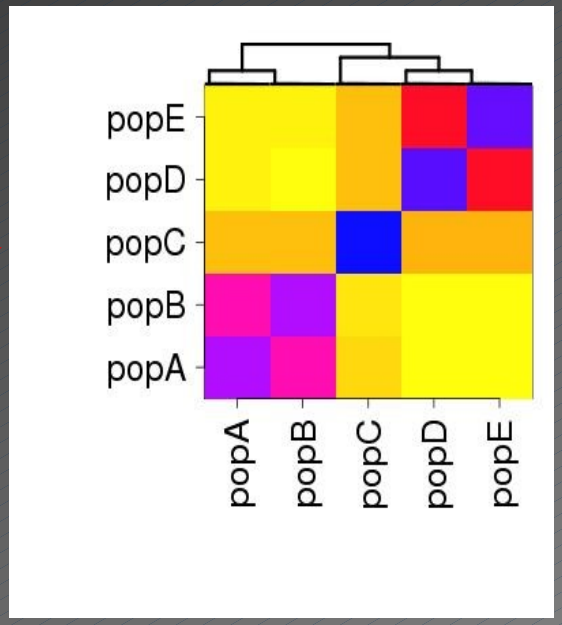
*Step 1: Similarity matrix*

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



*Step 2: Clustering*

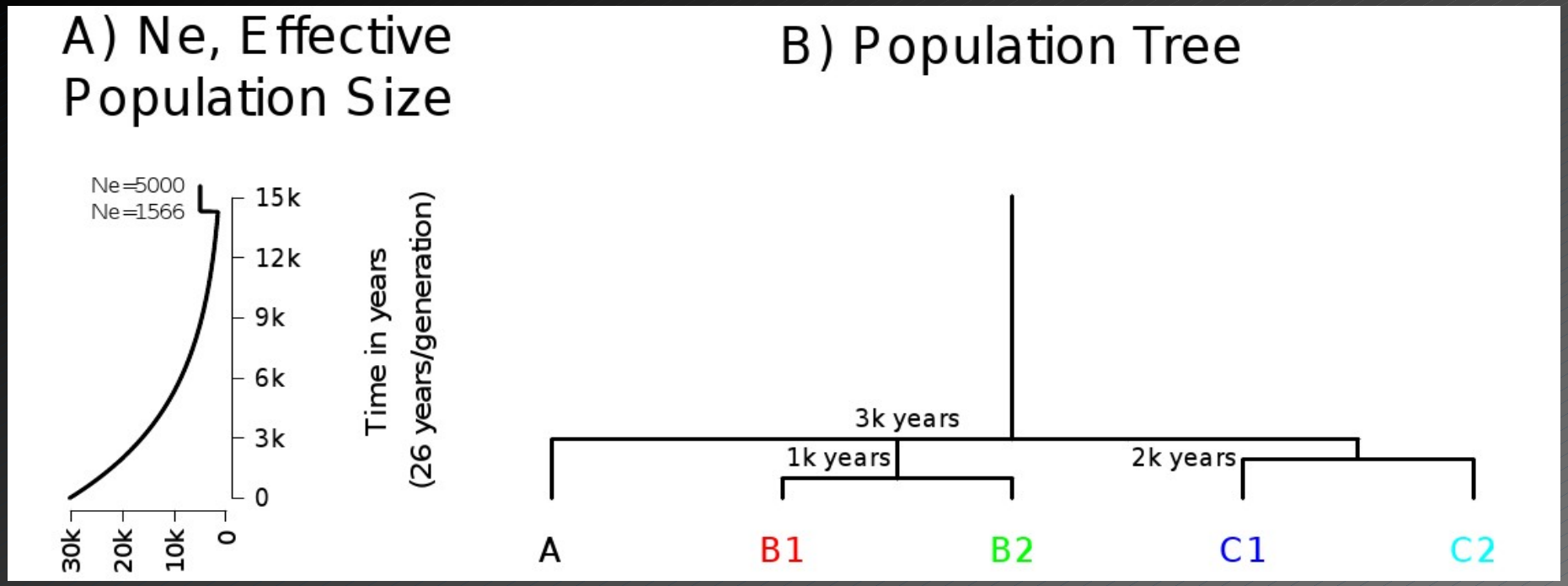
Populations $O(10^{2-3})$




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

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

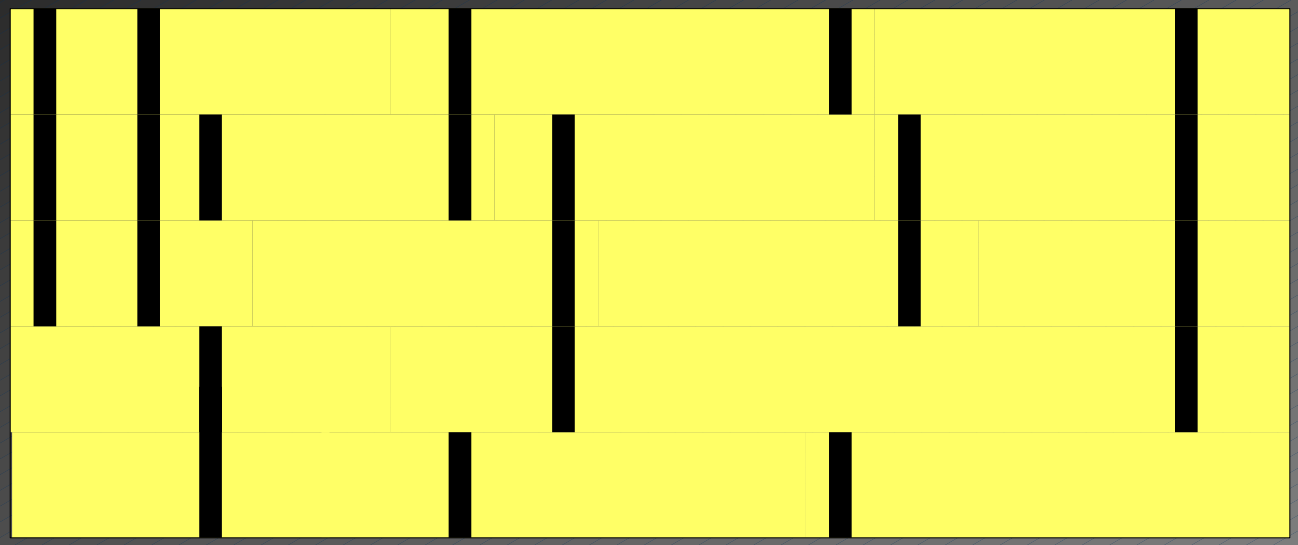
# Simulated data




- Simulate data using 'real' conditions
- Sequence data including linkage disequilibrium, random mating within populations, 'complex' demography
- Change how much data we show the models

# Part 1. Generic Approaches

- Treat SNPs as independent features
- Cluster individuals as independent samples from some clustering distribution on $Y_{il}$



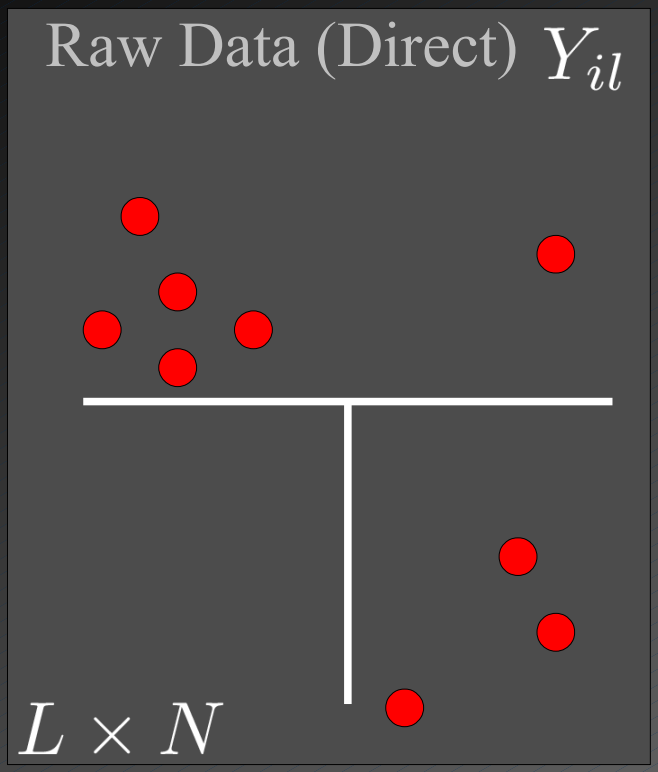
Individual Number $i$

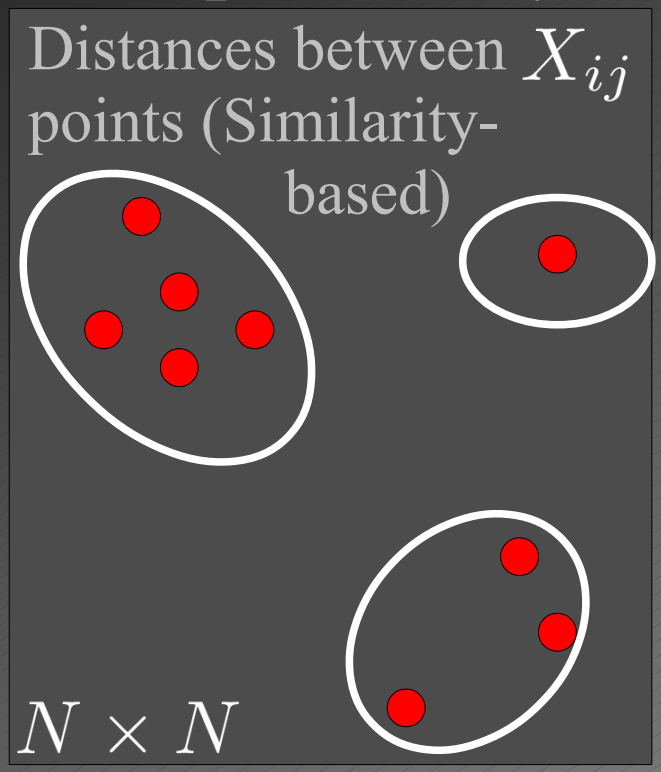
$Y_{il}$

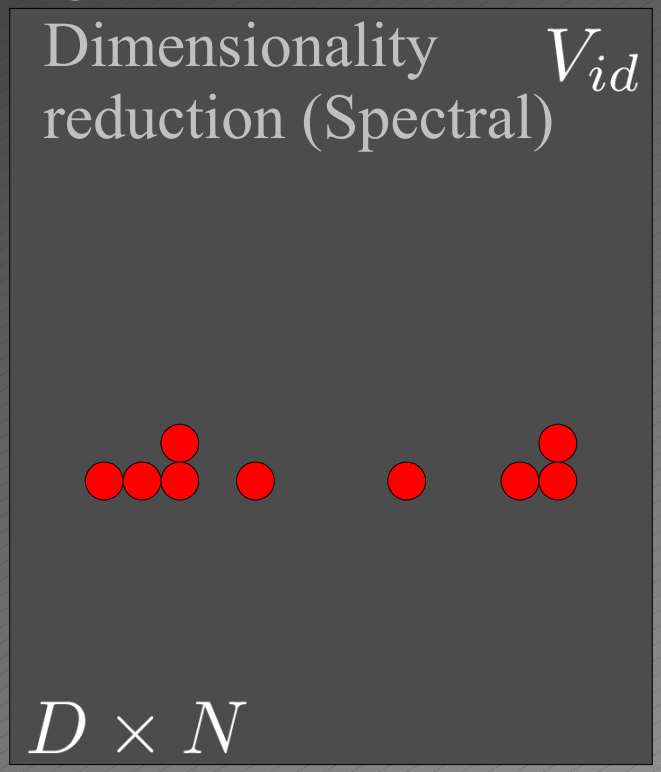
SNP location $l$

# Clustering

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




$L \gg N \gg D \quad O(k) = O(d)?$

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

# Similarity Measures (between SNPs)

- Almost infinite number of choices... e.g.

Cosine distance

$$\frac{\mathbf{Y_i} \cdot \mathbf{Y_j}}{|\mathbf{Y_i}||\mathbf{Y_j}|}$$

TF-IDF (term Frequency, inverse document frequency)

$$Y'_{il} = \left( \sum_l Y_{il}/L \right) \log(1/f_l)$$

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

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

Covariance

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

Exponential

$$e^{-\sum_l (1 - |Y_{il} - Y_{jl}|)}$$

Normalised covariance

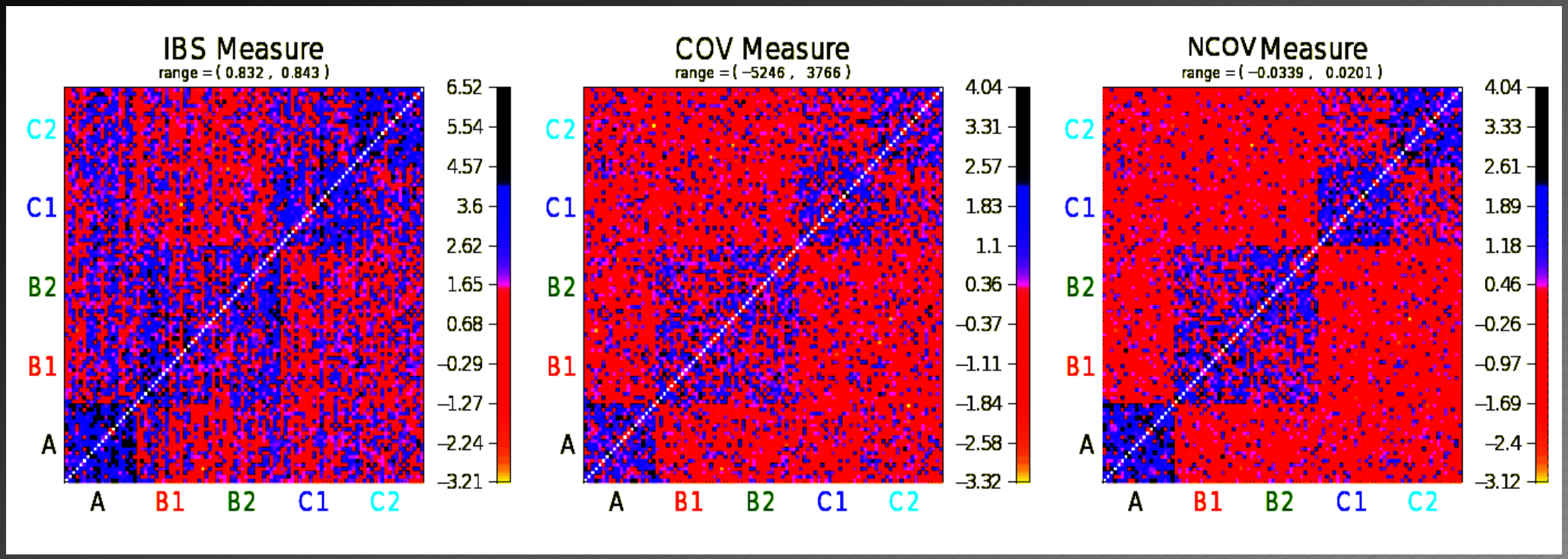
$$\sum_l \frac{(Y_{il} - f_l)(Y_{jl} - f_l)}{f_l(1 - f_l)}$$

Normalised count of shared alleles

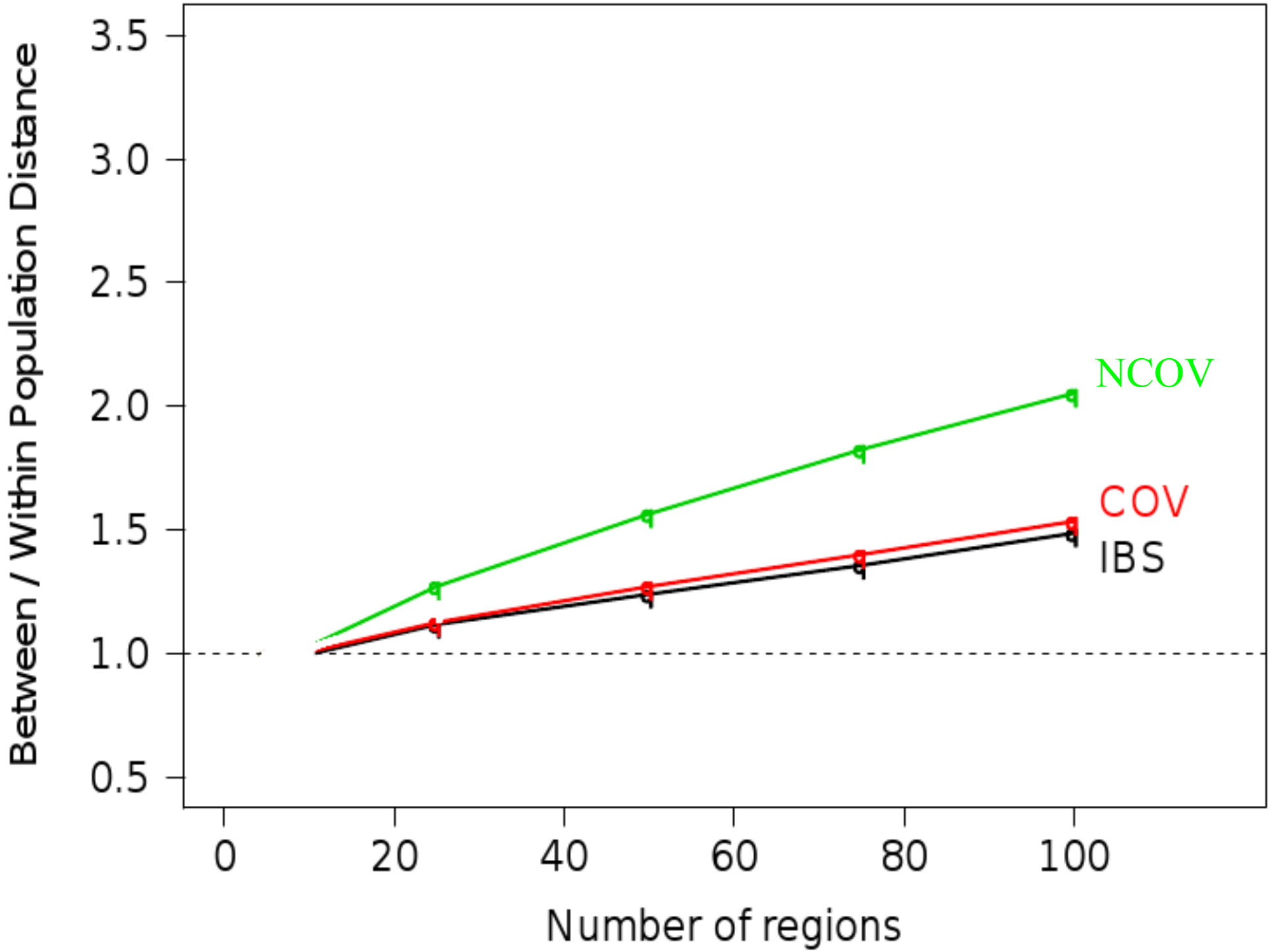
$$\sum_l \left[ \frac{Y_{il} Y_{jl}}{\hat{f}_l} + \frac{(1 - Y_{il})(1 - Y_{jl})}{1 - \hat{f}_l} \right]$$

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

# Similarity measures of simulated data




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


Between / Within Population Distance
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
20
40
60
80
100
Number of regions
NCOV
COV
IBS

# Dimensionality reduction

SVD of $Y_{il}$

$=$ PCA of $\mathrm{cov}(Y_{il})$ (suitably normalised)

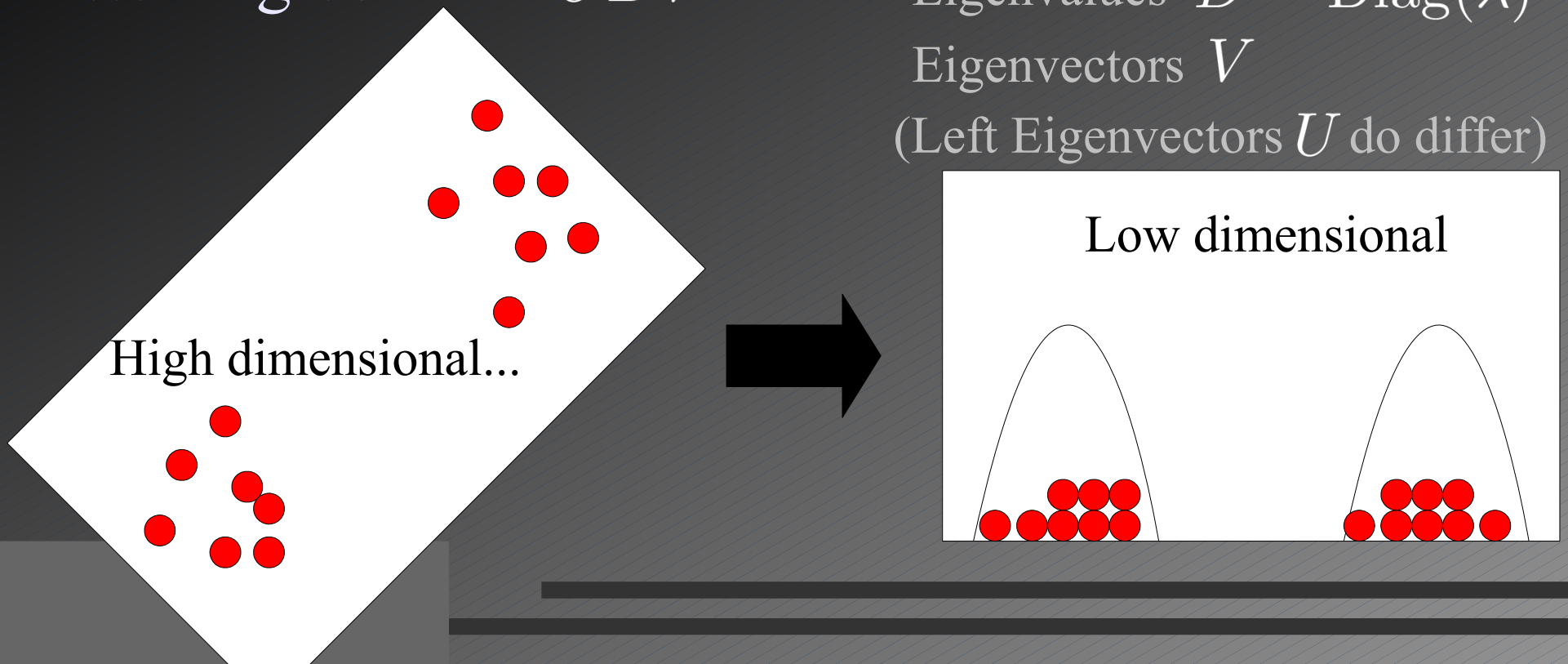
$=$ MDS on $Y_{il}$ (suitably normalised)

… all give $UDV^T$

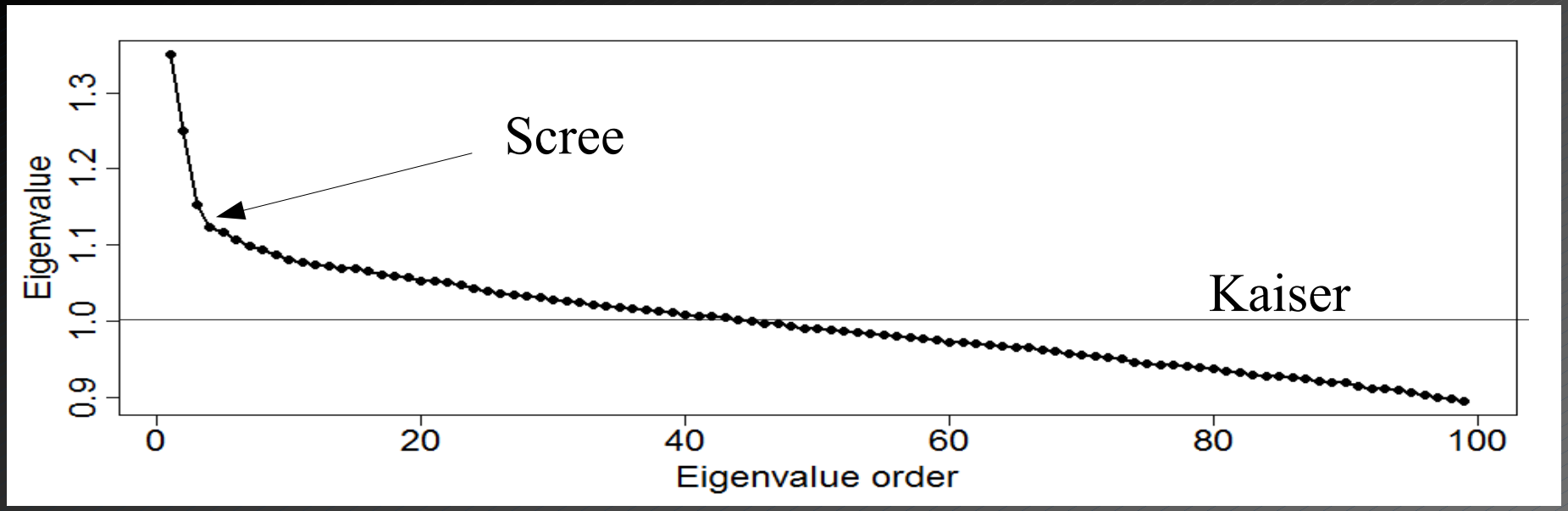
Eigenvalues $D = \mathrm{Diag}(\lambda)$

Eigenvectors $V$

(Left Eigenvectors $U$ do differ)



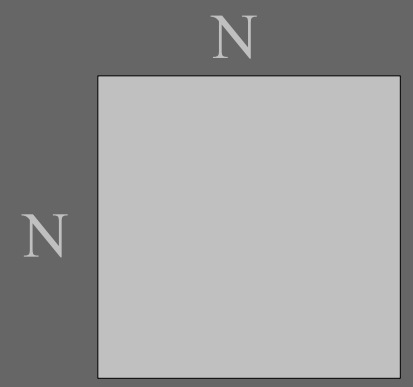
# How many dimensions?




- All Eigenvalue orientated...
  - Kaiser (1960) criterion (EV>1)
  - Scree test (Cattell 1966, “large jump in EV spectrum”)
  - Velicer's MAP criterion
  - Horn's Parallel Analysis (PA) criterion
  - Tracy-Widom distribution

} Also consider Eigenvectors

# How many dimensions?



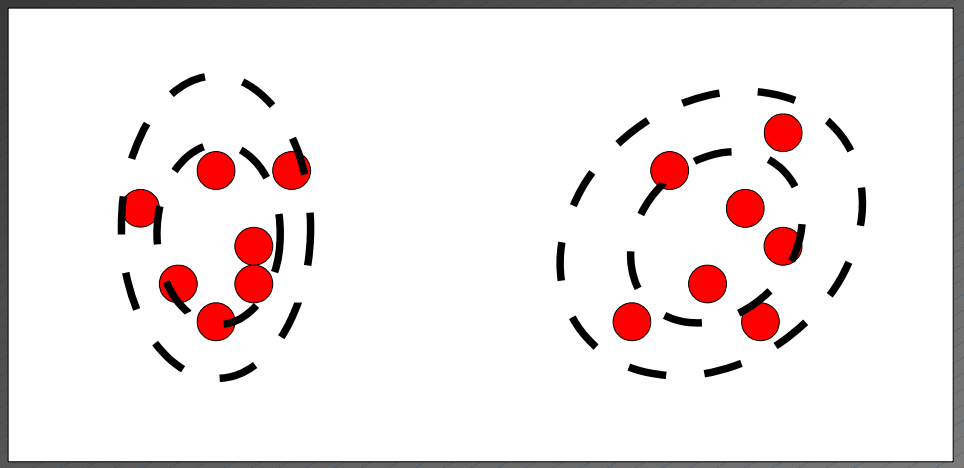

- MAP (Minimum Average Partial Correlation), Velicer 1976
  - Remove largest eigenvalue
  - Compute (partial) correlation between remaining eigenvectors and the data (accounting for previous eigenvectors)
  - Repeat

- 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

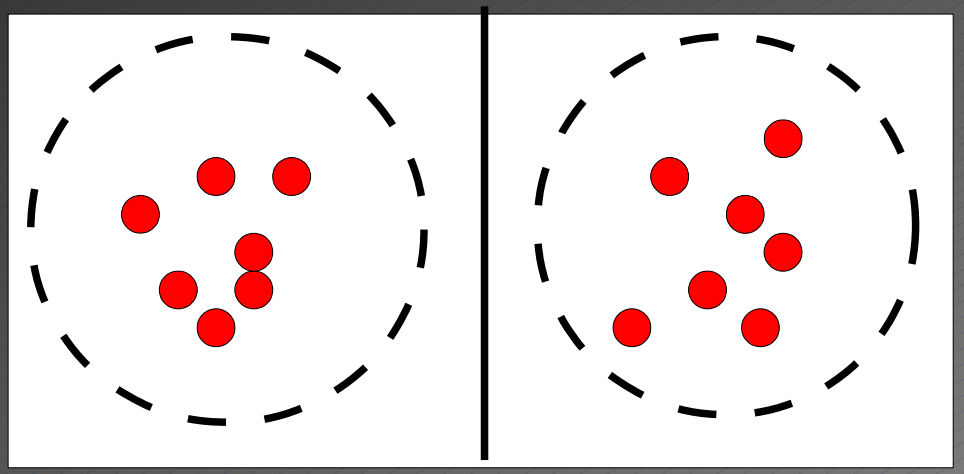
# Clustering: MVN

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



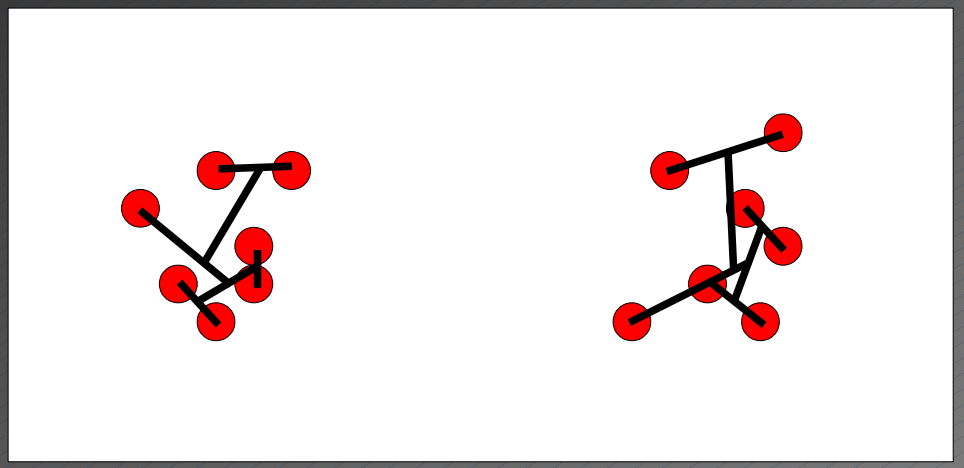
# Clustering: K-Means

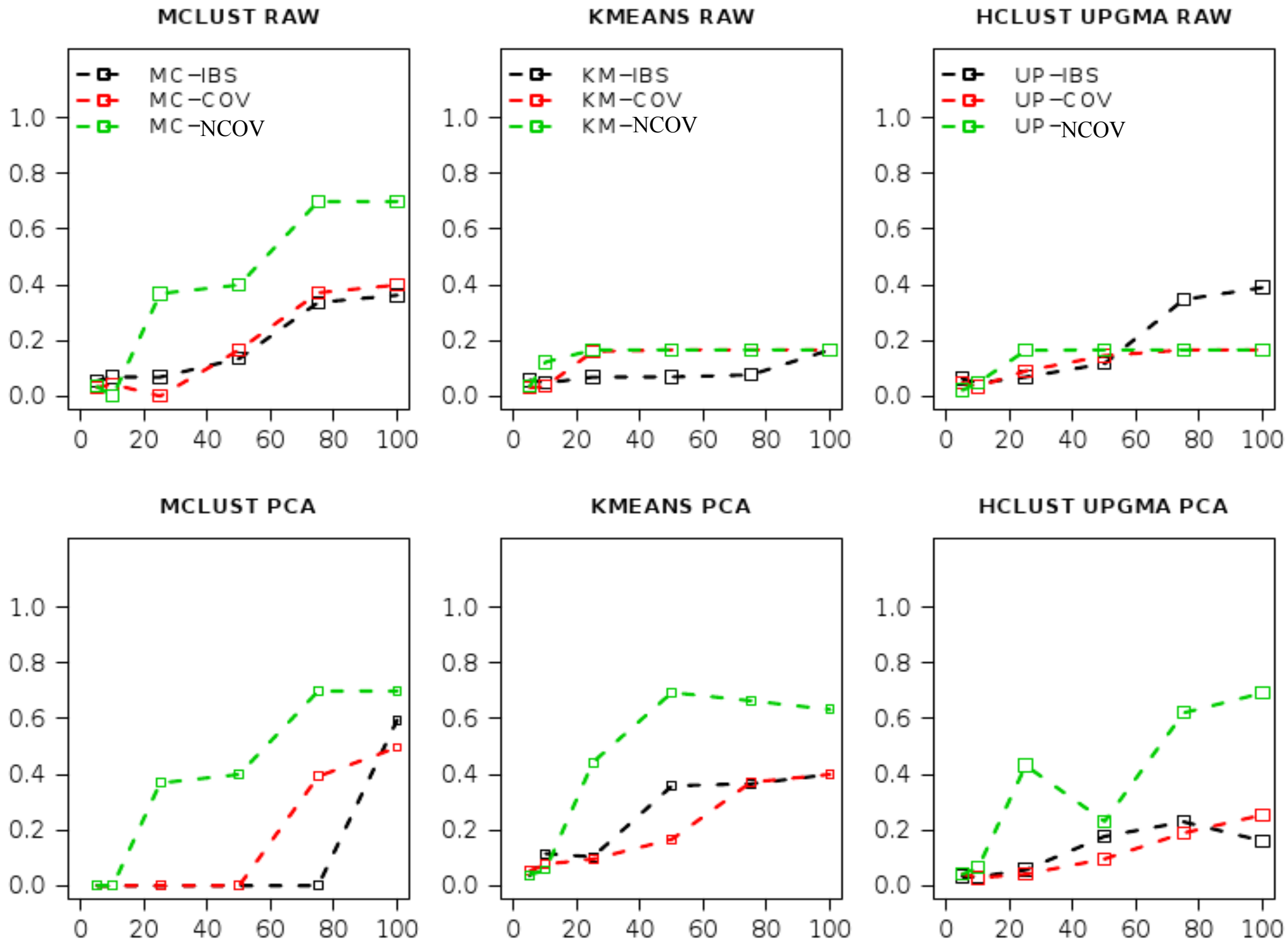
- Minimise Euclidean distance to cluster centers
- "Hard K-Means" (as uses the same distance penalty as MVN, but imposes a strict boundary)
- *K* estimated using the Calinski (1974) criterion (comparing variance within clusters to that between clusters)



# Clustering: Hierarchical methods

- Successively merge "closest" samples (or successively split... lots of ways to define "close")
- e.g. UPGMA (Unweighted-Pair Group Method with Arithmetic Mean). "Close" here is distance to the centroid of the clusters
- (e.g. Ward's (1963) minimum variance criterion)
- *K* estimated using the Calinski (1974) criterion




MCLUST RAW
MC−IBS
MC−COV
MC−NCOV
KMEANS RAW
KM−IBS
KM−COV
KM−NCOV
HCLUST UPGMA RAW
UP−IBS
UP−COV
UP−NCOV
MCLUST PCA
KMEANS PCA
HCLUST UPGMA PCA
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100

# An algorithm to generate papers from Student Projects: Roll 3 dice and refer to the table:

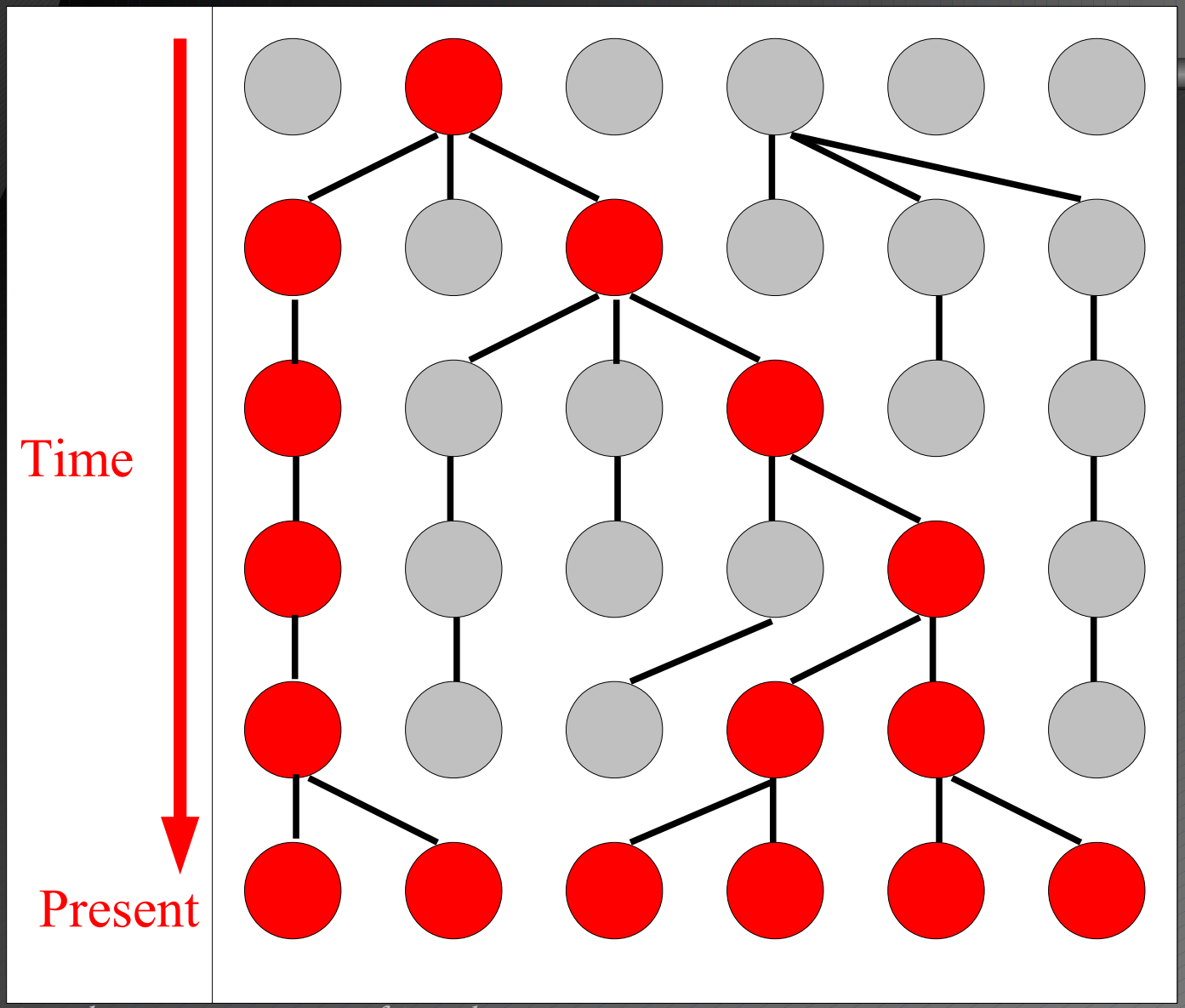
| Dice roll | Similarity Measure | Dimensionality Reduction | Clustering Algorithm |
|---|---|---|---|
| 1 | IBD/ASD | None | MVN |
| 2 | Covariance | PCA - MAP | K-Means |
| 3 | Normalised Covariance | PCA - Parallel Analysis | Hierarchical (standard) |
| 4 | Something from Document clustering | PCA - Tracy-Widom | Hierarchical (iteratively modifying data) |
| 5 | Something model-based | Spectral Graph Theory | Something from CS literature |
| 6 | Something else... | Something from image analysis | ??? |

# Part 2: Genetics models

- Similarity measures matter more than clustering model
- MVN model seems best
- On PCA, all models do similarly well
- No similarity measure is good enough

- Time to understand why...

# Ancestry process




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

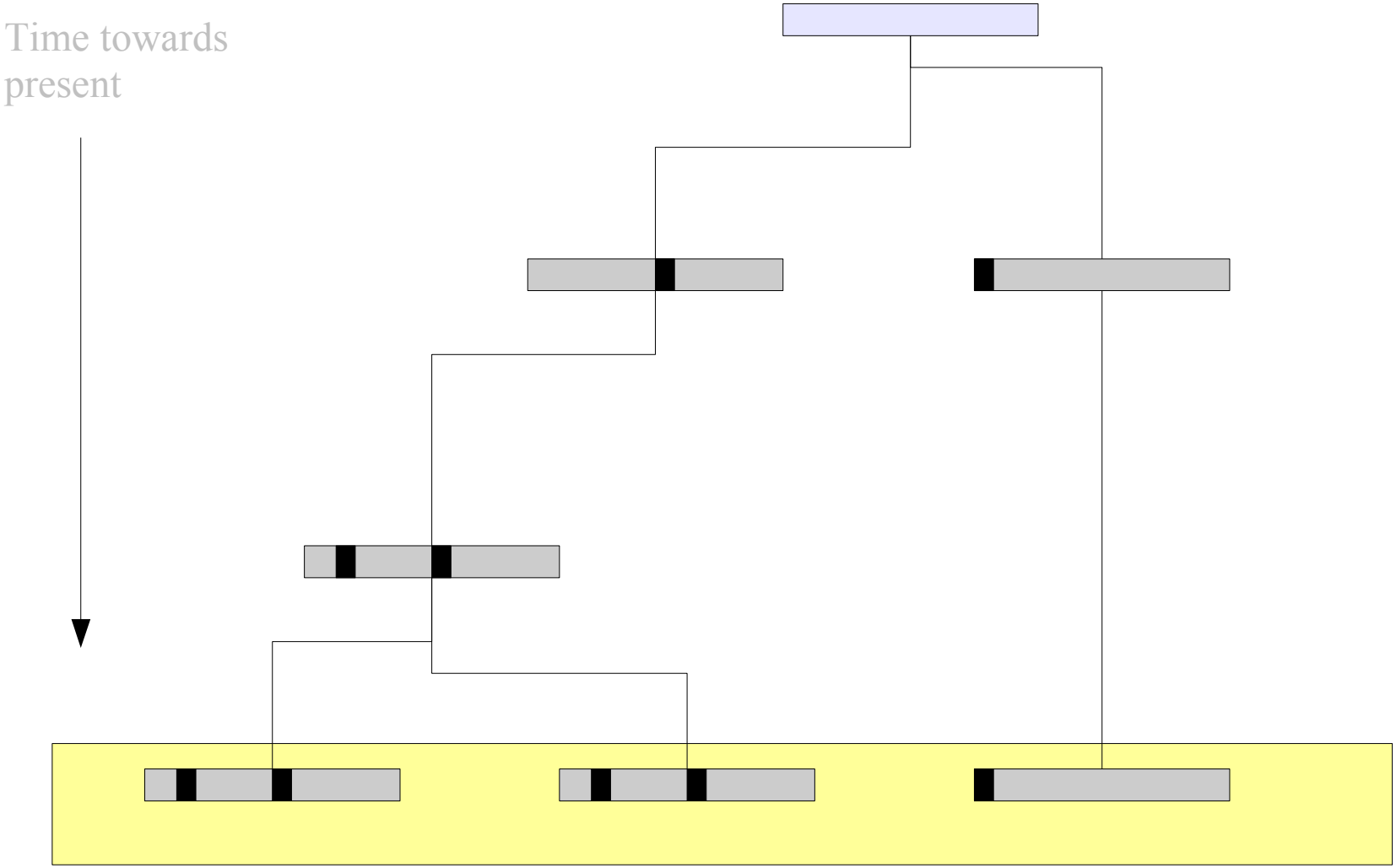
$$N \to \infty$$

keeping $N/T$ constant

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

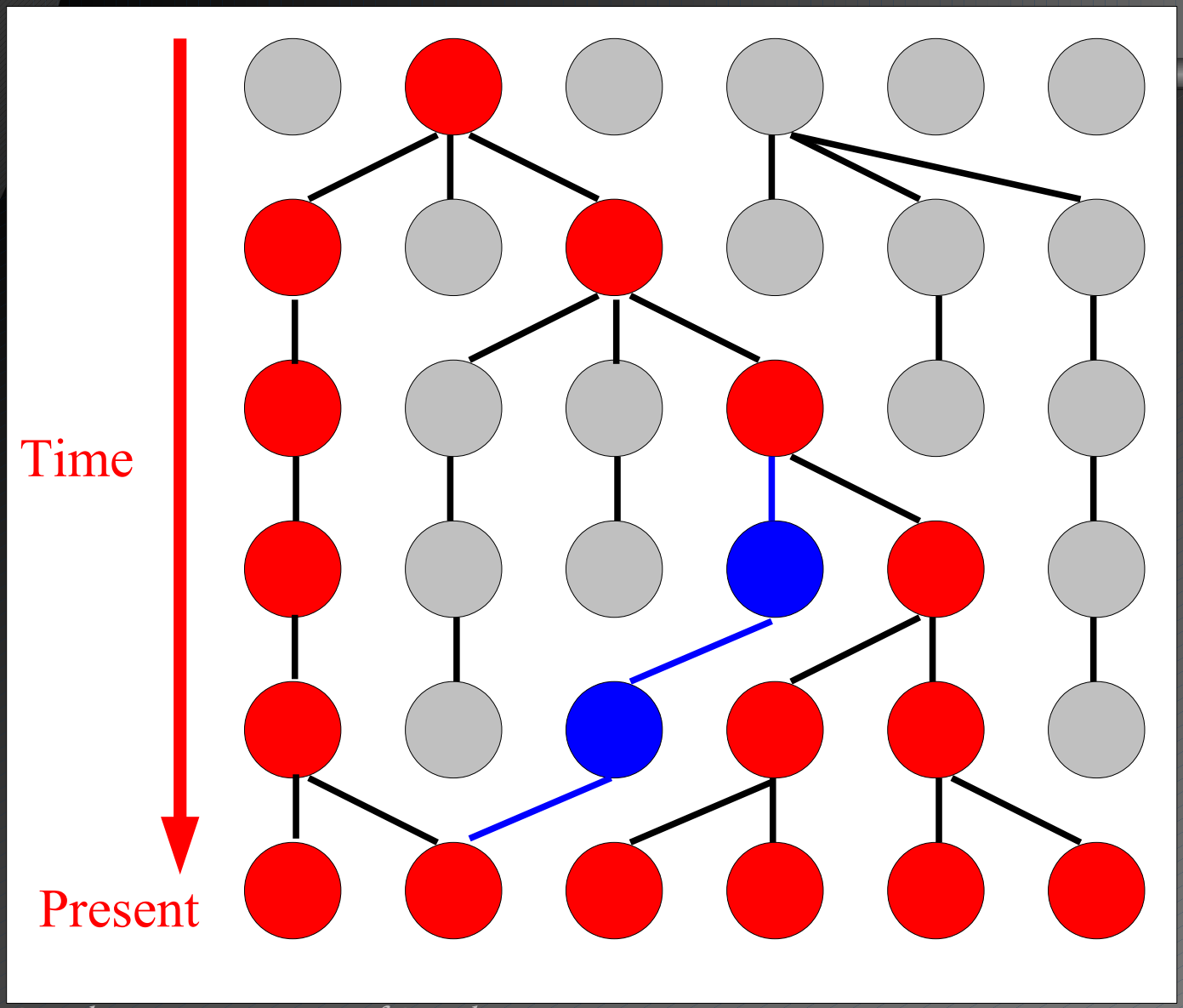
*Daniel Lawson, University of Bristol*
*dan.lawson@bristol.ac.uk*

# Genetic model - Ancestral Tree




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

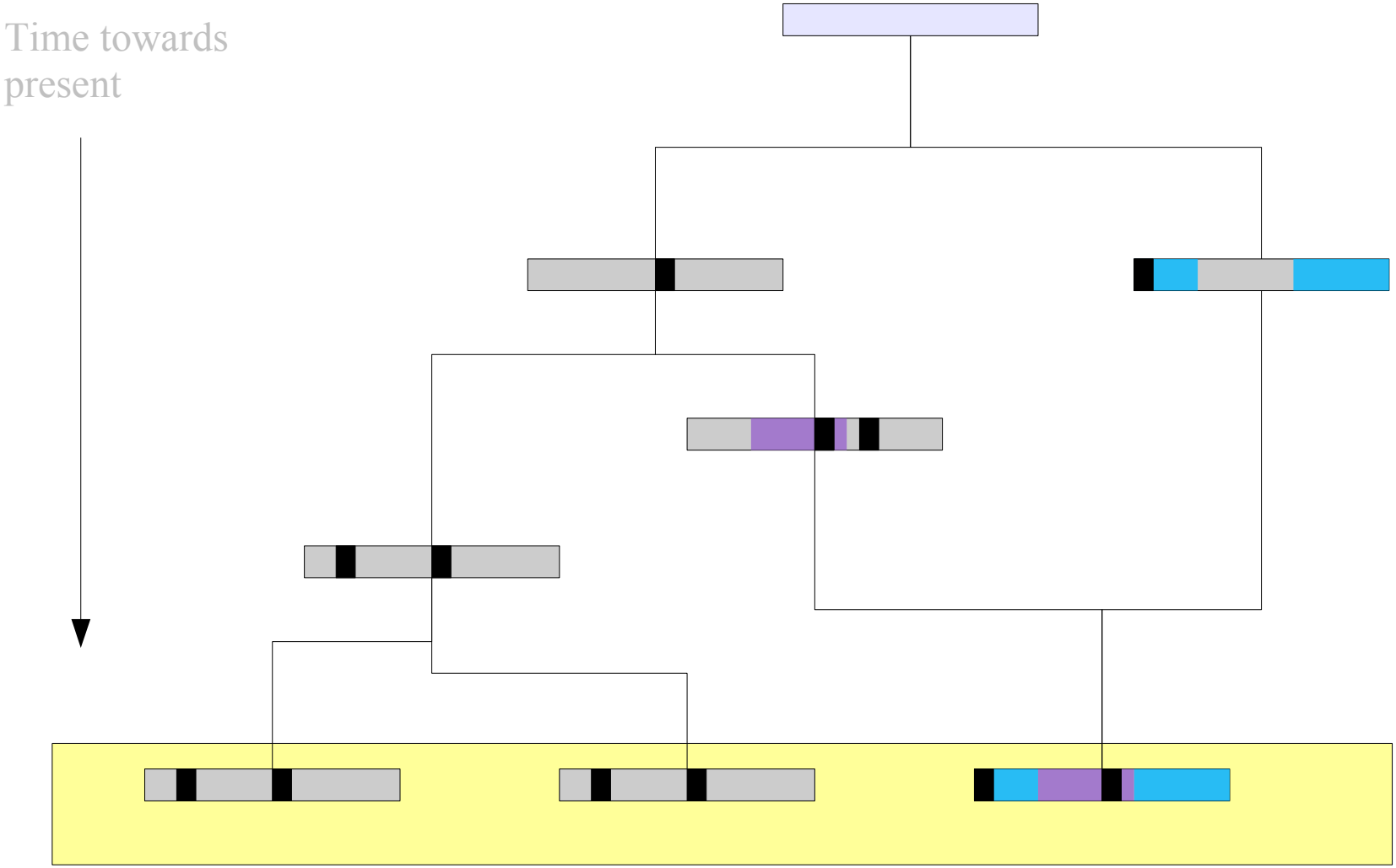
# Ancestry process with recombination




- Probability $\rho$ 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*

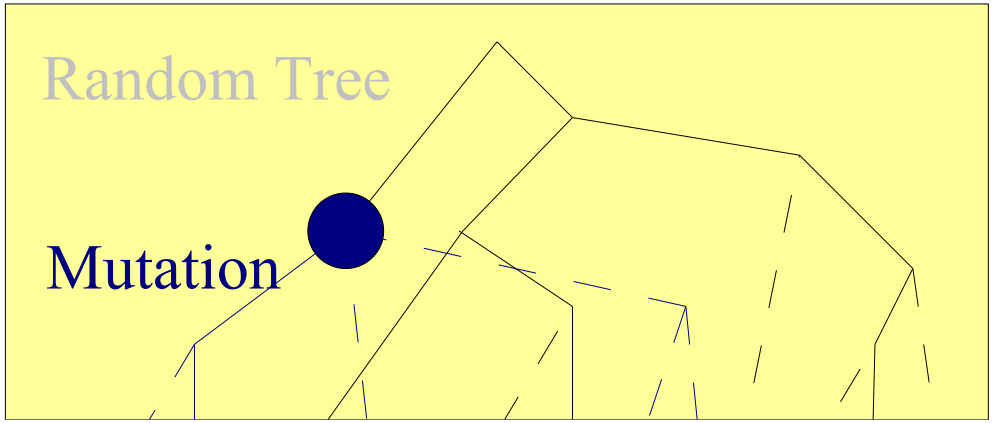
# Ancestral Recombination Graph - Summary

- 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

# Time to Most Recent Common Ancestor




- Recall that each SNP has a random tree
- McVean (2009) showed that

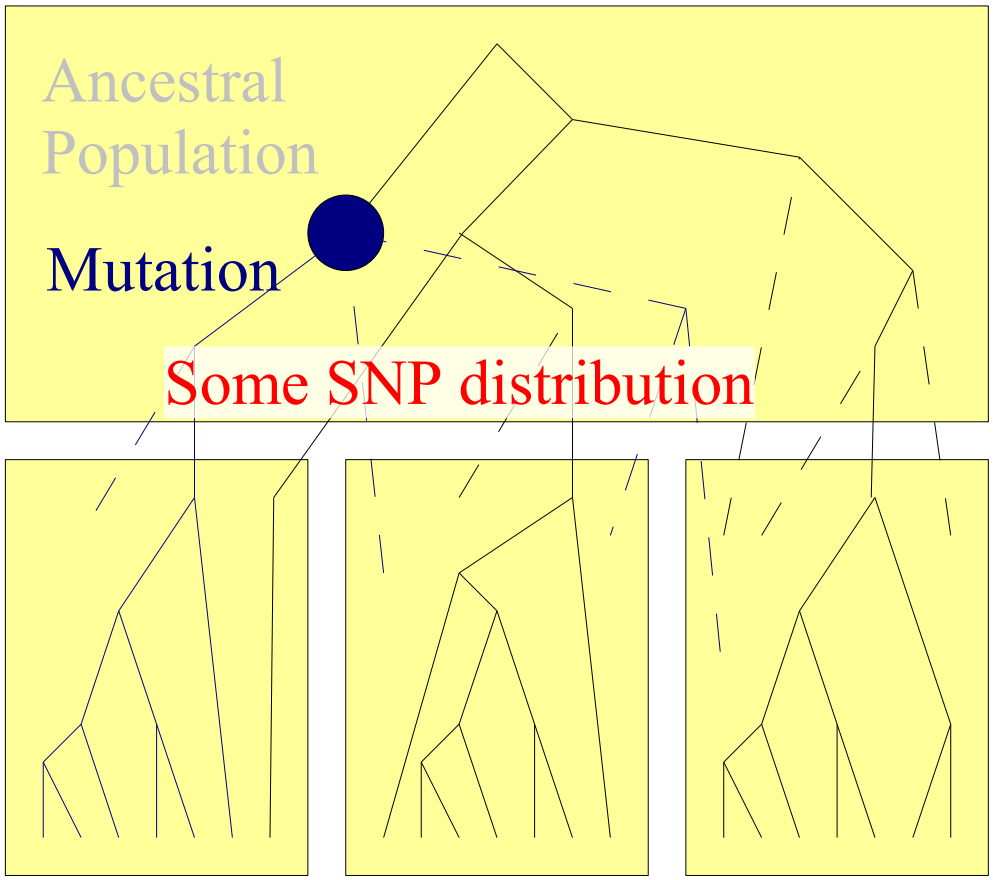
$$\mathbb{E}(t_{ij}) = f(\sum_l |x_{il} - x_{jl}|)$$

(where *f* is simple and known)

- i.e. Counting identical SNPs captures all the information in the tree
- How do times in tree relate to population structure?

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

# 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

$$f_k \sim \mathrm{Beta}(f_0, \nu)$$

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

$$f_k \sim \mathrm{N}\left(f_0, \sigma^2 f_0(1 - f_0)\right)$$

(Recall relationship to diffusion)

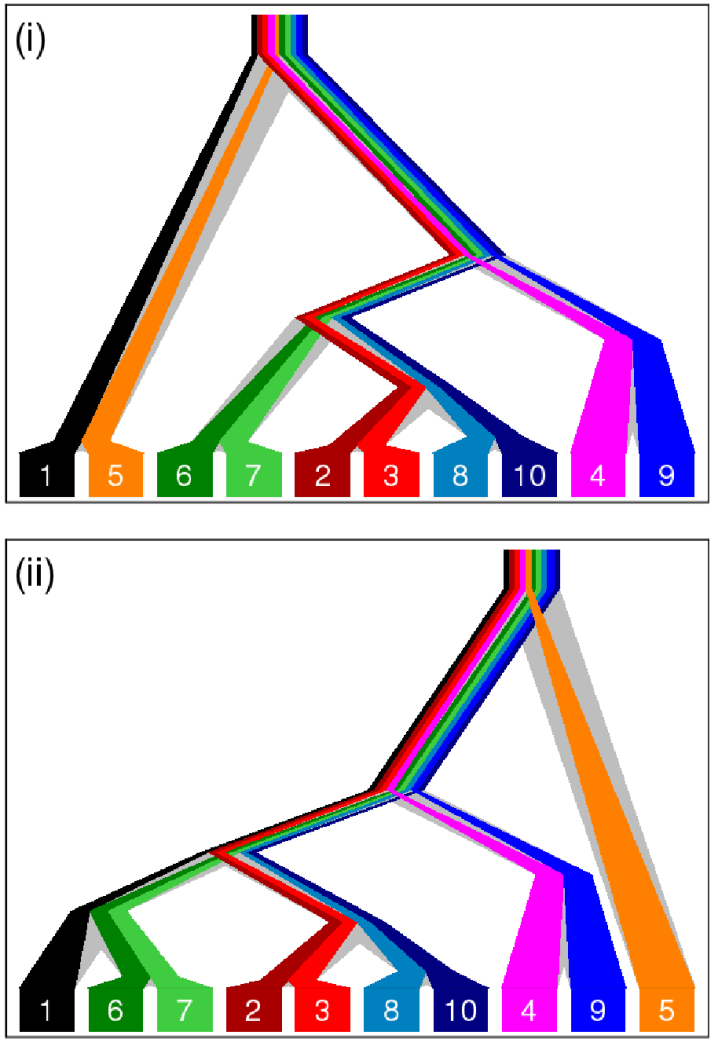
# Direct Model-based clustering

- Population model:
  - Beta distribution for SNP frequencies in each
  - Assume individuals exchangeable within populations
- Gives likelihood for frequency of SNPs
  - Binomial distribution
  - Assume no linkage (linkage approximations exist)
- Gives popular STRUCTURE* model
  - Still can't cope with large datasets
- Can we do this well on genomic (linked) data?

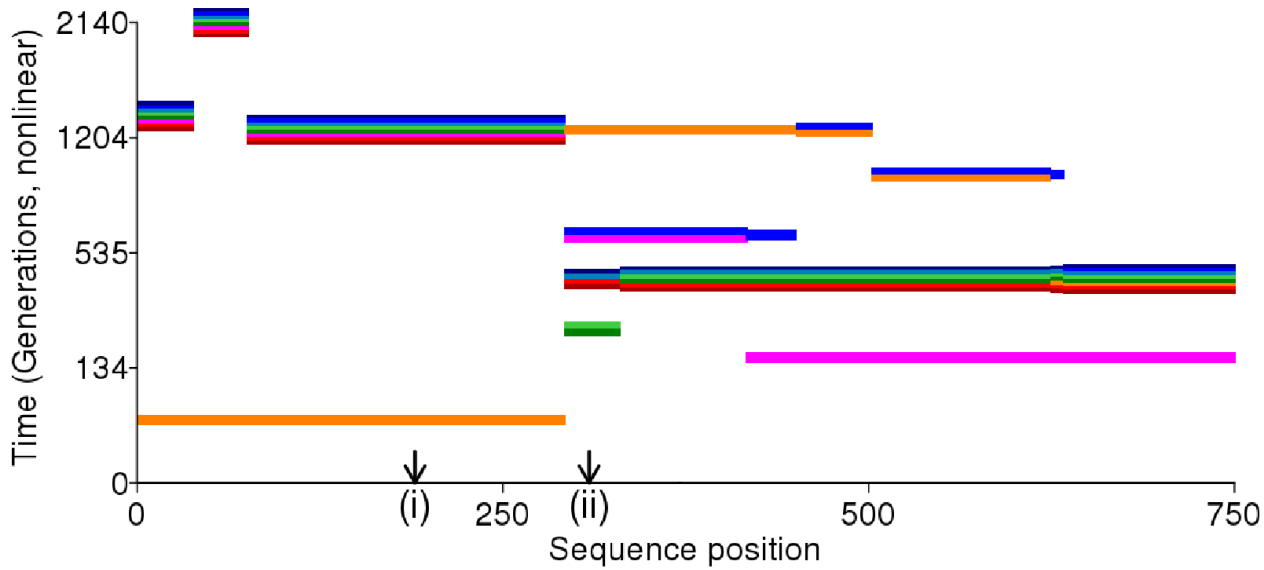
*Daniel Lawson, University of Bristol*
*dan.lawson@bristol.ac.uk*

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

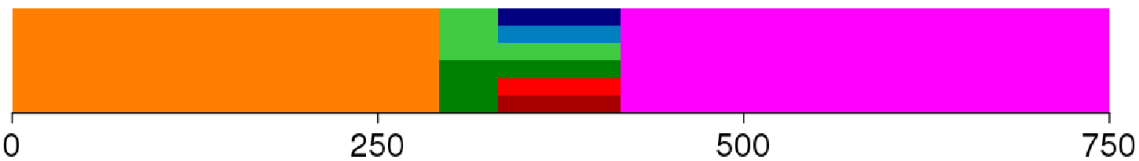
Local genealogies




Time to MRCA with haplotype 1




True 'nearest neighbour' distribution of haplotype 1



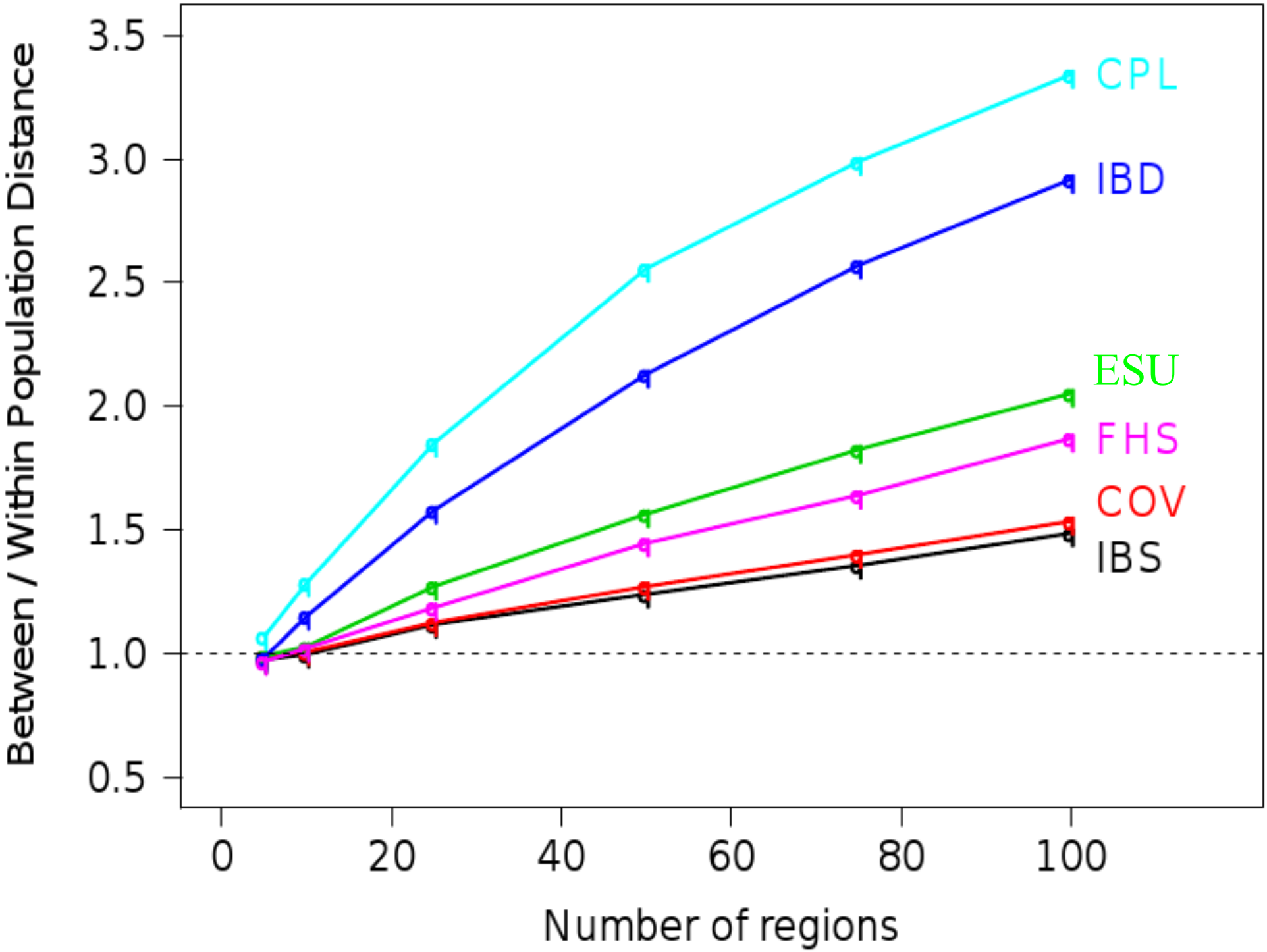

Mean painting of haplotype 1


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

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


Between / Within Population Distance
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
20
40
60
80
100
Number of regions
CPL
IBD
ESU
FHS
COV
IBS

# 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*

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

# Probability of a partition

- Dirichlet Process prior for partition $\eta$ :

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

- Rows of $P_{ab}$ (i.e.$G_0$ ) are Dirichlet *(containing hidden biological parameters)*…

- … so conjugate, and we integrate out $P_{ab}$

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

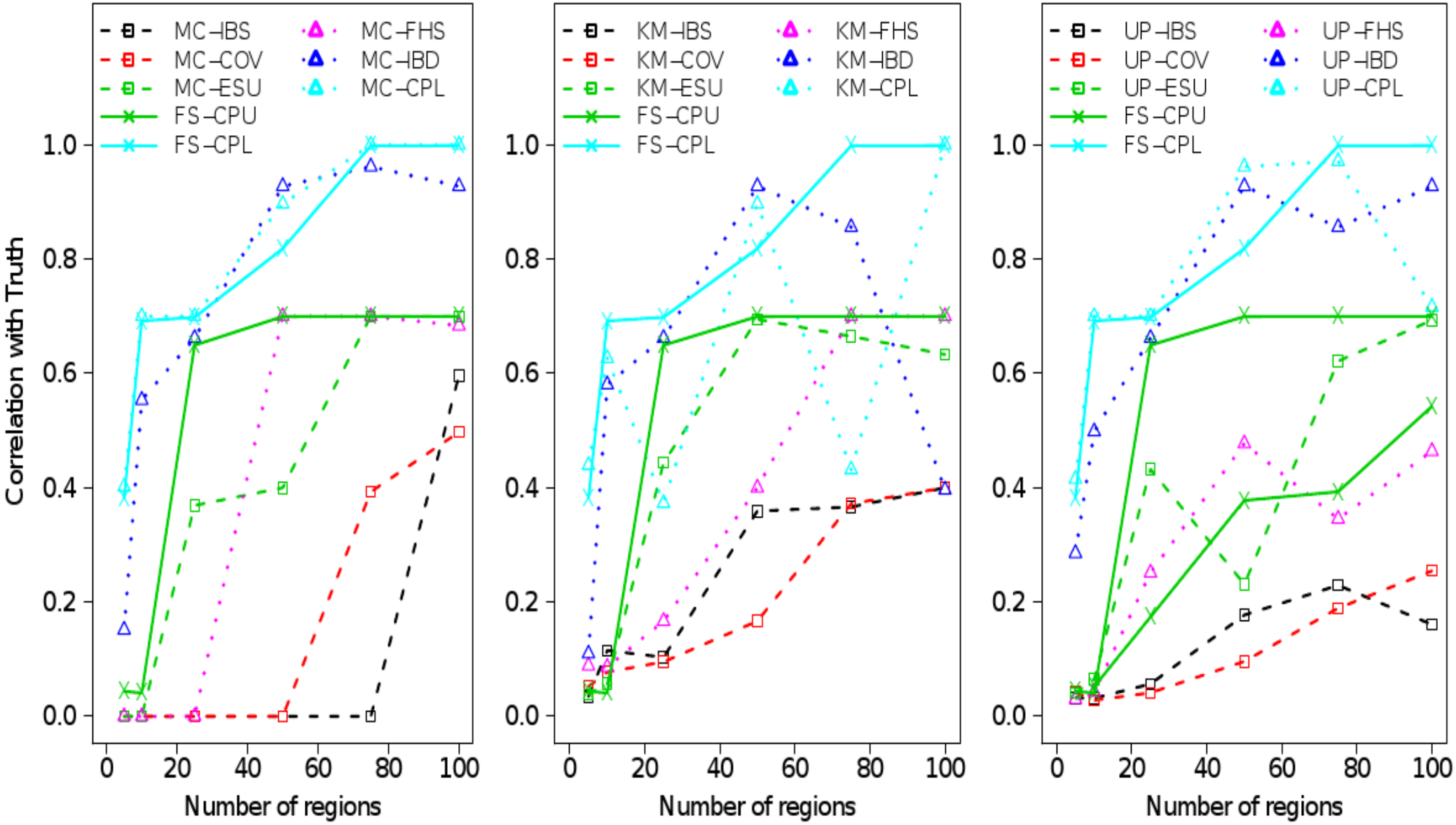
# Proven theoretical results

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

And the MVN likelihood with a structured covariance...

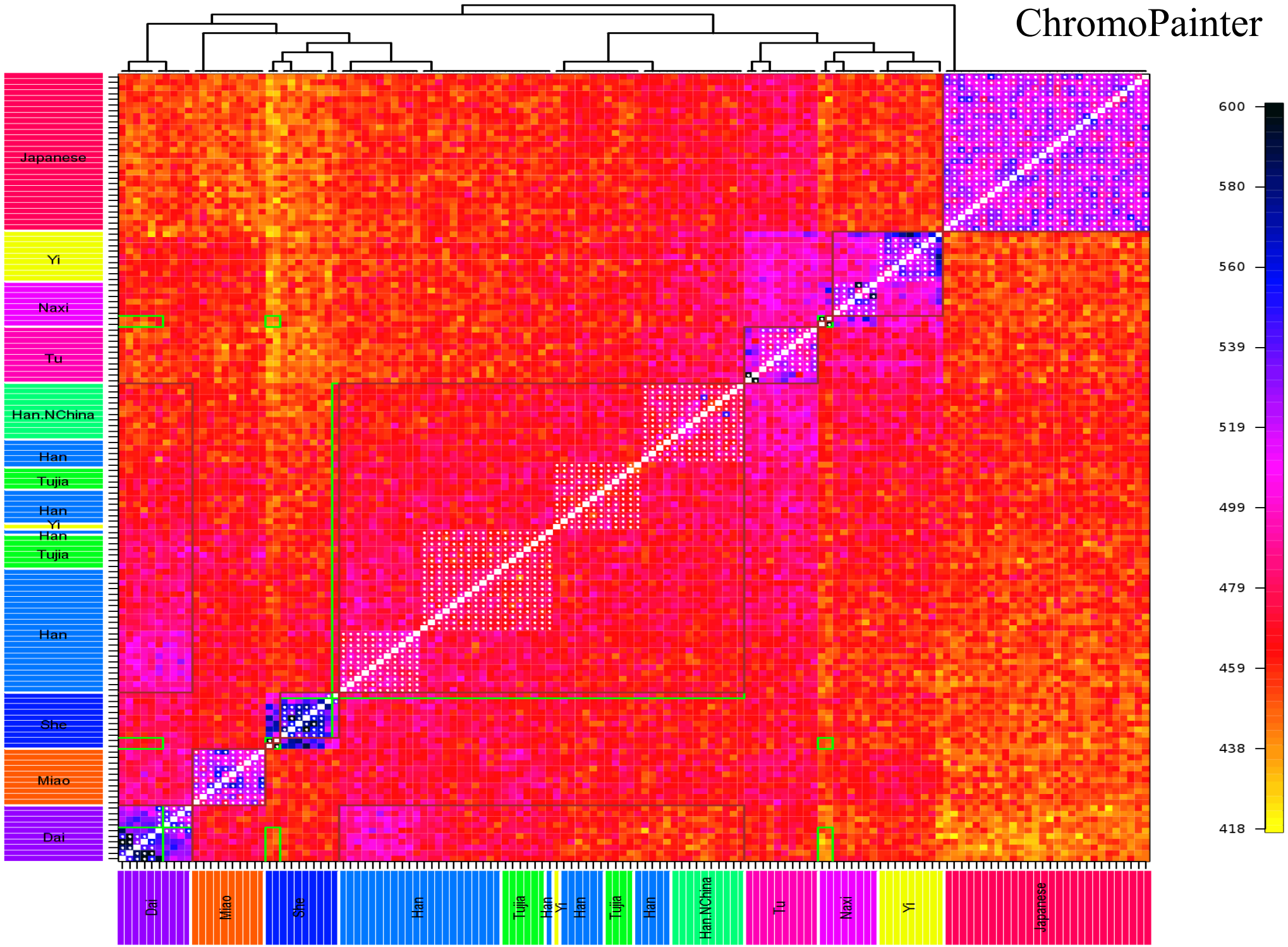
**Pritchard, Stephens and Donnelly, Genetics, 155:945-959, 2000*
*Calculations due to Simon Myers*

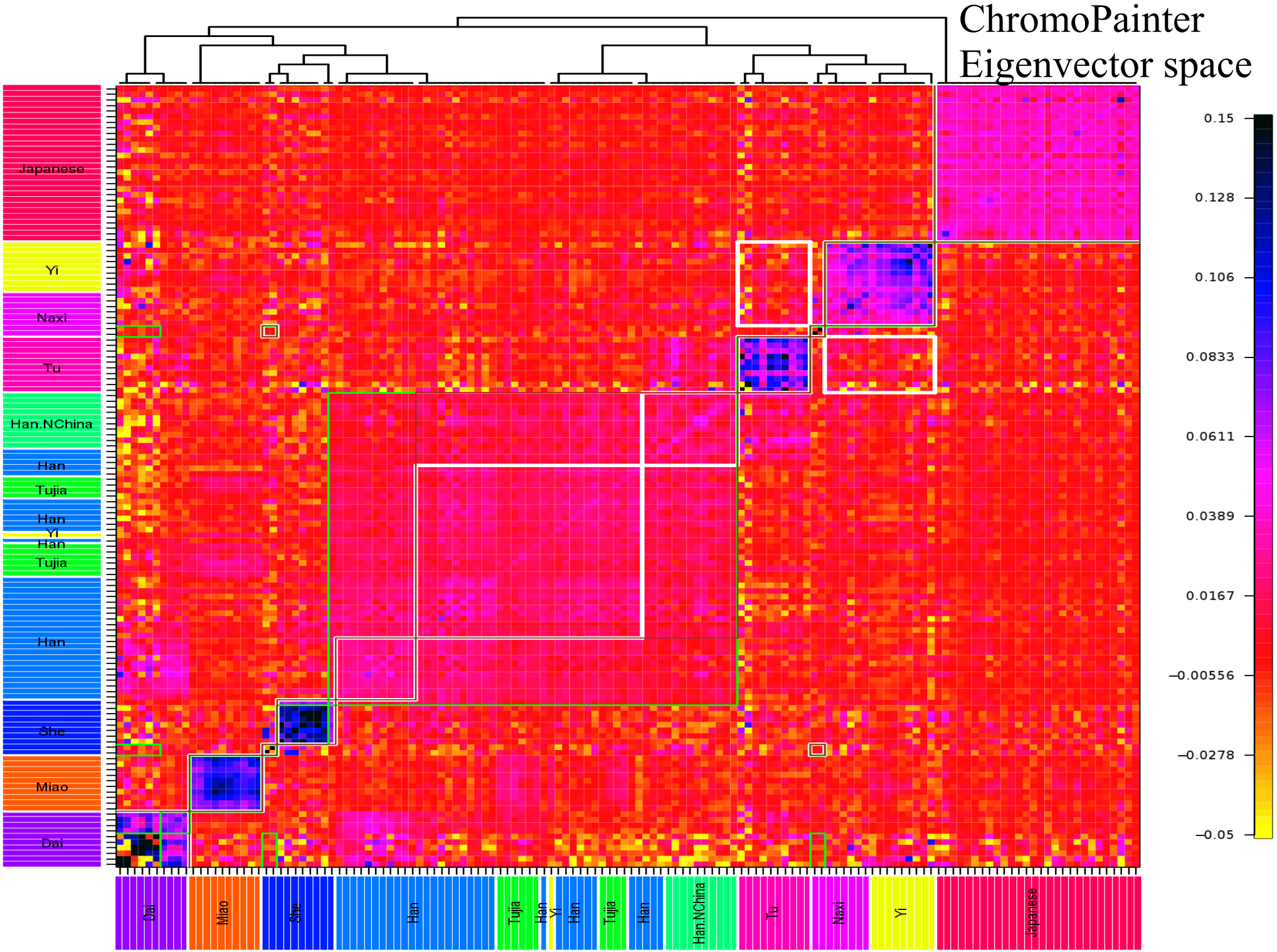
# Comparison of linked methods

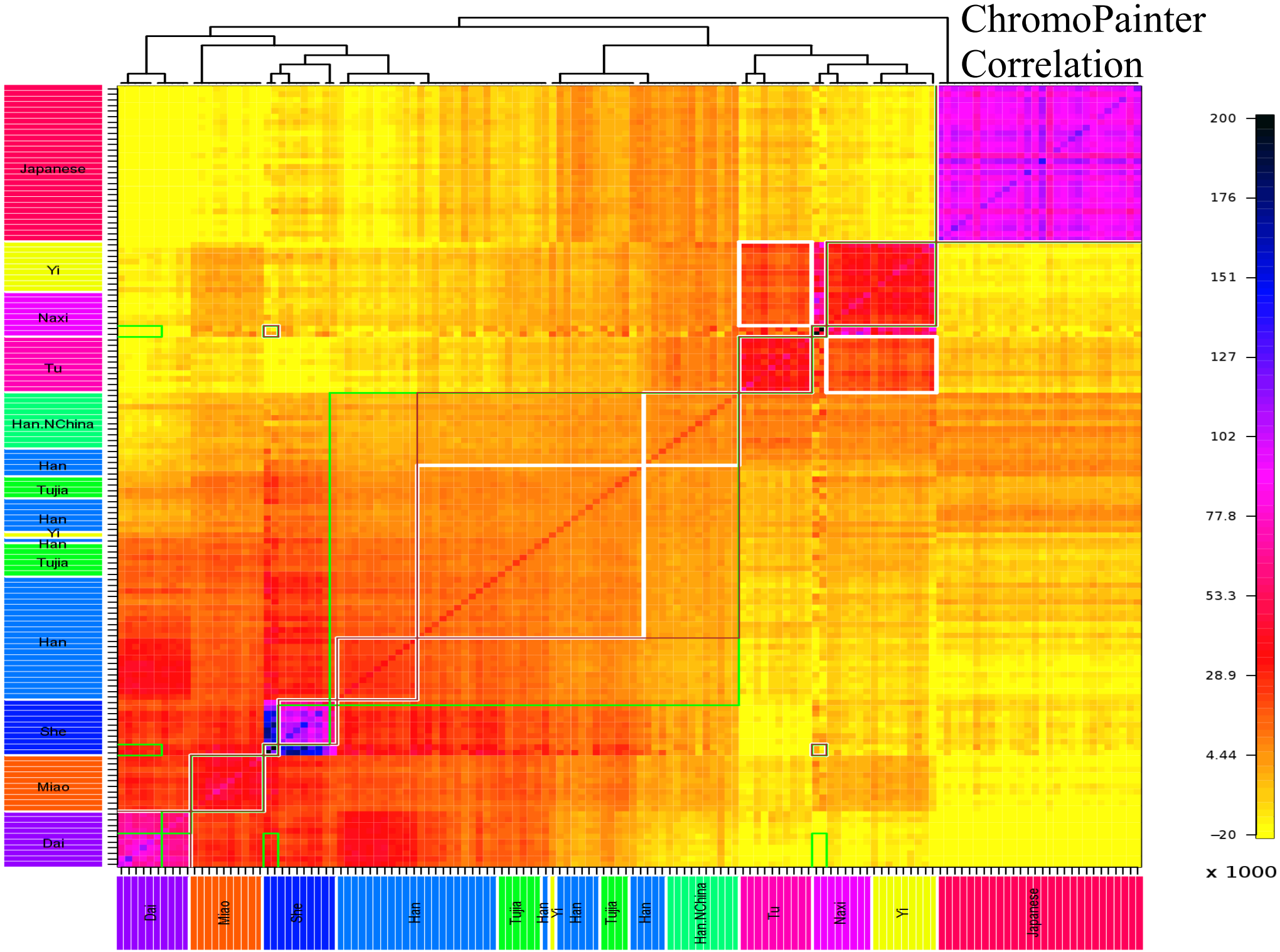


# HGDP data

- 938 Individuals worldwide
- 650k SNPs, linked (but relatively weakly)
- Known to contain structure at all scales, but previous models missed this
- Similarity approaches can analyse the whole data
- Focus on East Asian individuals


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


ChromoPainter
Eigenvector space
Japanese
Yi
Naxi
Tu
Han.NChina
Han
Tujia
Han
Yi
Han
Tujia
Han
She
Miao
Dai
0.15
0.128
0.106
0.0833
0.0611
0.0389
0.0167
−0.00556
−0.0278
−0.05
Dai
Miao
She
Han
Tujia
Han
Yi
Han
Tujia
Han
Han.NChina
Tu
Naxi
Yi
Japanese


ChromoPainter
Correlation
Japanese
Yi
Naxi
Tu
Han.NChina
Han
Tujia
Han
Yi
Han
Tujia
Han
She
Miao
Dai
200
176
151
127
102
77.8
53.3
28.9
4.44
−20
x 1000
Dai
Miao
She
Han
Tujia
Han
Yi
Han
Tujia
Han
Han.NChina
Tu
Naxi
Yi
Japanese

# Conclusions

- In General
  - Models matter
  - Summaries are part of the model
  - 'Model-Free' approaches are still making assumptions!
- Genetics
  - Normalising variance is important – other unlinked measures are all worse
  - Linked models extract more information
  - No 'correct' linked model at this stage!
  - ChromoPainter/FineSTRUCTURE pipeline is the most robust option

# Acknowledgements:

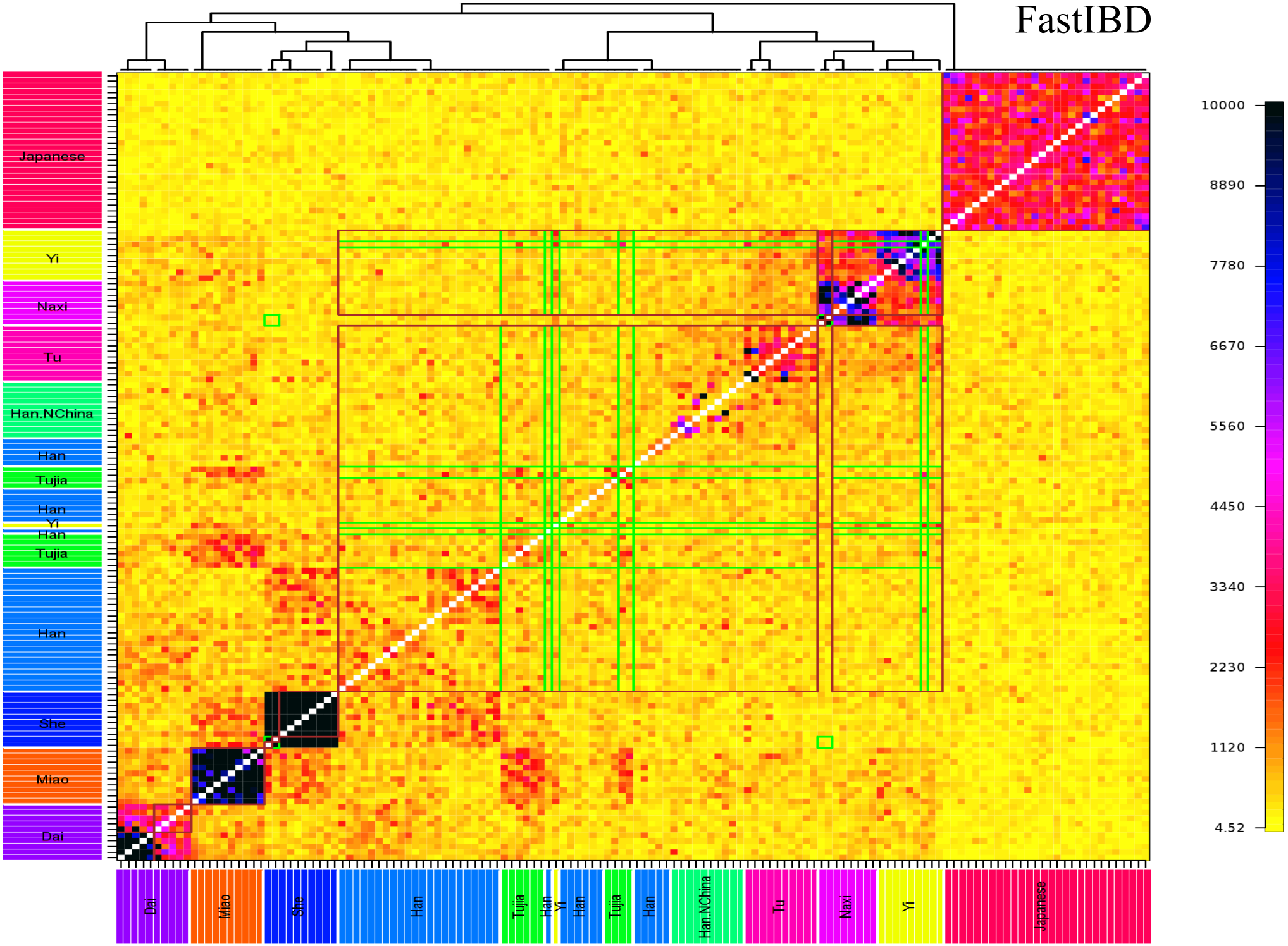
## fineSTRUCTURE

Garrett Hellenthal
(Oxford)
*(CP algorithm)*

Simon Myers
(Oxford)
*(theory)*

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

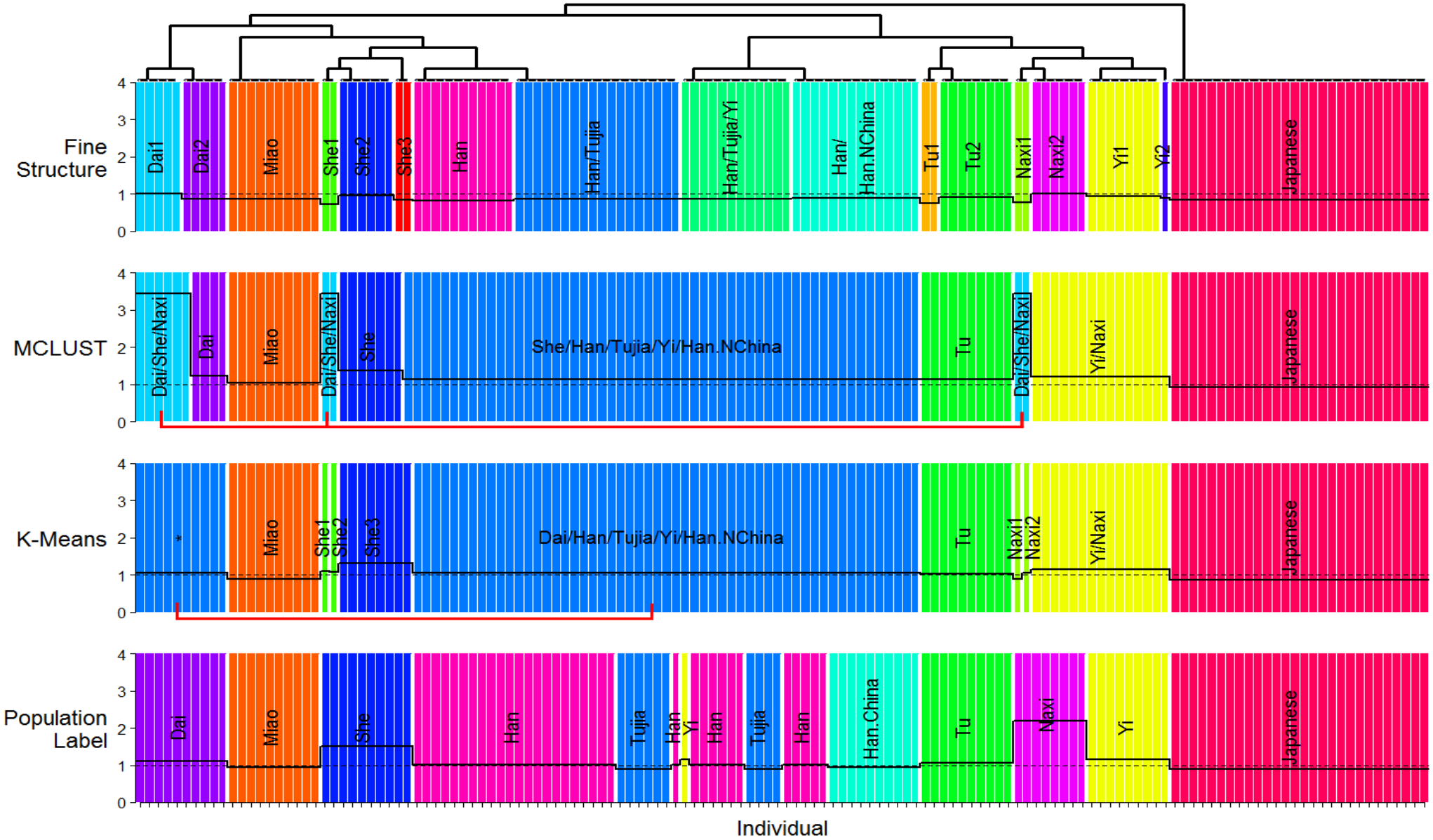
- Peter Green (Bristol) – Grant, support
- Bluecrystal HPC facilities @ Bristol
- FineSTRUCTURE Code & GUI: www.paintmychromosomes.com

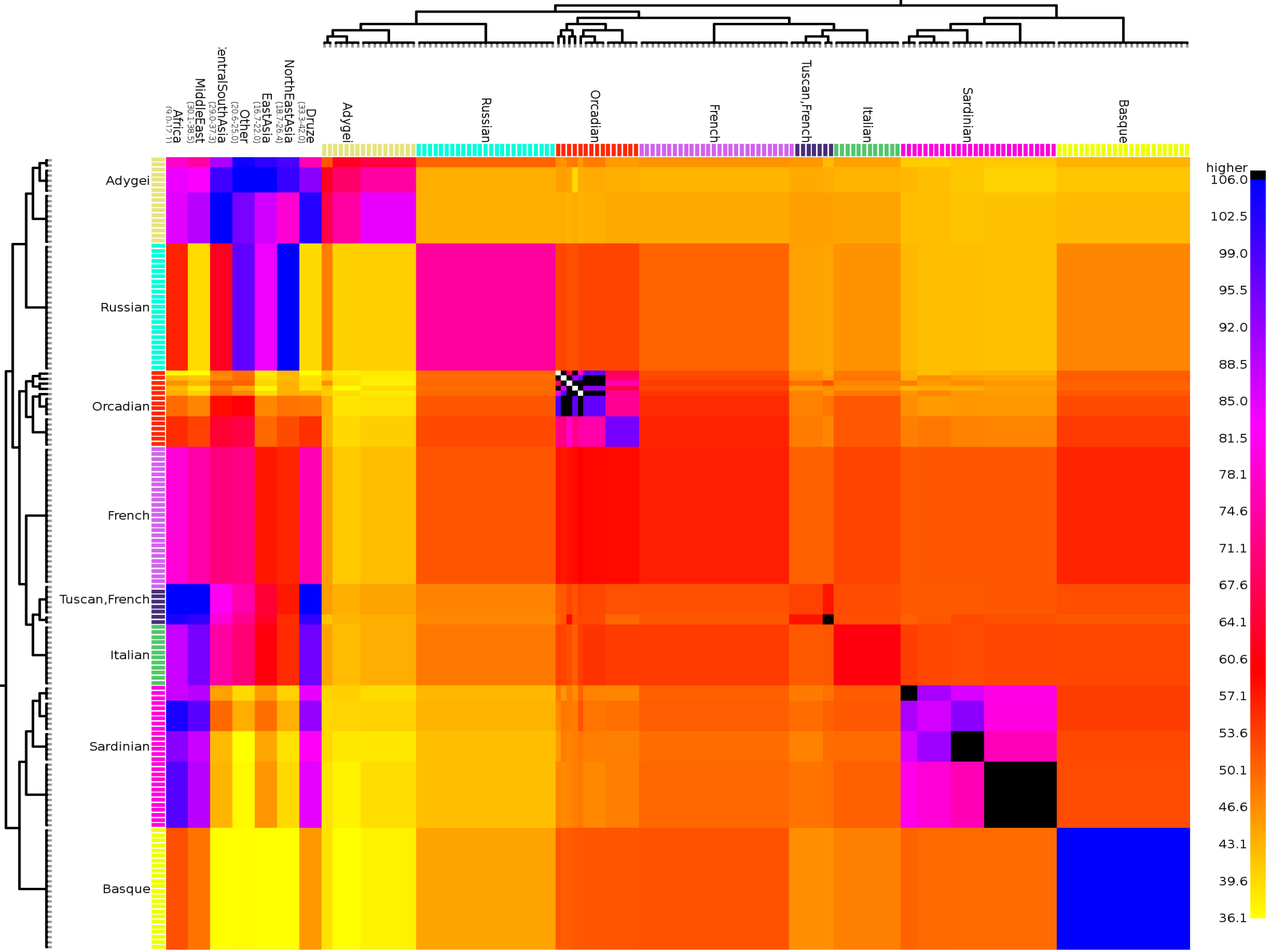

FastIBD
Japanese
Yi
Naxi
Tu
Han.NChina
Han
Tujia
Han
Yi
Han
Tujia
Han
She
Miao
Dai
Dai
Miao
She
Han
Tujia
Han
Yi
Han
Tujia
Han
Han.NChina
Tu
Naxi
Yi
Japanese
10000
8890
7780
6670
5560
4450
3340
2230
1120
4.52

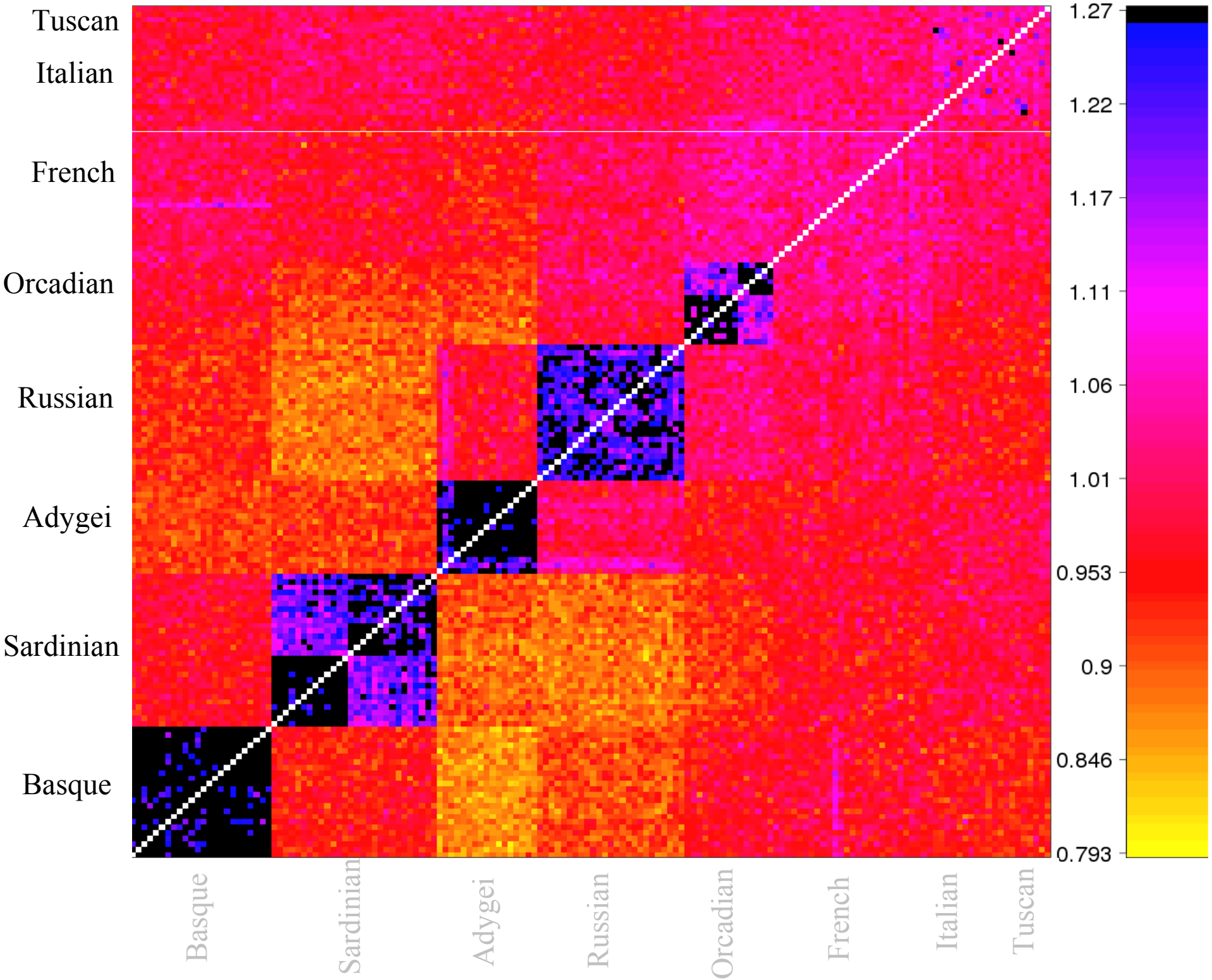
# FastIBD (Browning and Browning 2011)

- 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

# Comparison of clusterings

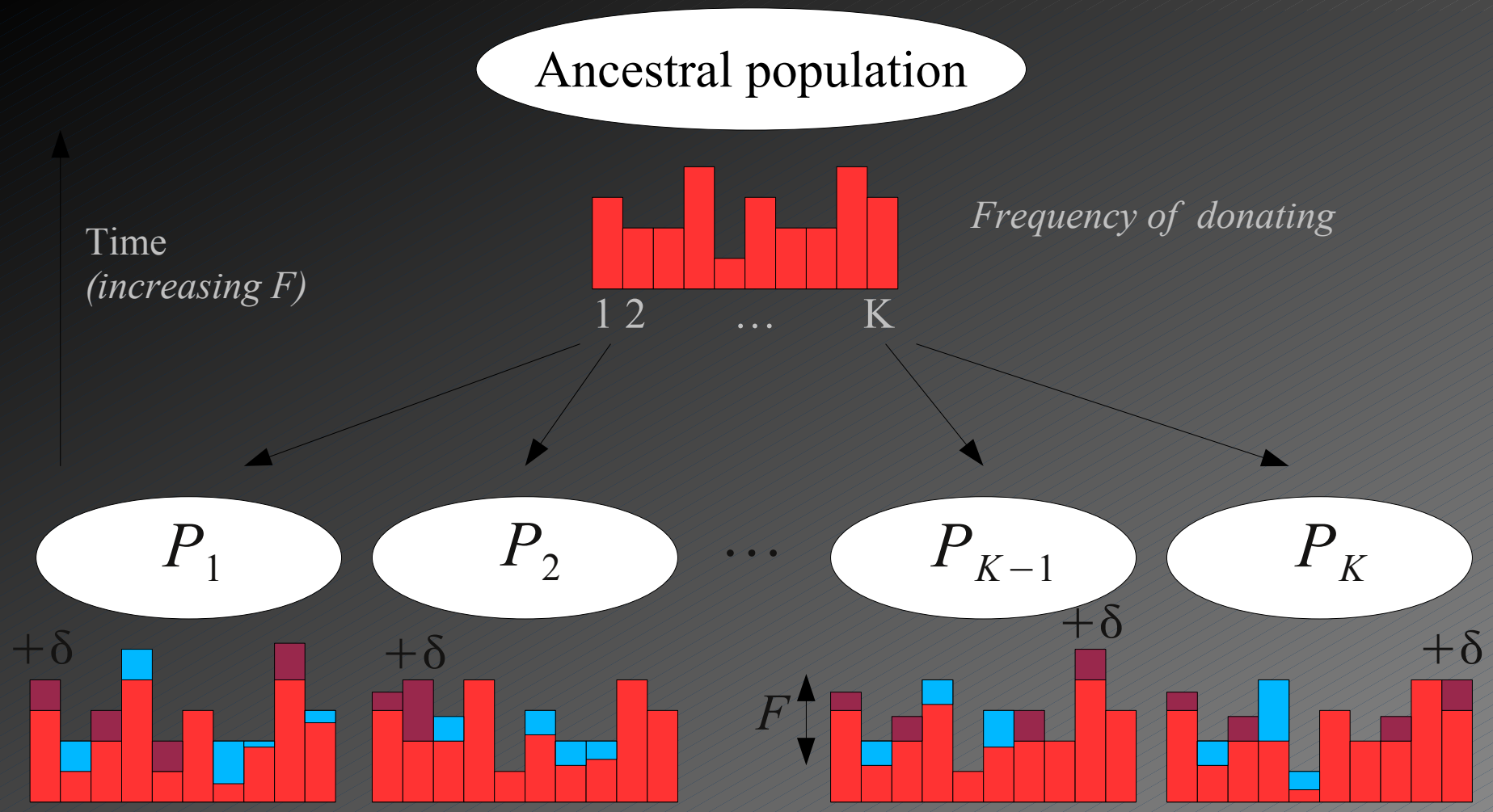



Africa (9.0-17.1)
MiddleEast (30.1-38.5)
CentralSouthAsia (29.0-37.3)
Other (20.6-25.0)
EastAsia (16.7-22.0)
NorthEastAsia (18.7-26.4)
Druze (33.3-42.0)
Adygei
Russian
Orcadian
French
Tuscan,French
Italian
Sardinian
Basque
higher
106.0
102.5
99.0
95.5
92.0
88.5
85.0
81.5
78.1
74.6
71.1
67.6
64.1
60.6
57.1
53.6
50.1
46.6
43.1
39.6
36.1


Tuscan
Italian
French
Orcadian
Russian
Adygei
Sardinian
Basque
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




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

# 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: $L \gg N \gg D$

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

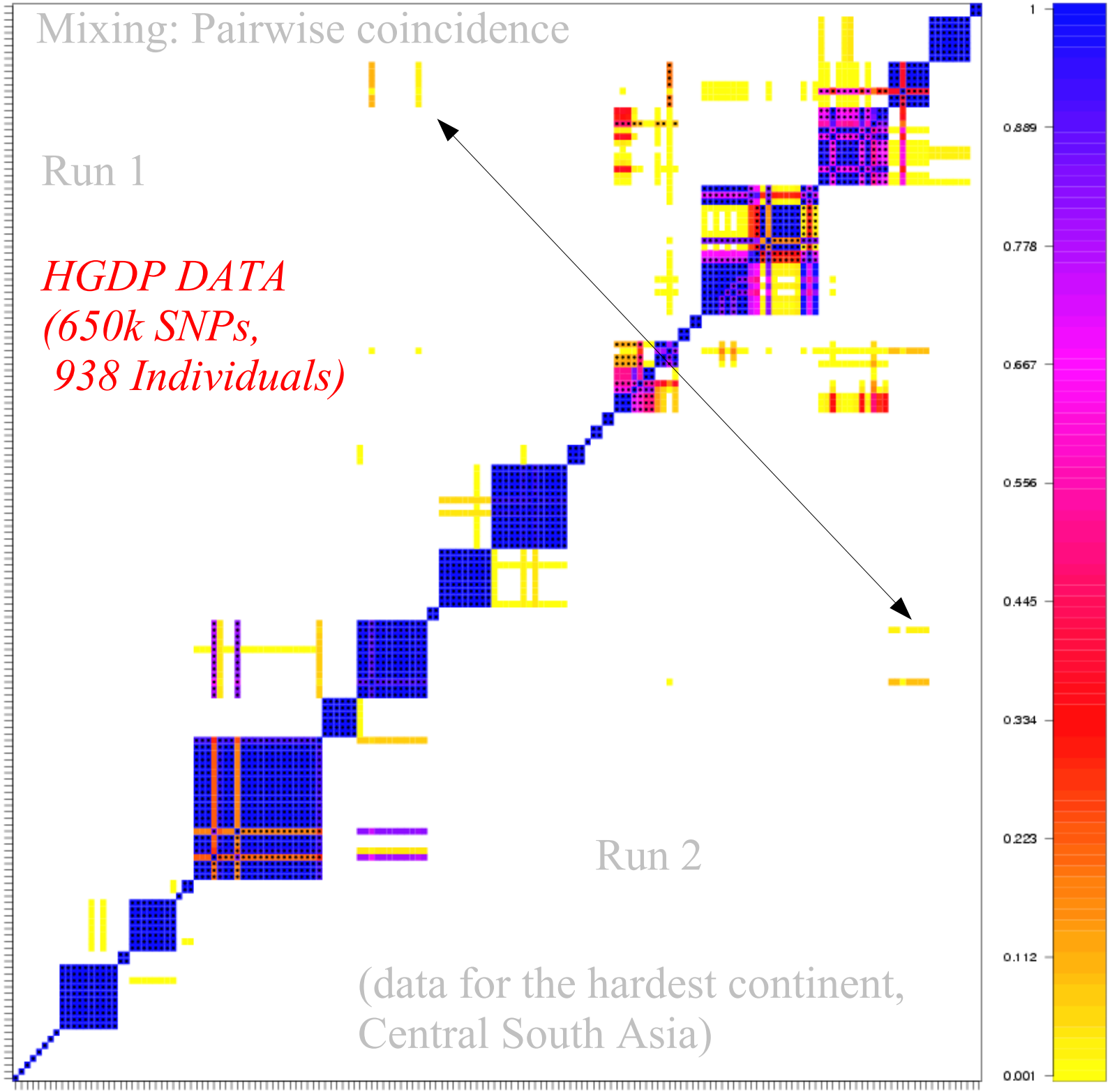
- Lowest dimension description usually best...
- Raw data approach terrible here (without good model)

# 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

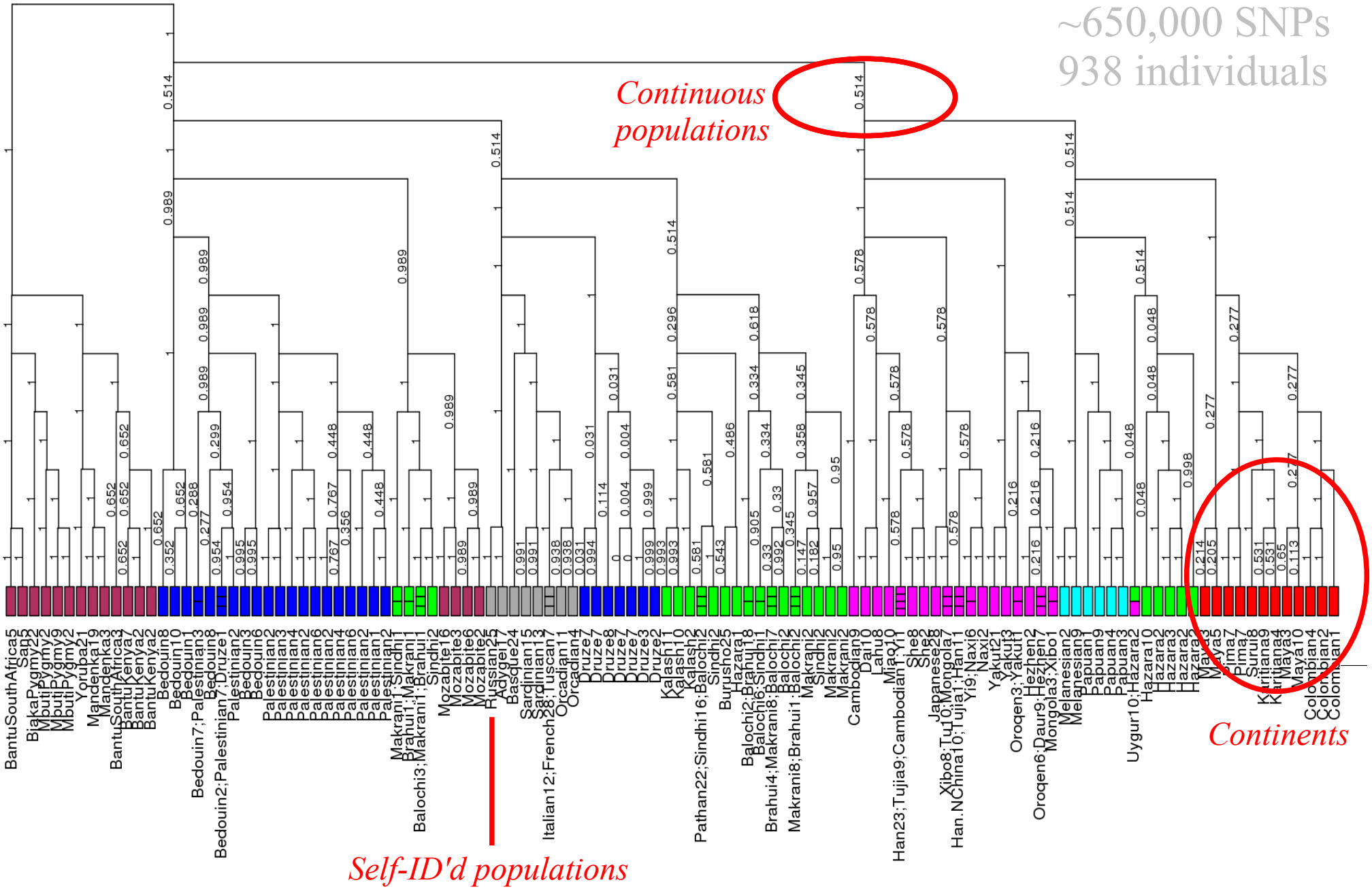
Daniel Lawson, University of Bristol
dan.lawson@bristol.ac.uk

*Wooding and Rogers, Genetics, 161:1641-1650, 2002
**Neal, in J. M. Bernardo, et al. (ed.), Bayesian Statistics 7, pp. 619-629, 2003


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

# MAP tree: whole world HGDP data

~650,000 SNPs
938 individuals



# 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

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

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

(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

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

*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 $X$)