SUPPLEMENTARY MATERIAL
Similarity matrices and clustering algorithms for population identification using genetic data

Daniel John Lawson*and Daniel Falush†

March 1, 2012

Contents

List of Figures

*Heilbronn Institute, Department of Mathematics, University of Bristol, Bristol, BS8 1TW, UK

†Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig Germany

S1 Clustering score based on the ChromoPainter matrix

The score measuring the ratio of observed to expected variance within a population is constructed by measuring the empirical variance for each pair of populations a and b. Consider the matrix of chunk counts donated to population a from population b, i.e. x_{ij} with $i \in a$ and $j \in b$. These pairs take values that are expected to be exchangable if there is no substructure. Moreover, the distribution is known from Propositions 1–4 of Lawson *et al.* (2012) and is effectively the FineSTRUCTURE model. The vector x_i is multinomial with probability p_{q_j} where $q_j = b$ is the population to which individual j belongs. Therefore (ignoring correlations between populations) x_{ij} is (marginally) a binomial distribution with probability parameter p_{ab} and number of samples $L_i = \sum_j x_{ij}$. In practise there is little value to allowing each individual to have its own L and so we instead use the model

$$x_{ij} \sim \text{Binomial}(\hat{p}_{ab}, L_a) \tag{S1}$$

where $\hat{p}_{ab} = x_{ab}/L_a$, the number of samples $L_a = \sum_b x_{ab}/(n_a^* n_b^*)$, and $x_{ab} = \sum_{i \in a, j \in b} x_{ij}$. n_a^* is the number of individuals in population a, and n_b^* is the number of individuals in population b (minus one if $a = b$ as we exclude the elements where $i = j$). The expected variance under this model is $V_{ab}^E = L_a \hat{p}_{ab}(1 - \hat{p}_{ab})$. This can be compared to the empirical observed variance V_{ab}^O of the x_{ij}.

We form an N by N matrix of these score ratios V^O/V^E, i.e. for each pair of individuals. The reported score for each individual is the average score for each row of this matrix (and is the same for all individuals in a population). We note that this test will identify substructure if the distribution of x_{ij} does not fit the model. It will not identify substructure when the distribution fits the model but neverless contains block structure, even though FineSTRUCTURE may identify substructure in this case.

Note that the application of this score to any non-ChromoPainter matrix has no intrisic meaning as there is no 'effective number of chunks' to define the overall normalisation of the matrix. However, note that rescaling a similarity matrix Z to Z^{c^2} divides the score by c for all pairs of populations. Therefore we can still evaluate the score, but normalise the 'effective number of chunks' in the binomial to make the average score 1 (as is done for IBD in Figure 5 of the main text). Although the score now has no absolute meaning, it still describes the relative variance within the population to between populations, and as is observed from the figure this still captures some types of clustering error.

S2 Additional approaches

Although a comprehensive review is not practical given the breadth of subjects considered here, it is worth listing some approaches that have been considered but that we have not evaluated directly.

1. i2ppca Intarapanich *et al.* (2009): we did try to evaluate this program but did could not identify more than 2 populations with 50 simulated data regions. This does not mean that it cannot be made to work on our dataset however. The approach uses the covariance matrix, and then Spectral decomposition using either the Tracy-Widom or 'Eigendev' Limpiti *et al.* (2011) criterion for choosing the number of Eigenvectors. Clustering is done via a varient of K-Means ('fuzzy K-Means'). The approach uses a unique type of hierarchical analysis, which removes sections of the covariance matrix that are 'distant'. Although this will add robustness against outliers or strong signals in other parts of the data, it will also discard information.

2. The approach of Biswas *et al.* (2009), which uses the normalised covariance matrix and PCA decomposition, with the Tracy-Widom statistic. This is in order to select the SNPs (i.e. thinning) that most correlate with population structure (i.e. the large PCs), for application of STRUCTURE Pritchard *et al.* (2000).

3. AWClust Gao and Starmer (2007) uses the Allele Sharing Distance, and then applies the Ward (1963) criterion directly to the similarity matrix. The number of populations is estimated by the gap statistic Tibshirani *et al.* (2001). We did not evaluate the gap statistic here, but it is not likely Lee *et al.* (2009) to be better than the BIC used by MCLUST.

4. The approach of Reeves and Richards (2009) uses the Jaccard similarity measure which for haploids is $Y_{il}Y_{jl}/f_l + (1-Y_{il})(1-Y_{jl})/(1-f_l)$ per SNP l. This falls between the covariance and normalised covariance matrix. They then clusters all components using the MODECLUS procedure in SAS 9.1 (SAS Institute, Cary, NC), which is one of many 'density valley finding' techniques.

5. The approach of Liu and Zhao (2006) uses a diffent similarity measure, the cosine similarity $Y_i' \cdot Y_j'/(\|Y_i'\|\|Y_j'\|$. This is *first* normalised in a non-standard way for population genetics, by as $Y_{il}' = (\sum_l Y_{il}/L) * \log(1/f_l)$, called 'term frequency - inverse document frequency' (tf-idf) for these respective components. See Manning *et al.* (2008) for more details of these and many other measures used in information retreaval problems. For reading this literature it is useful to think of individuals as 'documents' and SNPs

as 'terms'. SVD is applied to the similarity matrix, and both an MCLUST-like and K-Means clustering is obtained for the full set of Eigenvalues. Gap statistics are used to estimate K.

6. POPSTRUCT of Nakamura *et al.* (2005) uses the Allele Sharing Distance and applies hierarchical k-means based on Ward's criterion to the raw distance matrix. The number of clusters K is determined by cross-validation.

7. Gao and Martin (2009) provide a list of papers using Identity-By-State and/or Allele Sharing Distance, and develop some theory for clustering on the ASD.

S3 HGDP Results

For application of Spectral approaches to the CPL data, we find that the MAP criterion keeps 12 Eigenvalues, PA keeps 5, and Tracy-Widom keeps 14. Direct application of MCLUST and K-means finds only 2 populations (Japanese vs non-Japanese). Perhaps surprisingly Spectral MCLUST with all choices of Eigenvalue retained finds a different (though correlated) set of 9 populations, and Spectral K-Means finds 10. For the CPU data, the MAP criterion keeps 9 Eigenvalues, PA keeps 4, and Tracy-Widom 10. As on the CPL data, all Spectral MCLUST methods find 9 populations on CPU data, Spectral K-Means find 10 populations but the direct method finds only noise. On the IBD matrix, the MAP criterion keeps 11 EVs, the PA keeps 5, and the TW keeps 35 (which is indicative of a problem in the assumptions, and indeed clsutering is very poor this measure). Although some populations are robustly found, there is no overall consistency and all clusterings differ on a substantial point, both for different criteria on the same similarity matrix and across similarity matrices.

Figure S5 shows the results for the unlinked Coancestry matrix, which should be compared to the linked Coancestry matrix from Figure 4 of the main paper. Again the fineSTRUCTURE results are clear, with a slightly different breakdown of the Han and fewer population splits identified. The MCLUST results are again questionable, for related reasons. Whilst some Han individuals are incorrectly placed in with the Tu, again the most serious mistake is to again place the 2 Naxi and 2 She individuals together. As with the linked matrix K-Means merges the Han inappropriately due to admixture, this time with the She.

Figure S7 shows the Spectral IBD clustering results for the Tracy-Widom criterion ($m = 35$), which performs very poorly indeed. MCLUST incorrectly splits the Japanese and splits the Tu and K-Means arbitrarily splits the Yi/Naxi cluster. Both make mistakes with other individuals. Figure S8 shows the Spectral IBD clustering results for the Horn criterion ($m = 5$), which is better. MCLUST

makes a few mistakes (one Yi individual is clearly wrong) and K-Means incorrectly merges the Yi and Han/Tujia/Han.NChina groups, again due to admixture.

S3.1 HGDP Spectral Correlation

Comparison of Figures S6 and S9 is illustrative of how information might be lost by use of naive Spectral methods. Focusing on the She1/She4 pair, which are always clustered with the Naxi3/Naxi6 pair, the correlation between the Eigenvalues (Figure S6) shows that these individuals are distinguishable by being poorly correlated with any individuals in the sample. However, the correlation between the their similarity scores (Figure S9) finds them to be strongly correlated with their true populations (the remaining She and Naxi respectively). It is therefore only plausible to cluster these individuals together in the Spectral representation.

Note however that the features that distinguish each pair of individuals as a cluster (their high relatedness) are always present, and a 'corrected' spectral clustering algorithm might still be able to distinguish these pairs as clusters in their own right. The Eigenvalues are not theoretically relevant for clustering as they do not contain any information about the individuals. They simply scale the acceptable variance in each component, a feature that we believe is important in the determination of what we believe a population really is in this case.

S4 Details of algorithms

The data were simulated using SFS_CODE Hernandez (2008) as described in detail in Lawson *et al.* (2012). The file format was converted first with a bespoke script into PHASE format, then converted to BEAGLE Browning and Browning (2011) format using the script phase2beagle.pl. This and several other useful scripts can be downloaded from the `www.paintmychromosomes.com` website. A bespoke R script was used to convert to PLINK Purcell *et al.* (2007) format, which could be used to combine output files. The similarity measures were generated in the following way:

- BEAGLE FastIBD (IBD) was run with the command:

  ```
  beagle.sh missing=? fastibd=true unphased=$inputfile.bgl
  markers=$inputfile.markers ibdpairs=$outputfile.ibdpairs
  out=$outputfile
  ```

  then each pair was summed over all datafiles and all SNPs using a bespoke script.

- FastPHASE (FHS) was run with the command

```
fastPHASE -K20 -T1 -S$seed -Pp -B -H-3 -n -o$outfile.part1 $phasefile
fastPHASE -K20 -T1 -S$seed -Pzp -H-3 -n -S-3 -C0 -I$outfile.part1
-o$outfile $phasefile
```

which is appropriate because the simulated data is pre-phased (otherwise a slightly more lengthly phasing step replaces the first command). The pairwise distance given in Supplementary Section 4.8 Jakobsson *et al.* (2008) was computed with a bespoke script, which was very time consuming to run.

- ChromoPainter (CPL and CPU) matrices were generated as detailed in Lawson *et al.* (2012), and additionally described on the `www.paintmychromosomes.com` website.

- PLINK Purcell *et al.* (2007) was used to generate the main IBS measure using

```
plink-1.07-x86_64/plink --file $file --out ${file}ids3
--cluster --matrix
```

- COV, NCOV(a) and IBS(a) measures were computed from the equations given in Table 1 of the main text.

- PLINK PED/MAP files were converted to EIGENSTRAT format using the 'convertf' program in the EIGENSTRAT package. For the linked (ESL) model the smartpca program was run directly using "altnormstyle: NO, numoutevec: 100 and nsnpldregress: 10". For the standard model (ESU) the "smartpca.perl" script was run with the options "-m 5 -t 100 -s 6.0 -k 100".

The non-standard clustering algorithms were run in the following way:

- FineSTRUCTURE was run as detailed in Lawson *et al.* (2012), and additionally described on the `www.paintmychromosomes.com` website.

- AWClust Gao and Starmer (2007) could not read in the whole dataset of 2.5M SNPs in one go. We therefore performed thinning on the SNPs first (removing SNPS that are highly correlated with other SNPs) using the PLINK "–indep-pairwise 50 10 0.05" option, which keeps around one quarter of all SNPs. This was converted to AWClust input format using a bespoke script. The gap statistic estimation is too costly on this dataset so we 'cheated' by choosing the K that maximised the correlation with the truth.

- i2ppca Intarapanich *et al.* (2009) was run similarly to AWClust, i.e. on the thinned data converted by a bespoke script. It would not run with more than 50 regions, and on 50 regions it found $K = 2$ and had a low correlation with the truth. We therefore did not explore this method further.

References

Biswas, S., L. B. Scheinfeldt, and J. M. Akey, 2009 Genome-wide Insights into the Patterns and Determinants of Fine-Scale Population Structure in Humans. Am J Hum Genet. **15:** 641–650.

Browning, B. L. and S. R. Browning, 2011 A Fast, Powerful Method for Detecting Identity by Descent. Am. J. Hum. Genet. **88:** 173–182.

Gao, X. and E. R. Martin, 2009 Using Allele Sharing Distance for Detecting Human Population Stratification. Hum Hered. **68:** 182–191.

Gao, X. and J. Starmer, 2007 Human population structure detection via multilocus genotype clustering. BMC Genetics **25:** 34.

Hernandez, R., 2008 A flexible forward simulator for populations subject to selection and demography. Bioinformatics **24:** 2786–2787.

Intarapanich, A., P. J. Shaw, A. Assawamakin, P. Wangkumhang, C. Ngamphiw, K. Chaichoompu, J. Piriyapongsa, and S. Tongsima, 2009 Iterative pruning PCA improves resolution of highly structured populations. BMC Bioinformatics **10:** 382.

Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H.-C. Fung, Z. A. Szpiech, J. H. Degnan, K. Wang, R. Guerreiro, J. M. Bras, J. C. Schymick, D. G. Hernandez, B. J. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. M. Cann, J. A. Hardy, N. A. Rosenberg, and A. B. Singleton, 2008, February)Genotype, haplotype and copy-number variation in worldwide human populations. Nature *451*(7181)**:** 998–1003.

Lawson, D. J., G. Hellenthal, S. Myers, and D. Falush, 2012 Inference of population structure using dense haplotype data. PLoS Genetics **8:** e100245.

Lee, C., A. Abdool, and C.-H. Huang, 2009 PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics **10:** S73.

Limpiti, T., A. Intarapanich, A. Assawamakin, P. J. Shaw, P. Wangkumhang, J. Piriyapongsa, C. Ngamphiw, and S. Tongsima,

2011 Study of large and highly stratified population datasets by combining iterative pruning principal component analysis and structure. BMC Bioinformatics **12:** 255.

Liu, N. and H. Zhao, 2006 A non-parametric approach to population structure inference using multilocus genotypes. Hum Genomics **2:** 353–64.

Manning, C. D., P. Raghavan, and H. Schtze, 2008 *Introduction to Information Retrieval*. UK: Cambridge University Press.

Nakamura, T., A. Shoji, H. Fujisawa, and N. Kamatani, 2005 Cluster analysis and association study of structured multilocus genotype data. J. of Hum. Genet. **50:** 53–61.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of Population Structure Using Multilocus Genotype Data. Genetics **155:** 945–959.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. de Bakker, M. Daly, and P. Sham, 2007 PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics **81:** 559–75.

Reeves, P. A. and C. M. Richards, 2009 Accurate inference of subtle population structure (and other genetic discontinuities) using principal coordinates. PLoS One **4:** e4269.

Tibshirani, R., G. Walther, and T. Hastie, 2001 Estimating the number of clusters in a data set via the gap statistic. J. R. Statist. Soc. B **63:** 411–423.

Ward, J. H., 1963 Hierarchical grouping to optimize an objective function. Journal of Amer. Statist. Assoc. **58:** 236–244.

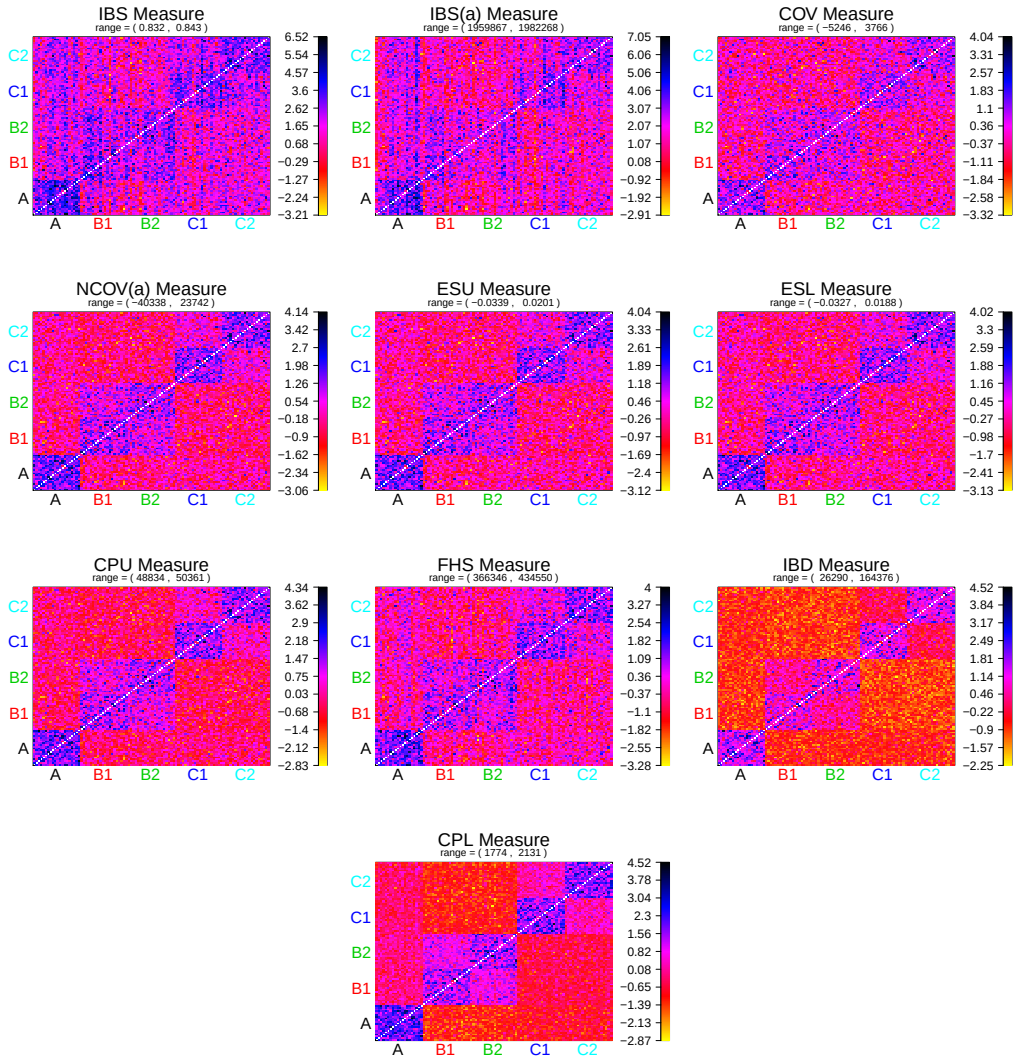

Figure S1: Visualisation of the similarity matrices as an image for one hundred 5Mb regions of simulated data. On the top row from left to right are: IBS (Identity-by-state as computed by PLINK), IBSa (as computed using the equation from Table 1 of the text) and COV (Raw Covariance). On the second row is NCOVa (as computed using the equation from Table 1 of the main text), ESU (Eigenstrat 'unlinked'), and ESL (Eigenstrat 'linked', i.e. using regression correction with $K = 10$). The third row is: CPU (ChromoPainter unlinked), FHS (FastPHASE Haplotype Sharing), IBD (FastIBD Identity By Descent). Finally CPL (ChromoPainter Linked) is on the fourth row. Each matrix has the diagonal removed as this is not informative about population structuring. The raw range is given above each matrix, but all plots are normalised by removing the diagonal, subtracting the row means and making the standard deviation 1.

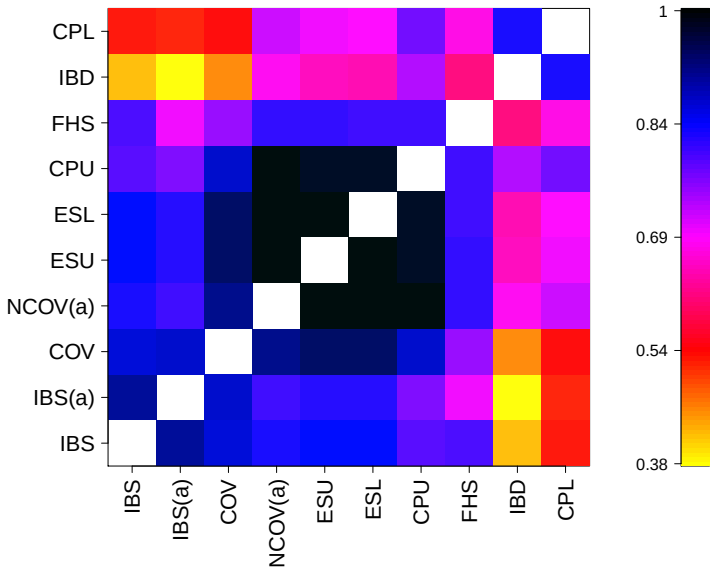

Figure S2: Correlation between the similarity measures considered here for fifty regions of simulated data. The labels are as given in Figure S1. There is clear block structure to this matrix and only a single choice from each class of unlinked methods is considered.

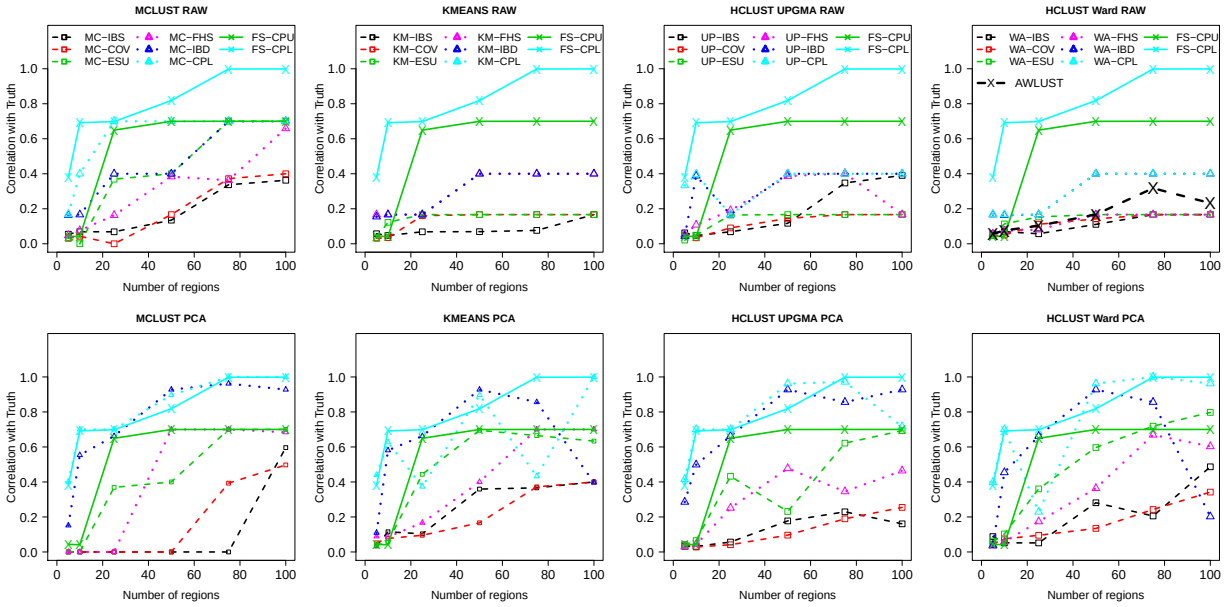

Figure S3: Correlation with the truth as a function of the number of 5Mb simulated data regions, for MCLUST (MC, left), K-Means(KM) and UPGMA (UP) and Ward's minimum variance criterion (WA, right). The top row shows the results for application to the 'raw' data. The bottom row shows the Spectral (PCA) results using the Tracy-Widom criterion, repeating some of the information from Figure 3 of the main text. Shown are the clustering performance based on different similarity matrices: IBS (Identity-by-state), COV (Covariance), ESU (Eigenstrat Unlinked), FHS (FastPHASE Haplotype Sharing), IBD (FastIBD Identity By Descent) and CPL (ChromoPainter Linked). FineSTRUCTURE is applied directly to the coancestry matrix only (FS-CPU for unlinked and FS-CPL for the linked ChromoPainter Coancestry matrix), and is repeated on each plot for reference. AWClust Gao and Starmer (2007) results are shown in the 'raw Ward' figure as it relates to this approach.

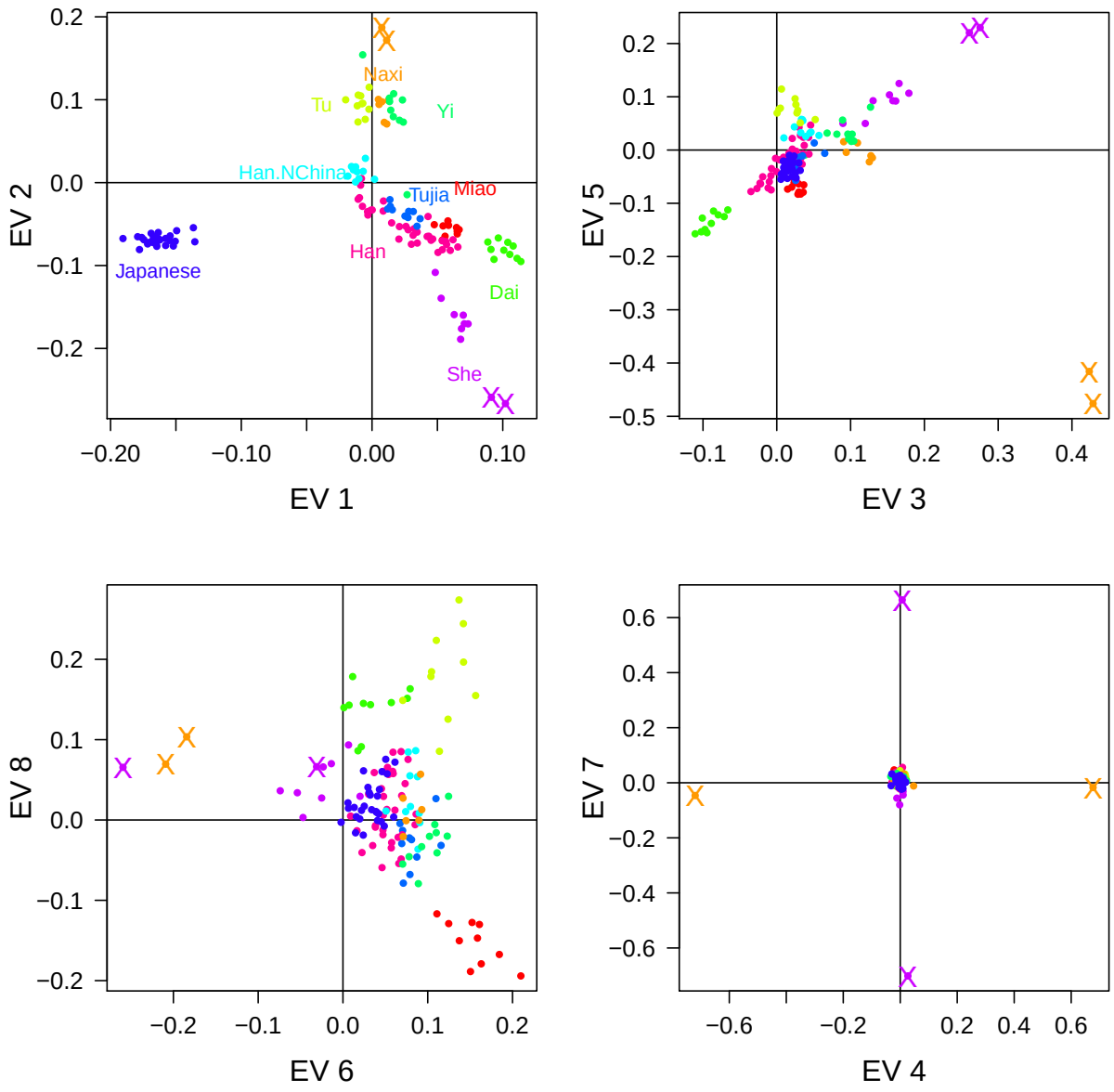

Figure S4: PCA decomposition of the first 8 Eigenvalues for the East Asian HGDP ChromoPainter Linked (CPL) similarity matrix. The individuals marked with a cross are the misplaced Naxi/She relative pairs. The colour coding of populations is consistent throughout the plots. Note that the EVs (Eigenvectors) are not plotted in order; EV's 4 and 7 have been plotted together as they simply separate the related pairs.

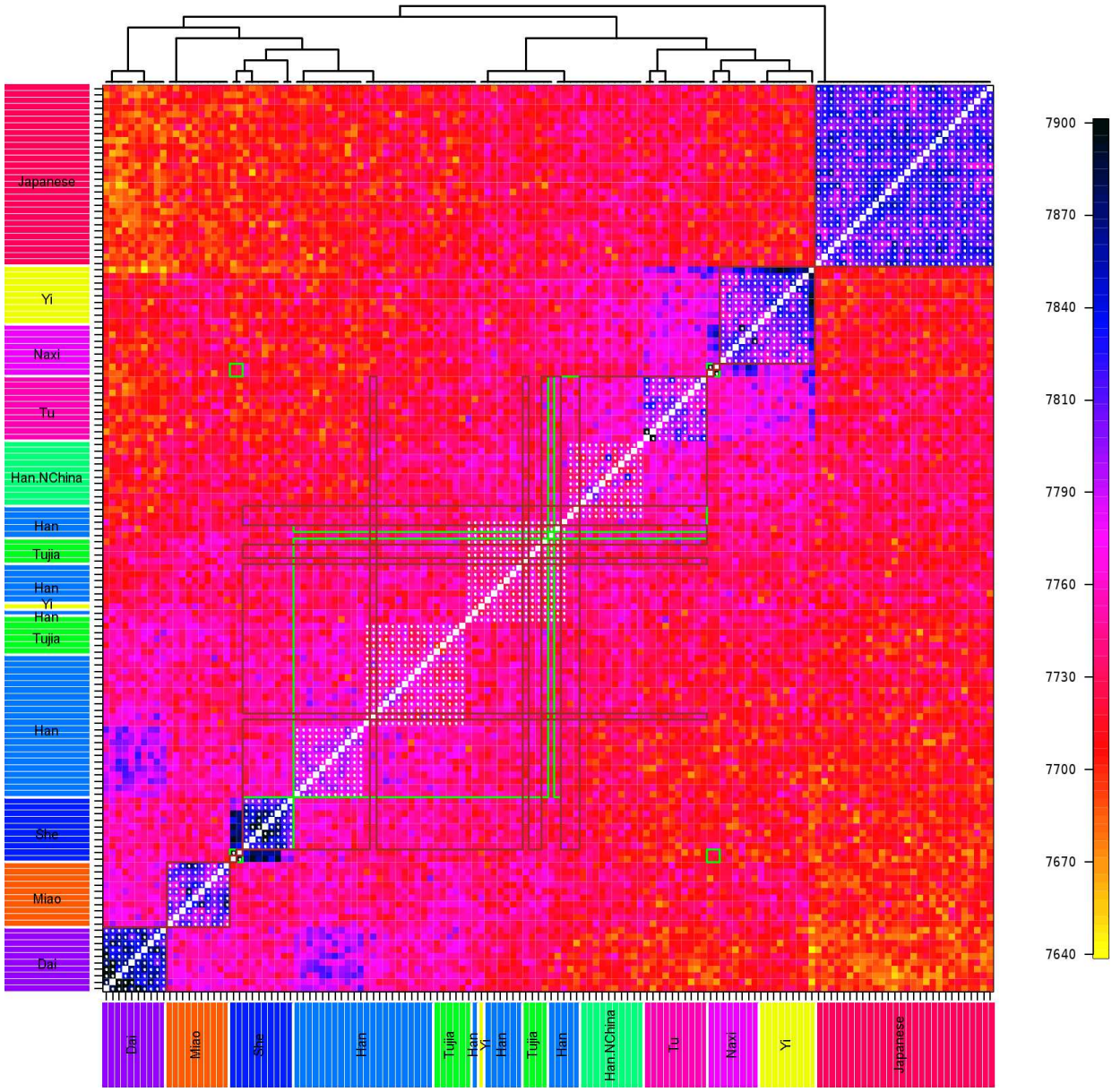

Figure S5: HGDP clustering results and coancestry matrix for the ChromoPainter unlinked (CPU) dataset. The main image shows the Coancestry matrix. The boxes (white for FineSTRUCTURE, green for Spectral MCLUST and brown for Spectral K-means both using the Tracy-Widom criterion) show the pairwise coincidence of the Maximum Aposteriori clustering for each method. The ordering has been chosen so as to minimise off-diagonal elements for the MCLUST method whilst respecting the FineSTRUCTURE clustering. The Coancestry matrix has been capped at 7900 to maximise contrast.

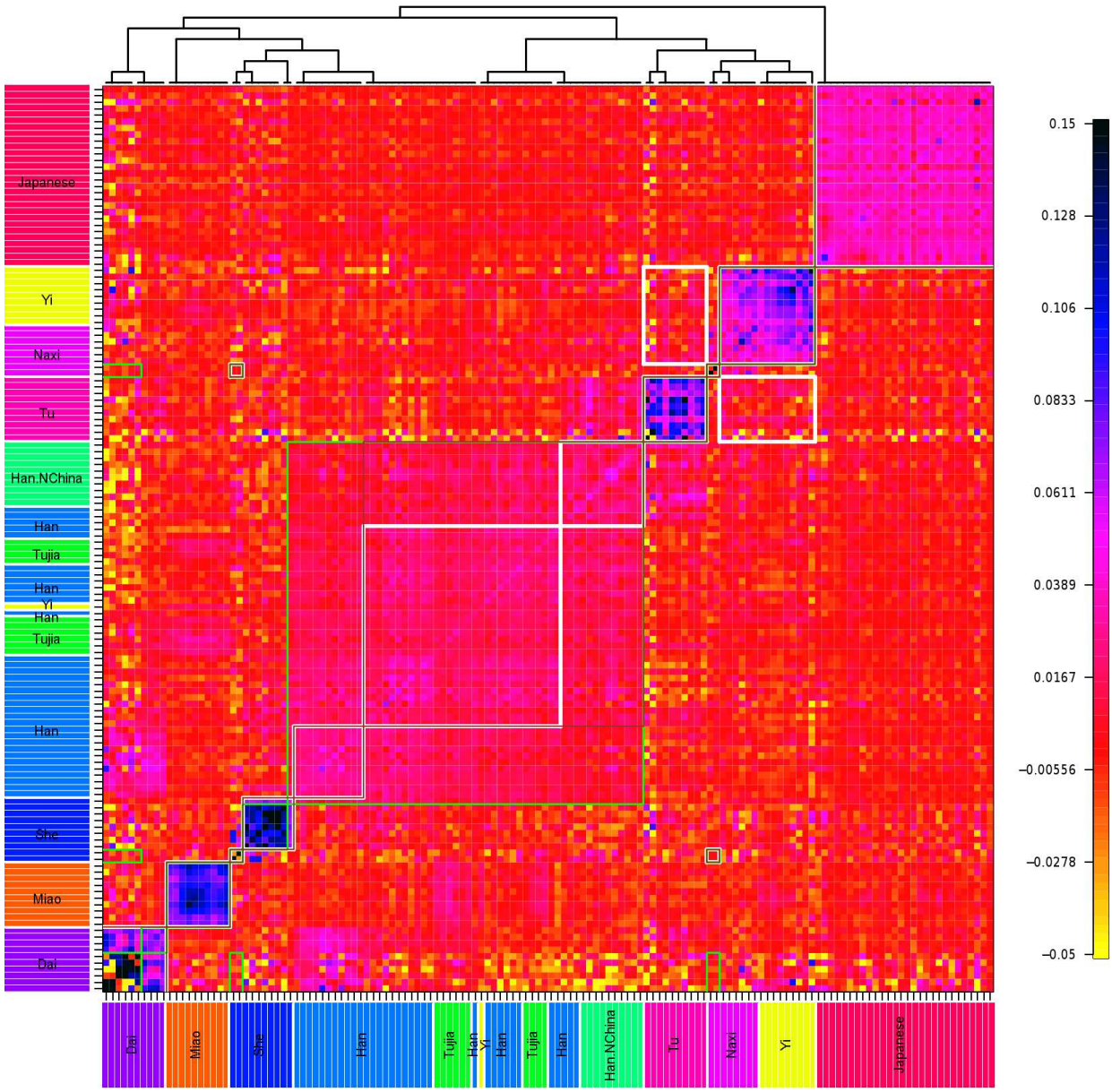

Figure S6: East Asian HGDP ChromoPainter Linked (CPL) similarity matrix MCLUST clustering results, with the Spectral correlation matrix for the ChromoPainter linked (CPL) dataset. The main image shows the Spectral correlation matrix, formed by taking the product EE^T of the $m = 14$ retained Eigenvectors E (with m chosen by the Tracy-Widom criterion). The boxes reflect pairwise coincidence of the MCLUST Maximum Aposteriori clustering for different choices of m: green for the Tracy-Widom criterion with $m = 14$, white for the PA criterion with $m = 5$, and brown for the MAP criterion with $m = 12$. The correlation is capped at a minimum of -0.05 and a maximum of 0.15 to maximise contrast.

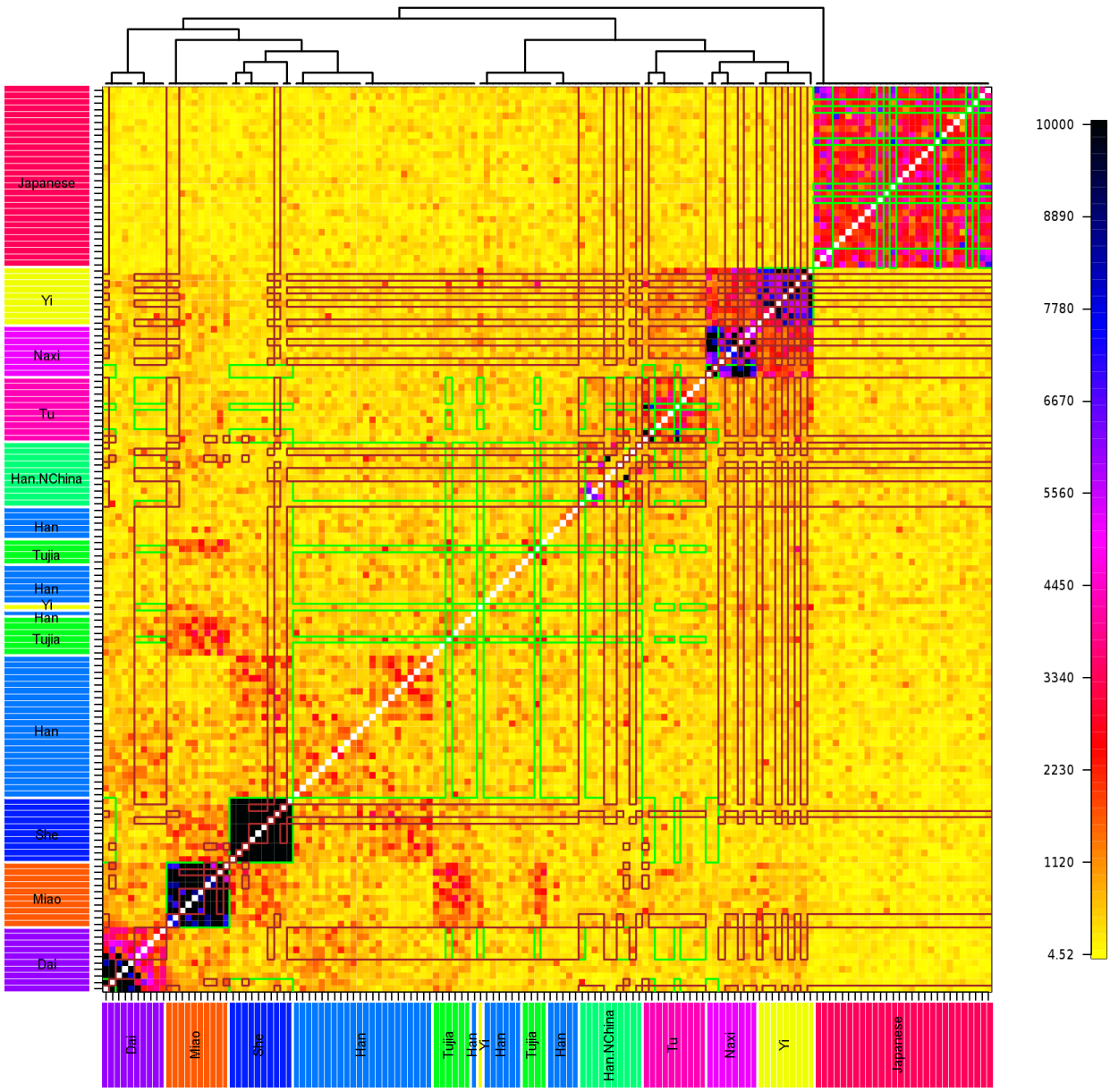

Figure S7: HGDP similarity matrix for the FastIBD dataset. The main image shows the FastIBD similarity measure applied to the East Asian individuals, as shown in Figure 6 of the main text. Spectral results using the Tracy-Widom ($m = 35$) statistic are shown; these should be compared to Supplementary Figure S8. Green boxes are drawn around pairs found which coincide in the Spectral MCLUST populations, and brown boxes for Spectral K-means. The ordering is formed from the FineSTRUCTURE tree. The similarity matrix has been capped at 10000 to maximise contrast.

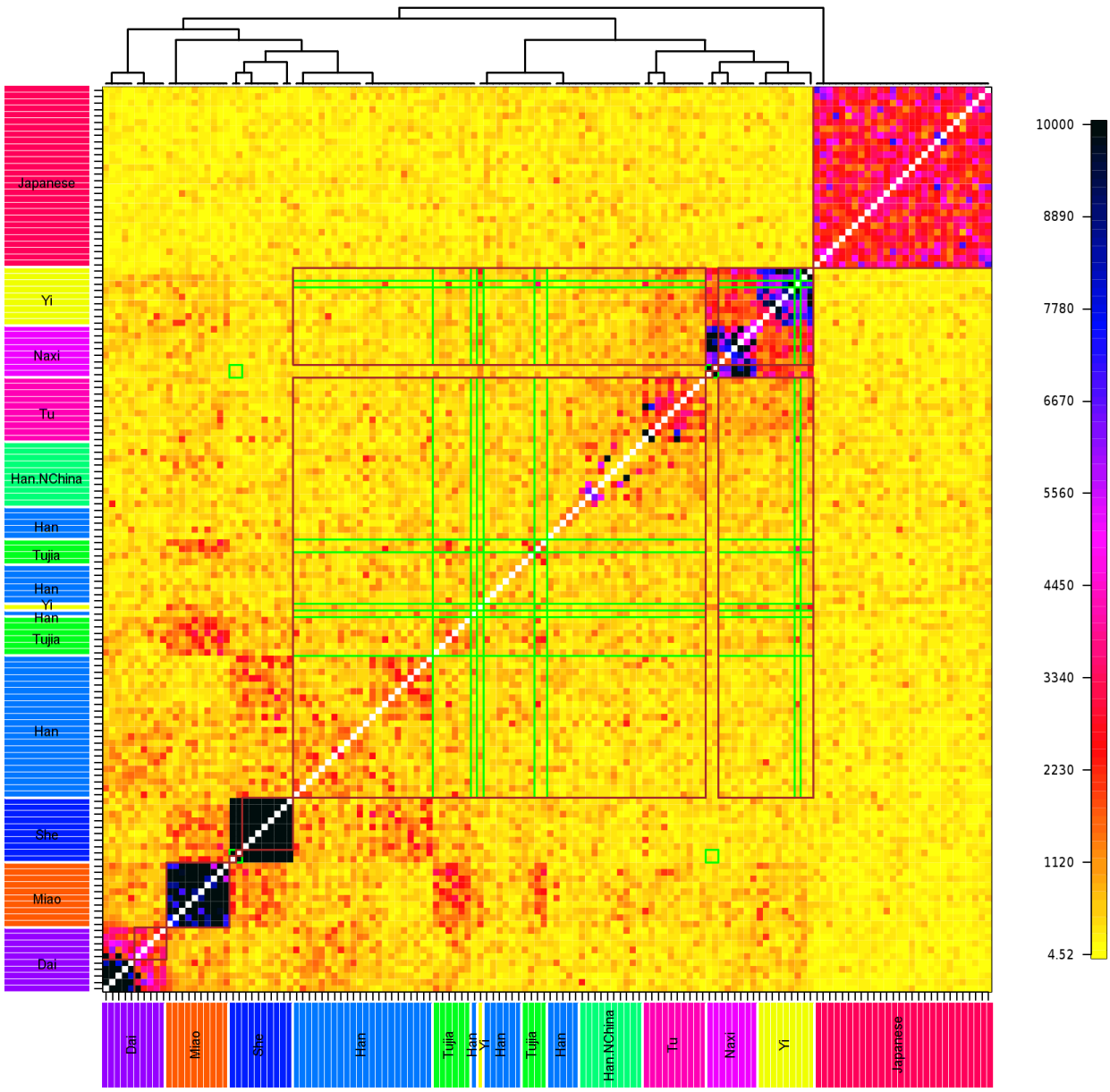

Figure S8: HGDP similarity matrix for the FastIBD dataset. The main image shows the FastIBD similarity measure applied to the East Asian individuals, as shown in Figure 6 of the main text. Spectral results using the PA criterion ($m = 7$) are used as these were subjectively the best; these should be compared to Supplementary Figure S7. Green boxes are drawn around pairs found which coincide in the Spectral MCLUST populations, and brown boxes for Spectral K-means. Other details are as Supplementary Figure S7.

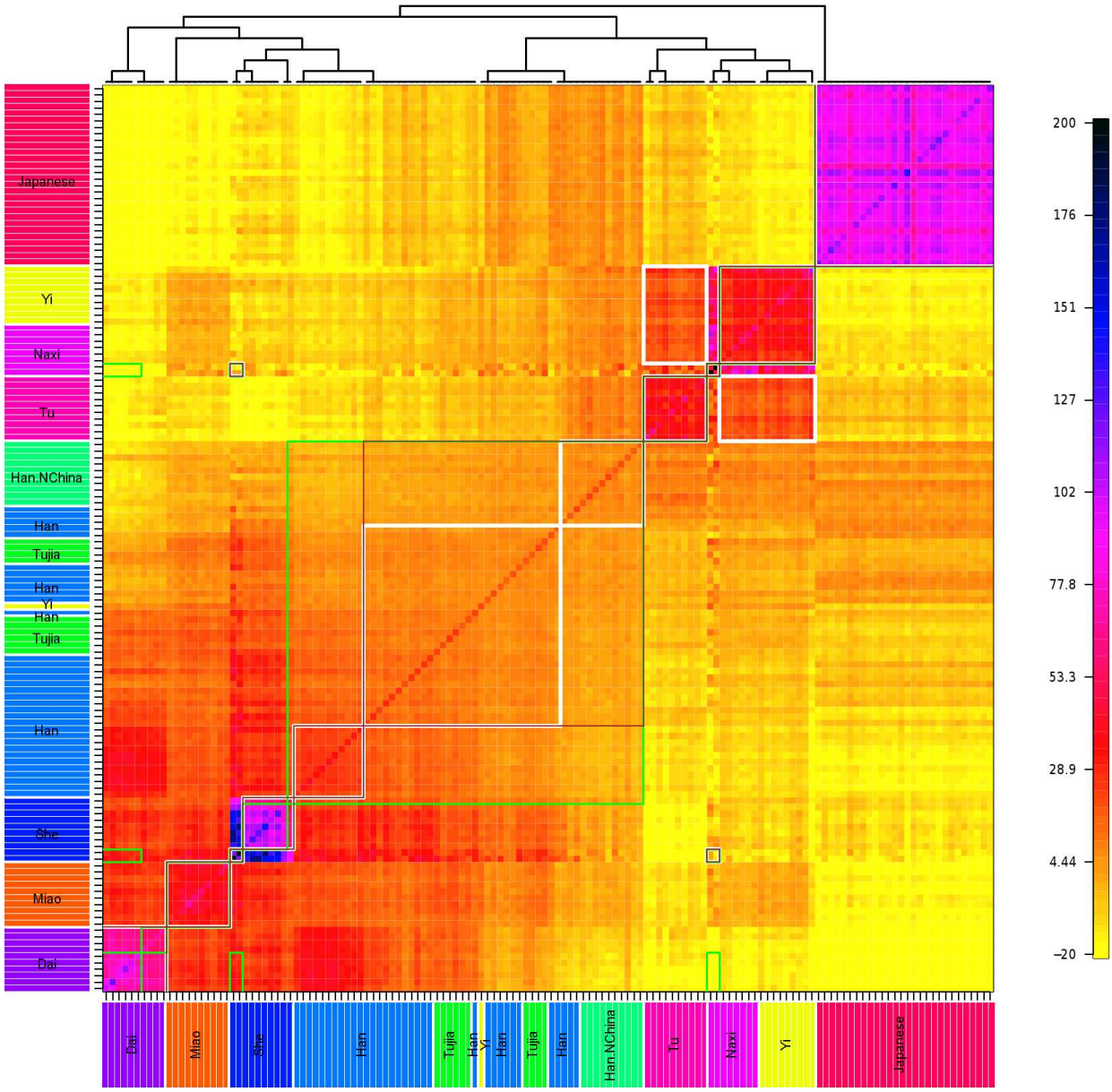

Figure S9: HGDP MCLUST clustering results and similarity correlation matrix for the ChromoPainter unlinked (CPU) dataset. The main image shows the 'PCA reconstructed similarity' correlation matrix, formed by taking the product $E\text{Diag}(\lambda)E^T$ of the $m = 14$ retained Eigenvectors E (with m chosen by the Tracy-Widom criterion). The information content is identical to that of Figure S6, but the eigenvalues have been used to scale the EVs. The boxes reflect the same Spectral MCLUST clusterings in Figure S6. Correlation is capped at a minimum of -20 and a maximum of 200 to maximise contrast.

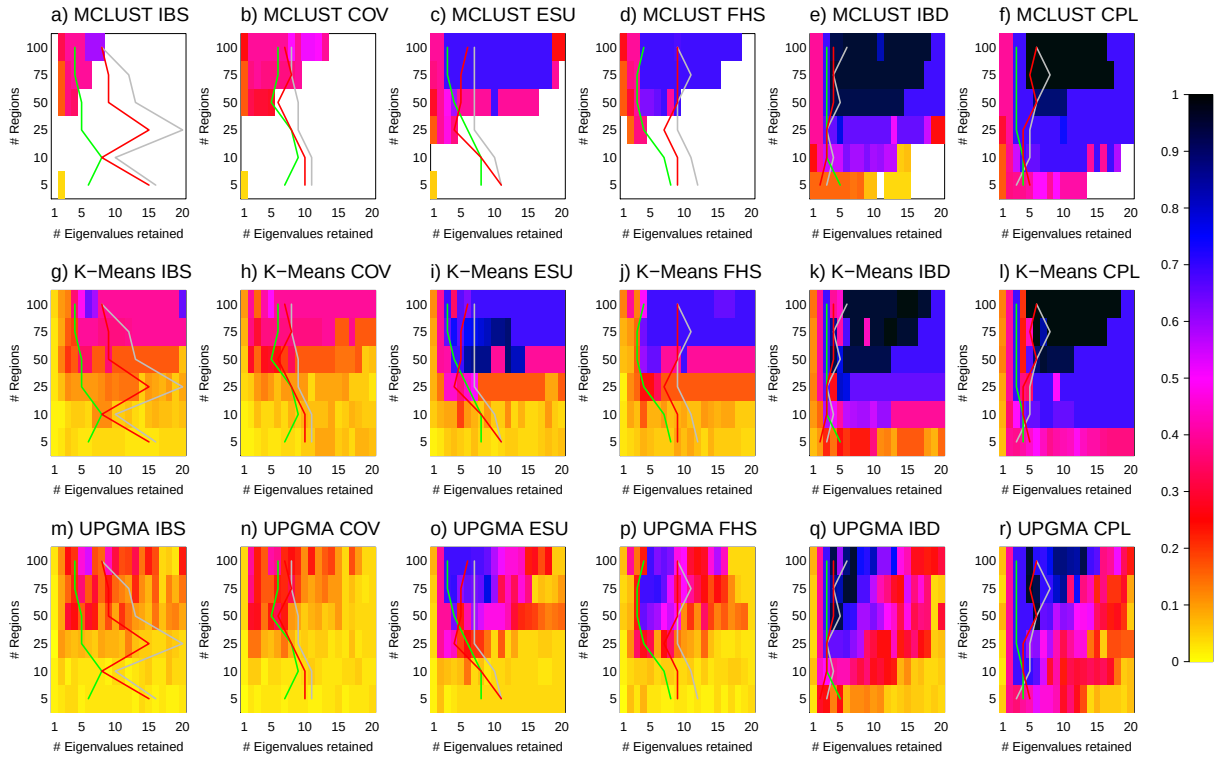

Figure S10: Correlation with the truth as a function of both the number of regions, and the number of Eigenvalues retained for fitting, for (top) the MCLUST model, (centre) the K-means model, and (bottom) the UPGMA model. MCLUST runs that failed to estimate K are white. From left to right are: IBS (Identity-by-state), COV (Covariance), ESU (Eigenstrat Unlinked), FHS (FastPHASE Haplotype Sharing), IBD (Identity By Descent) and CPL (ChromoPainter Linked). The grey line corresponds to the number of Eigenvectors according to the MAP criterion, green lines to the PA criterion, and red lines to the Tracy-Widom criterion.