Supporting information for:

CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across *K*

Naama M Kopelman, Jonathan Mayzel, Mattias Jakobsson, Noah A Rosenberg, Itay Mayrose

Supplementary Note 1 – the relation between similarity scores of replicate runs to values of SD_{RUNS} and SD_{INDIVS} and to the assigned dynamic threshold

Table S1 presents the mean values of similarity scores between simulated replicate runs belonging to the major mode at K=3, for different combinations of values of SD_{RUNS} and SD_{INDIVS}. For example, the mean similarity score between simulated runs of the same mode for SD_{RUNS}=0.1 and SD_{INDIVS}=0.1 was 0.80. For comparison, for the dataset presented in Figure 4 (main text), the mean similarity score for runs in the major modes ranged from 0.95 to 0.98, depending on the K value (K=2, 3, 4, 5, 6), corresponding to data simulated with both SD values less than or equal to 0.025.

In a second empirical example, we checked the mean similarity scores for combined samples from the HGDP-CEPH (Li *et al.* 2008) and HapMap (Consortium *et al.* 2010) studies of worldwide human samples. This dataset of 2,055 individuals from 64 *a priori* populations consisted of 938 unrelated individuals from the H952 subset (Rosenberg 2006) and 1,117 unrelated individuals from the HAP1117 subset (Pemberton et al. 2010). Following quality control, we obtained a merged set of 2,055 individuals from 64 populations and 486,592 autosomal SNPs that the two datasets shared in common (Kopelman 2014). For STRUCTURE analyses, a smaller computationally feasible set of 5,233 markers was chosen such that adjacent markers were separated by at least 500kb (Kopelman 2014). We ran STRUCTURE with the admixture model 40 times for each value of K from 2 to 6, with a burn-in period of length 10,000 iterations followed by 20,000 additional iterations. The mean similarity scores obtained by CLUMPP ranged from 0.939 to 0.999, depending on the K value.

Table S2 presents the mean values of the dynamic threshold assigned by CLUMPAK to simulated sets of unimodal runs at K=3, for different combinations of values of SD_{RUNS} and SD_{INDIVS} (see Figure 2 for simulation results). For example, the threshold for SD_{RUNS}=0.1 and SD_{INDIVS}=0.1 was 0.80. Similarly, Table S3 presents the mean values of the dynamic threshold for simulated sets of bimodal runs at K=3 and f=0.25 (i.e. 25% of the runs assigned to the minor mode), for different combinations of values of SD_{RUNS} and SD_{INDIVS} (see Figure 3 for simulation results). Edges whose weights are smaller than the threshold are removed, and the weights of the remaining edges are shifted downward by the value of the threshold. CLUMPAK explores a range of possible threshold values, searching for the largest threshold for which the fraction of singleton clusters is smaller than 0.1 and the mean node degree is at least 50% of the total number of vertices.

Table S1. Mean CLUMPP similarity scores for the unimodal simulation scenario at *K*=3, under different choices of values of SD_{RUNS} and SD_{INDIVS}. For each choice of SD_{RUNS} and SD_{INDIVS}, 30 simulations were tested. For each simulation, the mean was calculated across all pairs of runs in that simulation ((40×39)/2=780 pairs). The mean values obtained for the same settings were then averaged across simulations (30 simulations for each setting).

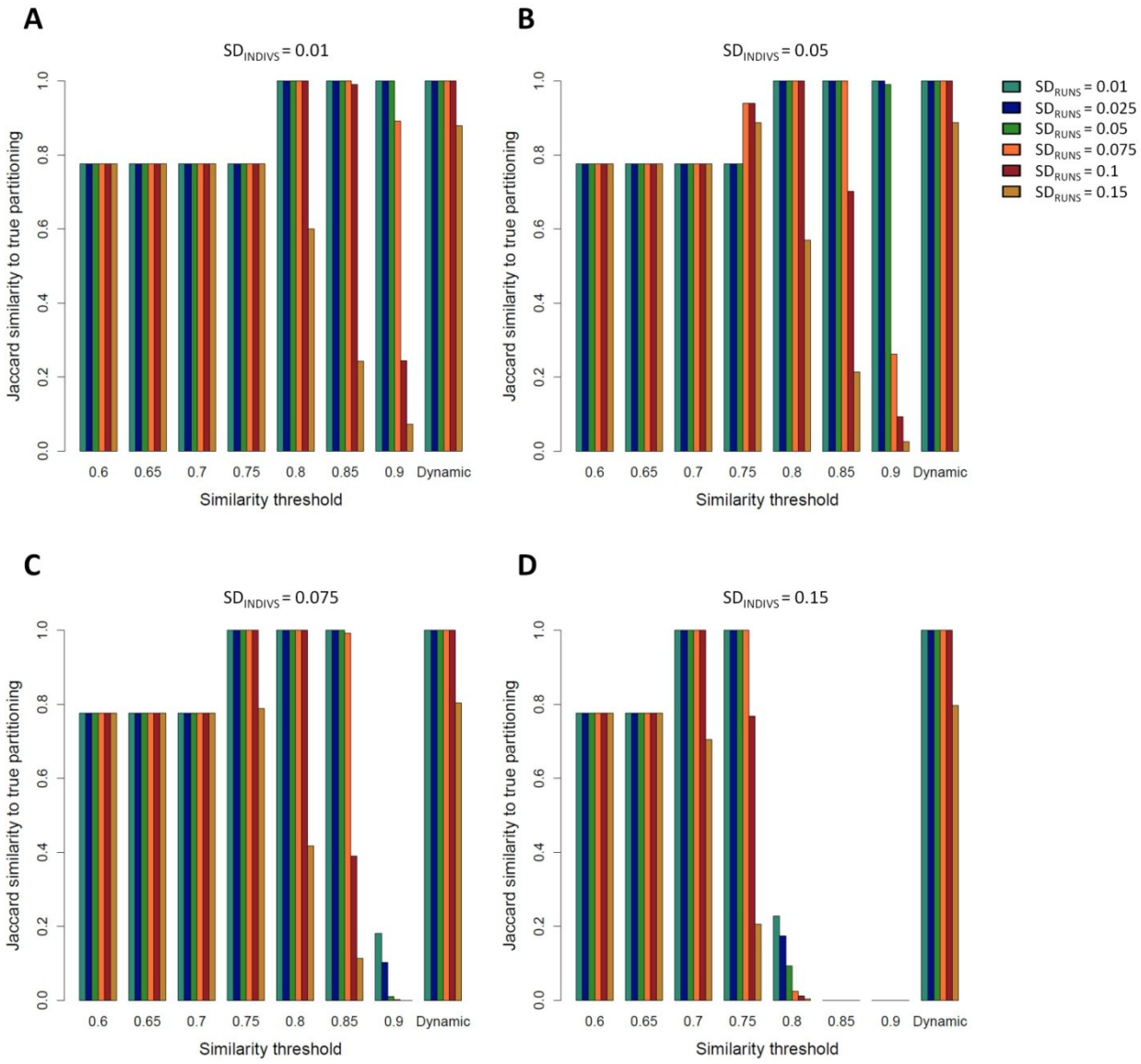
SD_{RUNS} \ SD_{INDIVS}	**0.01**	**0.025**	**0.05**	**0.075**	**0.1**	**0.15**
0.01	0.98	0.96	0.93	0.89	0.85	0.78
0.025	0.96	0.95	0.92	0.88	0.85	0.78
0.05	0.93	0.92	0.90	0.87	0.84	0.77
0.075	0.90	0.89	0.87	0.85	0.82	0.76
0.1	0.86	0.86	0.84	0.82	0.80	0.74
0.15	0.79	0.79	0.78	0.76	0.75	0.70

Table S2. Mean value of the dynamic threshold for the unimodal simulation scenario at K=3, under different choices of values of SD_{RUNS} and SD_{INDIVS}. For each choice of SD_{RUNS} and SD_{INDIVS}, 30 simulations were tested. The values obtained for the same settings were averaged across simulations (30 simulations for each setting).

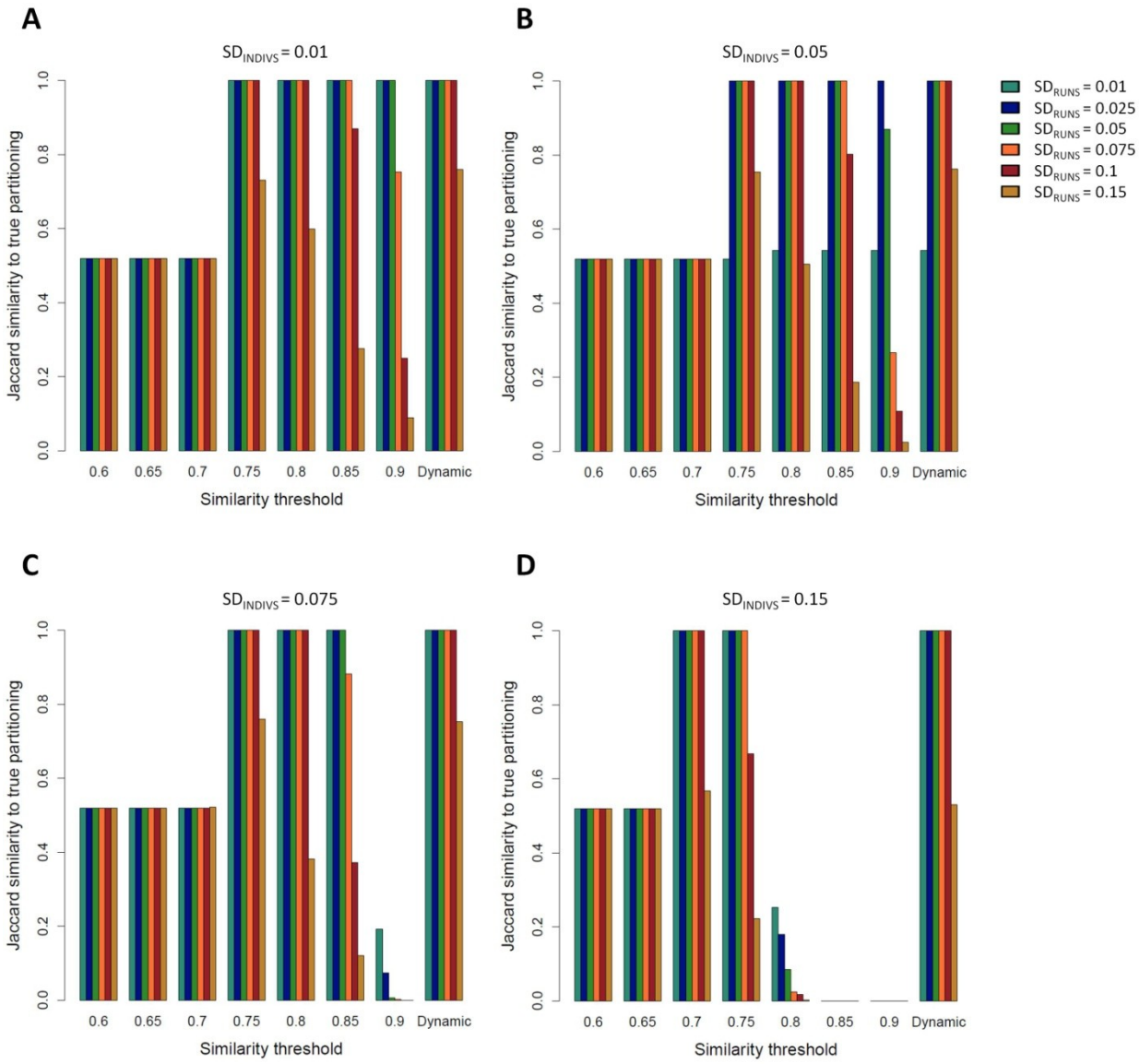
SD_{RUNS} \ SD_{INDIVS}	**0.01**	**0.025**	**0.05**	**0.075**	**0.1**	**0.15**
0.01	0.98	0.96	0.92	0.88	0.85	0.78
0.025	0.96	0.95	0.91	0.88	0.84	0.77
0.05	0.93	0.92	0.90	0.87	0.83	0.76
0.075	0.89	0.89	0.87	0.85	0.82	0.75
0.1	0.86	0.86	0.84	0.82	0.80	0.74
0.15	0.79	0.79	0.78	0.76	0.75	0.70

Table S3. Mean value of the dynamic threshold for the bimodal simulation scenario at K=3, under different choices of values of SD_{RUNS} and SD_{INDIVS}. Simulations were carried out with f=0.25 of the runs assigned to the minor mode. For each choice of SD_{RUNS} and SD_{INDIVS}, 30 simulations were tested. The values obtained for the same settings were averaged across simulations (30 simulations for each setting).

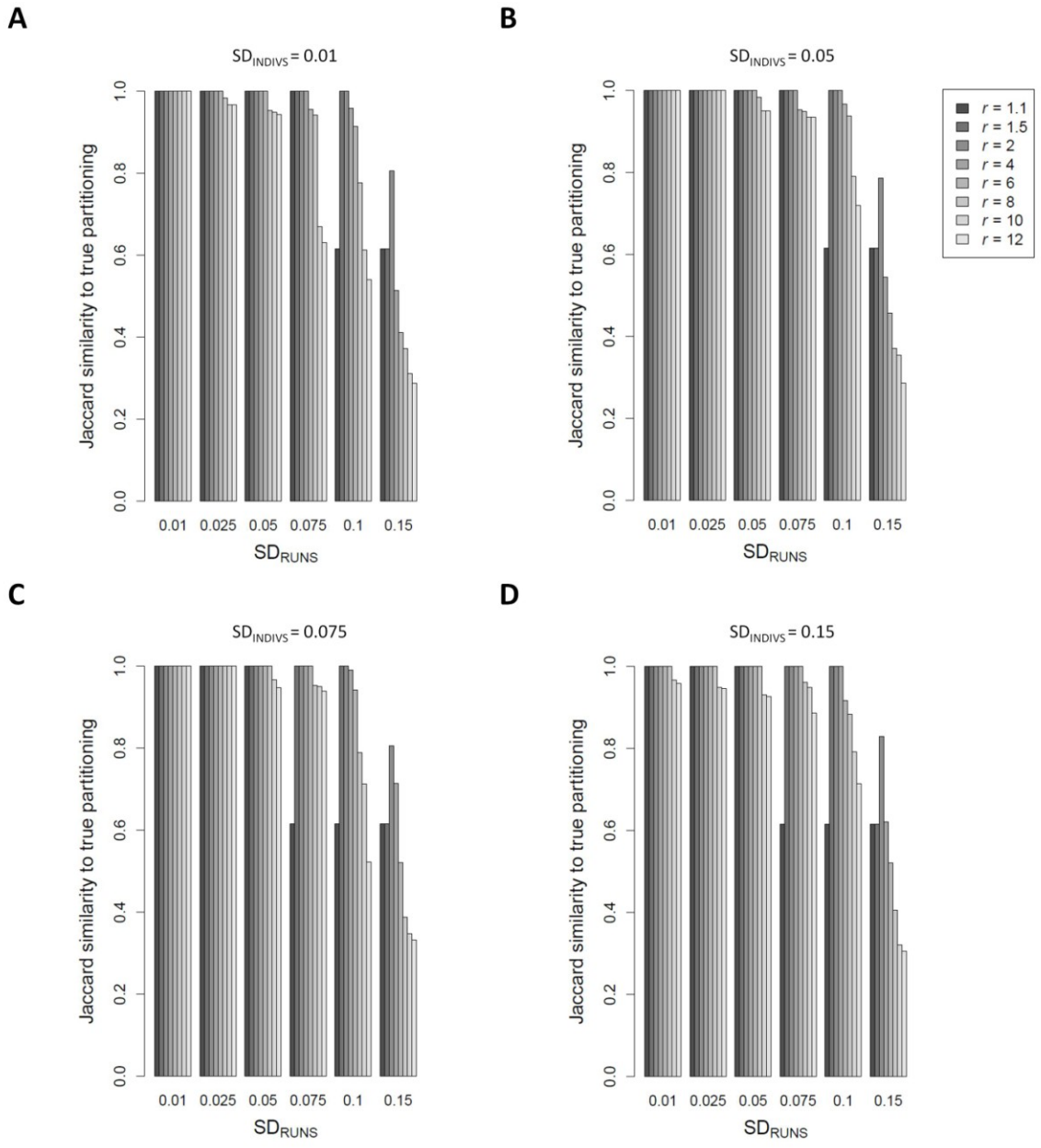
SD_{RUNS} \ SD_{INDIVS}	**0.01**	**0.025**	**0.05**	**0.075**	**0.1**	**0.15**
0.01	0.97	0.95	0.92	0.88	0.84	0.76
0.025	0.95	0.94	0.91	0.87	0.83	0.76
0.05	0.90	0.90	0.88	0.85	0.82	0.75
0.075	0.86	0.85	0.84	0.82	0.79	0.73
0.1	0.82	0.81	0.80	0.79	0.76	0.71
0.15	0.75	0.75	0.74	0.73	0.72	0.67



<u>Figure S1.</u> Jaccard similarity scores between the clustering solution obtained by CLUMPAK and the "true" (simulated) partitioning in a bimodal case, as a function of of SD_{RUNS}, SD_{INDIVS}, and either a fixed threshold value or a dynamic threshold. Simulations were carried out with *f*=0.125 of the runs assigned to the minor mode. (A) SD_{INDIVS}=0.01. (B) SD_{INDIVS}=0.05. (C) SD_{INDIVS}=0.075. (D) SD_{INDIVS}=0.15.



<u>Figure S2.</u> Jaccard similarity scores between the clustering solution obtained by CLUMPAK and the "true" (simulated) partitioning in a bimodal case, as a function of SD_{RUNS}, SD_{INDIVS}, and either a fixed threshold value or a dynamic threshold. Simulations were carried out with a fraction *f*=0.375 of the runs assigned to the minor mode. (A) SD_{INDIVS}=0.01. (B) SD_{INDIVS}=0.05. (C) SD_{INDIVS}=0.075. (D) SD_{INDIVS}=0.15.



<u>Figure S3.</u> Jaccard similarity scores between the clustering solution obtained by CLUMPAK and the "true" (simulated) partitioning in a bimodal case, as a function of SD_{RUNS}, SD_{INDIVS}, and the value of the MCL parameter *r*. Simulations were carried out with a fraction *f*=0.25 of the runs assigned to the minor mode, and the default procedure of dynamically detecting an optimal similarity threshold. (A) SD_{INDIVS}=0.01. (B) SD_{INDIVS}=0.05. (C) SD_{INDIVS}=0.075. (D) SD_{INDIVS}=0.15.

Appendix 1: Seed vectors used in the simulations

Six 3-component "seed" vectors, for the major and minor modes of three populations, were taken from the same dataset used to make Figure 4 (main text). We used the ancestry coefficients of the major and minor modes of the Mozabite, Bedouin, and Druze populations at K=3 (see Figure 4). These populations were chosen because bimodality was evident at K=3. The three seed vectors used in the simulations, representing the major mode are (0.459, 0.036, 0.506), (0.547, 0.385, 0.068), and (0.743, 0.236, 0.021), corresponding to the Mozabite, Bedouin, and Druze populations, respectively. The three simulated seed vectors that correspond the minor mode are: (0.459, 0.535, 0.006), (0.585, 0.398, 0.017), and (0.660, 0.071, 0.269). Thus, for example, if the fraction of runs to be assigned to the minor mode was f=0.25, then for population 1 (Mozabite), in each simulation of 40 runs, 30 were generated based on the seed vector representing the major mode (0.459, 0.036, 0.506), and 10 runs were generated based on the seed vector representing the minor mode (0.459, 0.535, 0.006). In the unimodal simulations conducted using f=0.0, representing a single mode, all runs were sampled from the major mode.

In addition to this set of seed vectors, two other sets of seed vectors were tested, and the results were very similar to those presented in Figures 2 and 3 (results not shown).

Appendix 2: The procedure for setting the parameters of a Dirichlet distribution to obtain predetermined expectations and variances

Let $P=(p_1, p_2, p_3)$ be the target 3-element membership vector, such that $0 \leq p_i \leq 1$, and $p_1 + p_2 + p_3 = 1$. We use the Dirichlet distribution to sample around P, varying the level of variance around the target vector using a concentration parameter α. Let $X=\{x_i\} \sim Dir(\alpha \cdot p_1, \alpha \cdot p_2, \alpha \cdot p_3)$. Following the properties of a Dirichlet distribution, $E(x_i)=p_i$ for i=1, 2, 3. Thus, the expected values are determined by the target vector P and not by α. However, by increasing α, the sample variance decreases, so that the samples are more centered around the expected values.

Specifically, given a desired variance v (and standard deviation SD), we wish to find α such that $v = Var(x_i) = \frac{p_i(1-p_i)}{\alpha+1}$, i=1,2,3. This strategy, however, leads to an overdetermined system of equations because we have three constraints (one for each i) for the single α parameter. We therefore set the value of α according to the desired variance of the first component, so that $\alpha = \frac{p_1 \cdot (1-p_1)-v}{v}$. We note that although α is adjusted only according to the first component, it acts as a concentration parameter for all three components, as higher α values decrease the variance of all components.

REFERENCES

International HapMap Consortium, Altshuler DM, Gibbs RA *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52-58.

Kopelman NM, 2014 The complex genealogy of jewish populations, pp. 148. *The Department of Zoology*. Tel Aviv University, Tel Aviv.

Li JZ, Absher DM, Tang H *et al.* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100-1104.

Pemberton TJ, Wang C, Li JZ, Rosenberg NA (2010) Inference of unexpected genetic relatedness among individuals in HapMap phase III. *American Journal of Human Genetics*, **87**, 457-464.

Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of Human Genetics*, **70**, 841-847.