PopCluster: an algorithm to identify genetic variants with ethnicity-dependent effects

Supplementary Data

Anastasia Gurinovich[1*], Harold Bae[2], John J. Farrell[3], Stacy L. Andersen[4], Stefano Monti[5], Annibale Puca[6,7], Gil Atzmon[8,9], Nir Barzilai[9], Thomas T. Perls[4], Paola Sebastiani[10]

1 Bioinformatics Program, Boston University, Boston, MA, USA
2 College of Public Health and Human Sciences, Oregon State University, Corvallis, OR, USA
3 Division of Genetics, Department of Medicine, Boston University School of Medicine, Boston, MA, USA
4 Geriatrics Section, Department of Medicine, Boston University School of Medicine & Boston Medical Center, Boston, MA, USA
5 Section of Computational Biomedicine, Boston University School of Medicine, Boston, MA, USA
6 Department of Medicine and Surgery, University of Salerno, Fisciano, SA, Italy
7 Cardiovascular Research Unit, IRCCS MultiMedica, Sesto San Giovanni, MI, Italy
8 Department of Human Biology, Faculty of Natural Sciences, University of Haifa, Israel
9 Institute for Aging Research, Albert Einstein College of Medicine, Bronx, NY, USA
10 Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA

* agurinov@bu.edu

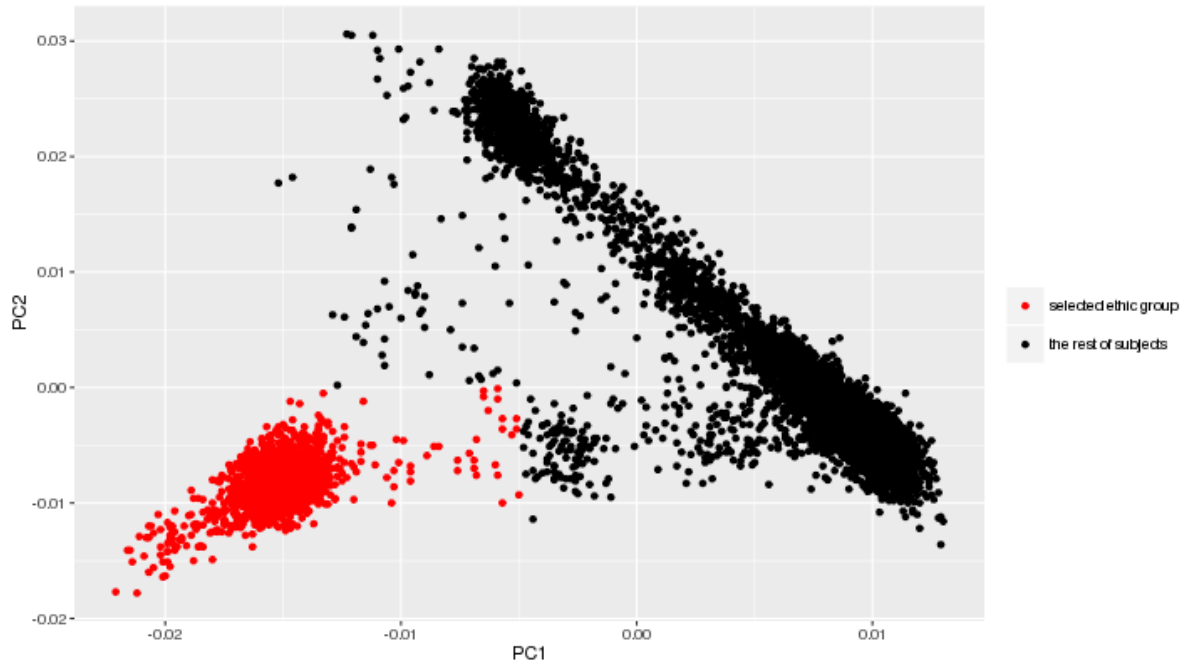

Figure S1: To evaluate the TPR and the FPR we simulated an allele A to be associated with a phenotype of interest but not associated with principal components in a selected ethnic group: subjects with $PC_1 \leq -0.005$ and $PC_2 \leq 0$ (red dots). For the rest of the subjects (black dots), the allele was simulated to be significantly associated with PC_1 and PC_2.

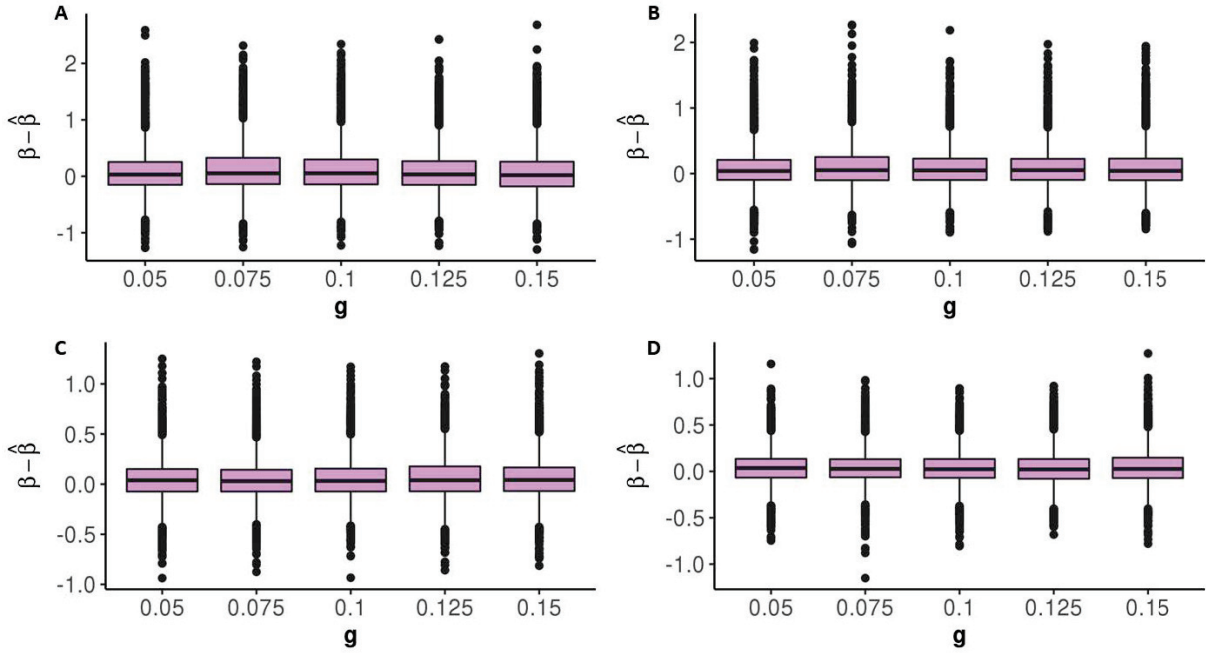

Figure S2: Boxplots of the differences between true effect β and estimated effect $\hat{\beta}$ in the clusters where allele A was simulated to be associated with EL. Each set of boxplots corresponds to the boxplots in Figure 5.

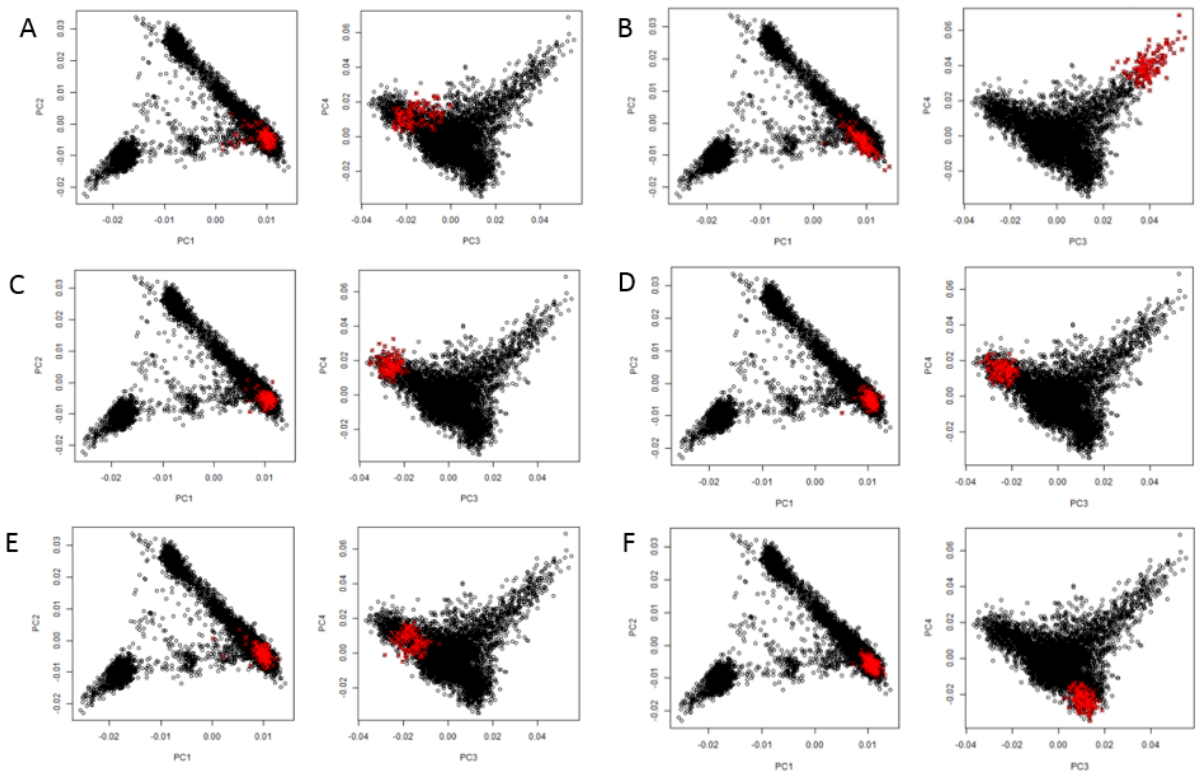

Figure S3: **Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL.** Subjects are colored red if they belong to (A): cluster 118, (B): cluster 126, (C): cluster 128, (D): cluster 129, (E): cluster 133, (F): cluster 134. Subjects not belonging to respective clusters are colored black in every plot.

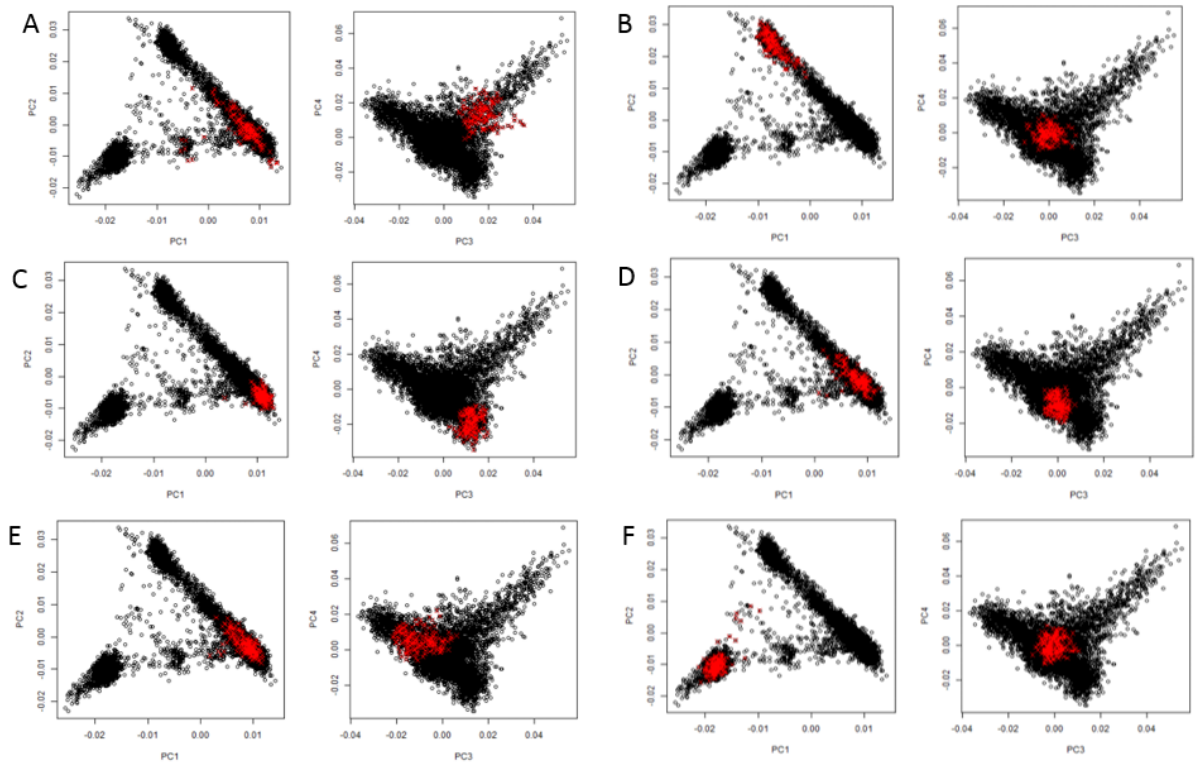

Figure S4: **Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL.** Subjects are colored red if they belong to (A): cluster 141, (B): cluster 148, (C): cluster 153, (D): cluster 160, (E): cluster 170, (F): cluster 176. Subjects not belonging to respective clusters are colored black in every plot.

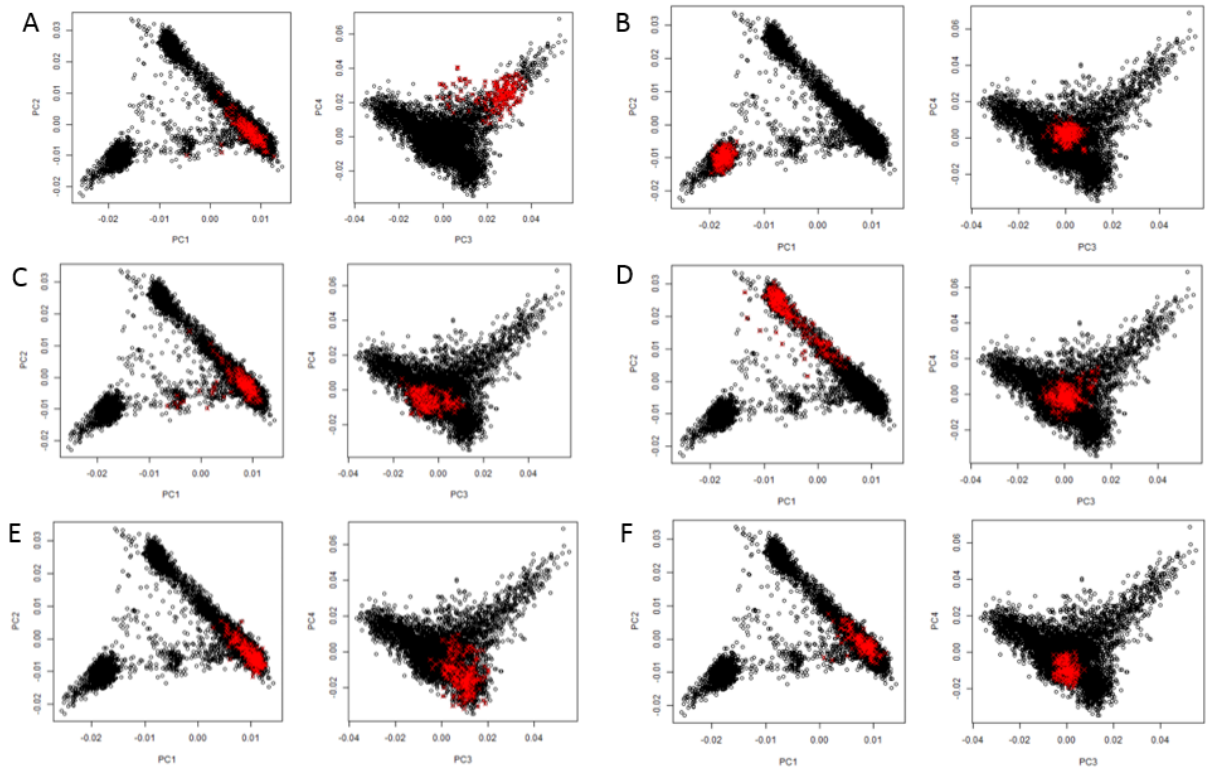

Figure S5: **Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL.** Subjects are colored red if they belong to (A): cluster 180, (B): cluster 193, (C): cluster 194, (D): cluster 240, (E): cluster 249, (F): cluster 253. Subjects not belonging to respective clusters are colored black in every plot.

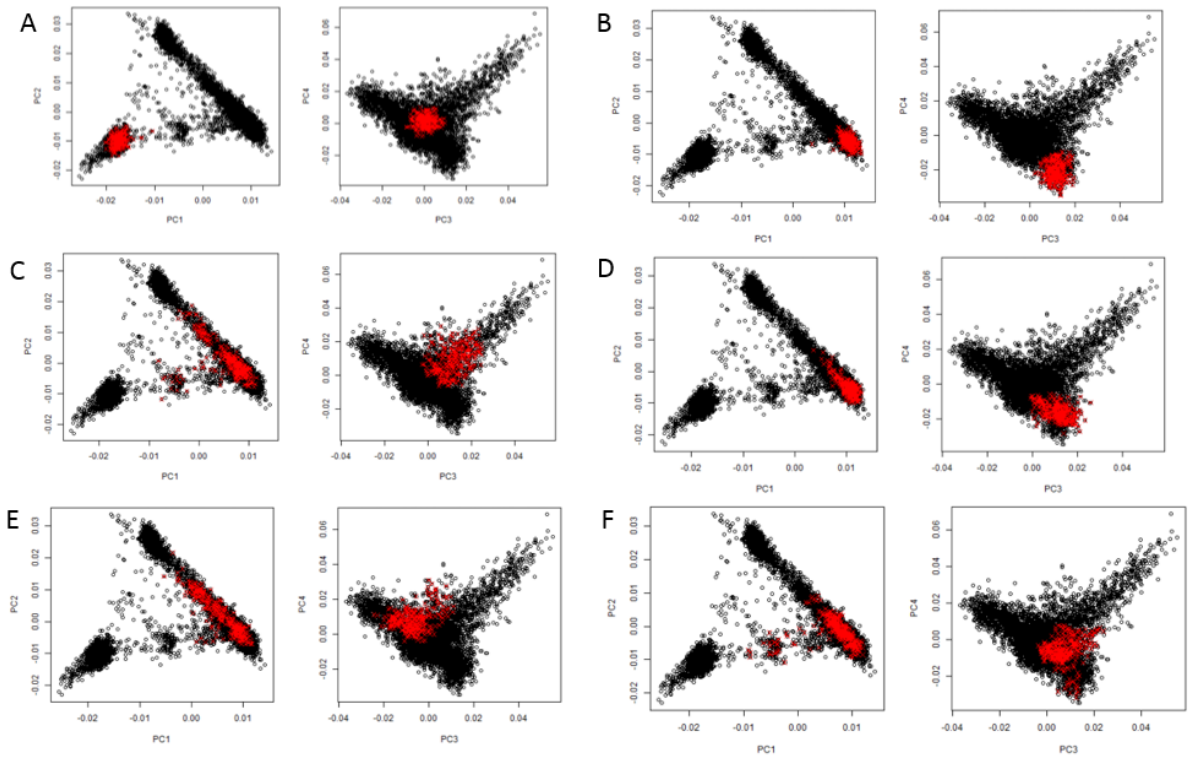

Figure S6: **Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL.** Subjects are colored red if they belong to (A): cluster 274, (B): cluster 287, (C): cluster 290, (D): cluster 296, (E): cluster 297, (F): cluster 316. Subjects not belonging to respective clusters are colored black in every plot.

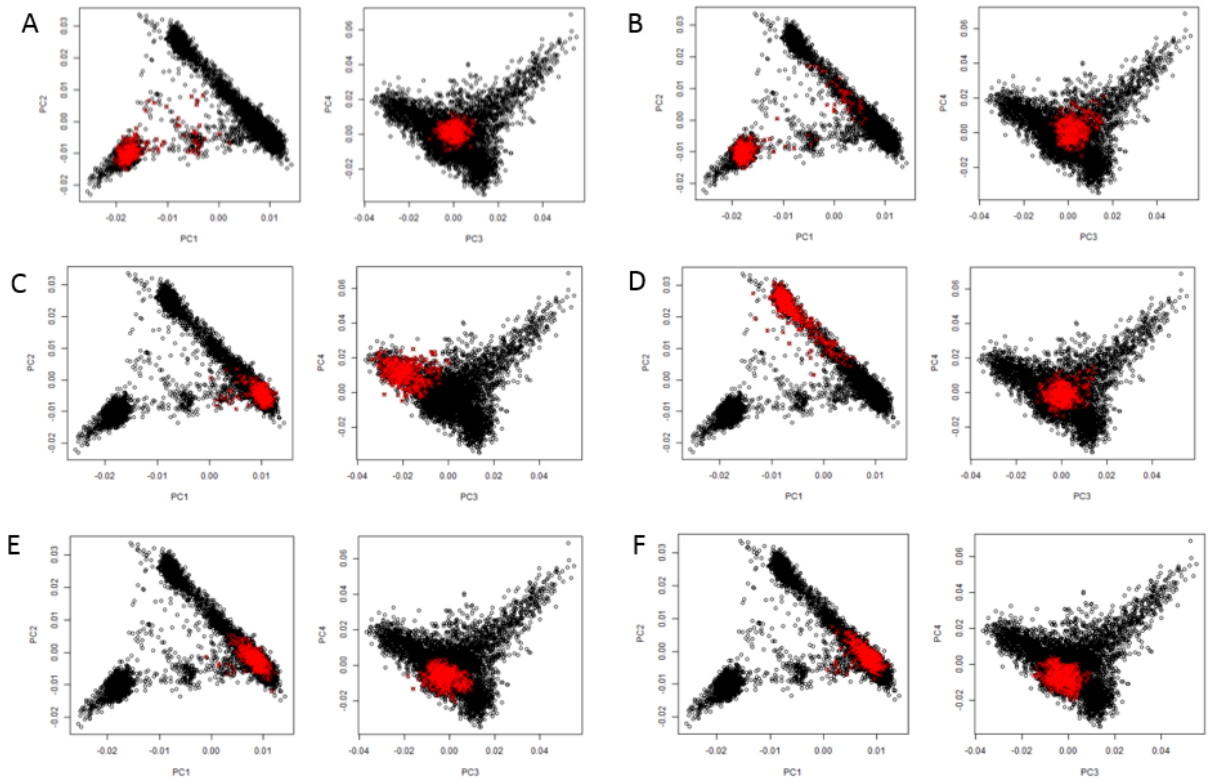

Figure S7: **Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL.** Subjects are colored red if they belong to (A): cluster 330, (B): cluster 348, (C): cluster 380, (D): cluster 388, (E): cluster 396, (F): cluster 413. Subjects not belonging to respective clusters are colored black in every plot.

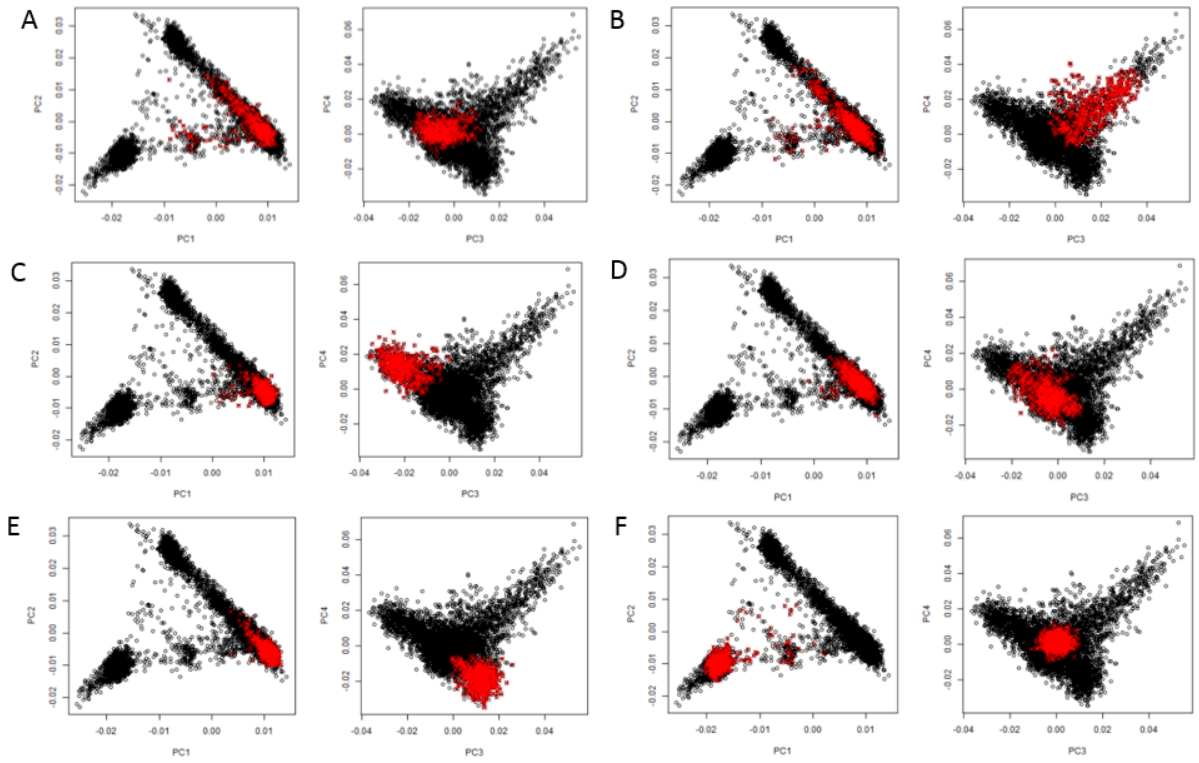

Figure S8: **Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL.** Subjects are colored red if they belong to (A): cluster 416, (B): cluster 470, (C): cluster 508, (D): cluster 566, (E): cluster 583, (F): cluster 604. Subjects not belonging to respective clusters are colored black in every plot.

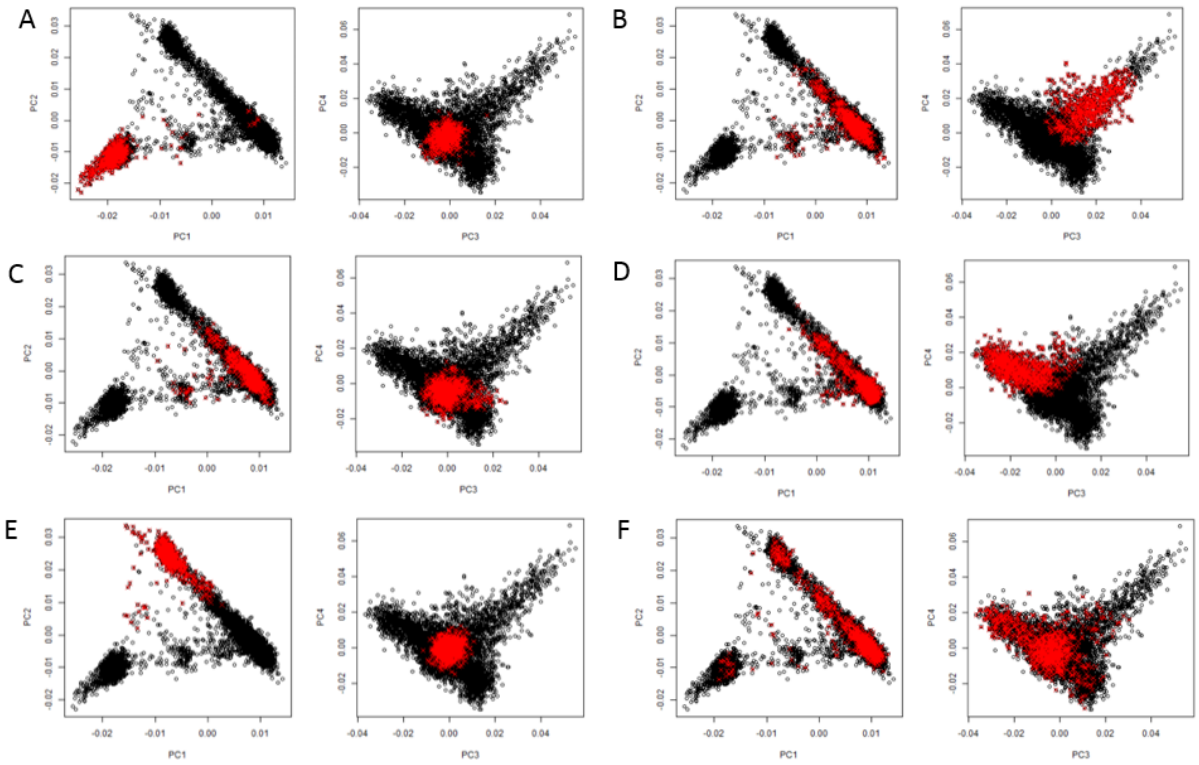

Figure S9: **Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL.** Subjects are colored red if they belong to (A): cluster 606, (B): cluster 611, (C): cluster 721, (D): cluster 805, (E): cluster 818, (F): cluster 828. Subjects not belonging to respective clusters are colored black in every plot.

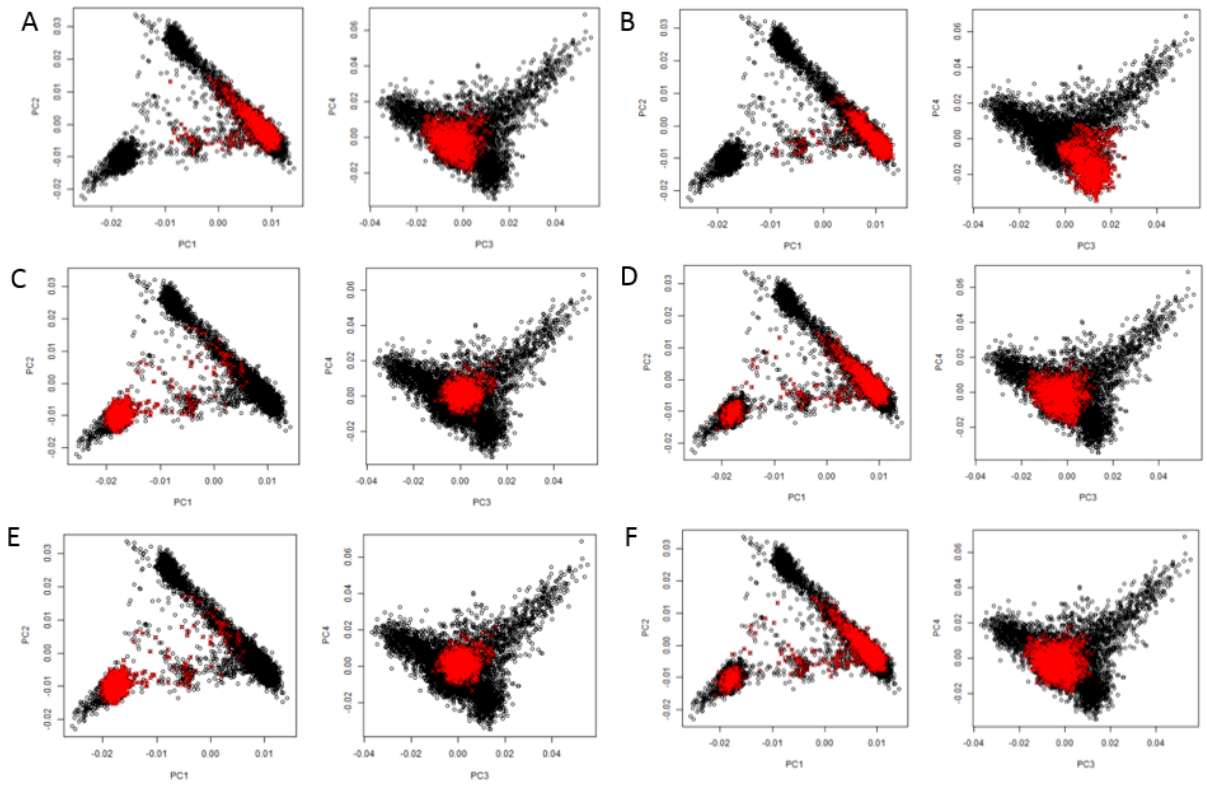

Figure S10: **Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL.** Subjects are colored red if they belong to (A): cluster 829, (B): cluster 899, (C): cluster 952, (D): cluster 1005, (E): cluster 1145, (F): cluster 1199. Subjects not belonging to respective clusters are colored black in every plot.

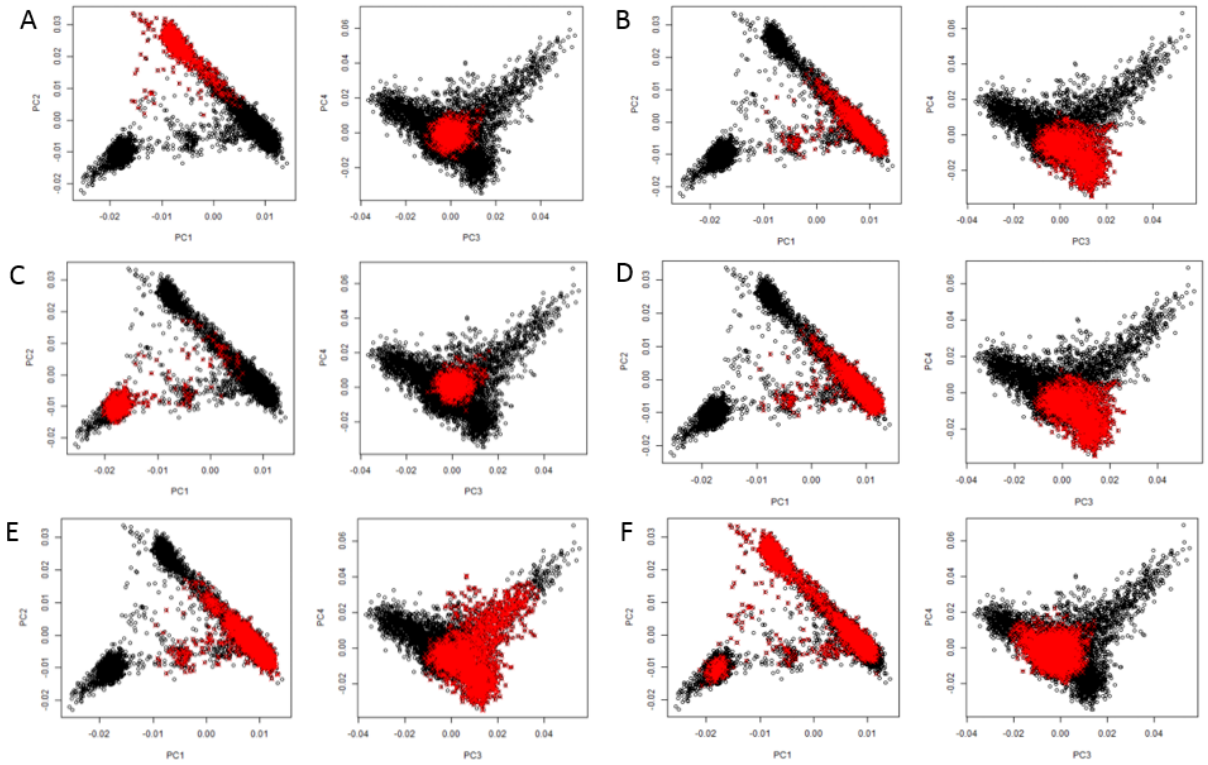

Figure S11: **Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL.** Subjects are colored red if they belong to (A): cluster 1206, (B): cluster 1620, (C): cluster 1765, (D): cluster 1869, (E): cluster 2480, (F): cluster 2971. Subjects not belonging to respective clusters are colored black in every plot.

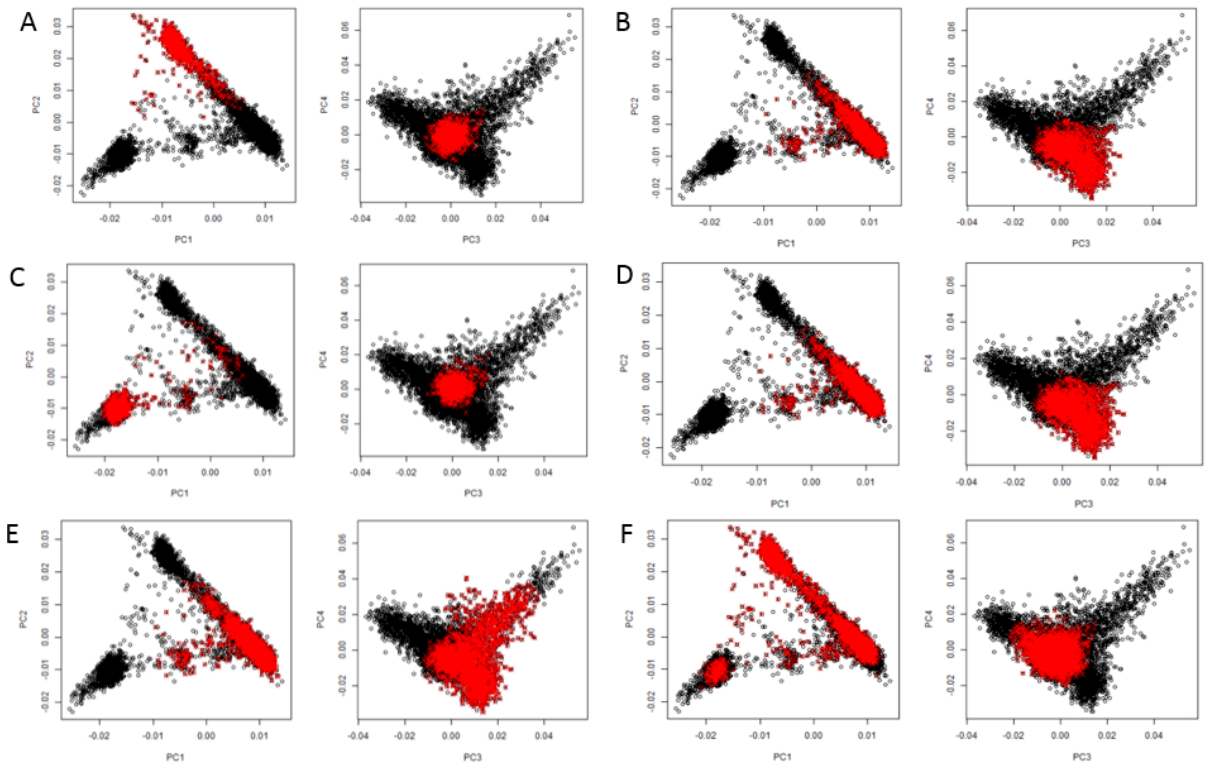

Figure S12: **Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the study of EL.** Subjects are colored red if they belong to (A): cluster 3776, (B): cluster 4921, (C): cluster 7401, (D): cluster 8355, (E): cluster 8961. Subjects not belonging to respective clusters are colored black in every plot.

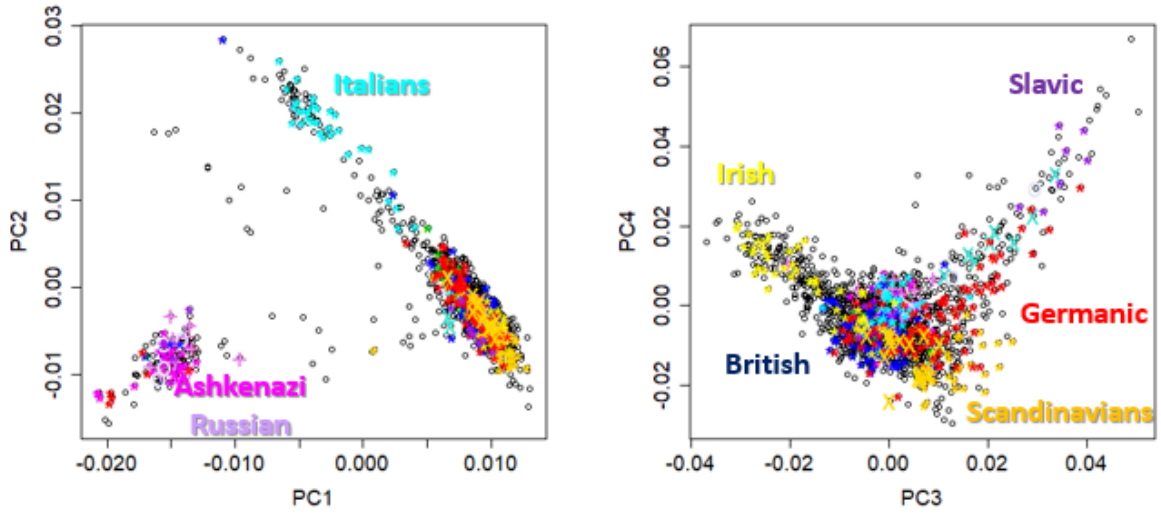

Figure S13: Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of subjects in the NECS. Subjects are labeled by ethnicity using the information about mother tongue and places of birth of subjects and their parents *(Sebastiani et al., 2012)*.

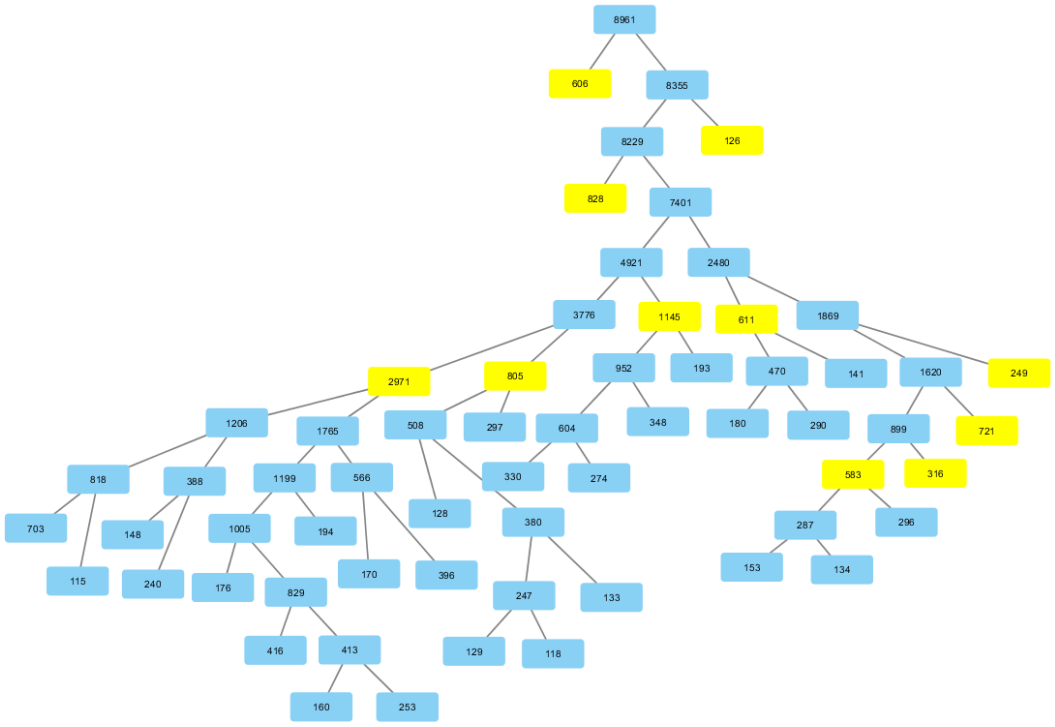

Figure S14: **Full hierarchical tree structure returned for the analysis of EL dataset before pruning of redundant clusters step.** Highlighted in yellow are the clusters that were identified by PopCluster as having ethnic-specific effects of SNP rs3764814 on EL. Numbers inside the nodes represent the number of subjects in each cluster. Visualization was done using Cytoscape *(Shannon et al., 2003)*.

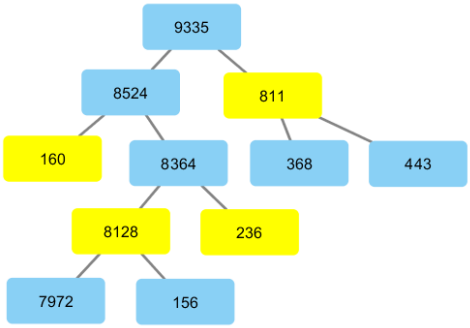

Figure S15: **Full hierarchical tree structure returned for the analysis of survival past age 90 HRS dataset before pruning of redundant clusters step.** Highlighted in yellow are the clusters that were identified by PopCluster as having ethnic-specific effects of SNP rs3764814 on surviving past age 90. Numbers inside the nodes represent the number of subjects in each cluster. Visualization was done using Cytoscape *(Shannon et al., 2003)*.

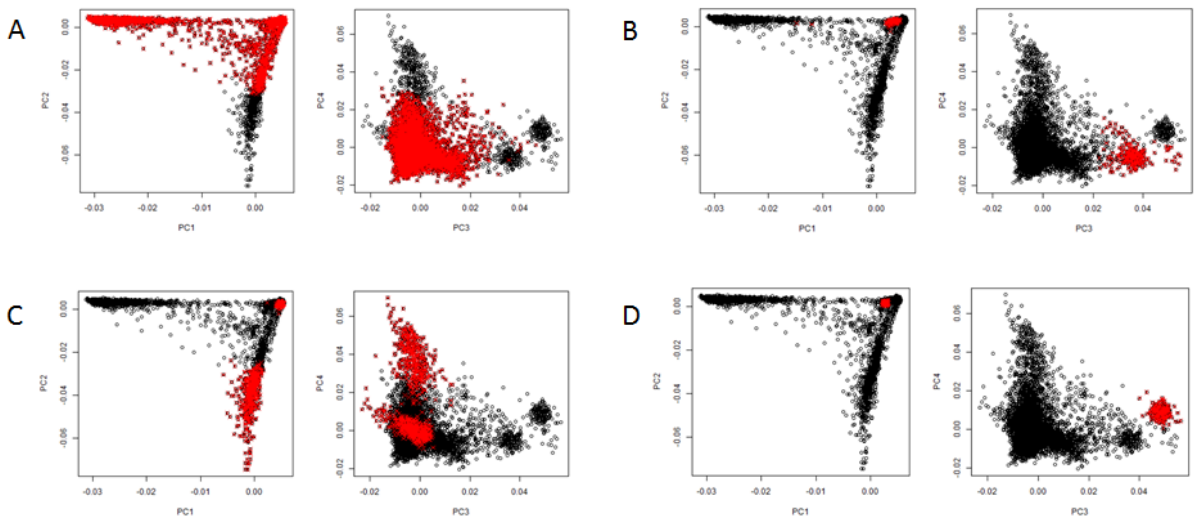

Figure S16: **Scatter plots of PC1-PC2 and PC3-PC4 calculated using genome-wide genotype data of all subjects in the HRS.** Subjects are colored red if they belong to (A): cluster 8128, (B): cluster 236, (C): cluster 811, (D): cluster 160. Subjects not belonging to respective clusters are colored black in every plot.

Table S1: **Percentage of simulation runs (FPR) that returned n number of significant associations (Clusters) ($n = \{0, 1, 2, 3\}$).**

Clusters	Simulation 1	Simulation 2	Simulation 3	Simulation 4
0	83.8	83.2	84.1	83.8
1	15.6	15.8	15.1	15.4
2	0.6	0.3	0.8	0.8
3	–	0.02	–	–

For each of the four different simulations (when the permutation was done on a whole dataset) to evaluate the FPR of PopCluster, we report the information on how many significant clusters each run returned. We define significant clusters here as clusters in which the association between the simulated phenotype and the SNP has a p-value less than 0.05 divided by the total number of clusters returned. By design, any association returned by the algorithm is a false positive, because we reshuffle the labels of cases and controls. As expected, the majority of runs returned no significant associations. Simulation 1: "Original data shuffled" (original dataset with random reshuffling of cases and controls). Simulation 2: "Original data even" (original dataset with equal number of cases and controls randomly generated). Simulation 3: "No-relatedness data shuffled" (as "original data shuffled" after we removed related individuals). Simulation 4: "No-relatedness data even" (as "original data even" after we removed related individuals).

Table S2: **Percentage of simulation runs (TPR) that returned at least one cluster with more than 80% subjects in the region simulated to have an association between allele A and the phenotype.**

p_1/g	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
0.05	99.4	99.9	100	100	99.9
0.1	95.8	100	99.8	99.9	99.9
0.25	86.5	96.4	99.4	99.7	99.9
0.5	82.4	94	99.5	100	100

Percentage of simulation runs (TPR) that returned at least one cluster with more than 80% subjects in the region simulated to have an association between allele A and the phenotype for each of the 20 simulations, in which allele A was simulated based on various combinations of probabilities p_1 and $p_2 = p_1 + g$. p_1 is the probability of an allele A in cases in the region where allele A was simulated to be associated with the phenotype. p_2 is the probability of an allele A in controls in the same region. g is the difference in allele frequencies A between cases and controls in this region.

Table S3: **Average number of returned clusters with more than 80% subjects in the region simulated to have an association between allele A and the phenotype for the scenario 1 of the TPR evaluation.**

p_1/g	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
0.05	2.4	2.7	2.8	3.0	3.2
0.1	2.3	2.6	2.8	2.9	3.0
0.25	2.0	2.4	2.5	2.7	2.8
0.5	1.9	2.2	2.5	2.6	2.8

Average number of returned clusters with more than 80% subjects in the region simulated to have an association between allele A and the phenotype for each of the 20 simulations, in which allele A was simulated based on various combinations of probabilities p_1 and $p_2 = p_1 + g$. p_1 is the probability of an allele A in cases in the region where allele A was simulated to be associated with the phenotype. p_2 is the probability of an allele A in controls in the same region. g is the difference in allele frequencies A between cases and controls in this region.

Table S4: **Percentage of simulation runs (TPR) of PopCluster that returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset.**

p_1/g	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
0.05	39.4	39.5	36.1	39	38
0.1	30.5	31	31.8	29.9	30.3
0.25	26.5	26.5	20.7	25.7	25.7
0.5	27.3	22.6	26.5	22.6	22

Percentage of simulation runs (TPR) of PopCluster that correctly returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset for each of the 20 simulations, in which allele A was simulated based on various combinations of probabilities p_1 and $p_2 = p_1 + g$. p_1 is the probability of an allele A in cases. p_2 is the probability of an allele A in controls. g is the difference in allele frequencies A between cases and controls.

Table S5: **Percentage of simulation runs (TPR) of PopCluster that returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset with re-shuffled case/control labels.**

p_1/g	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
0.05	38.9	41.4	41.9	40.3	39.7
0.1	29.1	28.8	27	25.4	26.2
0.25	25.7	21.9	23.9	23.3	21.6
0.5	25.6	24.2	21.9	21.4	20.3

Percentage of simulation runs (TPR) of PopCluster that correctly returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset with reshuffled case/control labels for each of the 20 simulations, in which allele A was simulated based on various combinations of probabilities p_1 and $p_2 = p_1 + g$. p_1 is the probability of an allele A in cases. p_2 is the probability of an allele A in controls. g is the difference in allele frequencies A between cases and controls. Among those single clusters, 100% of them were found to have a significant association between the simulated allele and phenotype. For the simulation runs, that returned more than one cluster as a result, the average number of clusters that were found significant was 58%.

Table S6: **Percentage of simulation runs (TPR) of PopCluster that returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset with balanced number of re-shuffled case/control labels.**

p_1/g	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
0.05	30.3	24.9	26	27	25.5
0.1	22.8	22.5	23.9	22.2	22.6
0.25	24.5	22.9	25.8	21.1	22.2
0.5	22.9	22.8	21.7	20.8	22.5

Percentage of simulation runs (TPR) of PopCluster that correctly returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset with balanced number of reshuffled case/control labels for each of the 20 simulations, in which allele A was simulated based on various combinations of probabilities p_1 and $p_2 = p_1 + g$. p_1 is the probability of an allele A in cases. p_2 is the probability of an allele A in controls. g is the difference in allele frequencies A between cases and controls. Among those single clusters, 100% of them were found to have a significant association between the simulated allele and phenotype. For the simulation runs, that returned more than one cluster as a result, the average number of clusters that were found significant was 67%.

Table S7: **Percentage of simulation runs (TPR) of PopCluster that returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset with balanced number of re-shuffled case/control labels in each cluster.**

p_1/g	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**
0.05	31.4	25.2	24.5	26	26
0.1	23.9	22.3	23.2	20.4	22.9
0.25	23.9	21.2	21.9	23.2	21.5
0.5	23.7	22.6	23	21.3	20

Percentage of simulation runs (TPR) of PopCluster that correctly returned only one top cluster in simulations where the allele is associated with phenotype in a whole dataset with balanced number of reshuffled case/control labels in each cluster for each of the 20 simulations, in which allele A was simulated based on various combinations of probabilities p_1 and $p_2 = p_1 + g$. p_1 is the probability of an allele A in cases. p_2 is the probability of an allele A in controls. g is the difference in allele frequencies A between cases and controls. Among those single clusters, 100% of them were found to have a significant association between the simulated allele and phenotype. For the simulation runs, that returned more than one cluster as a result, the average number of clusters that were found significant was 67%.

Table S8: **Complete list of clusters detected by PopCluster for 371 SNPs and EL.**

https://open.bu.edu/handle/2144/29809

Following the link you will find *El-results.csv* file with all the results for the analysis of PopCluster on 371 SNPs and EL. Column *Cluster* contains labels for the clusters, all of which are visualized on the PCA plots in Figures S4-S13. Labels reflect cluster sizes, e.g. cluster labeled *118* has 118 subjects, and sorted by their size. Legend for the table in the *EL-results.csv* file: OR: odds ratio for EL in carriers of the allele; P-value: p-value of the association.

Table S9: **Complete list of clusters detected by PopCluster for 11 SNPs and HRS.**

https://open.bu.edu/handle/2144/29809

Following the link you will find *HRS-results.csv* file with all the results for the analysis of PopCluster on 11 SNPs and surviving past age 90. Column *Cluster* contains labels for the clusters. Only 3 out of 11 SNPs (Table 2 in the paper) have significant associations with a phenotype of surviving past the age of 90: rs72834698, rs5882, and rs405509. Legend for the table in the *HRS-results.csv* file: OR: odds ratio for surviving past 90 in carriers of the allele; P-value: p-value of the association.