A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots

Daniel Falush, Dan Lawson,

Lucy van Dorp

The identification of genetically homogeneous groups of individuals is a long standing issue in population genetics. A recent Bayesian algorithm implemented in the software STRUCTURE allows the identification of such groups. However, the ability of this algorithm to detect the true number of clusters (*K*) in a sample of individuals when patterns of dispersal among populations are not homogeneous has not been tested. The goal of this study is to

There are also biological reasons to be careful interpreting *K*. The population model that we have adopted here is obviously an idealization. We anticipate that it will be flexible enough to permit appropriate clustering for a wide range of population structures. However, as we pointed out in our discussion of data set 3 (*Choice of K for simulated data*), clusters may not necessarily correspond to "real" populations. As another example, imagine a species that lives on a continuous plane, but has low dispersal rates, so that allele frequencies vary continuously across the plane. If we sample at *K* distinct locations, we might infer the presence of *K* clusters, but the inferred number *K* is not *biologically* interesting, as it was determined purely by the sampling scheme. All that sometimes depend on the model used. The *F* model is in general more permissive of additional populations being fitted to a data set, as it permits the existence of two or more populations with very similar allele frequencies (particularly if the prior on *F* is chosen to favor small values). Consequently, $P(X|K)$ is sometimes maximized for a higher value of *K* than under the uncorrelated model. This cuts to the heart of one of the principal reasons why inferring *K* is so difficult and why estimates for *K* should be treated with caution: the number of populations supported by the data may depend on how different one would expect allele frequencies in the different populations to be *a priori*, which is often difficult to specify.

For some data sets, higher estimates of *K* obtained using the *F* model may reflect deviations from random assortment that are not caused by genuine population subdivision. Table 1A shows model likelihoods esti-

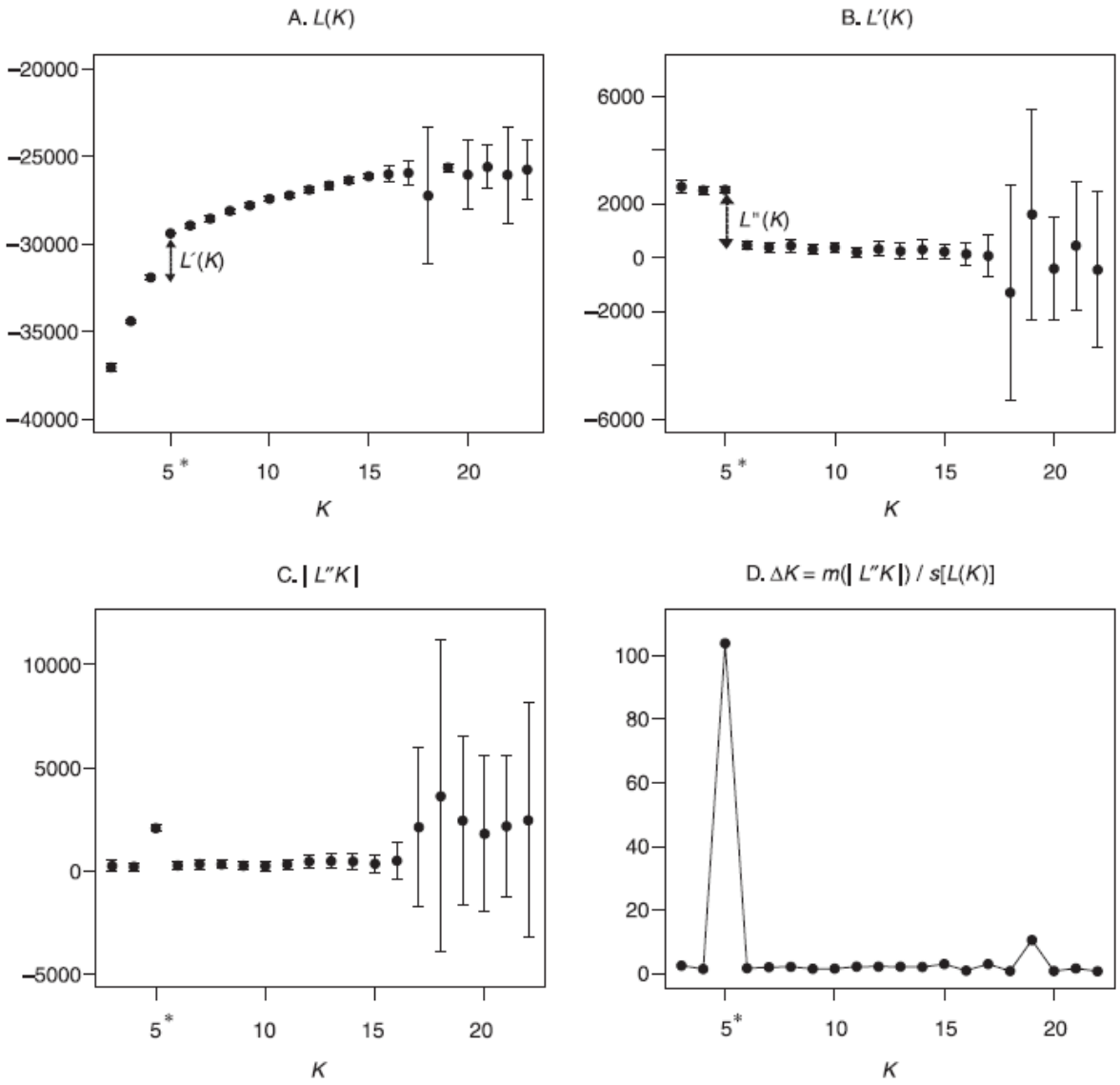
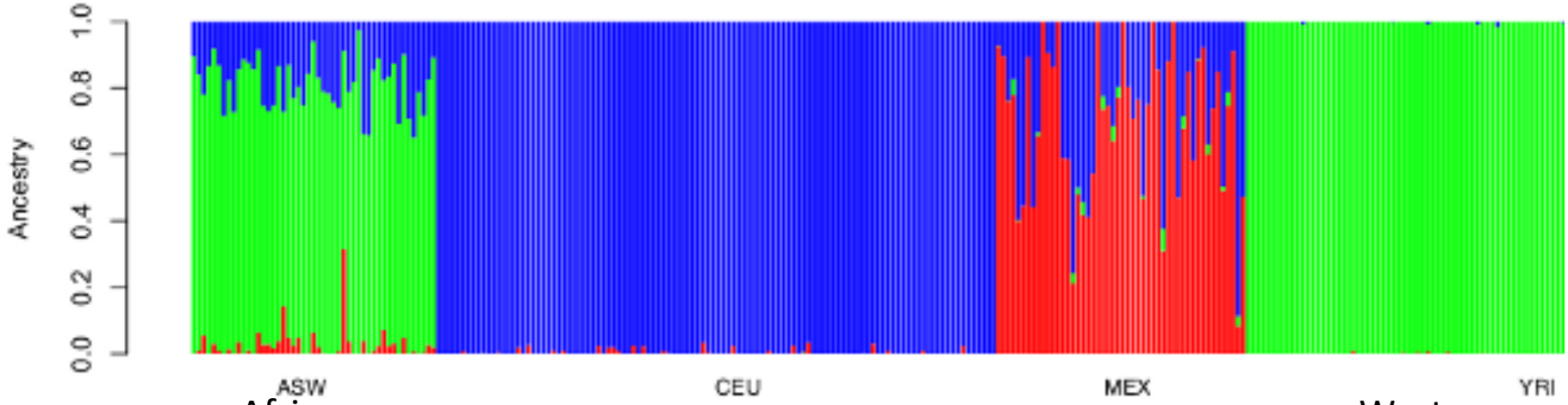


Fig. 2 Description of the four steps for the graphical method allowing detection of the true number of groups K^*. (A) Mean L(K) (± SD) over 20 runs for each K value. The model considered here is a hierarchical island model using all 100 individuals per population and 50 AFLP loci. (B) Rate of change of the likelihood distribution (mean ± SD) calculated as $L'(K) = L(K) - L(K-1)$. (C) Absolute values of the second order rate of change of the likelihood distribution (mean ± SD) calculated according to the formula: $|L''(K)| = |L'(K+1) - L'(K)|$. (D) ΔK calculated as $\Delta K = m|L''(K)| / s[L(K)]$. The modal value of this distribution is the true K(*) or the uppermost level of structure, here five clusters.

STRUCTURE/ADMIXTURE bar plots represent ancestry proportions after recent admixture



African Americans

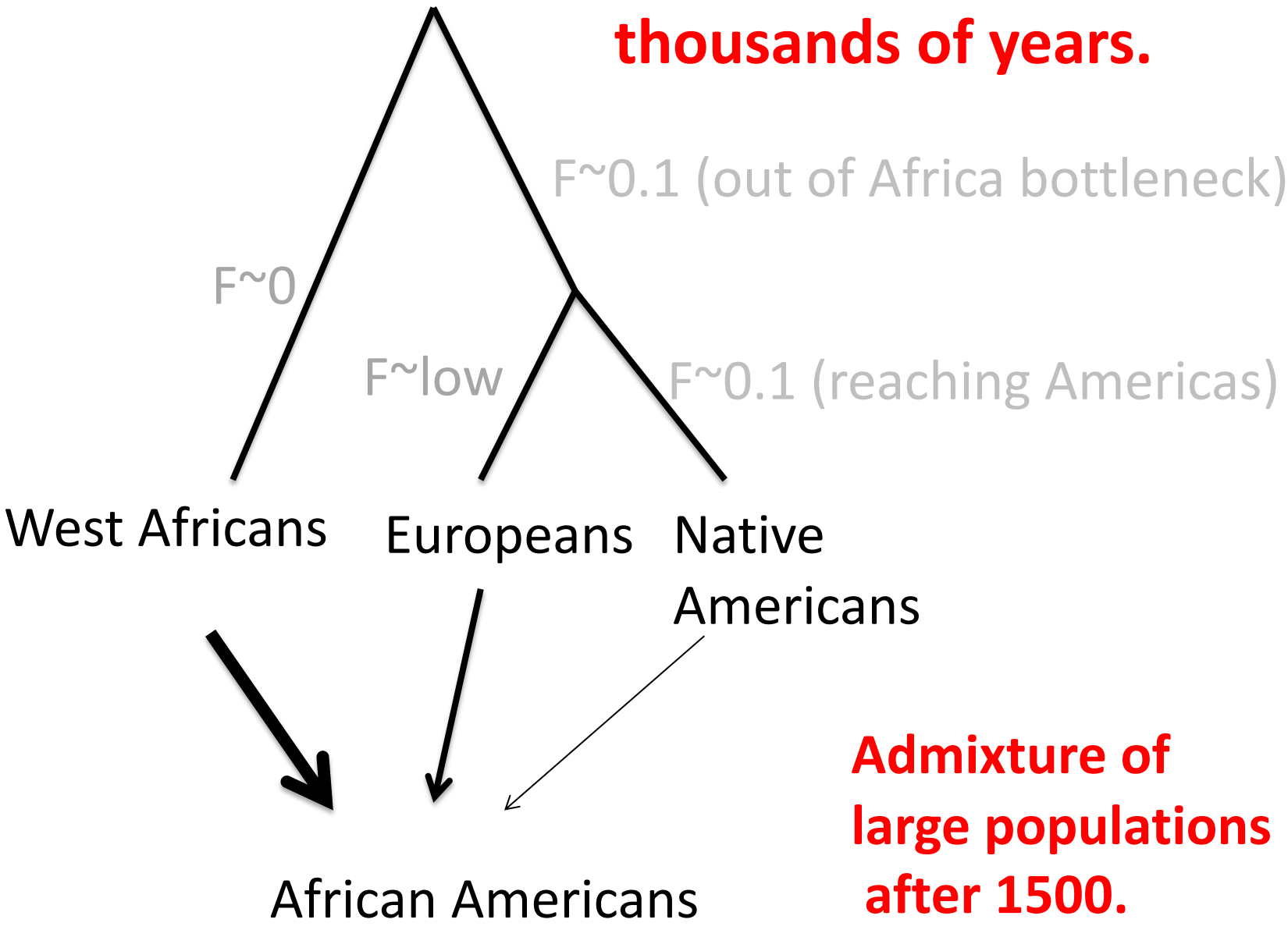
Europeans

Hispanics

West Africans

1000 genomes project

Genetic drift over thousands of years.



Admixture of large populations after 1500.

Procedure for (over) interpreting STRUCTURE results

(1) Estimate *K* using a refined statistical procedure.

(2) Assume that at this is the true value of *K*.

(3) Assume each of the *K* ancestral population existed at some point in the past.

(4) Assume that modern individuals were produced by recent mixing of these ancestral populations.

Treating ancestral population as atomic units of inheritance

(3a) Do not ask how the inferred ancestral populations are related to each other.

(3b) Neglect the possibility an ancestral population might itself be admixed.

(3c) Neglect substructure within the inferred ancestral populations.

(3d) Label ancestral populations based on the locations they are currently most frequent in.

PLOS | GENETICS

RESEARCH ARTICLE

Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference

Lucy van Dorp[1,2], David Balding[1,3], Simon Myers[4], Luca Pagani[5,6], Chris Tyler-Smith[5], Endashaw Bekele[7], Ayele Tarekegn[8], Mark G. Thomas[9], Neil Bradman[8], Garrett Hellenthal[1]*

1 University College London Genetics Institute (UGI), University College London, London, United Kingdom, **2** Centre for Mathematics and Physics in the Life Sciences and EXperimental Biology (CoMPLEX), University College London, London, United Kingdom, **3** Schools of BioSciences and of Mathematics & Statistics, University of Melbourne, Melbourne, Australia, **4** Department of Statistics, University of Oxford, Oxford, United Kingdom, **5** The Wellcome Trust Sanger Institute, Hinxton, United Kingdom, **6** Department of Archaeology and Anthropology, University of Cambridge, Cambridge, United Kingdom, **7** Addis Ababa University, Addis Ababa, Ethiopia, **8** Henry Stewart Group, London, United Kingdom, **9** Research Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

* g.hellenthal@ucl.ac.uk

Citation: van Dorp L, Balding D, Myers S, Pagani L, Tyler-Smith C, Bekele E, et al. (2015) Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference. PLoS Genet 11(8): e1005397. doi:10.1371/journal.pgen.1005397

Editor: Anna Di Rienzo, University of Chicago, UNITED STATES

Received: December 16, 2014

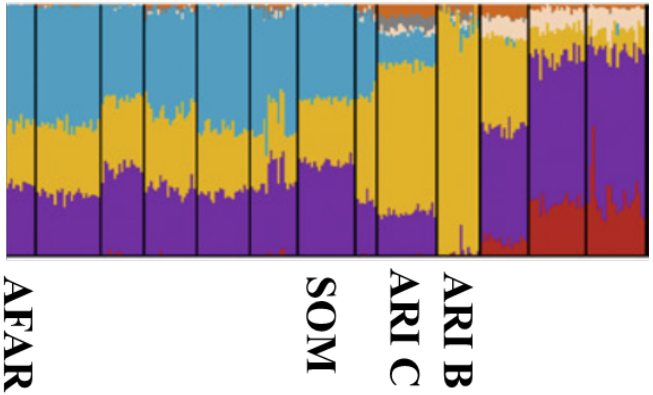
Accepted: June 26, 2015

Published: August 20, 2015

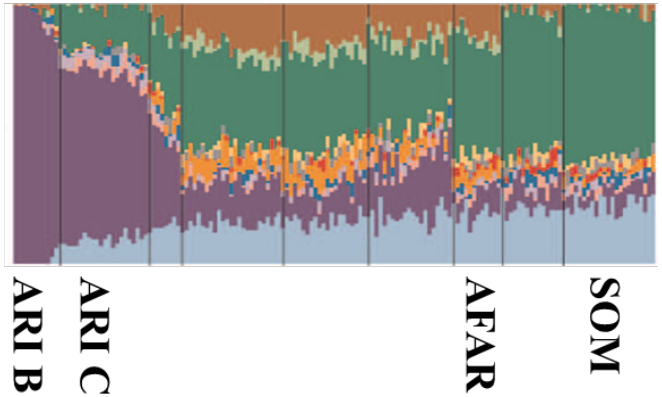
Abstract

The Ari peoples of Ethiopia are comprised of different occupational groups that can be distinguished genetically, with Ari Cultivators and the socially marginalised Ari Blacksmiths recently shown to have a similar level of genetic differentiation between them ($F_{ST} \approx 0.023 - 0.04$) as that observed among multiple ethnic groups sampled throughout Ethiopia. Anthropologists have proposed two competing theories to explain the origins of the Ari Blacksmiths as (i) remnants of a population that inhabited Ethiopia prior to the arrival of agri-

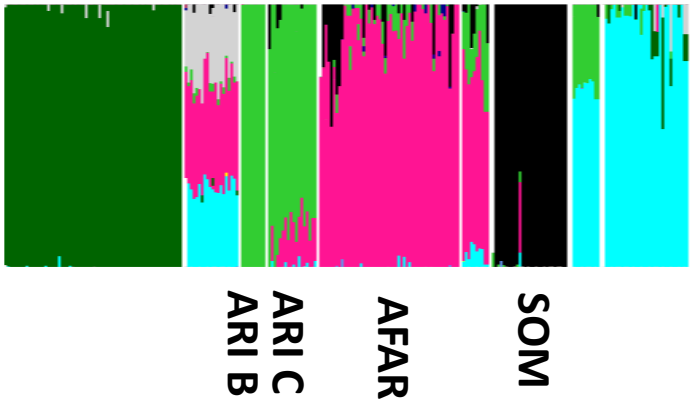
A



B



C



Three studies,
Three admixture plots,
Three colour schemes,
Three admixture histories,
Two interpretations.

One insight provided by the ADMIXTURE plot (Figure 1C) concerns the origin of the Ari Blacksmiths. This population is one of the occupational caste-like groups present in many Ethiopian societies that have traditionally been explained as either remnants of hunter-gatherer groups assimilated by the expansion of farmers in the Neolithic period or as groups marginalized in agriculturalist communities due to their craft skills.51 The prevalence of an Ethiopian-specific cluster (yellow in Figure 1C) in the Ari Blacksmith sample could favor the former scenario; the ancestors of this occupational group could have been part of a population that inhabited the area before the spread of agriculturalists.

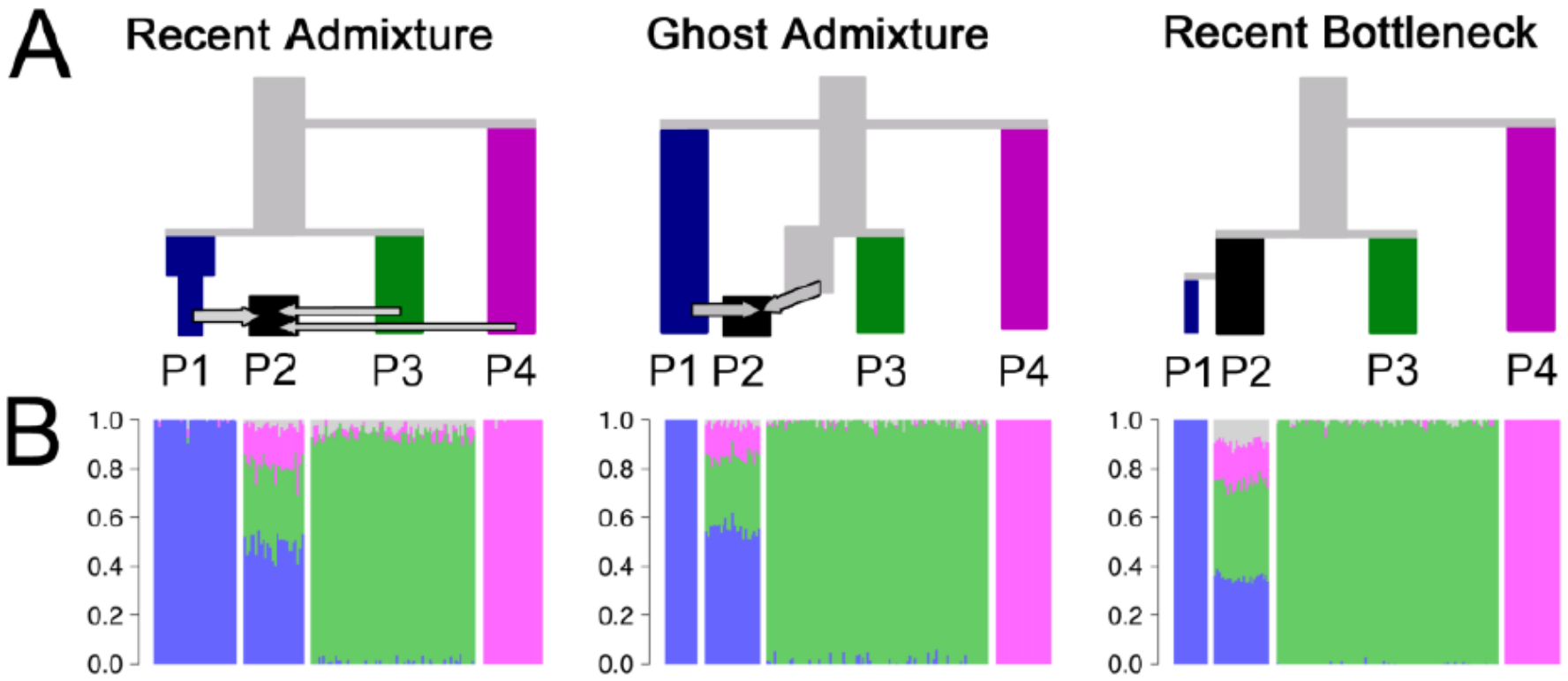
As the Ari Blacksmiths have negligible EthioSomali ancestry, it seems most likely that the Ari Cultivators are the descendents of a more recent admixture between a population like the Ari Blacksmiths and some other HOA population

ADMIXTURE results for three simulation scenarios

Recent admixture into Ari Cultivators

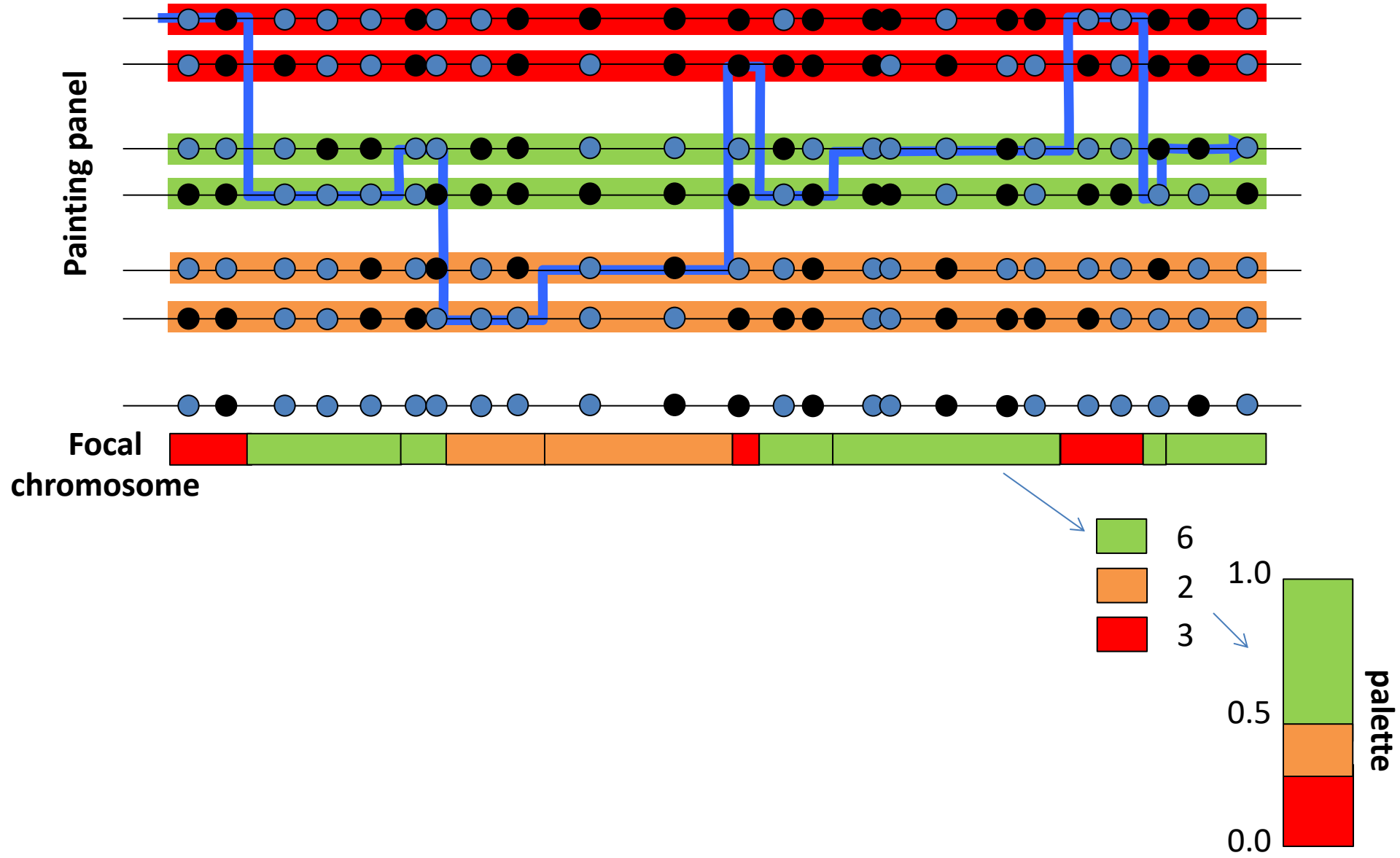
Ghost admixture into Ari Cultivators

Strong drift in Ari Blacksmiths

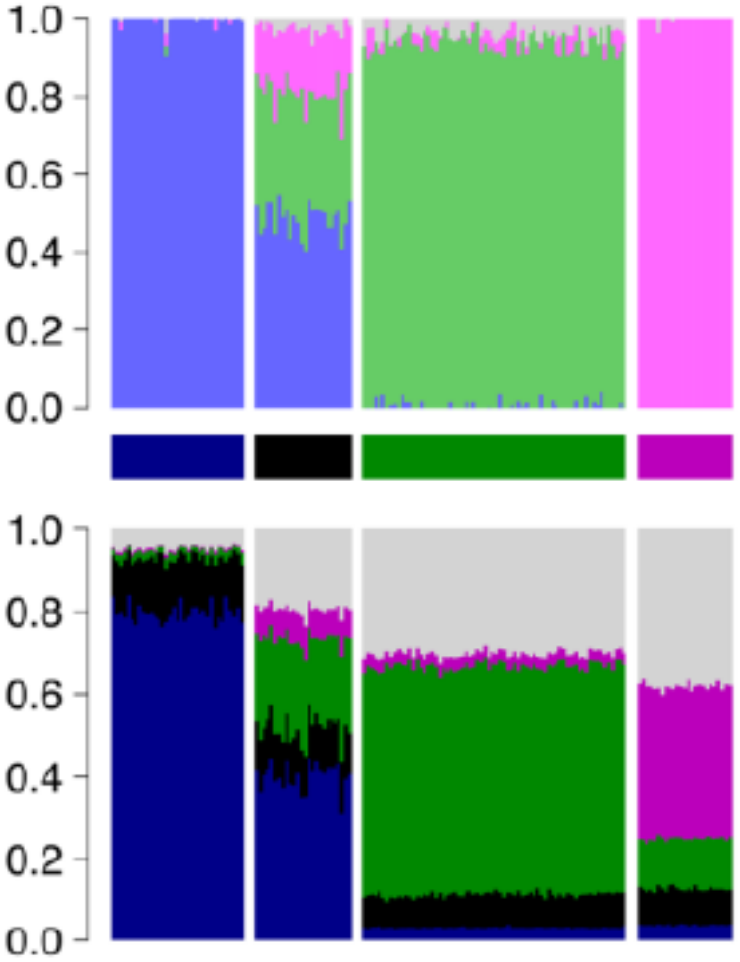


Chromosome painting

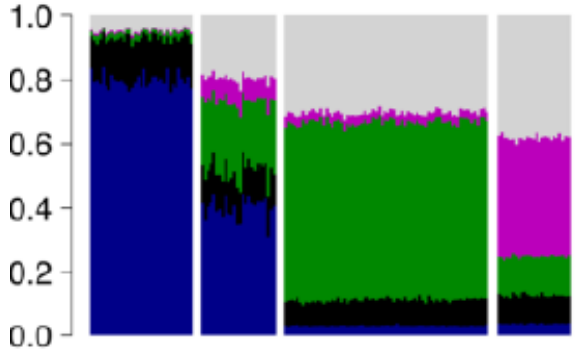
- (Lawson et al. 2012)



Idea: compare admixture profiles with painting palettes



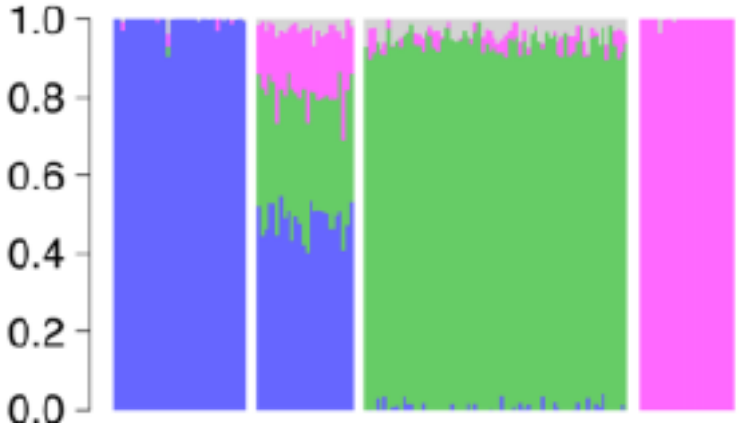
Under a recent admixture scenario, the palette of a admixed individual should be a mix of the palettes of non-admixed individuals



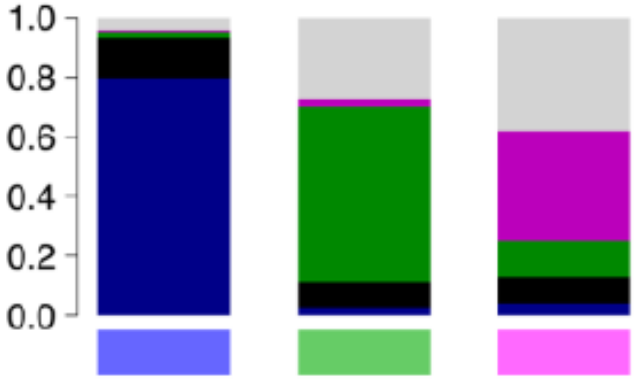
X=N*P
Individual palettes

N individuals
P Labelled populations
K Ancestral populations

M=K*P
Palettes for ancestral populations



A = N*K
Individual admixture proportions

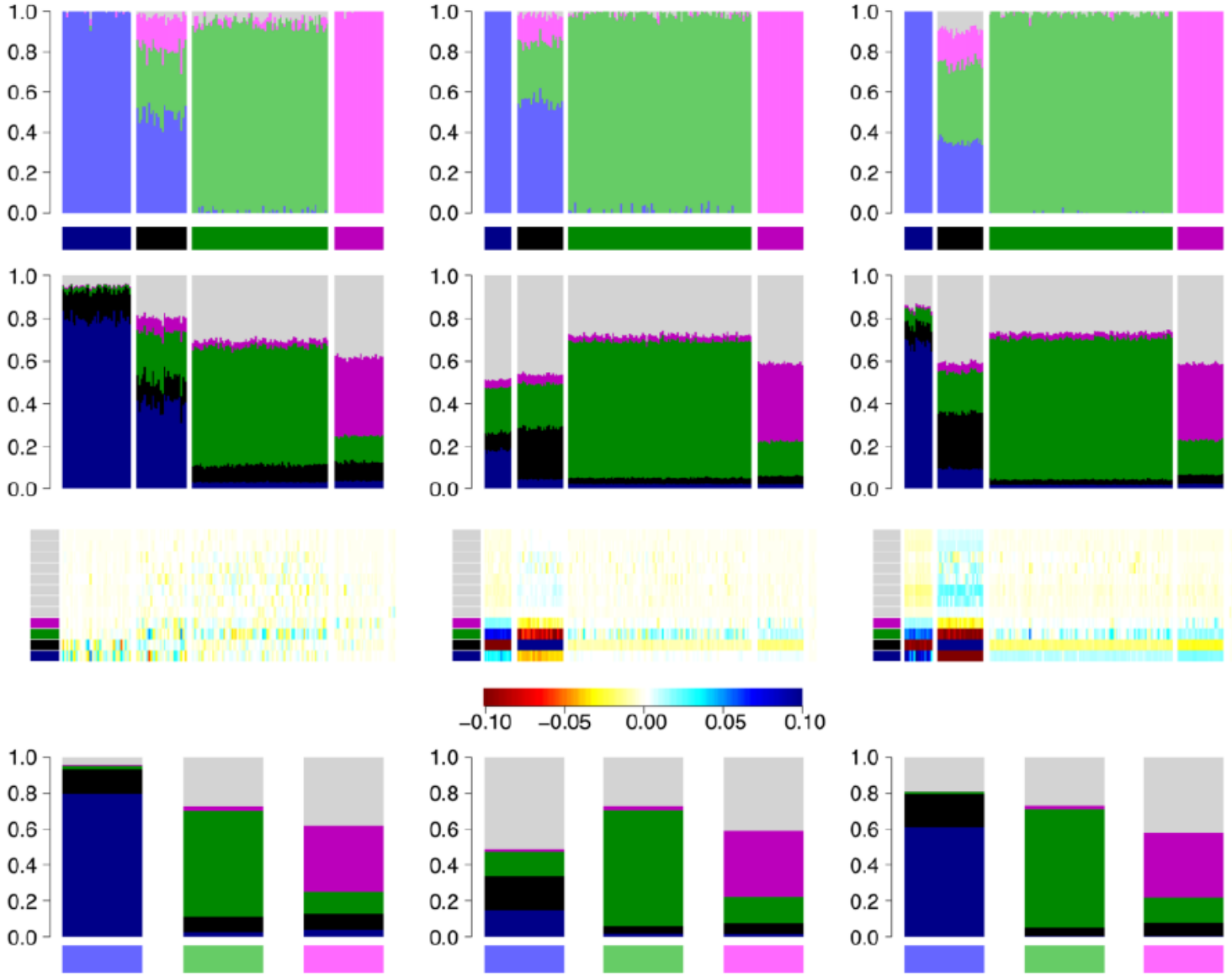


Choose M to minimize AM-X.

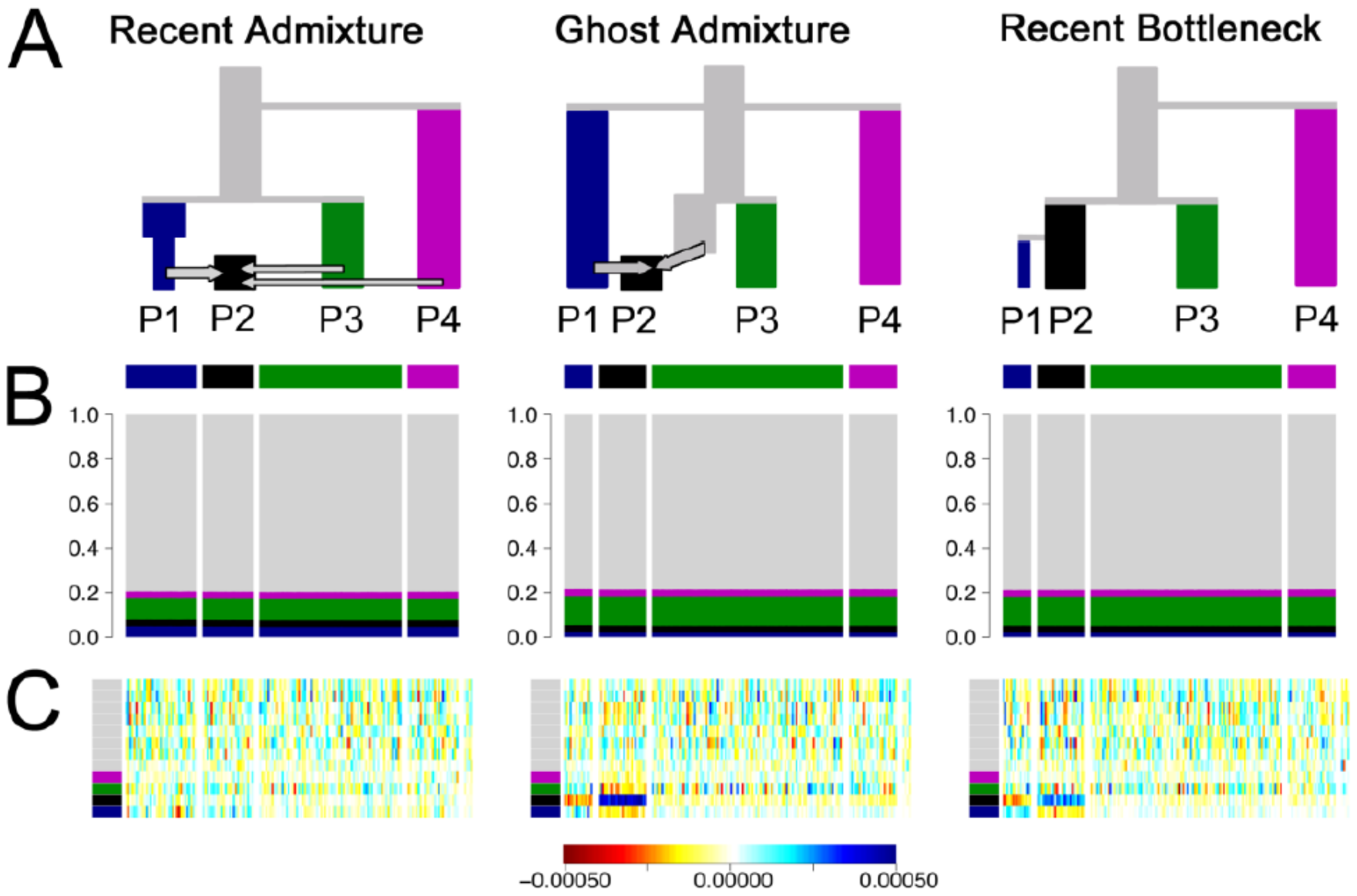
Recent admixture into Ari Cultivators

Ghost admixture into Ari Cultivators

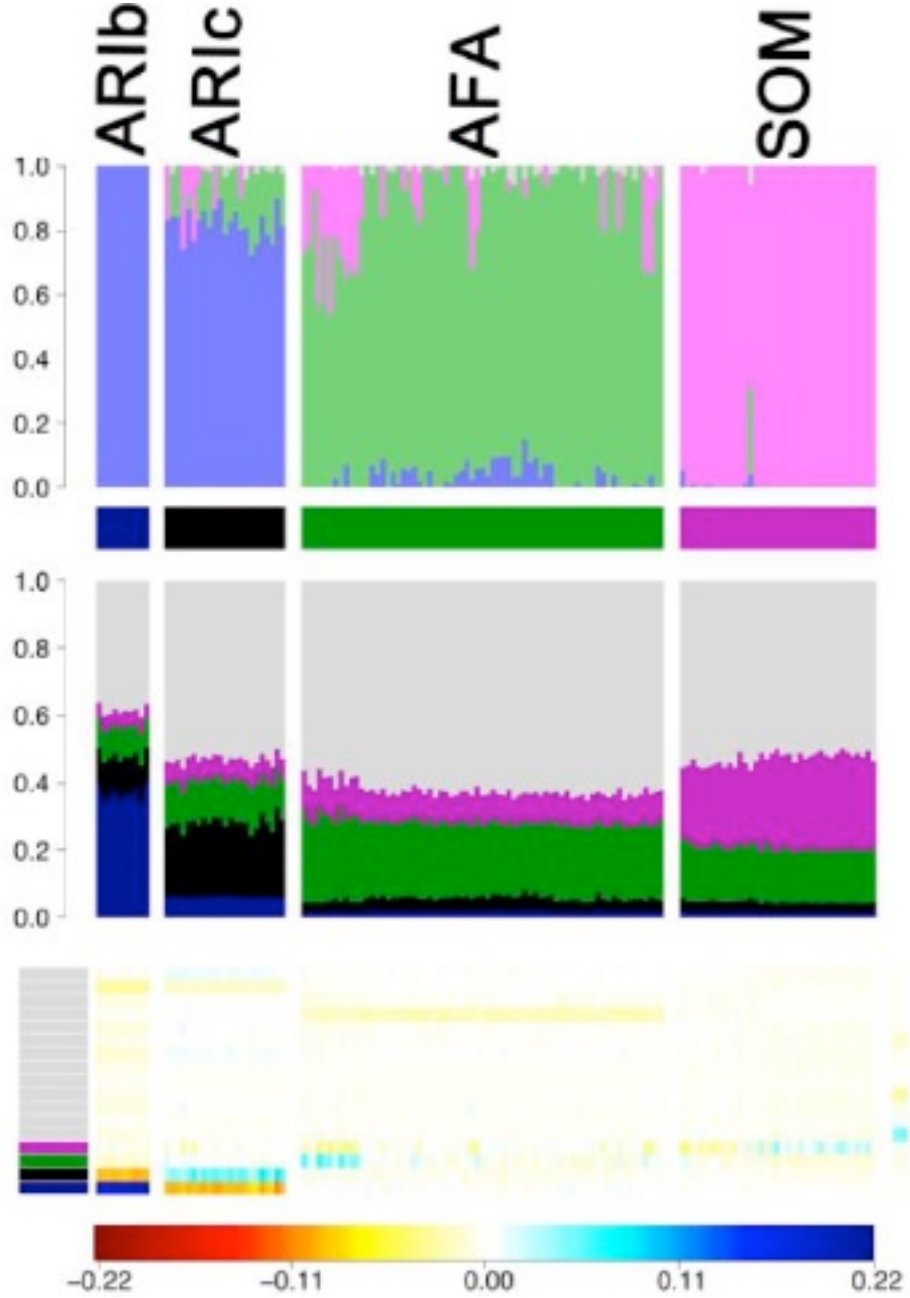
Strong drift in Ari Blacksmiths

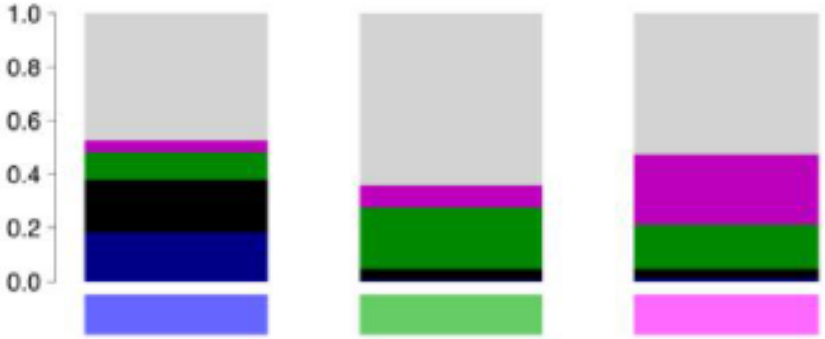


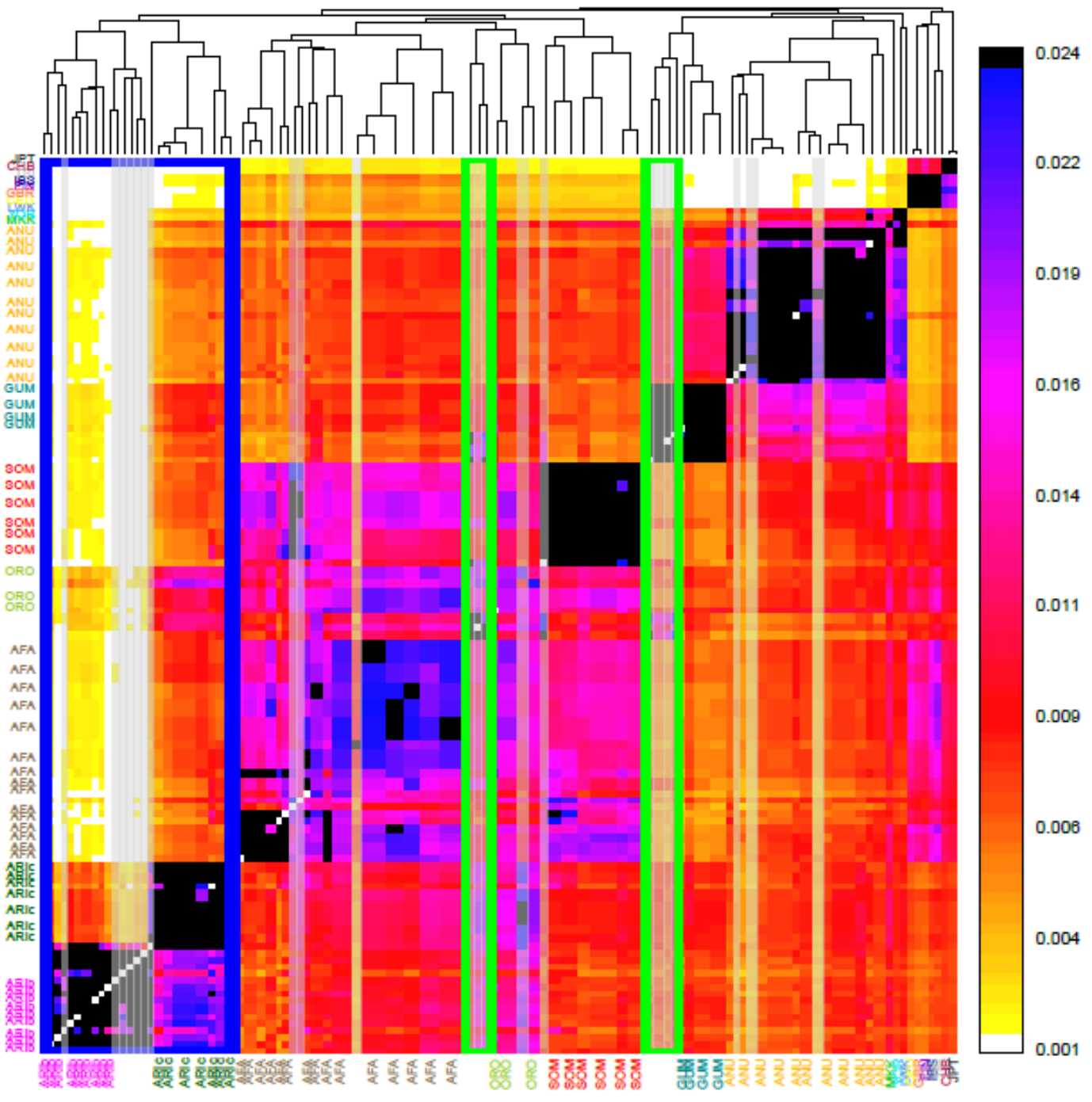
Painting while ignoring linkage

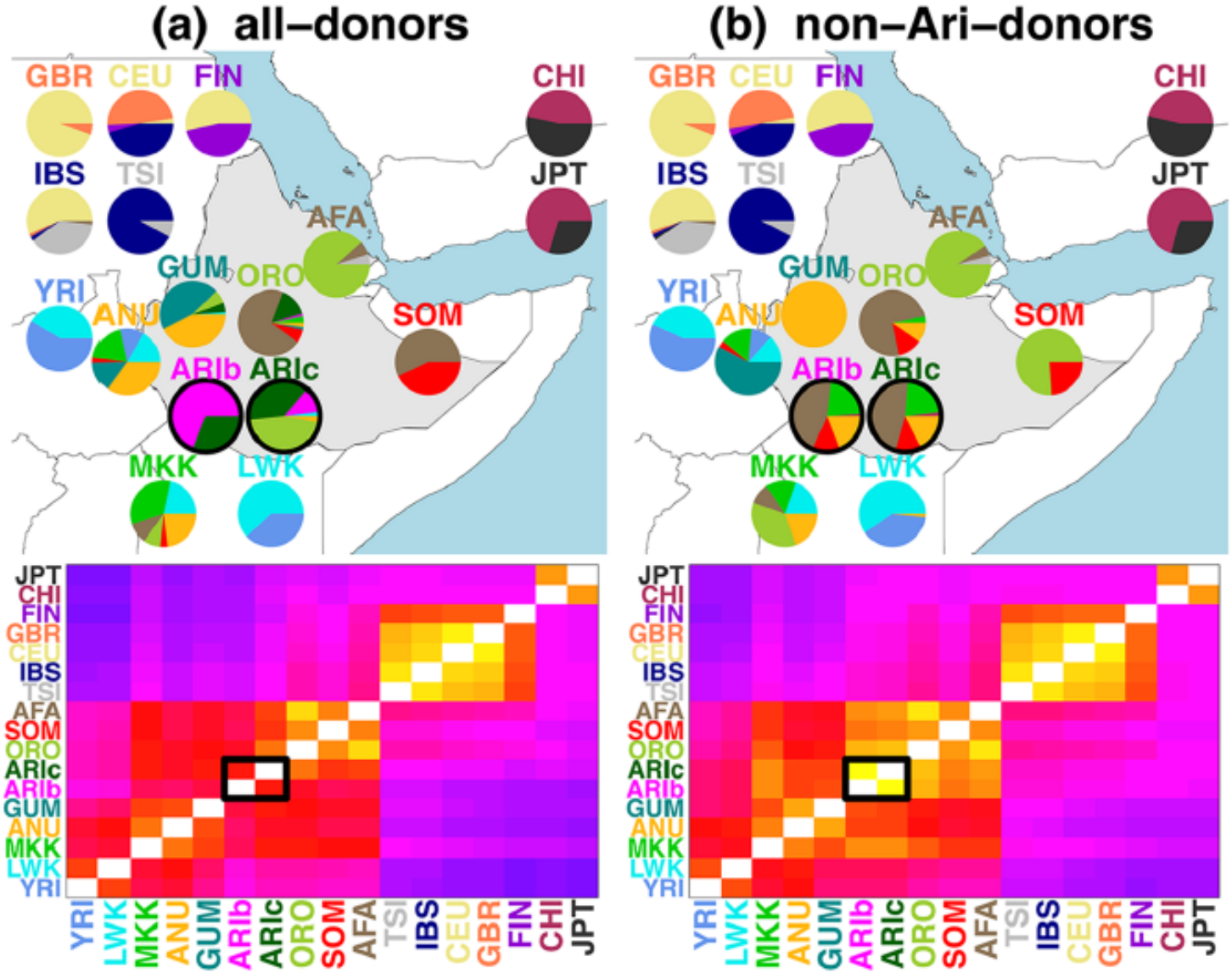


Real data


ARIb
ARIc
AFA
SOM
1.0
0.8
0.6
0.4
0.2
0.0
−0.22
−0.11
0.00
0.11
0.22

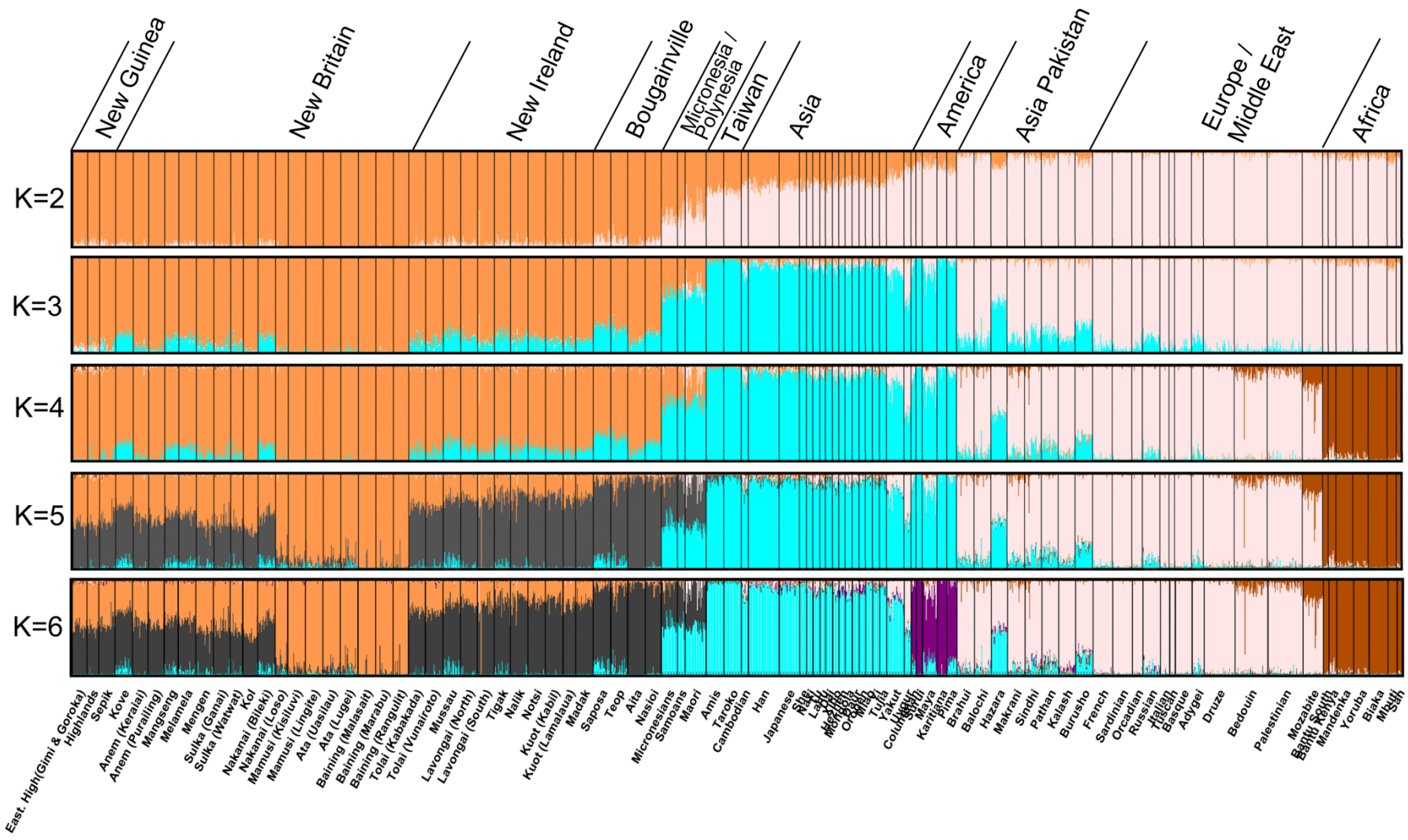

1.0
0.8
0.6
0.4
0.2
0.0


0.024
0.022
0.019
0.016
0.014
0.011
0.009
0.006
0.004
0.001

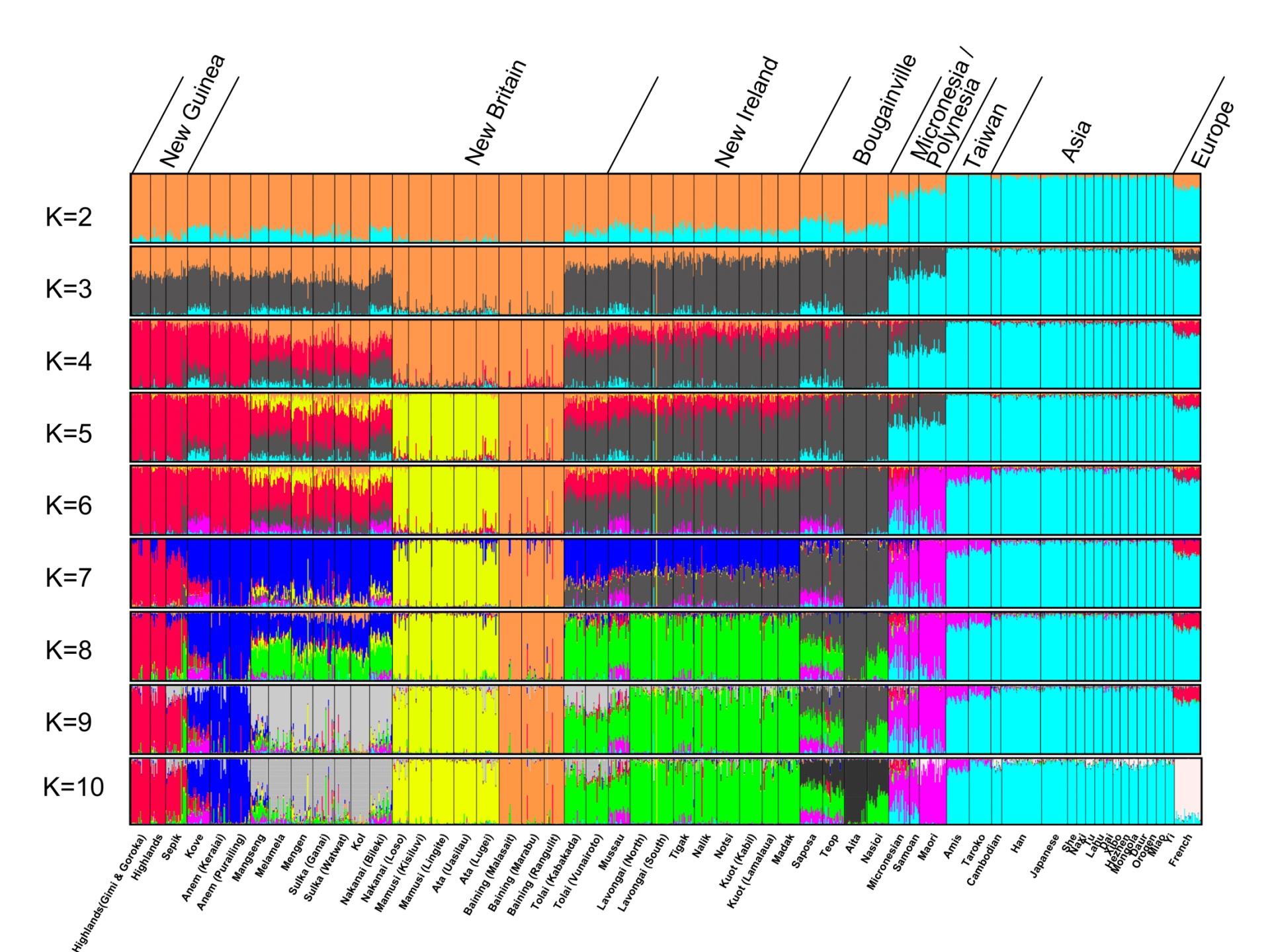

(a) all−donors
(b) non−Ari−donors
GBR
CEU
FIN
CHI
IBS
TSI
JPT
AFA
GUM
ORO
YRI
ANU
SOM
ARIb
ARIc
MKK
LWK
JPT
CHI
FIN
GBR
CEU
IBS
TSI
AFA
SOM
ORO
ARIc
ARIb
GUM
ANU
MKK
LWK
YRI

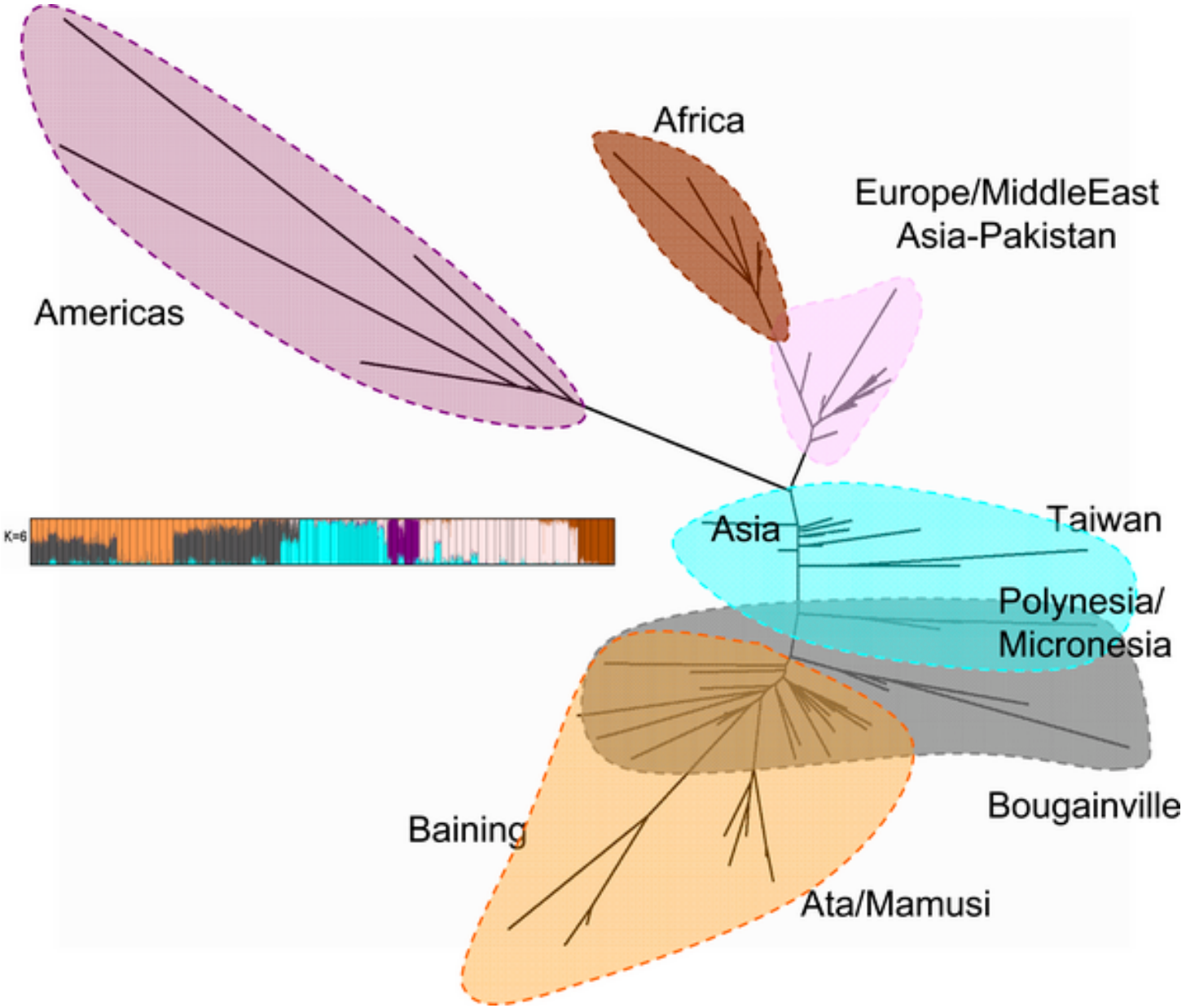
Fun with sampling

(0) Make sure to over-sample your favorite group.
(2a) If your favorite group does not have its own population, increase K until it does.



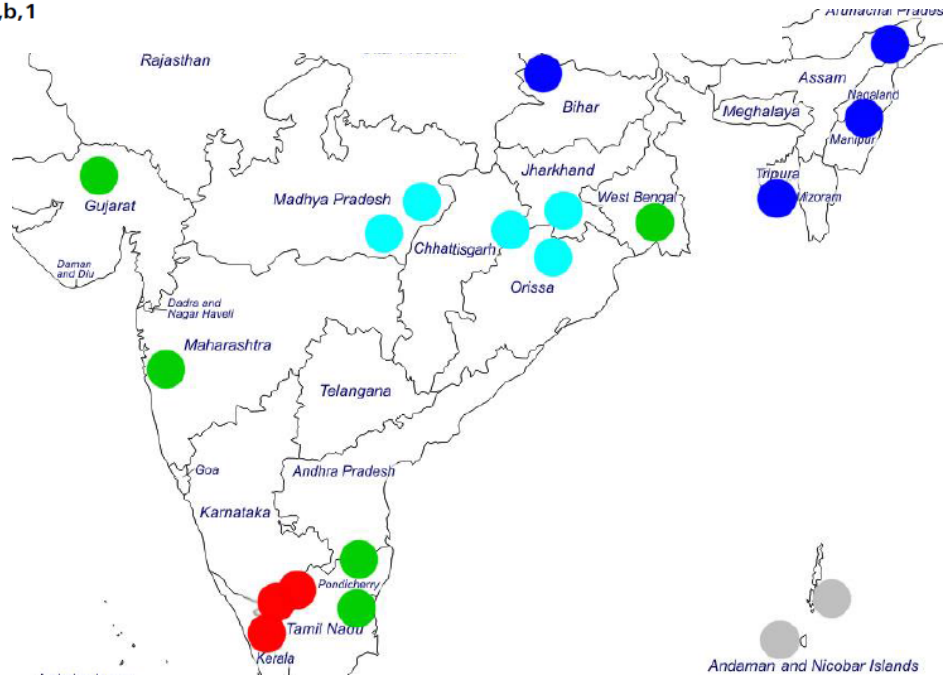
Friedlander 2008

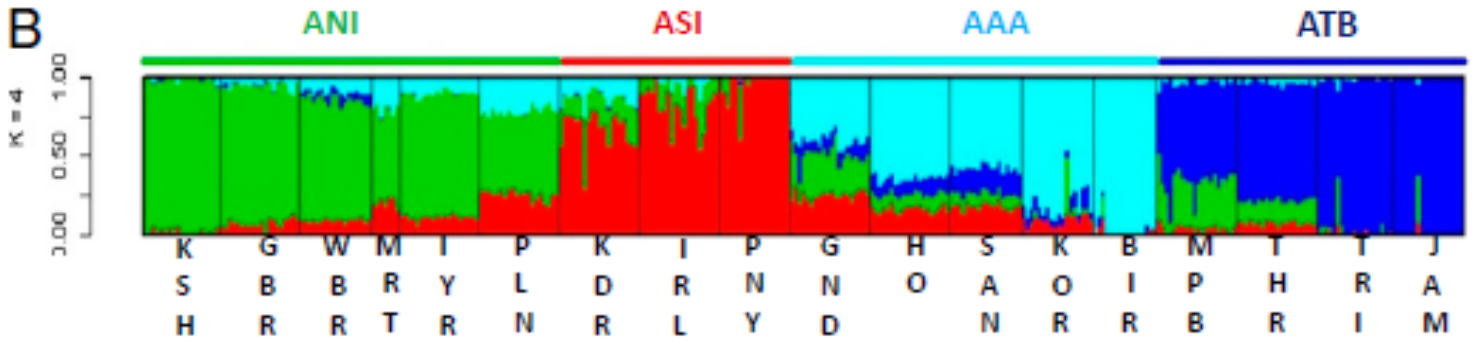

New Guinea
New Britain
New Ireland
Bougainville
Micronesia / Polynesia
Taiwan
Asia
Europe
K=2
K=3
K=4
K=5
K=6
K=7
K=8
K=9
K=10
Highlands(Gimi & Goroka)
Highlands
Sepik
Kove
Anem (Keraiai)
Anem (Purailing)
Mangseng
Melamela
Mengen
Sulka (Ganai)
Sulka (Watwat)
Kol
Nakanai (Bileki)
Nakanai (Loso)
Mamusi (Kisiluvi)
Mamusi (Lingite)
Ata (Uasilau)
Ata (Lugei)
Baining (Malasait)
Baining (Marabu)
Baining (Rangulit)
Tolai (Kabakada)
Tolai (Vunairoto)
Mussau
Lavongai (North)
Lavongai (South)
Tigak
Nalik
Notsi
Kuot (Kabil)
Kuot (Lamalaua)
Madak
Saposa
Teop
Arta
Nasioi
Micronesian
Samoan
Maori
Amis
Taroko
Cambodian
Han
Japanese
She
Naxi
Tu
Lahu
Dai
Xibo
Hezhen
Mongola
Daur
Oroqen
Miao
Yi
French

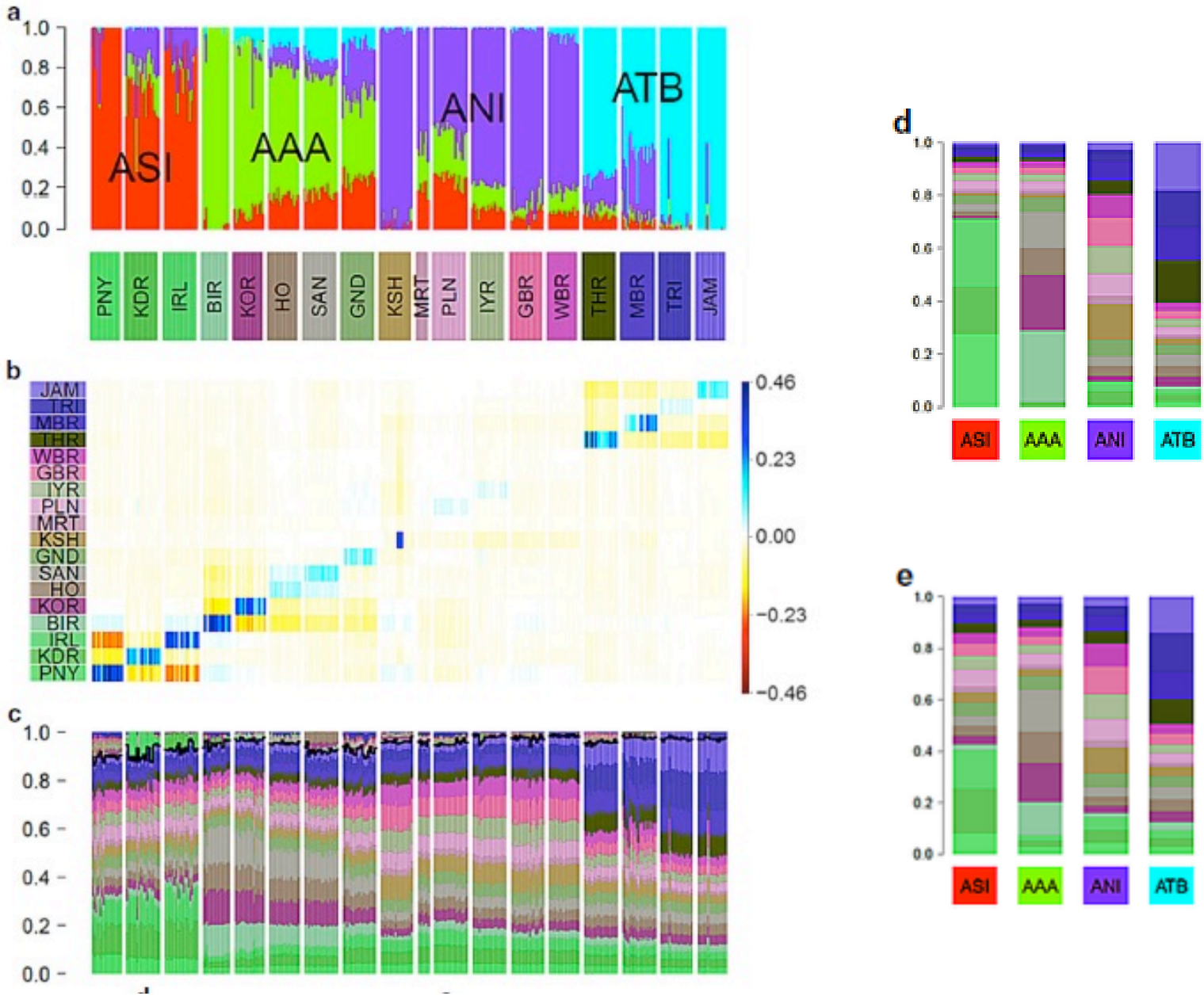

Africa
Europe/MiddleEast
Asia-Pakistan
Americas
K=6
Asia
Taiwan
Polynesia/
Micronesia
Bougainville
Baining
Ata/Mamusi

Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure

Analabha Basu[a,1], Neeta Sarkar-Roy[a], and Partha P. Majumder[a,b,1]






a
1.0
0.8
0.6
0.4
0.2
0.0
ASI
AAA
ANI
ATB
PNY
KDR
IRL
BIR
KOR
HO
SAN
GND
KSH
MRT
PLN
IYR
GBR
WBR
THR
MBR
TRI
JAM
b
JAM
TRI
MBR
THR
WBR
GBR
IYR
PLN
MRT
KSH
GND
SAN
HO
KOR
BIR
IRL
KDR
PNY
0.46
0.23
0.00
−0.23
−0.46
c
1.0
0.8
0.6
0.4
0.2
0.0
d
1.0
0.8
0.6
0.4
0.2
0.0
ASI
AAA
ANI
ATB
e
1.0
0.8
0.6
0.4
0.2
0.0
ASI
AAA
ANI
ATB

Conclusions

STRUCTURE/ADMIXTURE bar plots widely over-interpreted

Mixed ancestry profiles do not imply admixture

Recent genetic drift causes populations to be estimated as pure

Be alert for possibility of ghost admixture

Palettes can be used to visualize model fit and provide richer history

Provides a good **starting point** for population genetic analysis

Fitting and plotting procedure will be available from www.paintmychromosomes.com