Supplementary Text for "pong: a network-graphical approach for the analysis of population structure"

Aaron A. Behr [1,2,*], Gracie Liu-Fang[3,†], Katherine Z. Liu[2,†], Priyanka Nakka[1,4], and Sohini Ramachandran[1,4*]

[1]Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, USA
[2]Department of Computer Science, Brown University, Providence, RI, USA
[3]Computer Science Department, Wellesley College, Wellesley, MA, USA
[4]Center for Computational Molecular Biology, Brown University, Providence, RI, USA

Cluster similarity metrics

We implemented and tested several metrics for cluster similarity. The default metric used by pong, $\mathcal{J}$ (Equation 1), is derived from the Jaccard index used in set comparison. For a given pair of clusters $\{\vec{q}_{\bullet\, a}, \vec{r}_{\bullet\, b}\}$, let N^* be the set of indices for which at least one of $\{\vec{q}_{\bullet\, a}, \vec{r}_{\bullet\, b}\}$ has a nonzero entry; that is, $N^* = \{i \in \{1, \ldots, N\} : q_{ia} + r_{ib} > 0\}$. Then,

$$\mathcal{J}(\vec{q}_{\bullet\, a}, \vec{r}_{\bullet\, b}) = 1 - \sqrt{\frac{\sum_{i \in N^*} (q_{ia} - r_{ib})^2}{2|N^*|}} \tag{1}$$

$\mathcal{J}$ is designed to emphasize overlap in membership coefficients while ignoring overlap in nonmembership (i.e., individuals with membership coefficients of 0 in the clusters under comparison). Although we recommend using $\mathcal{J}$, pong implements other similarity metrics: G' (as used in *CLUMPP* Jakobsson and Rosenberg (2007)), the average sum of squared differences between $\vec{q}_{\bullet\, a}$ and $\vec{r}_{\bullet\, b}$ (subtracted from 1), and average Manhattan distance (subtracted from 1). pong's implementation is designed such that users familiar with Python and NumPy can add their own similarity metrics to the source code if desired.

Processing of 1000 Genomes Data

Variant calls for 1,019,196 genome-wide single-nucleotide variants (SNVs) in 2,504 individuals were extracted from the 1000 Genomes Project Phase 3 data repository `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/` (release date: Nov 6, 2014), using the command-line tool `tabix` (Li, 2011).

A total of 78 individuals were excluded from analysis based on relatedness: one individual from each pair of first- and second-degree relatives was removed, leaving a total of 2,426 individuals. Next, SNVs were pruned for linkage disequilibrium using the `--indep-pairwise` flag in PLINK (Purcell et al., 2007). We removed every SNV with $r^2 > 0.1$ with any other SNV within a 50-SNV sliding window (PLINK command-line parameters for `--indep-pairwise`: 50 10 0.01), leaving a total of 225,705 SNVs for analysis.

*To whom correspondence should be addressed: aaron_behr@alumni.brown.edu, sramachandran@brown.edu
†These authors contributed equally to this work.

ADMIXTURE (Alexander et al., 2009) was applied 10 times per value of K to these data, with K taking on values in the closed interval $[2, 10]$. The value of K that minimized cross-validation error was $K = 8$.

Supplementary Figure Captions

Figure S1: pong's main visualization of major modes in population structure in the 1000 Genomes (phase 3) with detailed population labels, $K = 2$ through $K = 10$.

Figure S2: CLUMPAK's (Kopelman et al., 2015) visualization of modes in population structure in the 1000 Genomes (phase 3), $K = 2$ through $K = 10$.

Figure S3: pong's visualization of population structure in the 1000 Genomes (phase 3), based on Q matrices from Consortium (2015), $K = 5$ through $K = 25$.

References

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, **19**(9), 1655–1664.

Consortium, T. . G. P. (2015). A global reference for human genetic variation. Nature, **526**(7571), 68–74.

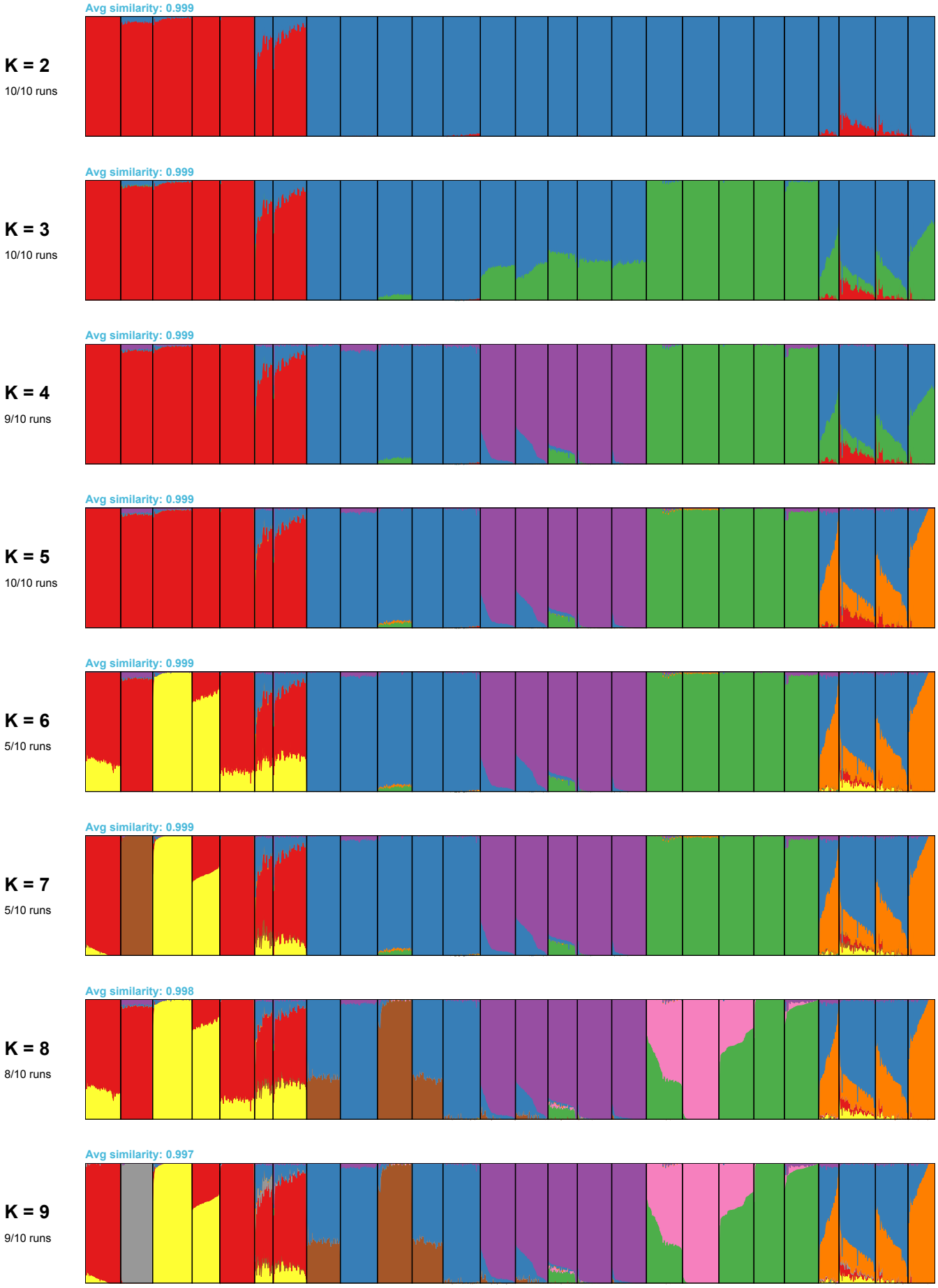
Jakobsson, M. and Rosenberg, N. a. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics (Oxford, England), **23**(14), 1801–6.

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., and Mayrose, I. (2015). C LUMPAK : a program for identifying clustering modes and packaging population structure inferences across K. Molecular Ecology Resources, pages doi: 10.1111/1755–0998.12387.

Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics, **27**, 718–9.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, **81**(3), 559–575.

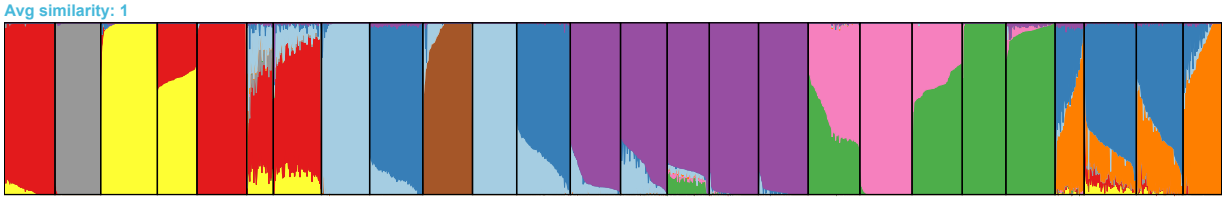
Figure S1


Avg similarity: 0.999
K = 2
10/10 runs
Avg similarity: 0.999
K = 3
10/10 runs
Avg similarity: 0.999
K = 4
9/10 runs
Avg similarity: 0.999
K = 5
10/10 runs
Avg similarity: 0.999
K = 6
5/10 runs
Avg similarity: 0.999
K = 7
5/10 runs
Avg similarity: 0.998
K = 8
8/10 runs
Avg similarity: 0.997
K = 9
9/10 runs

K = 10

4/10 runs

Avg similarity: 1

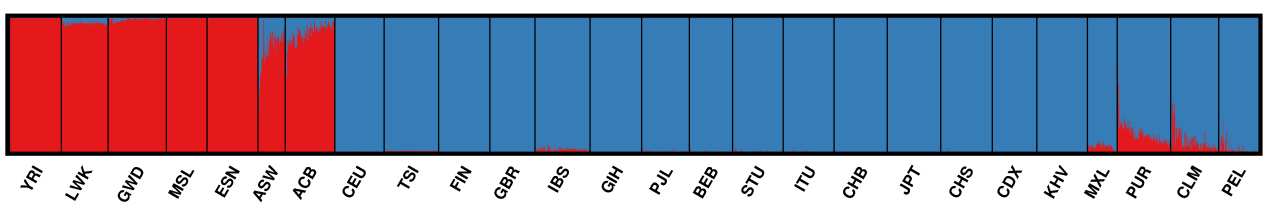


Yoruba in Ibadan, Nigeria
Luhya in Webuye, Kenya
Gambian in Western Divisions in the Gambia
Mende in Sierra Leone
Esan in Nigeria
Americans of African Ancestry in SW USA
African Caribbeans in Barbados
"Utah Residents (CEPH) with Northern and Western European Ancestry"
Toscani in Italia
Finnish in Finland
British in England and Scotland
Iberian Population in Spain
Gujarati Indian from Houston, Texas
Punjabi from Lahore, Pakistan
Bengali from Bangladesh
Sri Lankan Tamil from the UK
Indian Telugu from the UK
Han Chinese in Bejing, China
Japanese in Tokyo, Japan
Southern Han Chinese
Chinese Dai in Xishuangbanna, China
Kinh in Ho Chi Minh City, Vietnam
Mexican Ancestry from Los Angeles USA
Puerto Ricans from Puerto Rico
Colombians from Medellin, Colombia
Peruvians from Lima, Peru

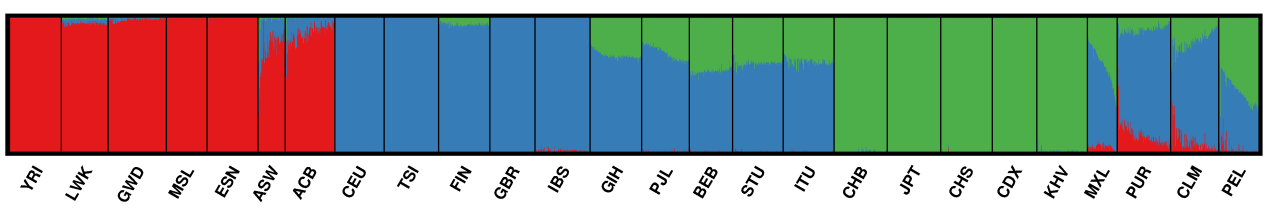
CLUMPAK main pipeline - Job 1439007297 summary

Major modes for the uploaded data:

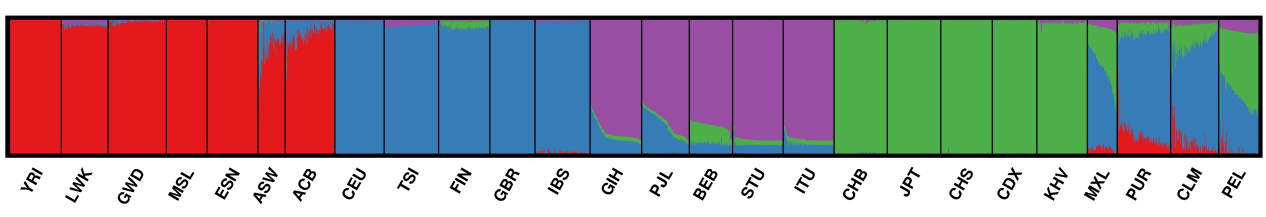
K=2



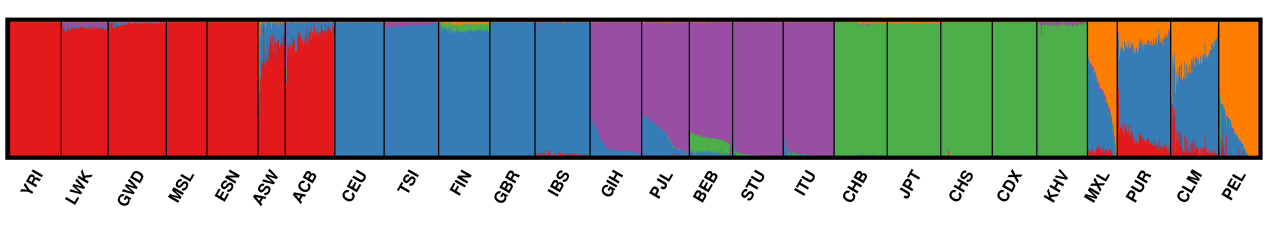
K=3



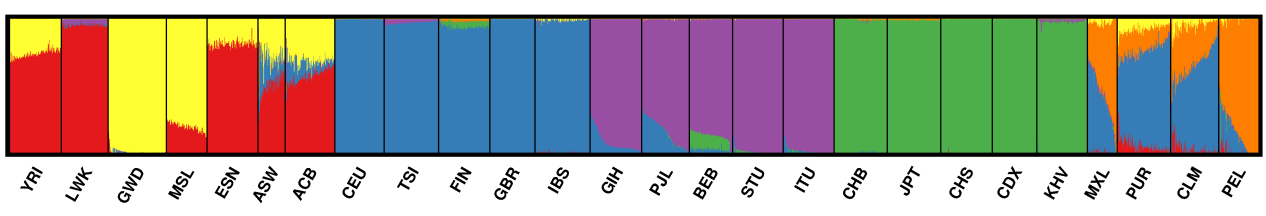
K=4



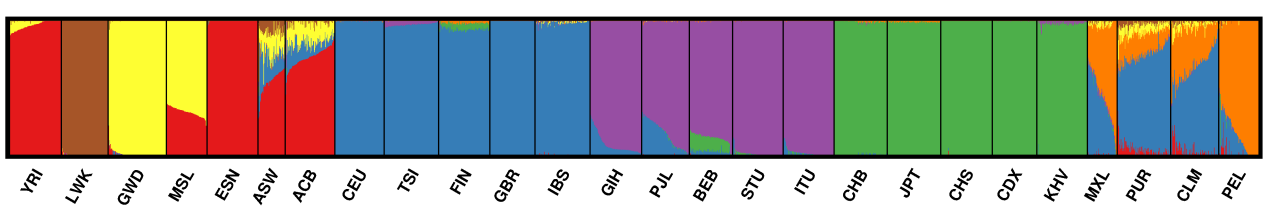
K=5



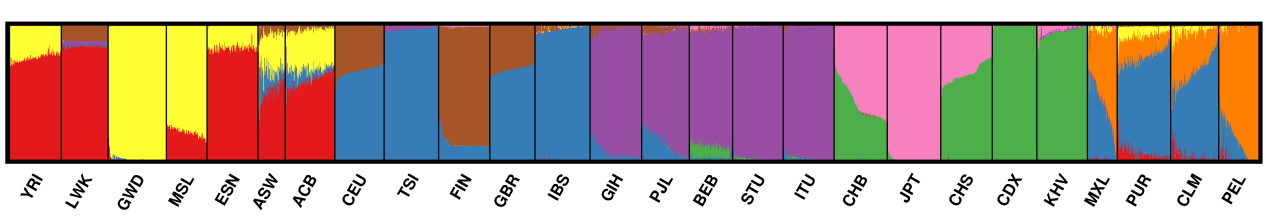
K=6



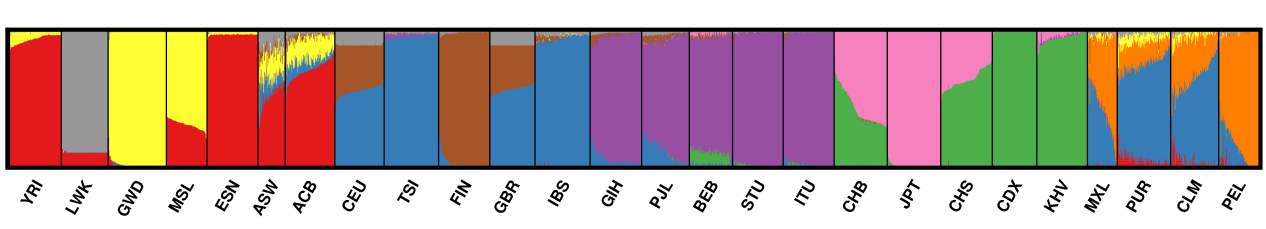
K=7



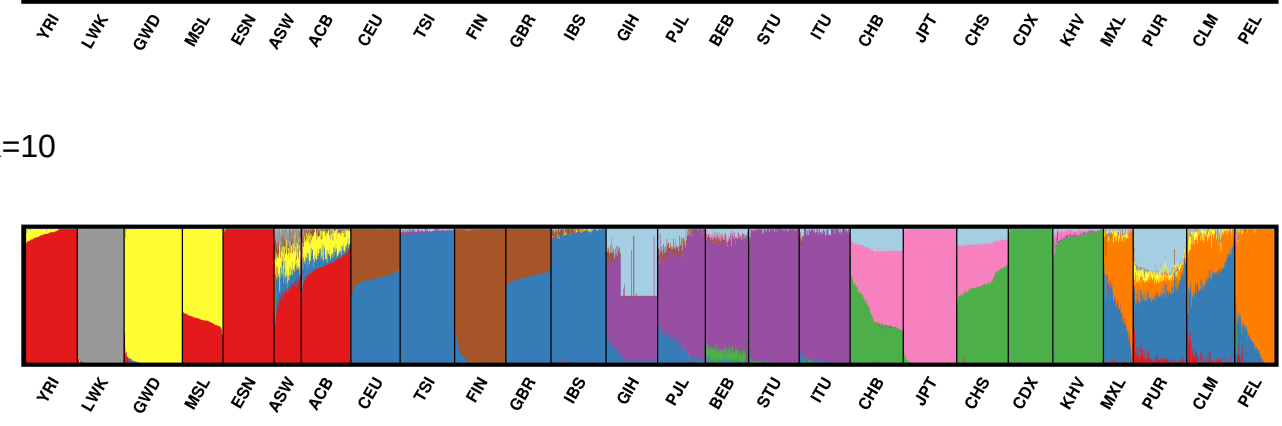
K=8



K=9

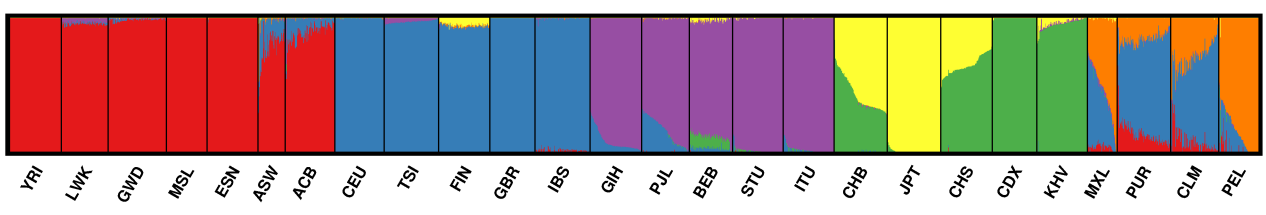


K=10

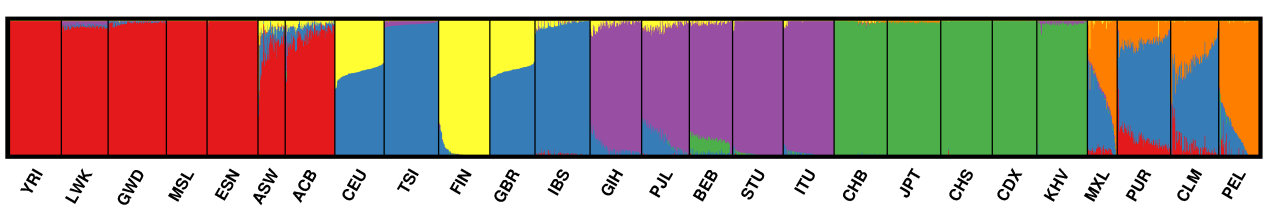


Minor modes for the uploaded data:

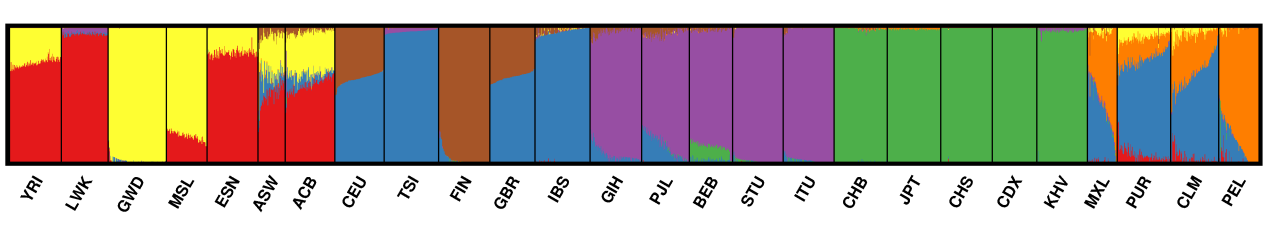
K=6 MinorCluster1



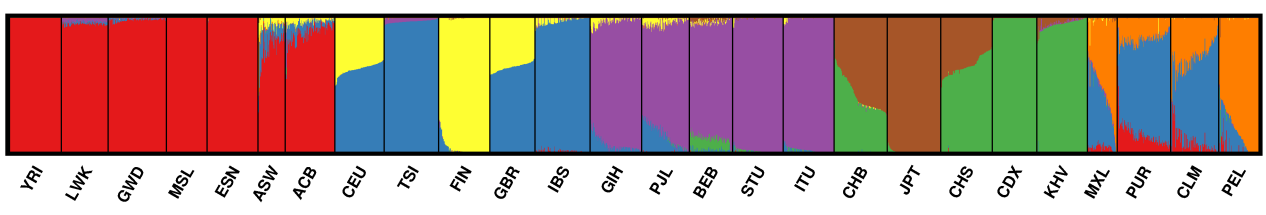
K=6 MinorCluster2



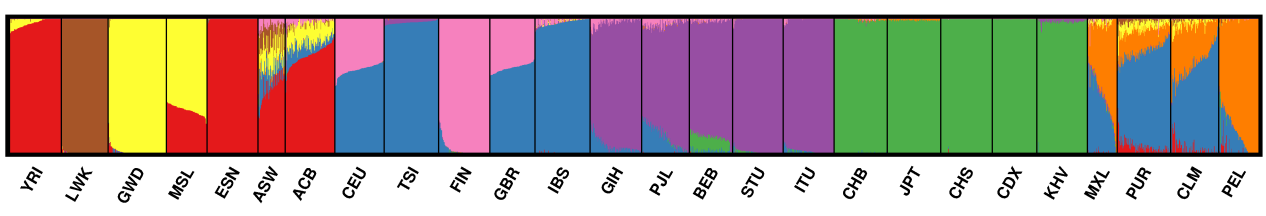
K=7 MinorCluster1



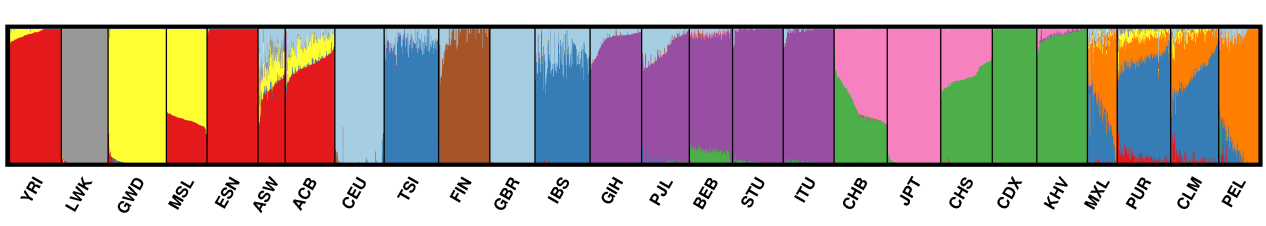
K=7 MinorCluster2



K=8 MinorCluster1



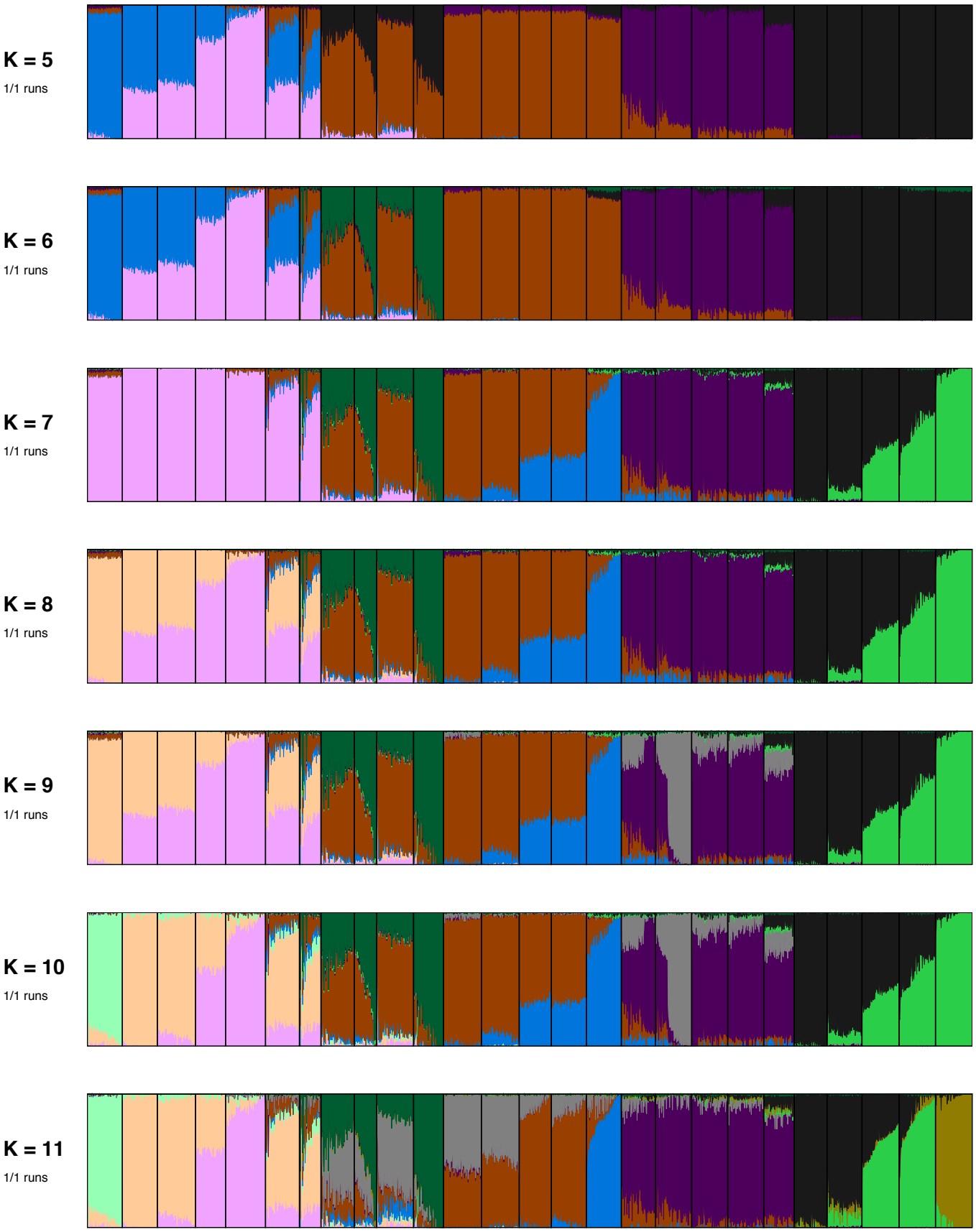
K=10 MinorCluster1



Division of runs by mode:

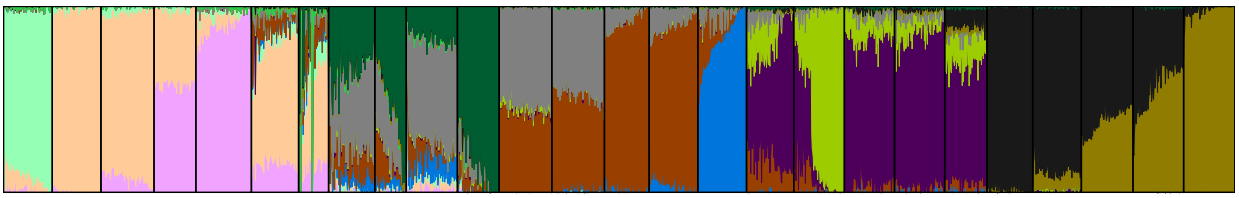
K=2	10/10
K=3	10/10
K=4	10/10
K=5	10/10
K=6	5/10, 3/10, 2/10
K=7	5/10, 3/10, 2/10
K=8	9/10, 1/10
K=9	10/10
K=10	6/10, 4/10

Figure S3


K = 5
1/1 runs
K = 6
1/1 runs
K = 7
1/1 runs
K = 8
1/1 runs
K = 9
1/1 runs
K = 10
1/1 runs
K = 11
1/1 runs

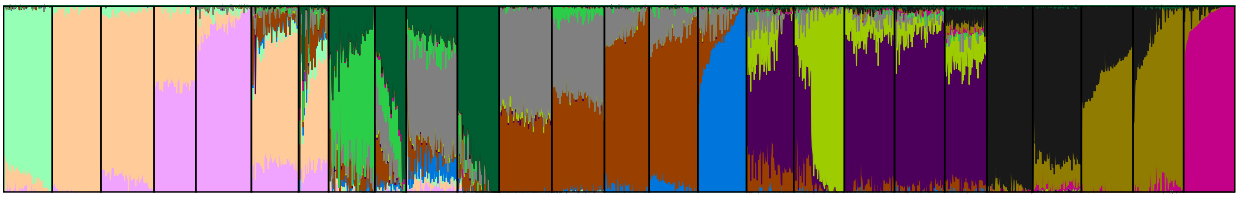
K = 12

1/1 runs



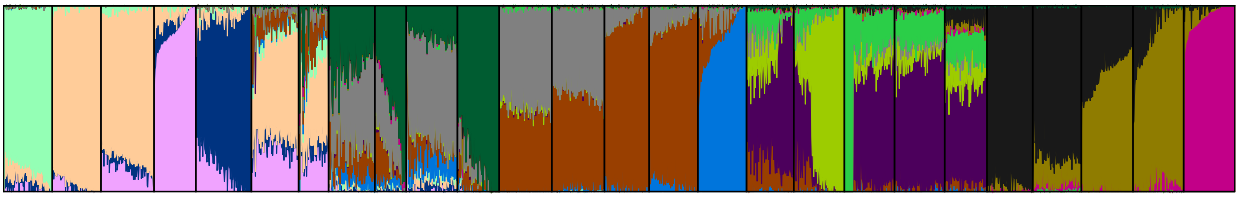
K = 13

1/1 runs



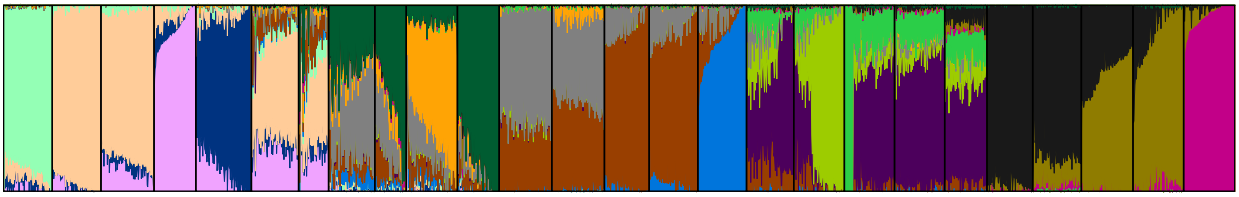
K = 14

1/1 runs



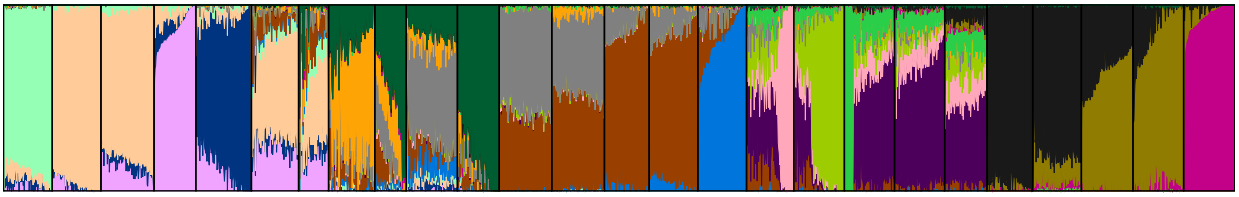
K = 15

1/1 runs



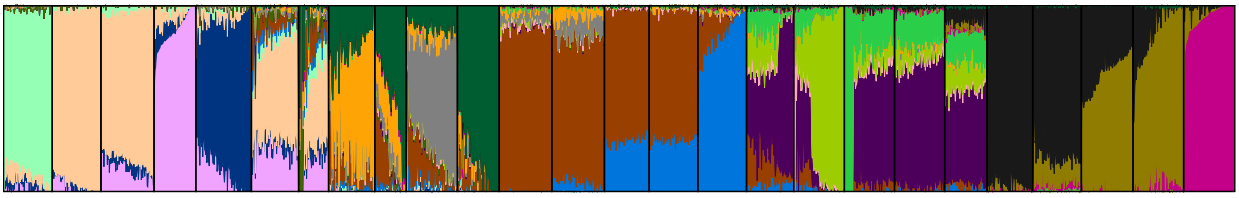
K = 16

1/1 runs



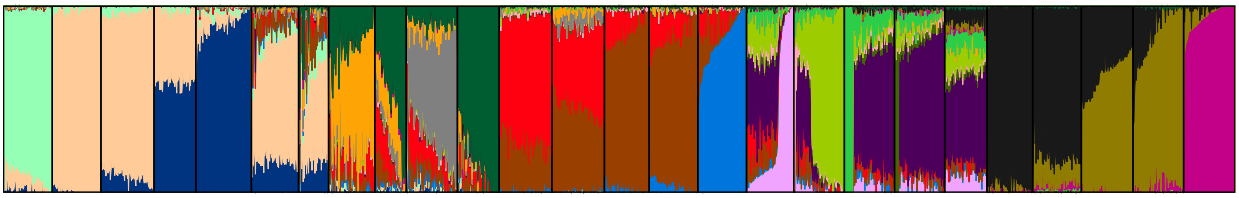
K = 17

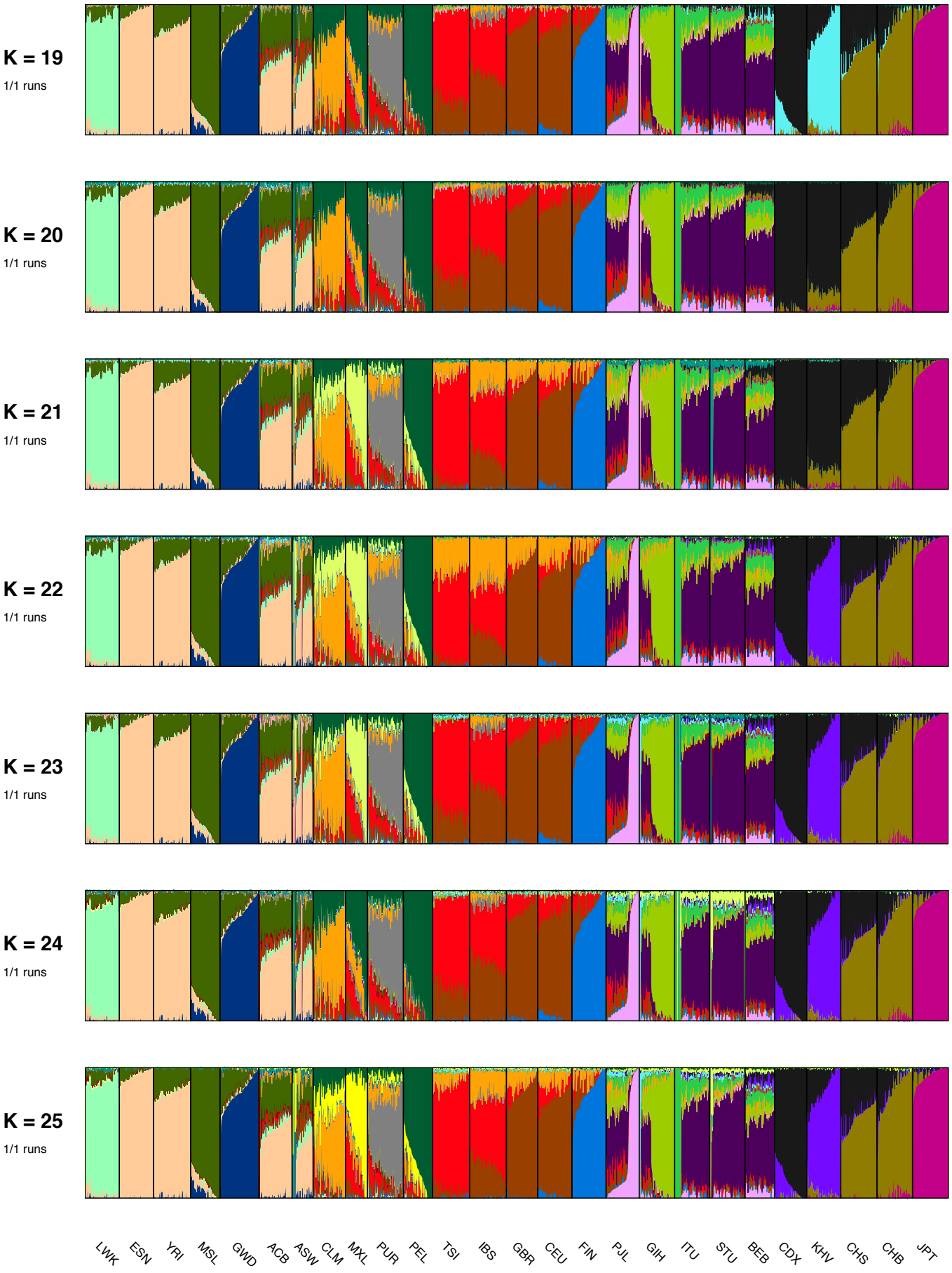
1/1 runs



K = 18

1/1 runs




K = 19
1/1 runs
K = 20
1/1 runs
K = 21
1/1 runs
K = 22
1/1 runs
K = 23
1/1 runs
K = 24
1/1 runs
K = 25
1/1 runs
LWK
ESN
YRI
MSL
GWD
ACB
ASW
CLM
MXL
PUR
PEL
TSI
IBS
GBR
CEU
FIN
PJL
GIH
ITU
STU
BEB
CDX
KHV
CHS
CHB
JPT