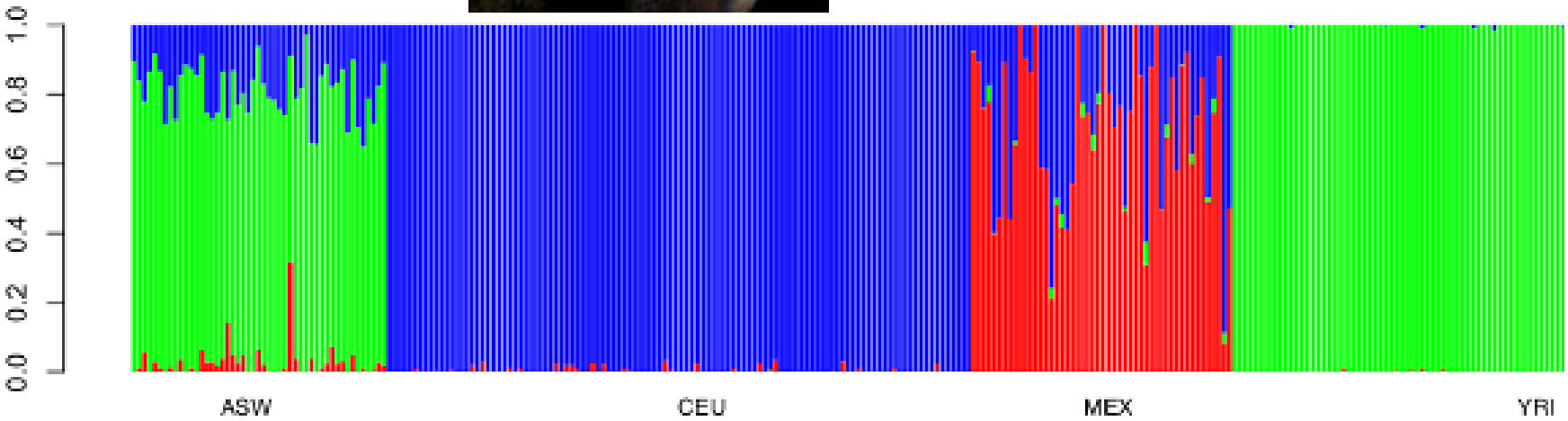
A tutorial on how to over-interpret structure/ADMIXTURE plots

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-



African Americans

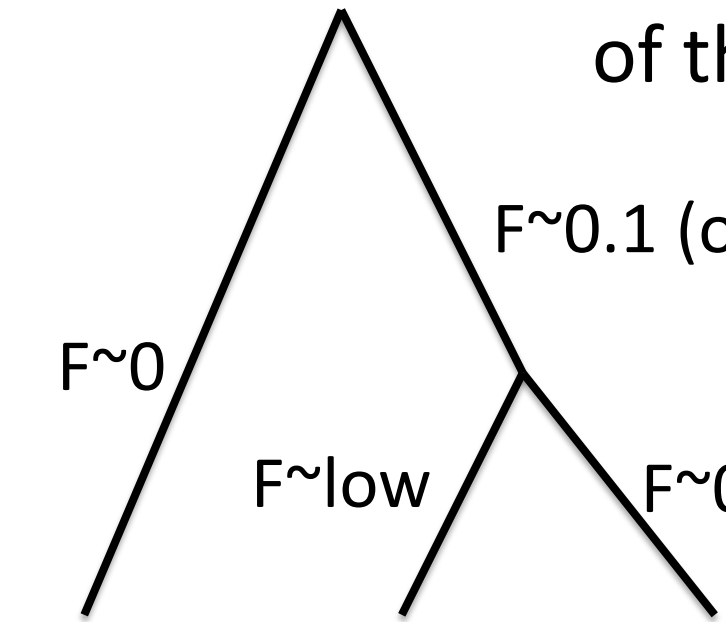
Europeans

Hispanics

West Africans

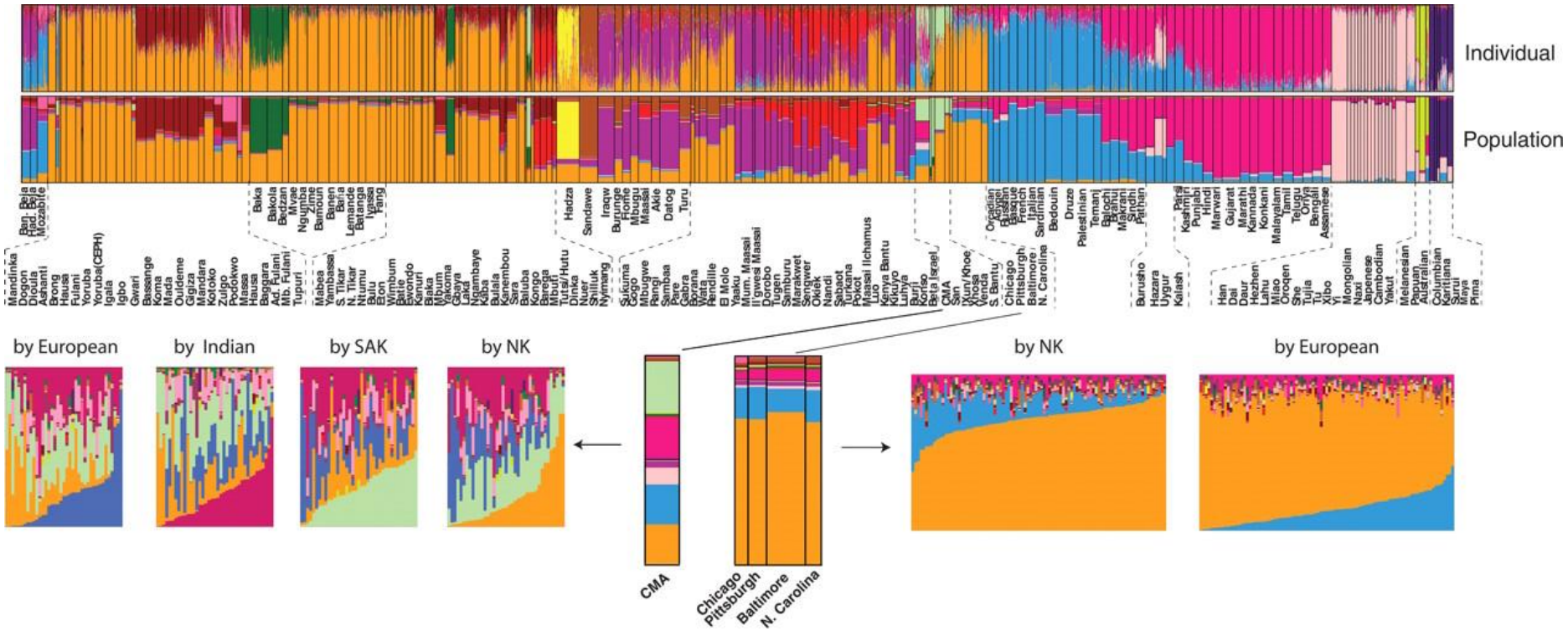
1000 genomes project

Genetic drift over 10s of thousands of years.



Admixture of large populations after 1600.

African Americans



Tishkoff 2009

Protocol

(1) Obsess over estimating K.

(2) Choose the highest K output by the various K estimating algorithms.

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

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

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

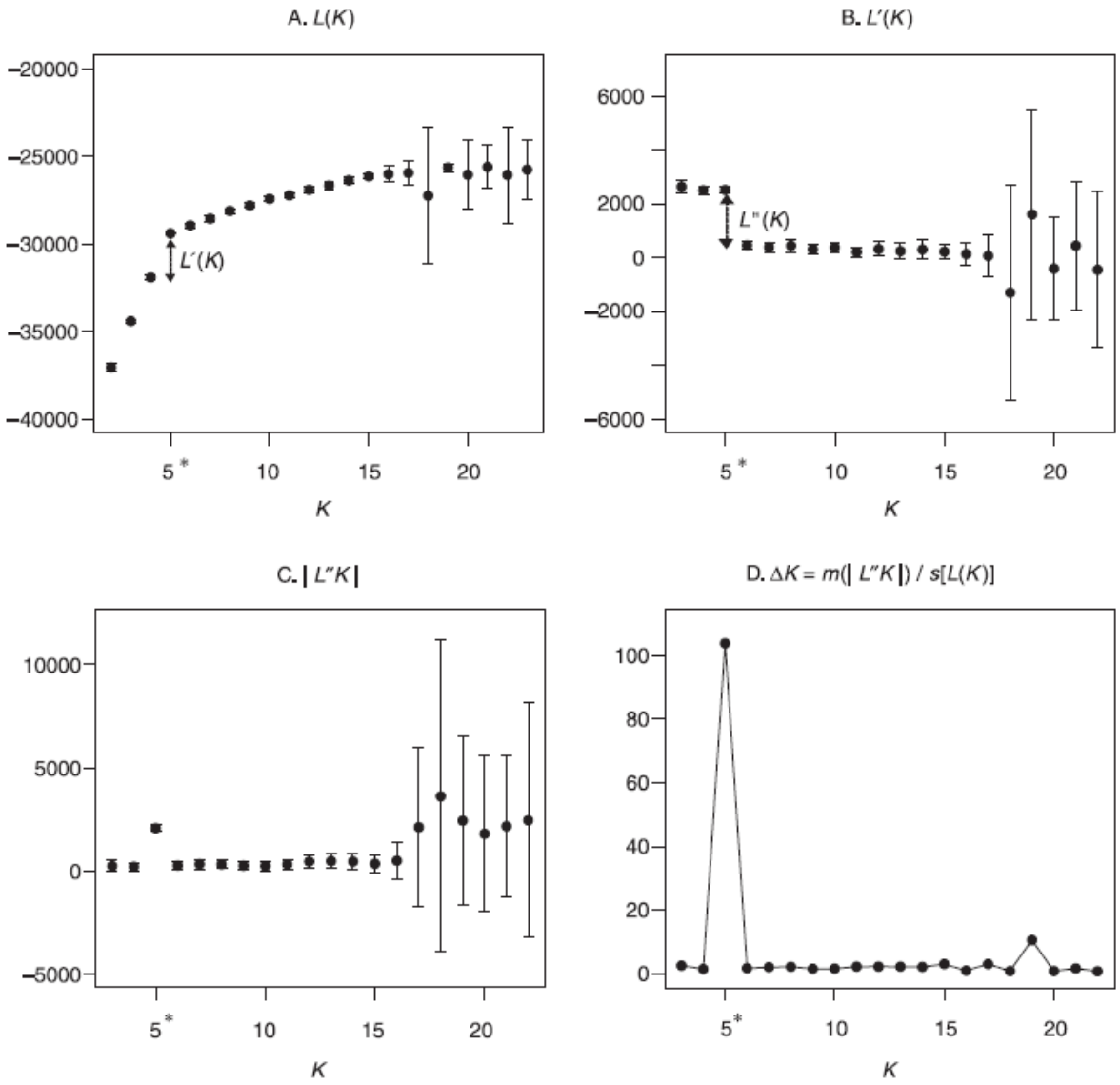
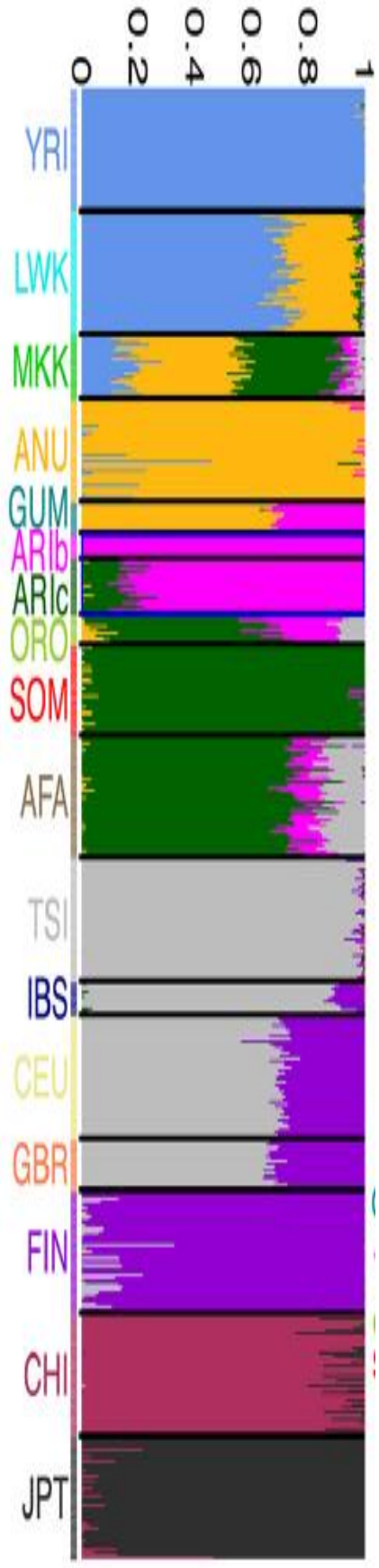
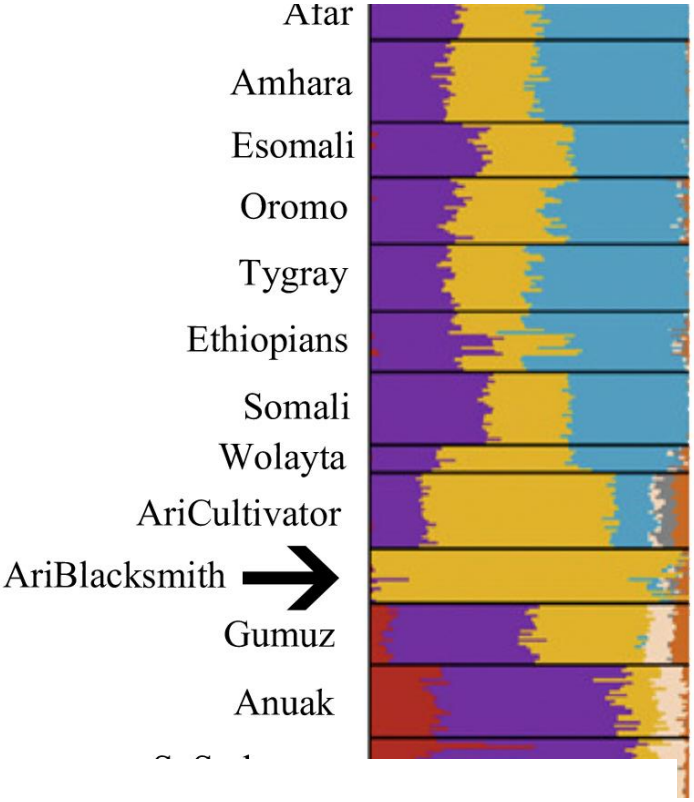
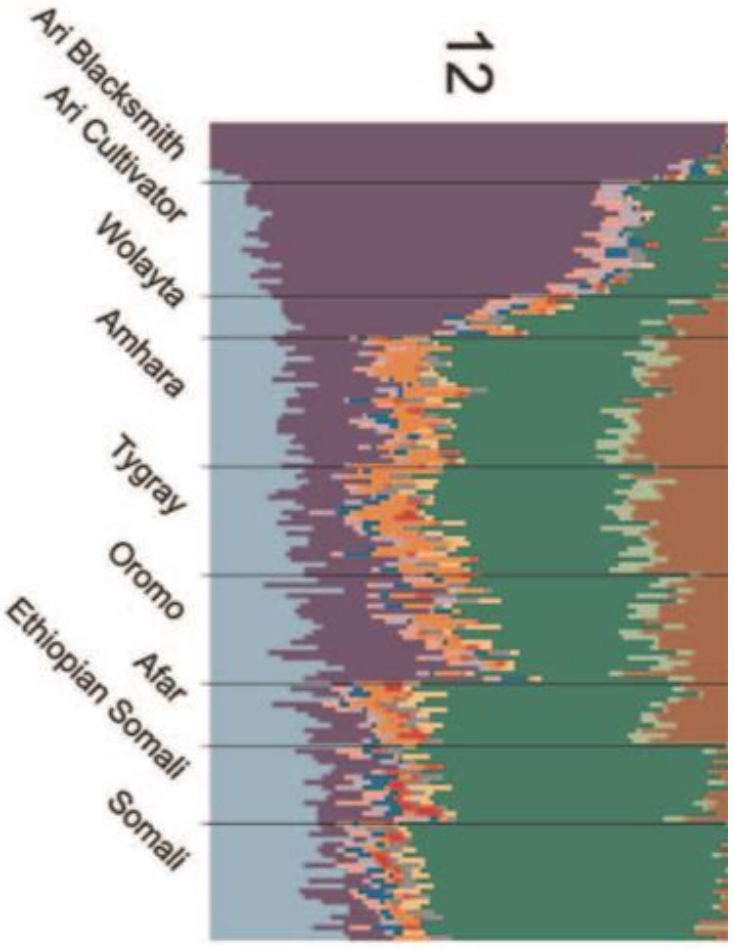


Fig. 2 Description of the four steps for the graphical method allowing detection of the true number of groups K^*. (A) Mean L(K) (± SD) over 20 runs for each K value. The model considered here is a hierarchical island model using all 100 individuals per population and 50 AFLP loci. (B) Rate of change 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.

Two studies followed the protocol and
suggest that ARIb group (blacksmiths)
is an unadmixed remnant of a hunter gatherer
population and that ARIc (cultivators) is a product
of admixture between the hunter gatherers and farmers

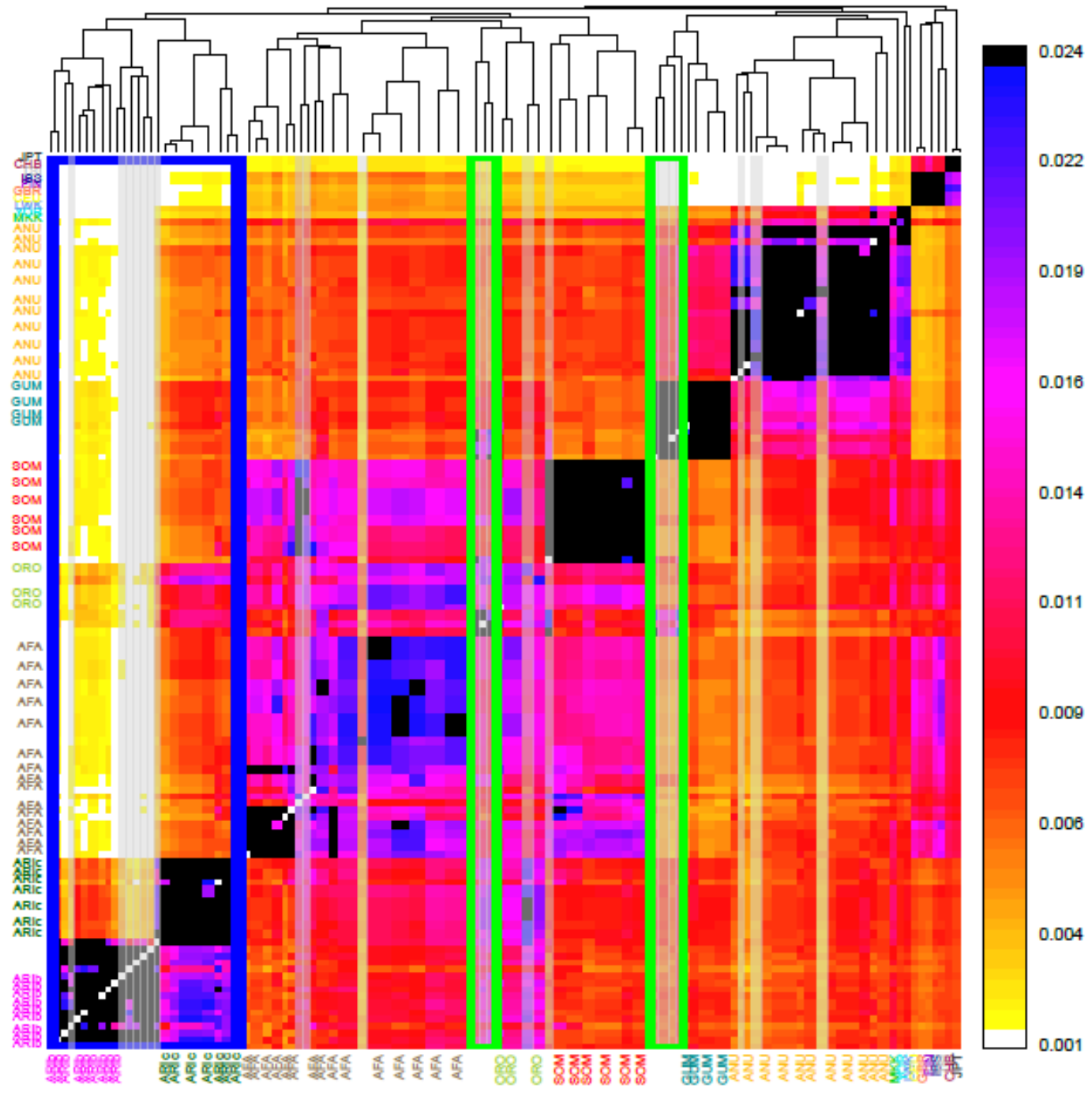

0
0.2
0.4
0.6
0.8
1
YRI
LWK
MKK
ANU
GUM
ARIb
ARIc
ORO
SOM
AFA
TSI
IBS
CEU
GBR
FIN
CHI
JPT

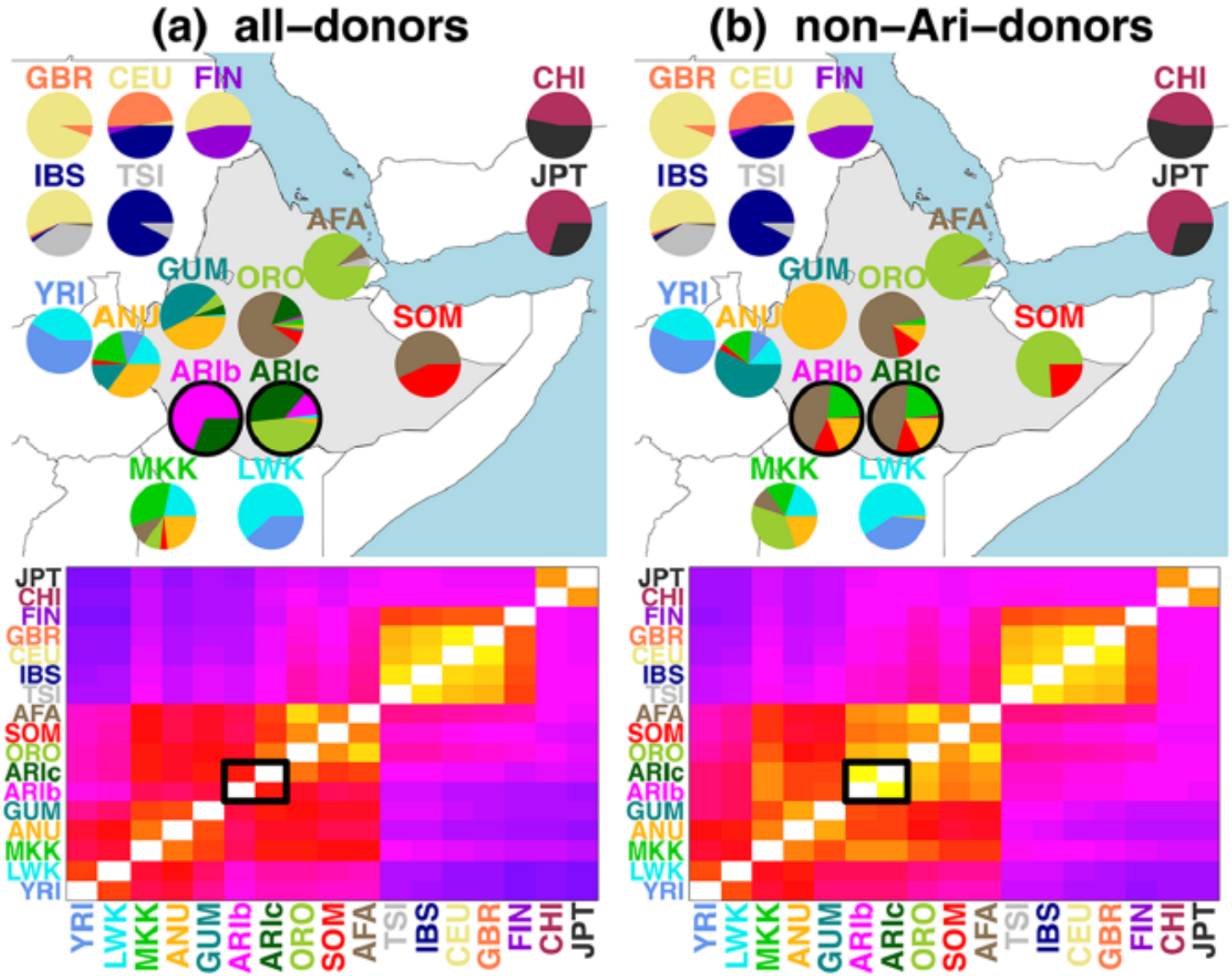

Afar
Amhara
Esomali
Oromo
Tygray
Ethiopians
Somali
Wolayta
AriCultivator
AriBlacksmith
Gumuz
Anuak

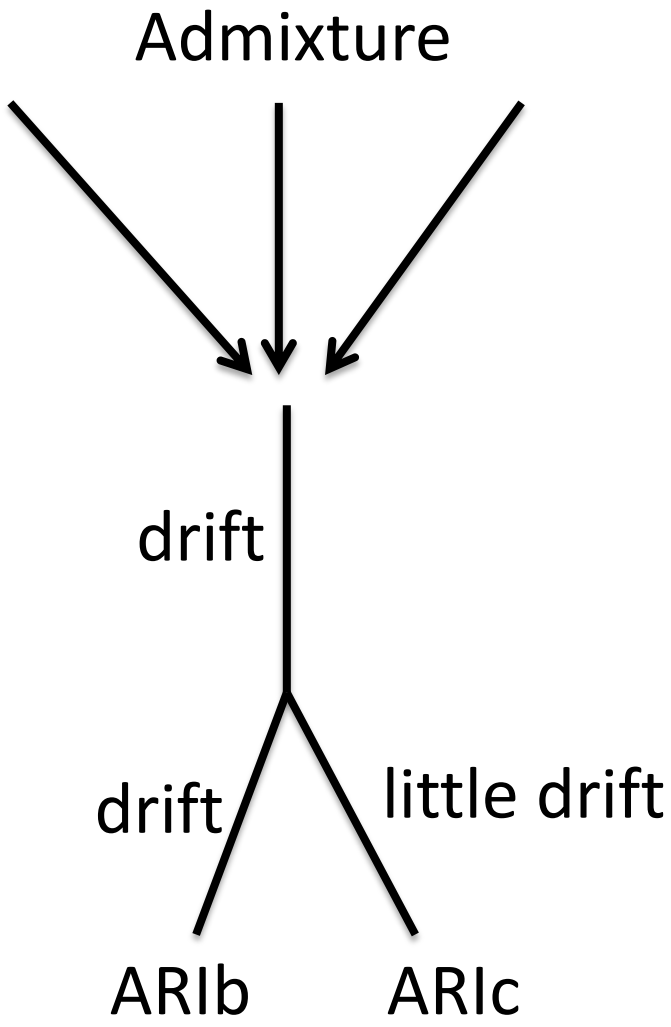

12
Ari Blacksmith
Ari Cultivator
Wolayta
Amhara
Tygray
Oromo
Afar
Ethiopian Somali
Somali

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

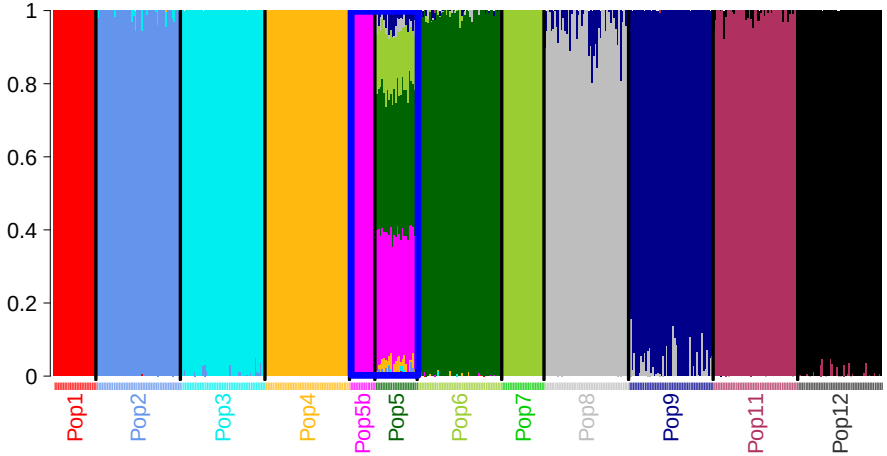

0.024
0.022
0.019
0.016
0.014
0.011
0.009
0.006
0.004
0.001


(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

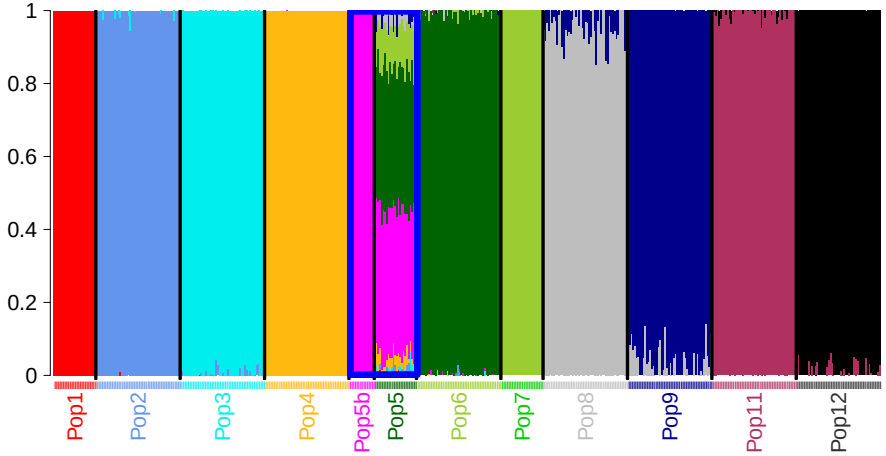

Admixture
drift
drift
little drift
ARIb
ARIc

Unsupervised ADMIXTURE applied to Simulations

Drift-last hypothesis



Admixture-last hypothesis



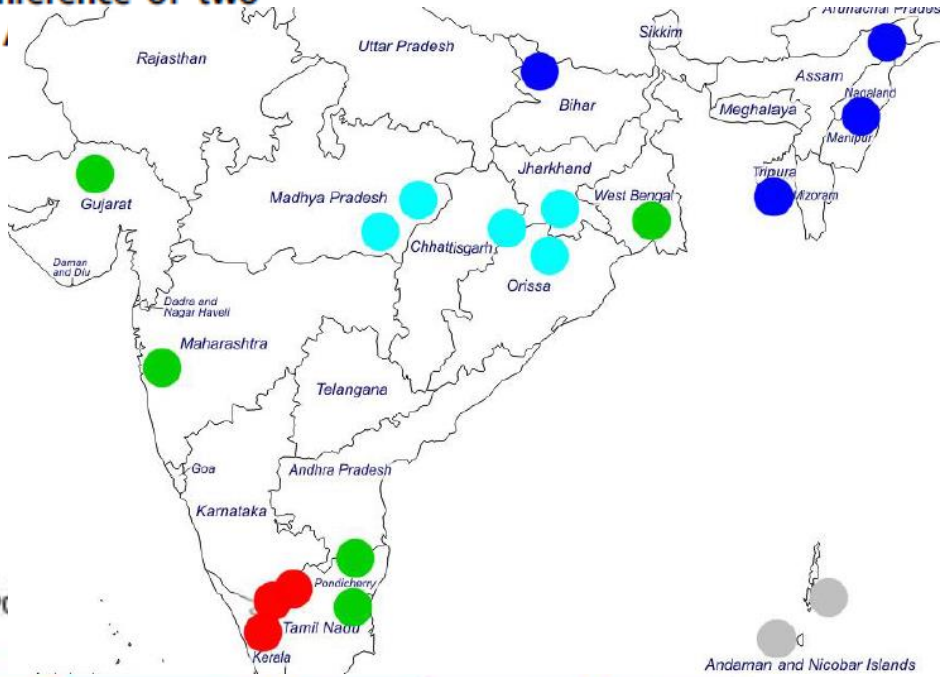
PLOS | GENETICS

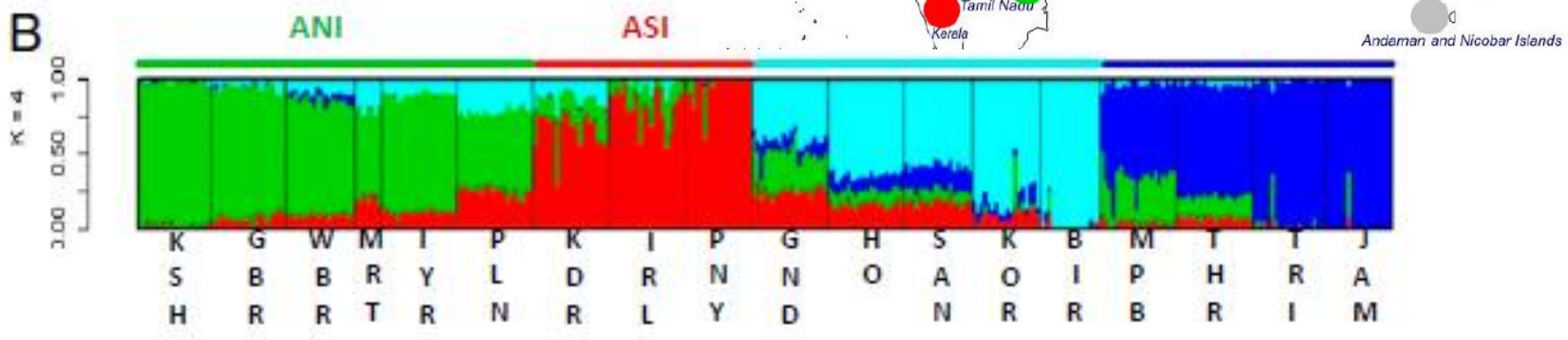
RESEARCH ARTICLE

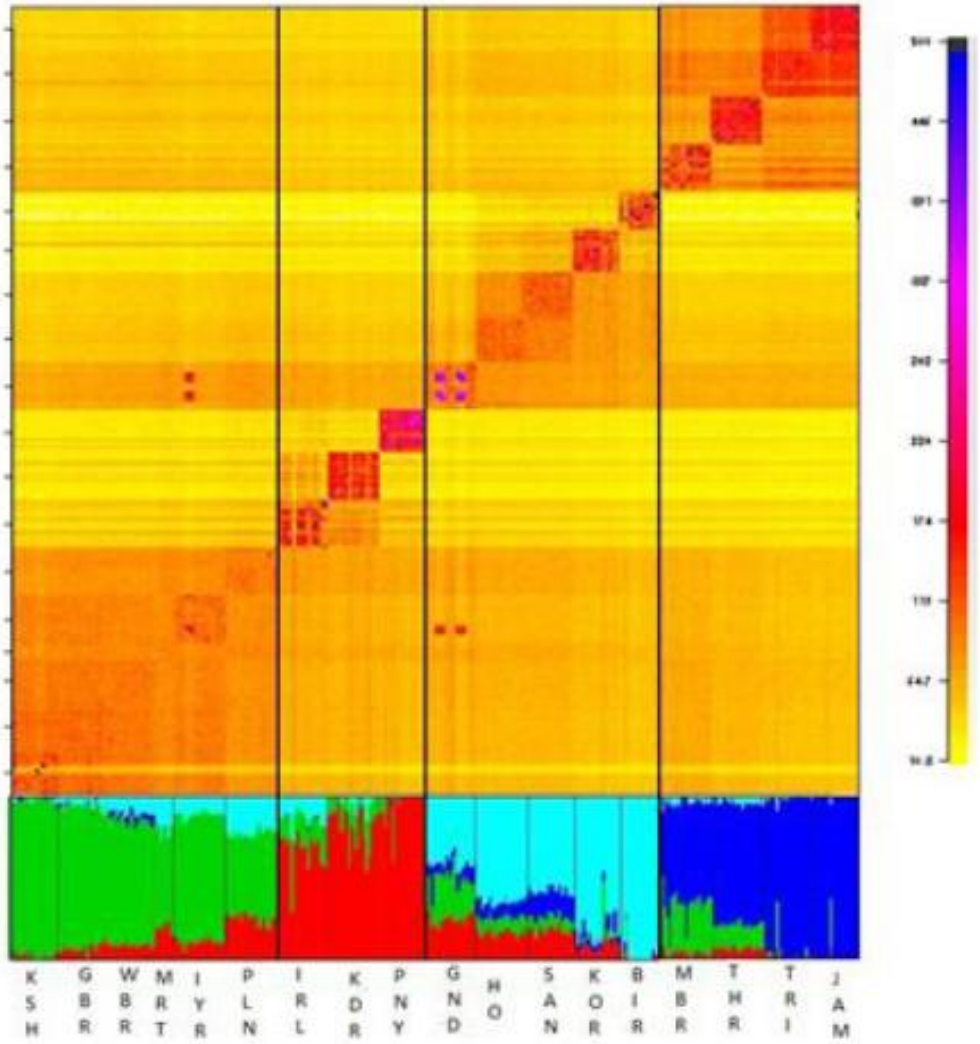
Evidence for a Common Origin of Blacksmiths and Cultivators in the Ethiopian Ari within the Last 4500 Years: Lessons for Clustering-Based Inference

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]*

India, occupying the center stage of Paleolithic and Neolithic migrations, has been underrepresented in genome-wide studies of variation. Systematic analysis of genome-wide data, using multiple robust statistical methods, on (*i*) 367 unrelated individuals drawn from 18 mainland and 2 island (Andaman and Nicobar Islands) populations selected to represent geographic, linguistic, and ethnic diversities, and (*ii*) individuals from populations represented in the Human Genome Diversity Panel (HGDP), reveal four major ancestries in mainland India. This contrasts with an earlier inference of two ancestries based on limited population sampling.






KSH
GBR
WBR
MRT
IYR
PLN
IRL
KDR
PNY
GND
HO
SAN
KOR
BIR
MBR
THR
TRI
JAM

Some extra tips for historical interpretation

(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.

(4a) Do not ask how the ancestral populations are related to each other.

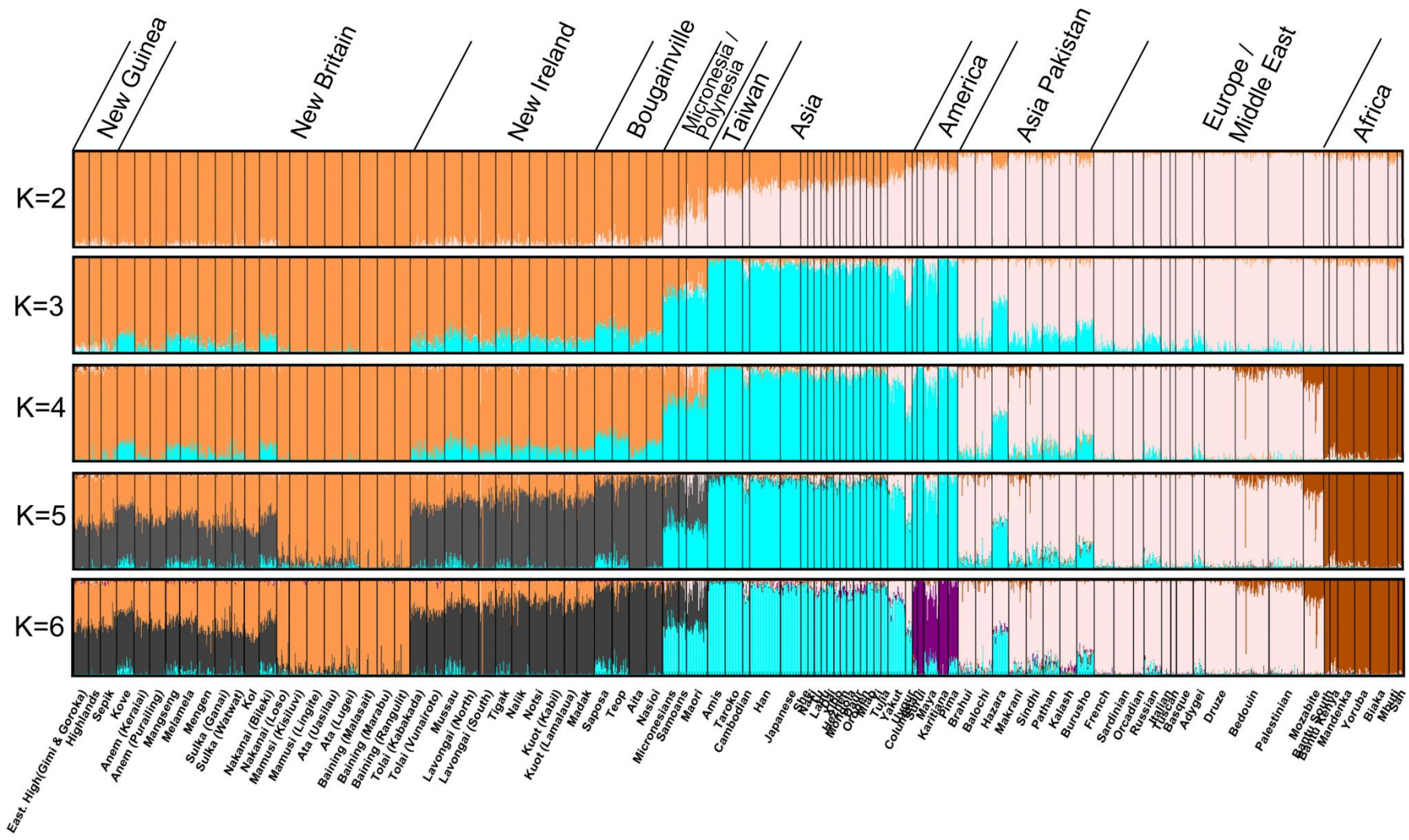
(4b) Neglect possibility an ancestral population might itself be admixed.

(4c) Label ancestral populations based on the locations they are currently most frequent in.

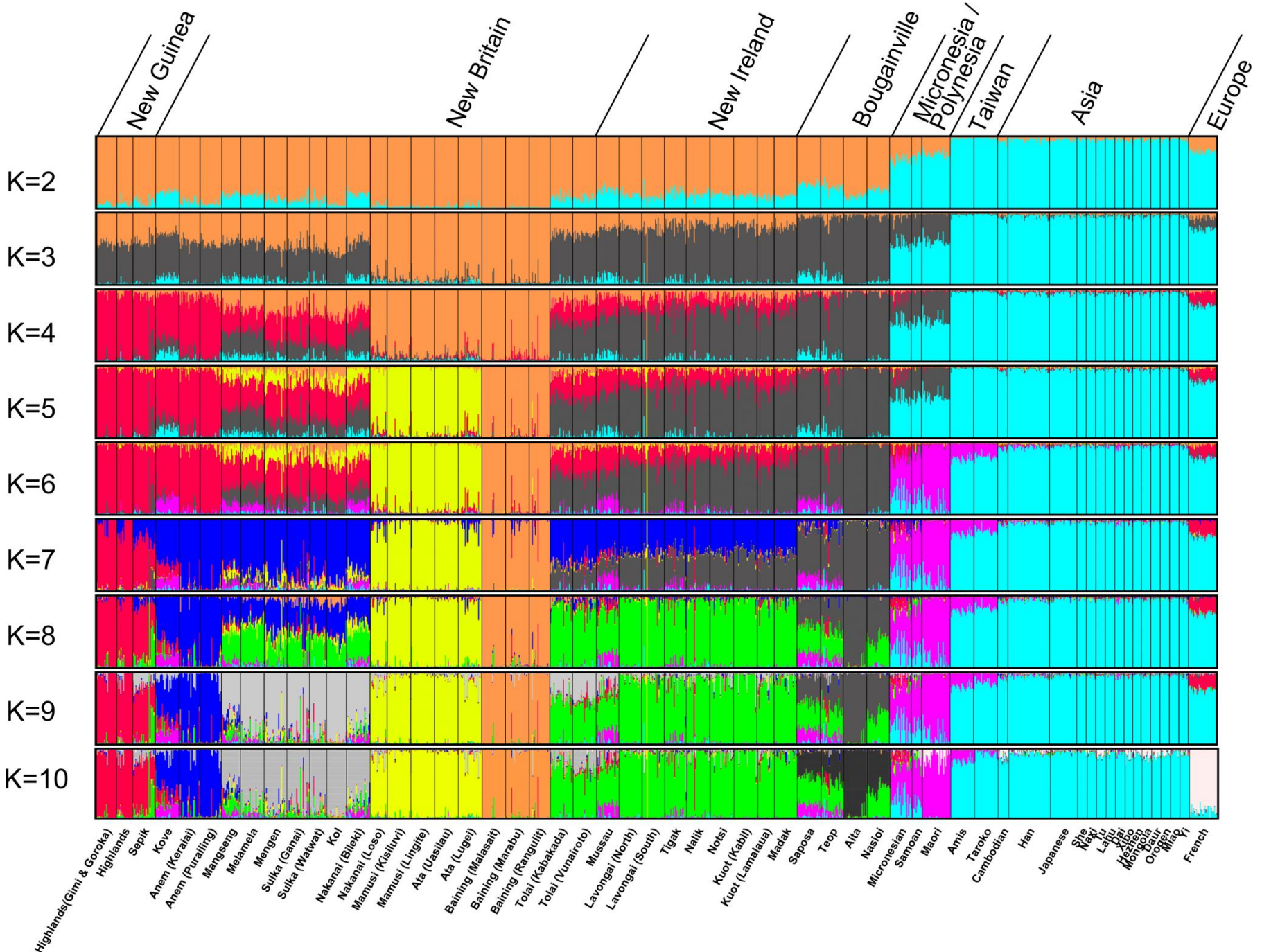
(5a) Assume that each admixture event happened as a single pulse.

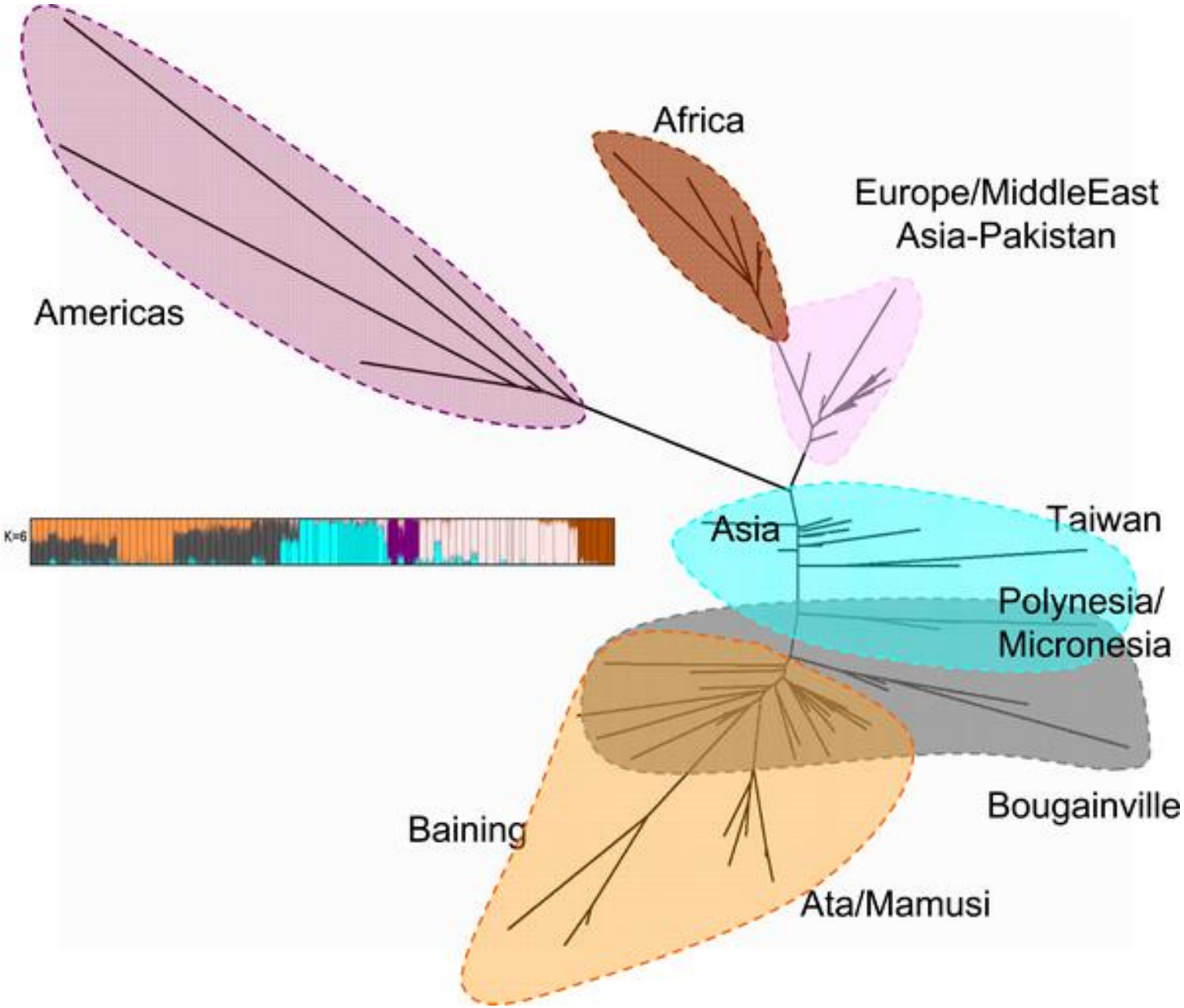
(5b) Note that a location is a "melting pot“ or possibly “cross roads” because it uses all K ancestral populations.

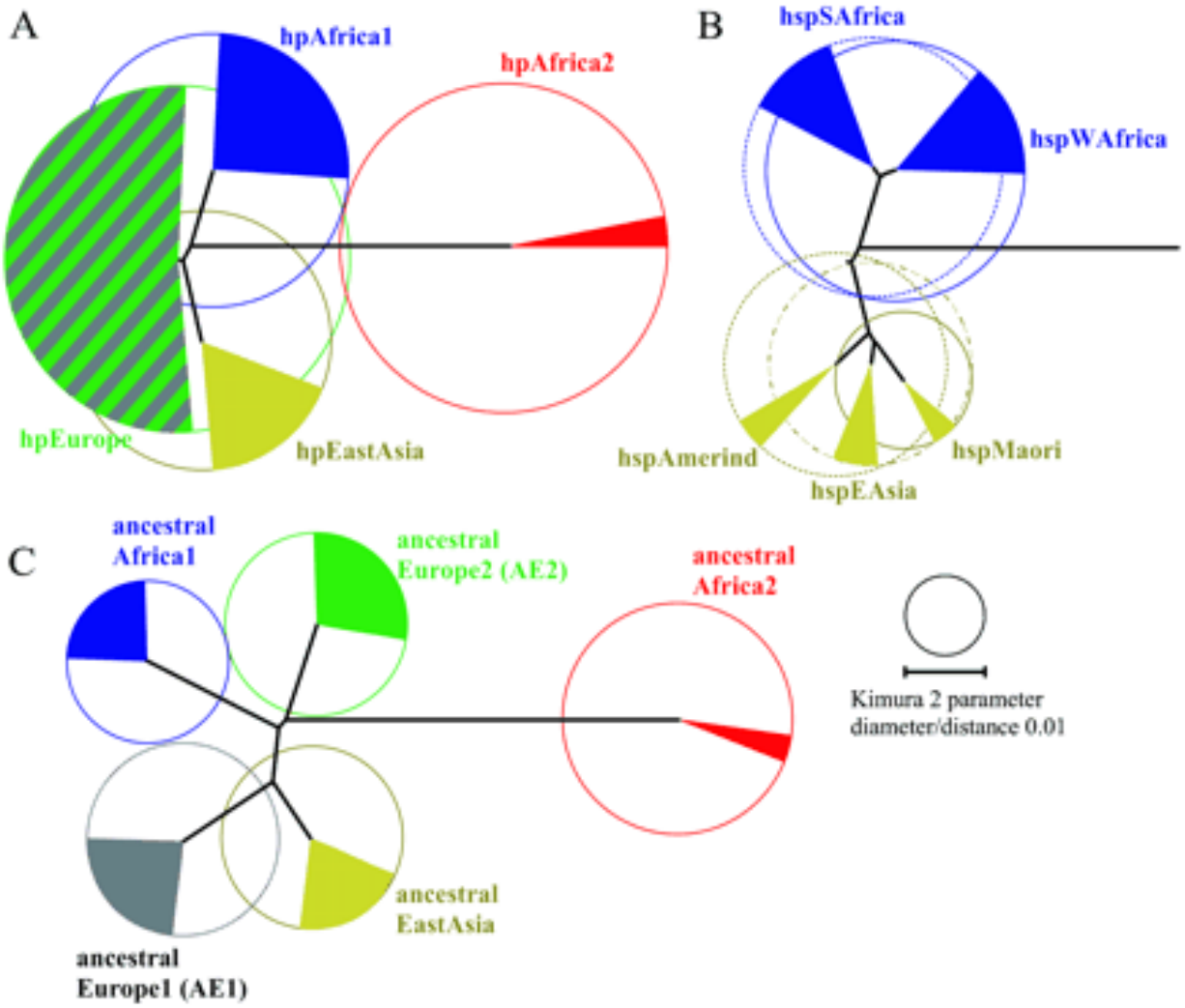
(6) Do not check these conclusions using other methods (or values of K).



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 (Keralai)
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
Aita
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


A
hpAfrica1
hpAfrica2
hpEurope
hpEastAsia
B
hspSAfrica
hspWAfrica
hspAmerind
hspMaori
hspEAsia
C
ancestral
Africa1
ancestral
Europe2 (AE2)
ancestral
Africa2
Kimura 2 parameter
diameter/distance 0.01
ancestral
EastAsia
ancestral
Europe1 (AE1)

Falush 2003