12-17-2019

Population Structure Analyses Provide Insight into the Source Populations Underlying Rural Isolated Communities in Illinois

Amanda C. Owings
University of Illinois at Urbana-Champaign, aowings2@illinois.edu

Samuel Bonfim Fernandes
University of Illinois at Urbana-Champaign, samuelf@illinois.edu

Marcus O. Olatoye
University of Illinois at Urbana-Champaign, omo@illinois.edu

Amanda J. Fogleman
Southern Illinois University School of Medicine, afogleman@siumed.edu

Whitney E. Zahnd
University of South Carolina, Columbia, SC, wzahnd@siumed.edu

See next page for additional authors

Recommended Citation

Owings, Amanda C.; Fernandes, Samuel Bonfim; Olatoye, Marcus O.; Fogleman, Amanda J.; Zahnd, Whitney E.; Jenkins, Wiley D.; Malhi, Ripan S.; and Lipka, Alexander E., "Population Structure Analyses Provide Insight into the Source Populations Underlying Rural Isolated Communities in Illinois" (2019). *Human Biology Open Access Pre-Prints*. 152.
https://digitalcommons.wayne.edu/humbiol_preprints/152

Authors
Amanda C. Owings, Samuel Bonfim Fernandes, Marcus O. Olatoye, Amanda J. Fogleman, Whitney E. Zahnd, Wiley D. Jenkins, Ripan S. Malhi, and Alexander E. Lipka

Population Structure Analyses Provide Insight into the Source Populations Underlying Rural Isolated Communities in Illinois

Amanda C. Owings,[1] Samuel Bonfim Fernandes,[2] Marcus O. Olatoye,[3] Amanda J. Fogleman,[4] Whitney E. Zahnd,[5] Wiley D. Jenkins,[4,6] Ripan S. Malhi,[1,2,7] and Alexander E. Lipka[3]*

[1]Program in Ecology, Evolution and Conservation Biology, University of Illinois Urbana–Champaign, Urbana, Illinois, USA.

[2]Carl R. Woese Institute for Genomic Biology, University of Illinois Urbana–Champaign, Urbana, Illinois, USA.

[3]Department of Crop Sciences, University of Illinois Urbana–Champaign, Urbana, Illinois, USA.

[4]Department of Population Science & Policy, Southern Illinois University School of Medicine, Springfield, Illinois, USA.

[5]Rural & Minority Health Research Center, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, USA.

[6]Epidemiology and Biostatistics, Southern Illinois University School of Medicine, Springfield, Illinois, USA.

[7]Department of Anthropology, College of Liberal Arts and Sciences, University of Illinois Urbana–Champaign, Urbana, Illinois, USA.

*Correspondence to: Alexander E. Lipka, Department of Crop Sciences, University of Illinois Urbana–Champaign, W-210A Turner Hall, 1102 S Goodwin Ave, Urbana, IL 61801 USA. E-mail: alipka@illinois.edu.

Short Title: Population Structure in Rural Communities

KEY WORDS: POPULATION STRUCTURE, GENOME-WIDE MARKERS, SOURCE POPULATIONS, FOUNDER EFFECTS, RURAL COMMUNITIES.

Abstract

We have previously hypothesized that relatively small and isolated rural communities may experience founder effects, defined as the genetic ramifications of small population sizes at the time of a community's establishment. To explore this, we used an Illumina Infinium Omni2.5Exome-8 chip to collect data from 157 individuals from four Illinois communities, three rural and one urban. Genetic diversity estimates of 999,259 autosomal markers suggested that the reduction in heterozygosity due to shared ancestry was approximately 0, indicating a randomly mating population. An eigenanalysis, which is similar to a principal component analysis but ran on a genetic coancestry matrix, conducted in the SNPRelate R package revealed that the majority of these individuals formed one cluster with a few putative outliers obscuring population variation. An additional eigenanalysis on the same markers in a combined data set including the 2,504 individuals in the 1000 Genomes database found that most of the 157 Illinois individuals clustered into one group in close proximity to individuals of European descent. A final eigenanalysis of the Illinois individuals with the 503 individuals of European descent (within the 1000 Genomes Project) revealed two clusters of individuals and likely two source populations; one British and one consisting of multiple European subpopulations. We therefore demonstrate the feasibility of examining genetic relatedness across Illinois populations and assessing the number of source populations using publicly available databases. When assessed, it becomes

possible for population structure information to contribute to the understanding of genetic history in rural populations.

Two key characteristics of many rural communities in the US Midwest is that they were founded several hundred years ago, and that little migration has occurred in comparison with similar communities in Africa, Asia, and Europe (described in JENKINS *et al.* 2016). While many non-genetic factors may explain a substantial amount of increased incidence of certain diseases in these rural communities (see BEFORT *et al.* 2012; HINES AND MARKOSSIAN 2012; and HENRY *et al.* 2014 for specific examples), quantification of a possible genetic predisposition to diseases in such communities could assist efforts to account for and minimize disease risk. It is therefore critical to compare and contrast genetic characteristics of rural populations to those from urban populations. This will particularly enable the testing of our hypothesis that small and isolated rural communities may experience genetic founder effects to a greater extent than their more urban peers (JENKINS *et al.* 2016). Such founder effects may influence disease susceptibility and have long lasting impacts (RUDAN 1999). We hypothesize that a small town, founded by a small number of individuals and relatively geographically isolated, can remain affected by the initial founder effect over hundreds of years. Similar examples have been observed previously (e.g. the island of Sardinia), where geography presents a physical barrier to travel (PORTAS *et al.* 2010).

Researchers can use genetic data to estimate how closely related individuals in a population are to each other, as well as to determine if members of a rural community have a single or multiple source population(s) (the location of the population's origin; FALUSH *et al.* 2003; WANG *et al.* 2007). Determining if there is more than one source population is an important step for examining population structure; multiple source populations would suggest higher initial genetic diversity than a single source population and minimize any impacts of a founder effect. The ability to use genetic data to quantify subpopulation structure is an important factor in population studies (WACHOLDER *et al.* 2000; THOMAS AND WITTE 2002; CAMPBELL *et* *al.* 2005). Population structure analyses can be performed with large numbers of single nucleotide polymorphisms (SNPs) using small amounts of DNA and commercially available SNP chips. Given the use of genetic data from such chips in previous research (VAAGS *et al.* 2012; TERAO *et al.* 2013; DE VIVO *et al.* 2014; MAYBA *et al.* 2014; MACHIELA *et al.* 2016), it appears that they are well-suited for quantifying subpopulation structure in rural isolated populations and hence provide insight into the impact of founder effects and isolation on current community genetic diversity.

Genome-wide marker obtained from SNP chips can also be used to obtain measures of genetic diversity. Such measures include average gene diversity over loci, which estimates overall population diversity (NEI 1987). Population similarity can be measured using Wright's indices including FIS, which examines the reduction in heterozygosity in a population due to shared ancestry (WRIGHT 1950). This measure can help estimate the relatedness of individuals within a population. Typical values of FIS in European population have been reported in German (-0.0010-0.0108) (STEFFENS *et al.* 2006) and several Iberian populations: Basques (0.0000), Navarre (0.015), Pass Valley (0.0144) (CARDOSO *et al.* 2017). Thus, measures of average gene diversity over loci and FIS could indicate if rural populations have less diversity and/or appear to exhibit genetic drift, including a genetic bottleneck or founder effect, compared to other world populations.

Beyond measuring average gene diversity and FIS, we speculate that another critical analysis leading to accurate quantification of subpopulation structure in rural populations would be to compare their genetic relationships with various populations throughout the world. Such an analysis could facilitate the identification of source populations and provide insight into the presence of founder effects. The undertaking of such an endeavor is now possible given the availability of whole-genome sequenced data sets such as those from the 1000 Genomes Project (1KGP), PopRes, Ancestry DNA, the Human Genome Diversity Project, and HapMap projects (ANCESTRYDNA ; CANN *et al.* 2002; INTERNATIONAL HAPMAP 2003; NELSON *et al.* 2008; GENOMES PROJECT *et al.* 2015). Genome-wide markers segregating in both the rural populations and these whole-genome sequenced data sets could then be analyzed to quantify genetic relationships and identify source populations. Approaches such as STRUCTURE (PRITCHARD *et al.* 2000), principal component analysis (PRICE *et al.* 2006), and ADMIXTURE (ALEXANDER *et al.* 2009) are adequate for using genome-wide markers to infer which subpopulations are present in the resulting combined data sets. However, advances in methodologies, including the eigenanalysis approach of ZHENG AND WEIR (2016), now make it possible to characterize which ancestral populations underlie the individuals living in rural communities by directly incorporating the probabiltiy of markers being identical by descent (IBD) into the calculations.

The purpose of this study was to examine whole-genome SNP data from individuals from three rural and one urban population in Illinois, USA and characterize their genetic properties, including genetic diversity and relatedness. To achieve this, we characterized the genetic properties of these individuals, and then compared them to the 1KGP database. We hypothesized that such an assessment could shed light on potential founder effects and suggest genetic differentiation from more urban populations.

Materials and Methods

Illinois IsoPop Data Set

The individuals comprising the Isolated Populations Project (IsoPop) data set, as well as the methods used to recruit them, have been described elsewhere (DEAN 2017). Briefly, 176 individuals were recruited from three rural communities (70 individuals from community #1, 30 from community #2, and 41 from community #3) and one urban community (35 individuals; community #4) in Illinois. These three rural communities were thought to have been settled in the past 300 years and are relatively isolated (JENKINS *et al.* 2016). The three rural communities were between 100 and 400 miles from each other, with the nearest urban centers to each community being located between 30 and 60 miles away (Wiley Jenkins, personal communication). In addition to providing genealogical information and saliva samples, the participants took surveys and engaged in community forums. The genealogy information was used to remove individuals that were first degree relatives with an already-recruited participant so as to not artificially inflate the degree of relatedness within the groups. This project was approved by the SIUSOM IRB (Springfield Committee for Research Involving Human Subjects; #15-328) and all participants provided informed consent.

DNA Extraction and Marker Identification of IsoPop Individuals

Extraction of DNA was carried out using an Oragene® prepIT-L2P kit (DNA Genotek) following the standard protocol with a few modifications. Incubation occurred in a heat block for between 2-24 hours (protocol suggested two hours of incubation). Rehydration of the DNA pellets occurred by incubating at 50° C for an hour or more as needed. Sample concentration was assessed using the Qubit™ assay (ThermoFisher). The average DNA concentration obtained was 89.43 µg/ml with a range of 0.281-500 µg/ml. All samples, their population, and DNA concentration are listed in Supplementary Table 1.

Samples were aliquoted into separate tubes and taken to the Keck Biotechnology Sequencing Center at the University of Illinois at Urbana-Champaign. A water sample was included in the run to assess contamination, had a call rate of 0.4522, and was removed from analyses. Next, DNA samples were run on Illumina Infinium Omni2.5Exome-8 Bead chips (Illumina Inc, San Diego, CA) according to the Illumina LCG Assay Protocol (Part#15023139, Rev. D). Sequencing was carried out on the Illumina iScan to genotype 2,612,357 markers from the human genome. Sample results were viewed in Genome Studio and the "positive/negative" column was exported using a Dell PC with 64GB RAM. We removed a total of 19 individuals that either had a call rate of less than 0.90 as suggested by other studies (VERDU *et al.* 2014), or were first degree relatives to another individual (as reported by genealogical data), resulting in a total of 157 IsoPop individuals that were analyzed.

1KGP Database

Genomic data from the 1KGP consists of 2,504 individuals from 26 subpopulations across five continents and has been previously described (BIRNEY AND SORANZO 2015; GENOMES PROJECT *et al.* 2015). In brief, the 1KGP investigators sampled adult, "legally competent" individuals who are not from vulnerable or identifiable populations, using protocols that were in accordance with standard ethical guidelines (internationalgenome.org). Individuals in the database were self-reported to be healthy, and gave their gender and ethnicity. The entirety of genomic data from the 1KGP contain 88 million variant sites (GENOMES PROJECT *et al.* 2015) and was collected using whole-genome sequencing.

Computational Methods

To quantify trends of population structure between and within the IsoPop population and the individuals in the 1KGP, we first obtained a subset of informative SNPs. The raw IsoPop data generated from Genome Studio were exported as tables. These tables were loaded into RStudio using the data.table package where insertions and deletions were removed, as well as genotypes with a call rate < 90% (RStudio® 2015). The 1KGP data were downloaded at ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/ (shown in Figure 1). In order to match IsoPop with 1KGP data set, only SNPs on the forward strand were kept. Additionally, support files provided by Illumina (https://support.illumina.com/downloads.html) were used to convert SNP IDs into the reference SNP ID number. None of the individuals exceeded a threshold of 10% missing data. This data set was then converted into HapMap format and TASSEL (BRADBURY *et al.* 2007) was used to convert these data to VCF format. Next, PLINK (PURCELL *et al.* 2007) was used to remove SNPs with more than two alleles or more than 5% missing data. The reference allele was converted to the reference genome GRCh37 using PLINK 2.0. The resulting IsoPop data set used for subsequent analysis was composed of 157 individuals and 999,259 autosomal SNPs.

Genetic Diversity Estimates

Using the HapMap formatted files generated in RStudio, TASSEL (BRADBURY *et al.* 2007) was used to convert the files to VCF file format, and finally PGDSpider (LISCHER AND EXCOFFIER 2012) was used to convert to Arlequin project format (EXCOFFIER AND LISCHER 2010). The program Arlequin version 3.5.2.2 was used to calculate F_{IS} and average gene diversity using the approach of NEI (1987) across each marker and averaged for each chromosome. This was done to assess how genetically related these populations are to each other and potentially parse out founder effects. These F_{IS} values were calculated and graphed along the chromosomes for both the IsoPop and 1KGP individuals using VCFtools (DANECEK *et al.* 2011).

Eigenanalysis Using EIGMIX

The procedure described in ZHENG AND WEIR (2016) was used to assess the presence of source populations in the IsoPop data set. In summary, this eigenanalysis differs from a traditional principal component analysis in that a coancestry matrix from the SNP data is used. This analysis was conducted on three different subsets of the data, the first being the data comprising of only the 157 IsoPop individuals. The procedure was conducted a second time on the combined IsoPop data set and the 2,504 individuals from the 1KGP data set. Finally, this procedure was repeated using the IsoPop data set and the subset of 503 individuals in the 1KGP data set from five European subpopulations. This analysis was conducted using the SNPRelate package in R. All scripts used for these analyses are publicly available at https://github.com/AmandaO8, and the coancestry matrix of all individuals used in this analysis is presented as Supplementary File 1 and visualized in Supplementary Figure 1.

Results

Genetic Diversity of IsoPop Individuals Are Comparable to Other European Populations

Estimates of genetic diversity in the IsoPop individuals for each chromosome can be found in Supplementary Table 2. The observed F_{IS} values were all near 0, with only chromosomes 8 and 9 having positive values, indicating that there has been random mating in these populations. By population, average gene diversity (using the approach described in NEI 1987) over loci values are all close to 0.3 for each chromosome. These values are similar to other European populations and suggest that the IsoPop individuals are as genetically diverse as a typical population of European descent.

Using the same 999,259 SNPs considered in the IsoPop data set, F_{IS} values were also calculated for the 1KGP individuals, and the results are graphed in Figure 2. This enabled the direct comparison of heterozygosity between the IsoPop individuals and the 1KGP individuals. The IsoPop populations had smaller F_{IS} values than the full set of 2,504 1KGP individuals, suggesting lower levels of heterozygosity. However, the results also show that the distribution of F_{IS} values among the 503 1KGB individuals from five European subpopulations was similar to those of the four IsoPop communities. This suggests that the IsoPop individuals are less genetically diverse than individuals in the 1KGP as a whole, but similar to the 1KGP subset of individuals from European-descended populations.

Comparison with 1KGP Data Suggest Multiple Source Populations from Europe

To test for the presence of observable founder effects among the IsoPop populations, we conducted an eigenanalysis of 999,259 autosomal genome-wide markers that segregated among these individuals (Figure 3; Supplementary Figures 2-5). The majority of IsoPop individuals were in close proximity to each other on the plot of the first two eigenvectors, with four individuals far removed from the main cluster of individuals. Thus, all but four of the individuals (two from community #1, and two from community # 3; both of these communities are rural) in the IsoPop data set cluster together, suggesting that the majority of individuals are descended from a single source population and remaining four individuals are likely from two other source populations. Even with the removal of these four observations, the majority of individuals still cluster with each other (Figure 3). The two sets of individuals outside of the main cluster that group with each other are respectively from the same communities, suggesting the possibility of there being some individuals who are related and did not report it or were unaware.

Unexpectedly, the urban population does not appear to be any more diverse than the rural populations.

To further assess the genetic relatedness between the IsoPop individuals, we next conducted an eigenanalysis on the same set of 999,259 markers using the IsoPop data set combined with the 2,504 individuals from the 1KGP database. The resulting plot of the first two eigenvalues (Figure 4) revealed that the majority of the IsoPop individuals formed one cluster. Additional plots from this analysis are included as Supplementary Figures 6-8. This cluster overlaps with the 1KGP European individuals and is furthest from the 1KGP individuals with Asian and African ancestry. This result suggests that IsoPop individuals are a) more closely related to each other than to other world populations, and b) that their source population is most likely Europe.

A final eigenanalysis was conducted with the IsoPop individuals and the 503 individuals of European descent from the 1KGP. The corresponding plot summarizing results from the first two eigenvalues (Figure 5) had three main groups and three outlier individuals. Additional plots from this analysis are included as Supplementary Figures 9-11. Many IsoPop individuals from each population cluster with those of Great Britain, including all individuals of the urban population (community #4). Additionally, many individuals from the rural populations (communities #1, #2, and #3) group with people of Northern and Western European ancestry living in Utah, Finland, Spain, and Tuscany (CEU, FIN, IBS, and TSI, respectively). These more refined results supersede the immediately previous findings from the original eigenanalysis by suggesting that the rural populations have multiple European source populations and likely had several founding groups. This also indicates that the urban population (community #4) only has Great Britain as a source population and might be less diverse than the rural populations.

Discussion

The use of genomic markers from high-throughput genotyping data to compare the relatedness between individuals in rural communities to those in publicly available databases could help identify founder effects and source populations. To assess the capability of such an approach, we analyzed genetic data from 157 individuals living in four communities in Illinois (three of which were rural) and used state-of-the-art statistical approaches to compare their genetic similarity to the 2,504 individuals comprising the 1KGP database. Given the novelty of these IsoPop data, these results provided an initial glance into the genetic diversity underlying these individuals. In particular, our first finding was that not only were the three rural communities indistinct from each other, but that they were also indistinct from the urban ‘control’ population. This indicates that genetic founder effects may not be present in these isolated rural communities, and that community endogamy is not so reduced in rural areas as to influence observable genetic differences compared to a more urban area.

We next examined the IsoPop data in relation to the globally-representative 1KGP data set. Our first finding was that the eigenanalysis primarily grouped the IsoPop individuals into one cluster (Figure 4) which was closest to the subset of 1KGP European individuals, suggesting the IsoPop are more closely related to Europeans than other groups. Our results are also consistent with our theoretical expectations based on the genealogical data suggesting that the majority of IsoPop individuals are descended from people of European ancestry. Further evidence of the presence of a single European source population is provided by the respective plots of the first two eigenvectors clustering most of the IsoPop individuals into their own group, suggesting that the vast majority of these IsoPop individuals are closely related (Figure 3, Supplementary Figures 3-5). However, the individuals situated distantly from the cluster could be obscuring some of the variation in these populations.

The plot of the first two eigenvectors from the eigenanalysis of the IsoPop and the 1KGP European subpopulation suggests genetic similarity with British, Finnish, Spanish, Tuscan, and people of Northern and Western European ancestry living in Utah (Figure 5). While Figure 4 (IsoPop + total 1KGP) suggests one European source population, Figure 5 (IsoPop + 1KGP European subset) suggests multiple European source populations underlying the majority of the IsoPop individuals. Thus, the tight clustering of the IsoPop individuals with these populations potentially rules out the possibility of a single source population. The genetic diversity estimates of the IsoPop population are similar to those found in other studies, in that the ranges of the estimates overlap, but the average values were different. For example, the F_{IS} values for the IsoPop range from -0.00652 to 0.00177 and have mostly negative values whereas those of German populations range from -0.0022 to 0.0108 and have mostly positive values (STEFFENS *et al.* 2006). These F_{IS} values indicate that the IsoPop individuals are no more or less closely related to each other than expected under the null model of random mating.

Using the combined marker data from the IsoPop data set and the 1KGP database, we were able to infer that most of the IsoPop individuals are descended from at least two source populations originally from Europe. This result could aid researchers studying the prevalence of diseases in the three rural Illinois communities included in the IsoPop data set by suggesting that any alleles among these individuals that cluster with one of the source populations could have similar levels of genetic predisposition. More broadly, our study serves as a proof-of-concept to demonstrate that it is possible to use an approach like an eigenanalysis to compare the genetic characteristics between a set of individuals and those from a public database, and moreover to show that it is possible to obtain biologically meaningful results.

In general, research into the risk of disease attributable to specific gene variants and combinations is often hindered by a low carrier frequency of specific mutations among the general population (SHERRY *et al.* 2001). While this study showed insignificant differences across the rural and urban communities, we did not examine specific loci known/thought to be associated with increased disease risk. Additional work would specifically examine and characterize such loci, as the identification of specific populations with naturally increased carrier frequencies of specific gene variants of interest would greatly justify the utility of ecological and historical studies of diseases (PELTONEN *et al.* 2000). This in turn could result in multiple studies of how individual genetic makeup may impact such important topics such as drug efficacy (ARBITRIO *et al.* 2019) and variable outcomes to environmental exposure (RYU *et al.* 2018).

There are several limitations to this work. First, the rural communities were chosen as a matter of feasibility and convenience. While the community size was based upon the work of PORTAS *et al.* (2010), true isolation is more difficult to ascertain objectively. Rigor in assessing isolation and randomization of selection would be needed for future work. Second, the choice of the urban 'control' is also based on convenience. While the urban population has a population exceeding 110,000, it is by no means a major metropolitan center as reflected in that its population appeared to be related to just one European subpopulation (i.e., British) and is therefore potentially problematic to use as an urban control population. This could be because the sample urban population was not fully representative of the whole population, or perhaps this particular urban population is not as genetically diverse as others. Future studies could use a larger (or multiple) urban community in order to circumvent this potential problem. Third, follow-up studies that trace the history of settlement of these communities could complement and potentially substantiate the findings of the work presented here.

Another important limitation of this study is with the genotyping technologies employed to obtain markers in the IsoPop and 1KGP data sets. In addition to the potential for ascertainment bias inherent in using arrays such as Illumina (described in LIPKA *et al.* 2015), additional bias could arise from the fact that an Illumina chip was used to call markers in the IsoPop data set while whole genome sequencing was used in the 1KGP data set. However, our results suggest that such an ascertainment bias could be minimal. For example, there is a close proximity between the IsoPop to 1KGP individuals in Figures 4-5. We also observed a similar distribution of rare and common SNPs IsoPop individual and the 503 1KGP individuals of European descent (Supplementary Figure 12 and Supplementary Table 3), as well as similar linkage disequilibrium patterns (Supplementary Figure 13). Nevertheless, future studies should use the same sequencing platforms to obtain markers in all data sets that are evaluated. Finally, we encourage future studies to compare data from rural isolated communities from the US Midwest with marker data from other publicly available data sets besides the 1KGP data set that include more than just the five subpopulations of European descent, such as PopRes (NELSON *et al.* 2008). Such a comparison could shed further light on the number of source populations underlying these isolated communities.

Conclusions

This study utilized nearly one million high-quality SNPs, and to the best of our knowledge is the first to use SNP data to examine both population structure and founder effects in non-religious rural isolate populations in the Midwest US. The potential impact of founder effects on the genetic diversity of rural communities over hundreds of years could be the source of future studies. For example, these studies could consider advanced statistical approaches for quantifying such effects, and moreover parse out these effects on the population over multiple generations, from the founding of the population to the present day. Lastly, other SNP chips or whole-genome sequencing could be used to obtain a larger marker set (and thus capture an even greater amount of genomic diversity) and be used in a combined analysis with these IsoPop individuals and other publicly available data sets.

Received 3 April 2019; accepted for publication 12 November 2019.

Literature Cited

Alexander, D. H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19:1,655–1,664.

AncestryDNA, pp.

Arbitrio, M., F. Scionti, E. Altomare et al. 2019. Polymorphic variants in *NR1I3* and *UGT2B7* predict taxane neurotoxicity and have prognostic relevance in breast cancer patients: A case-control study. *Clin. Pharmacol. Ther.* 106:422–431.

Birney, E., and N. Soranzo. 2015. Human genomics: The end of the start for population sequencing. *Nature* 526:52–53.

Bradbury, P. J., Z. Zhang, D. E. Kroon et al. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2,633–2,635.

Campbell, C. D., E. L. Ogburn, K. L. Lunetta et al. 2005. Demonstrating stratification in a European American population. *Nat. Genet.* 37:868–872.

Cann, H. M., C. de Toma, L. Cazes et al. 2002. A human genome diversity cell line panel. *Science* 296:261–262.

Cardoso, S., R. Sevillano, D. Gamarra et al. 2017. Population genetic data of 38 insertion-deletion markers in six populations of the northern fringe of the Iberian Peninsula. *Forensic Sci. Int. Genet.* 27:175–179.

Danecek, P., A. Auton, G. Abecasis et al. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2,156–2,158.

Dean, C., A. J. Fogleman, W. E. Zahnd et al. 2017. Engaging rural communities in genetic research: Challenges and opportunities. *J. Community Genet.* 8:209–219.

De Vivo, I., J. Prescott, V. W. Setiawan et al. 2014. Genome-wide association study of endometrial cancer in E2C2. *Hum. Genet.* 133:211–224.

Excoffier, L., and H. E. Lischer. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour*. 10:564–567.

Falush, D., M. Stephens, and J. K. Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 164:1,567–1,587.

International HapMap Consortium. 2003. The International HapMap Project. *Nature* 426:789–796.

Jenkins, W. D., A. E. Lipka, A. J. Fogleman et al. 2016. Variance in disease risk: Rural populations and genetic diversity. *Genome* 59:519–525.

Lipka, A. E., C. B. Kandianis, M. E. Hudson et al. 2015. From association to prediction: Statistical methods for the dissection and selection of complex traits in plants. *Curr. Opin. Plant Biol.* 24:110–118.

Lischer, H. E., and L. Excoffier. 2012. PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics* 28:298–299.

Machiela, M. J., W. Zhou, E. Karlins et al. 2016. Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome. *Nat. Commun.* 7:1–9.

Mayba, O., F. Gnad, M. Peyton et al. 2014. Integrative analysis of two cell lines derived from a non-small-lung cancer patient—a panomics approach. *Pac. Symp. Biocomput.* 75–86.

Nei, M. 1987. *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Nelson, M. R., K. Bryc, K. S. King et al. 2008. The Population Reference Sample, POPRES: A resource for population, disease, and pharmacological genetics research. *Am. J. Hum. Genet.* 83:347–358.

Peltonen, L., A. Palotie, and K. Lange. 2000. Use of population isolates for mapping complex traits. *Nat. Rev. Genet.* 1:182–190.

Portas, L., F. Murgia, G. Biino et al. 2010. History, geography and population structure influence the distribution and heritability of blood and anthropometric quantitative traits in nine Sardinian genetic isolates. *Genet. Res. (Camb.)* 92:199–208.

Price, A. L., N. J. Patterson, R. M. Plenge et al. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38:904–909.

Pritchard, J. K., M. Stephens, N. A. Rosenberg et al. 2000. Association mapping in structured populations. *Am. J. Hum. Genet.* 67:170–181.

Purcell, S., B. Neale, K. Todd-Brown et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81:559–575.

Rudan, I. 1999. Inbreeding and cancer incidence in human isolates. *Hum. Biol.* 71:173–187.

Ryu, D. H., H. T. Yu, S. A. Kim et al. 2018. Is chronic exposure to low-dose organochlorine pesticides a new risk factor of T-cell immunosenescence? *Cancer Epidemiol. Biomarkers Prev*. 27:1,159–1,167.

Sherry, S. T., M. H. Ward, M. Kholodov et al. 2001. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* 29:308–311.

Steffens, M., C. Lamina, T. Illig et al. 2006. SNP-based analysis of genetic substructure in the German population. *Hum. Hered.* 62:20–29.

Terao, C., H. Yoshifuji, A. Kimura et al. 2013. Two susceptibility loci to Takayasu arteritis reveal a synergistic role of the *IL12B* and *HLA-B* regions in a Japanese population. *Am. J. Hum. Genet*. 93:289–297.

Thomas, D. C., and J. S. Witte. 2002. Point: Population stratification: A problem for case-control studies of candidate-gene associations? *Cancer Epidemiol. Biomarkers Prev.* 11:505–512.

Vaags, A. K., A. C. Lionel, D. Sato et al. 2012. Rare deletions at the neurexin 3 locus in autism spectrum disorder. *Am. J. Hum. Genet.* 90:133–141.

Verdu, P., T. J. Pemberton, R. Laurent et al. 2014. Patterns of admixture and population structure in native populations of Northwest North America. *PLoS Genet.* 10:1–17.

Wacholder, S., N. Rothman, and N. Caporaso. 2000. Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of bias. *J. Natl. Cancer Inst*. 92:1,151–1,158.

Wang, S., C. M. Lewis Jr., M. Jakobsson et al. 2007. Genetic variation and population structure in Native Americans. *PLoS Genet.* 3:2,049–2,067.

Wright, S. 1950. Genetical structure of populations. *Nature* 166:247–249.

Zheng, X., and B. S. Weir. 2016. Eigenanalysis of SNP data with an identity by descent interpretation. *Theor*. *Popul. Biol.* 107:65–76.

1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* 526:68–74.

Supplementary Table S1. List of IsoPop Individuals by Population, and Quality Assessments Made by Qubit Concentration and the Illumina Call Rate

IsoPop population 04 is the urban population.

Population	Sample #	Extraction Concentration (ug/mL)	Illumina Call rate	Removed?	Reason
01	01-01	167	0.9974812		
01	01-02	120	0.9968473		
01	01-03	67	0.9977021		
01	01-04	88.7	0.9975508		
01	01-05	41.5	0.9975206		
01	01-06	82.9	0.9966888		
01	01-07	105	0.993197		
01	01-08	133	0.9932184		
01	01-09	36.8	0.9868		
01	01-10	500	0.9943691		
01	01-11	158	0.9943488		
01	01-12	34	0.9968446		
01	01-13	94.5	0.9961242		
01	01-14	67.6	0.9962689		
01	01-15	5.62	0.992184		
01	01-16	203	0.9957651		
01	01-17	97.4	0.9970754		
01	01-18	213	0.9961342		
01	01-19	102	0.9979115		
01	01-20	69.1	0.997129		
01	01-21	155	0.9970437		
01	01-22	121	0.9865428		

01	01-23	106	0.9966873		
01	01-24	248	0.9983436		
01	01-25	24.9	0.9932256		
01	01-26	75.4	0.9961414		
01	01-27	145	0.9953892		
01	01-28	102	0.9954677		
01	01-29	131	0.9963363		
01	01-30	68.4	0.9962432		
01	01-31	47.5	0.9961475		
01	01-32	43.9	0.993865		
01	01-33	86.5	0.9548411	Yes	Relative
01	01-34	174	0.9934438		
01	01-35	17.8	0.9947618		
01	01-36	69.9	0.9942707		
01	01-37	134	0.99441	Yes	Relative
01	01-38	48	0.9950902		
01	01-39	45	0.99222		
01	01-40	56.7	0.8760257	Yes	Call rate
01	01-41	56.9	0.9423677	Yes	Call rate
01	01-42	87.5	0.9892825		
01	01-43	84.9	0.9912366		
01	01-44	409	0.99547		
01	01-45	128	0.9942868	Yes	Relative
01	01-46	114	0.9944782		
01	01-47	171	0.995186		
01	01-48	146	0.9953731		
01	01-49	108	0.9492325	Yes	Relative

01	01-50	8.84	0.9892086		
01	01-51	40.1	0.9935024		
01	01-52	81.4	0.9930404	Yes	Relative
01	01-53	310	0.9948721		
01	01-54	32.4	0.9954964		
01	01-55	61	0.9954524		
01	01-56	55.1	0.9957517		
01	01-57	47.4	0.9940705		
01	01-58	24	0.9932877		
01	01-59	184	0.9935063	Yes	Relative
01	01-60	175	0.9949337		
01	01-61	169	0.9954049		
01	01-62	76.2	0.9958156		
01	01-63	59.8	0.9959779		
01	01-64	67.7	0.9959041		
01	01-65	395	0.8925361	Yes	Relative & Call rate
01	01-66	80.3	0.96083		
01	01-67	38.4	0.9958524		
01	01-68	42.1	0.9971187		
01	01-69	121	0.9972297		
01	01-70	69.5	0.9976466		
02	02-01	65.4	0.997423		
02	02-02	83.6	0.9969759		
02	02-03	314	0.9952985		
02	02-04	16.6	0.993889		
02	02-05	29.6	0.995648		

02	02-06	18.3	0.9946512		
02	02-07	121	0.9954558		
02	02-08	66.6	0.9955929		
02	02-09	39.1	0.9960828		
02	02-10	17.7	0.9968446		
02	02-11	15.3	0.9024593	Yes	Call rate
02	02-12	28	0.9545383		
02	02-13	47.5	0.9946432		
02	02-14	42	0.995638		
02	02-15	71.2	0.9962069		
02	02-16	84	0.9963803		
02	02-17	87.7	0.9970464		
02	02-18	76.5	0.9974223		
02	02-19	28.4	0.9941466		
02	02-20	87.5	0.9949123		
02	02-21	160	0.9940981		
02	02-22	78.4	0.9962968		
02	02-23	43.2	0.9956552		
02	02-24	18.5	0.9956419		
02	02-25	87.5	0.9957023		
02	02-26	55.4	0.9968711		
02	02-27	24.4	0.9957361		
02	02-28	84.6	0.9956943		
02	02-29	34.1	0.9955435		
02	02-30	48.6	0.9967592		
03	03-01	85.4	0.9965085		
03	03-02	423	0.9964718		

03	03-03	33.6	0.9955335		
03	03-04	81.6	0.9360753	Yes	Call rate
03	03-05	77.6	0.9858863		
03	03-06	71.4	0.9933891		
03	03-07	53.2	0.9937685		
03	03-08	89.4	0.9941551		
03	03-09	230	0.9947806		
03	03-10	78.4	0.9936226		
03	03-11	196	0.9966524		
03	03-12	151	0.9958245		
03	03-13	128	0.9950378		
03	03-14	32.4	0.9974203		
03	03-15	79	0.9958003		
03	03-16	63.3	0.9955305		
03	03-17	49.2	0.9947369		
03	03-18	218	0.9964174		
03	03-19	143	0.9961563		
03	03-20	74.3	0.9960966		
03	03-21	30.4	0.9953693		
03	03-22	113	0.9974211		
03	03-23	67.7	0.9970517		
03	03-24	21.2	0.9971091		
03	03-25	51.7	0.9974364		
03	03-26	67.8	0.9964224		
03	03-27	41.9	0.9965323		
03	03-28	8.54	0.9965678		
03	03-29	0.281	0.5624231	Yes	Call rate

03	03-30	131	0.9931116		
03	03-31	119	0.9953712		
03	03-32	131	0.9962873		
03	03-33	55.3	0.9954233		
03	03-34	120	0.9919245		
03	03-35	81.5	0.9948736		
03	03-36	45.1	0.9955803		
03	03-37	40.7	0.996419		
03	03-38	43.5	0.9955733		
03	03-39	43.7	0.9966946		
03	03-40	37.2	0.9947117		
03	03-41	23.9	0.9941126		
04	04-01	20	0.9950535		
04	04-02	11.9	0.9957716		
04	04-03	37.1	0.9969537		
04	04-04	58.4	0.8425893	Yes	Call rate
04	04-05	39.4	0.8553525	Yes	Call rate
04	04-06	122	0.9579346		
04	04-07	0.603	0.5075501	Yes	Call rate
04	04-08	68.1	0.9936019		
04	04-09	18.4	0.9952101		
04	04-10	52.7	0.99563		
04	04-11	38.1	0.9967926		
04	04-12	156	0.8372148	Yes	Call rate
04	04-13	50.2	0.9472346	Yes	Call rate
04	04-14	127	0.9930136		
04	04-15	13.1	0.9929558		

04	04-16	8.95	0.9929485		
04	04-17	26	0.9948089		
04	04-18	35.1	0.9964679		
04	04-19	144	0.9970908		
04	04-20	46.4	0.9937616		
04	04-21	170	0.9941689		
04	04-22	86.3	0.9912428		
04	04-23	81.3	0.9956288		
04	04-24	20.5	0.9811592		
04	04-25	11.9	0.990388		
04	04-26	77.4	0.9663185		
04	04-27	70.2	0.9928811		
04	04-28	81.7	0.9906759		
04	04-29	83.6	0.9911		
04	04-30	71.7	0.98737		
04	04-31	236	0.992705		
04	04-32	97	0.8530886	Yes	Call rate
04	04-33	142	0.9205905	Yes	Call rate
04	04-34	60.6	0.9830862		
04	04-35	64.2	0.9909006		
Control	H20		0.452228	Yes	Call rate

Supplementary Table S2. Genetic Diversity Statistics for Four IsoPop Populations and for All IsoPop Together

IsoPop population 04 is the urban population.

	ISOPOP1		ISOPOP2		ISOPOP3		ISOPOP4		ISOPOP
Chr	ADOL	Theta (PI)	ADOL	Theta (PI)	ADOL	Theta (PI)	ADOL	Theta (PI)	F_{IS}
1	0.2915+/-0.1386	2597.6083	0.2900+/-0.1392	2589.5705	0.2900+/-0.1385	2584.2541	0.2918+/-0.1402	2574.1903	-0.0025
2	0.2998+/-0.1425	2664.5911	0.3004+/-0.1442	2668.5602	0.2998+/-0.1432	2657.7313	0.3009+/-0.1445	2653.4396	-0.0016
3	0.3058+/-0.1454	2290.2422	0.3059+/-0.1468	2291.4852	0.3048+/-0.1456	2277.5045	0.3061+/-0.1470	2267.9935	-0.0021
4	0.3043+/-0.1447	1922.8321	0.3029+/-0.1454	1912.6884	0.3025+/-0.1446	1908.4705	0.3036+/-0.1458	1907.1312	-0.0004
5	0.3026+/-0.1439	1930.1287	0.3026+/-0.1452	1929.8379	0.3014+/-0.1440	1917.7922	0.3028+/-0.1455	1917.0143	-0.0014
6	0.2951+/-0.1403	2054.056	0.2938+/-0.1410	1358.8209	0.2917+/-0.1394	2024.5564	0.295+/-0.1417	2030.1247	0.0000
7	0.3089+/-0.1469	1762.7036	0.3097+/-0.1486	1769.9704	0.3092+/-0.1477	1762.2940	0.31+/-0.1489	1752.5753	-0.0031
8	0.3086+/-0.1467	1840.5437	0.3095+/-0.1486	1846.6304	0.3083+/-0.1473	1839.5108	0.3099+/-0.1488	1833.2201	0.0002
9	0.3088+/-0.1469	1601.6217	0.3094+/-0.1485	1604.9891	0.3061+/-0.1463	1589.3094	0.3072+/-0.1476	1580.4584	0.0018
10	0.2997+/-0.1425	1830.8045	0.3002+/-0.1441	1835.9401	0.2972+/-0.142	1813.7356	0.2988+/-0.1435	1807.1032	-0.0035
11	0.3002+/-0.1428	1689.5503	0.3000+/-0.1440	1688.0544	0.3011+/-0.1439	1690.2960	0.3011+/-0.1446	1677.9968	-0.0019
12	0.2992+/-0.1423	1606.7668	0.3010+/-0.1445	1617.9341	0.299+/-0.1429	1604.6450	0.2979+/-0.1431	1583.6214	-0.0028
13	0.2936+/-0.1396	1229.7744	0.2927+/-0.1405	1227.3454	0.2914+/-0.1393	1221.4985	0.2947+/-0.1416	1227.6227	-0.0042
14	0.2913+/-0.1386	1028.9169	0.2906+/-0.1395	1027.2317	0.2909+/-0.139	1027.0603	0.2929+/-0.1407	1028.8643	-0.0006
15	0.3013+/-0.1433	1015.2592	0.3028+/-0.1454	1019.9867	0.2998+/-0.1433	1007.4499	0.3011+/-0.1447	1003.2227	-0.0065
16	0.3061+/-0.1456	1078.9934	0.3068+/-0.1473	1082.6443	0.3044+/-0.1455	1070.3680	0.3074+/-0.1477	1070.6104	-0.0058
17	0.3016+/-0.1435	910.5357	0.3017+/-0.1449	910.4410	0.3029+/-0.1448	911.3420	0.3029+/-0.1456	905.4578	-0.0022
18	0.2966+/-0.1411	995.2315	0.2972+/-0.1427	993.9946	0.2931+/-0.1401	982.0236	0.2957+/-0.1421	983.0377	-0.0018
19	0.3014+/-0.1434	588.6017	0.3019+/-0.1450	589.8766	0.3013+/-0.1441	584.4372	0.3011+/-0.1447	575.9227	-0.0019
20	0.2991+/-0.1423	832.3287	0.3006+/-0.1443	834.9552	0.2989+/-0.1429	829.1955	0.3001+/-0.1442	828.8578	-0.0012
21	0.3126+/-0.1488	482.7003	0.3118+/-0.1498	481.9855	0.3127+/-0.1496	482.1951	0.315+/-0.1515	479.8110	-0.0047
22	0.2911+/-0.1386	405.5007	0.2937+/-0.1412	410.0502	0.2918+/-0.1396	404.6933	0.2903+/-0.1396	397.1279	-0.0043

Chr=Chromosome, ADOL=average diversity over loci, F_{IS}=inbreeding coefficient

Supplementary Table S3.

Contingency table showing the proportions of 999,259 markers that are common (minor allele frequency, MAF, greater than 0.05) and rare (MAF less than or equal to 0.05) among 157 IsoPop (denoted IsoPop; Rows) individuals and the 503 individuals in the 1000 Genomes database of European descent (denoted Eur; Columns).

		Eur	
IsoPop		MAF > 0.05	MAF ≤ 0.05
	MAF > 0.05	0.4413	0.0158
	MAF ≤ 0.05	0.0141	0.5288

Figure Captions

Figure 1. Computational methods workflow. Programs used are in dark gray, and unless otherwise noted, were performed in RStudio. The 1000 Genomes Project database has been abbreviated 1KGP.

Figure 2. plot of observed F_{IS} values (Y-axis) of 999,259 SNPs. The X-axis shows the populations: the 1000 Genomes Project individuals are listed as 1KGP, the individuals of European descent from the 1000 Genomes Project are listed as EUR, IsoPop represents all the Illinois individuals, followed by each population separately. The urban population is IsoPop4.

Figure 3. Eigen plot of 153 IsoPop individuals with the four distantly grouped individuals removed using 999,259 SNPs. The X-axis is the value of the first eigenvector, while the Y-axis is the value of the second eigenvector. Individuals from the four IsoPop communities are indicated with different symbols. The majority of the IsoPop individuals cluster into one group. The four IsoPop populations are represented by different symbols and are labeled as 01, 02, 03, and 04.

Figure 4. Eigen plot of 2,504 individuals from 1000 Genomes database and 157 IsoPop individuals using 999,259 SNPs. The X-axis is the value of the first eigenvector, while the Y-axis is the value of the second eigenvector. Individuals from the different subpopulations represented in the 1000 Genomes Project, as well as the four IsoPop communities are colored differently. The IsoPop individuals cluster into the individuals from the 1000 Genomes database that are of European descent. The world populations of Africa (AFR), Americas (AMR), East Asia (EAS), South Asia (SAS), and Europe (EUR), are plotted along with the IsoPop (IL) populations.

Figure 5. Eigen plot of 503 individuals from 1000 Genomes database of European descent and 157 IsoPop individuals using 999,259 SNPs. The X-axis is the value of the first eigenvector, while the Y-axis is the value of the second eigenvector. Individuals from the different European subpopulations represented in the 1000 Genomes Project, as well as the four IsoPop communities are colored differently. The IsoPop individuals cluster into two groups. The European (EUR) populations are plotted with the following abbreviations: Utah residents in CEPH (CEU), Finland (FIN), British in England and Scotland (GBR), Iberian populations in Spain (IBS), and Toscani in Italia (TSI), are plotted along with the IsoPop (IL) populations. The four IsoPop populations are represented by different symbols and are labeled as 01, 02, 03, and 04.

Supplementary Figure S1. Heatmap depicting values of the coancestry matrix for all 2,504 individuals from 1000 Genomes database and 157 IsoPop individuals. The actual numerical coancestry values between each pair of individuals are provided in Supplementary File 1.

Supplementary Figure S2. Scree plot for the eigenanalysis of all 157 IsoPop individuals. The X-axis indicates the index of eigenvalues, while the Y-axis indicates the numerical value of each eigenvalue.

Supplementary Figure S3. EIGMIX plot of all 157 IsoPop individuals. The four IsoPop populations are represented by different symbols and are labeled as 01, 02, 03, and 04. The X-axis is the value of the first eigenvector, while the Y-axis is the value of the second eigenvector.

Supplementary Figure S4. EIGMIX plot of all 157 IsoPop individuals. The four IsoPop populations are represented by different colors and are labeled as 01, 02, 03, and 04. The X-axis is the value of the first eigenvector, while the Y-axis is the value of the third eigenvector.

Supplementary Figure S5. EIGMIX plot of all 157 IsoPop individuals. The four IsoPop populations are represented by different colors and are labeled as 01, 02, 03, and 04. The X-axis is the value of the second eigenvector, while the Y-axis is the value of the third eigenvector.

Supplementary Figure S6. Scree plot for the eigenanalysis of all 2,504 individuals from 1000 Genomes database 157 IsoPop individuals. The X-axis indicates the index of eigenvalues, while the Y-axis indicates the numerical value of each eigenvalue.

Supplementary Figure S7. EIGMIX plot of all 2,504 individuals from 1000 Genomes database 157 IsoPop individuals. The four IsoPop populations are represented by different colors and are labeled as 01, 02, 03, and 04. The world populations of Africa (AFR), Americas (AMR), East Asia (EAS), South Asia (SAS), and Europe (EUR), are also indicated in different colors.

The X-axis is the value of the third eigenvector, while the Y-axis is the value of the first eigenvector.

Supplementary Figure S8. EIGMIX plot of all 2,504 individuals from 1000 Genomes database 157 IsoPop individuals. The four IsoPop populations are represented by different colors and are labeled as 01, 02, 03, and 04. The world populations of Africa (AFR), Americas (AMR), East Asia (EAS), South Asia (SAS), and Europe (EUR), are also indicated in different colors.

The X-axis is the value of the third eigenvector, while the Y-axis is the value of the second eigenvector.

Supplementary Figure S9. Scree plot for the eigenanalysis of 503 individuals from 1000 Genomes database of European descent and 157 IsoPop individuals The X-axis indicates the index of eigenvalues, while the Y-axis indicates the numerical value of each eigenvalue.

Supplementary Figure S10. Eigen plot of 503 individuals from 1000 Genomes database of European descent and 157 IsoPop individuals using EIGMIX. The X-axis is the value of the first eigenvector, while the Y-axis is the value of the third eigenvector. Individuals from each European subpopulation represented in the 1000 Genomes Project, as well as the each of the four IsoPop communities, are colored differently. The European (EUR) populations are plotted with the following abbreviations: Utah residents in CEPH (CEU), Finland (FIN), British in England and Scotland (GBR), Iberian populations in Spain (IBS), and Toscani in Italia (TSI), are plotted along with the IsoPop (IL) populations. The four IsoPop populations are represented by different symbols and are labeled as 01, 02, 03, and 04.

Supplementary Figure S11. Eigen plot of 503 individuals from 1000 Genomes database of European descent and 157 IsoPop individuals using EIGMIX. The X-axis is the value of the second eigenvector, while the Y-axis is the value of the third eigenvector. Individuals from each European subpopulation represented in the 1000 Genomes Project, as well as the each of the four IsoPop communities, are colored differently. The European (EUR) populations are plotted with the following abbreviations: Utah residents in CEPH (CEU), Finland (FIN), British in England and Scotland (GBR), Iberian populations in Spain (IBS), and Toscani in Italia (TSI), are plotted along with the IsoPop (IL) populations. The four IsoPop populations are represented by different symbols and are labeled as 01, 02, 03, and 04.

Supplementary Figure S12. Empirical density (Y-axis) of the differences in minor allele frequencies (MAFs) of 999,259 markers among the 157 IsoPop individuals and the 503 individuals in the 1000 Genomes database of European descent. The mode of this density is centered at 0, suggesting that the overwhelming majority of these markers have similar MAFs in both of these data sets.

Supplementary Figure S13. Linkage disequilibrium (LD) decay plots among the 2,504 individuals in the 1,000 Genomes data base (1KGP), the 503 individuals in the 1000 Genomes database of European descent (EUR), all 157 individuals of the IsoPop population (ISOPOP), as well as the IsoPop individuals subdivided into the four communities (ISOPOP1-ISOPOP4). For each graph, the Y-axis is the squared Pearson correlation coefficient (r^2) between marker pairs, and the X-axis depicts the physical distance between markers (kb). (A) the range of values in the X-axis is from 0 kb – 1,000 kb; (B) the range of values in the X-axis is from 0 kb – 300 kb. Note that the LD decay is higher for the urban population (ISOPOP4) compared to the three rural communities (ISOPOP1-ISOPOP3), which is the opposite of what would be expected should a founder effect exist in these rural communities.

Figure 1.

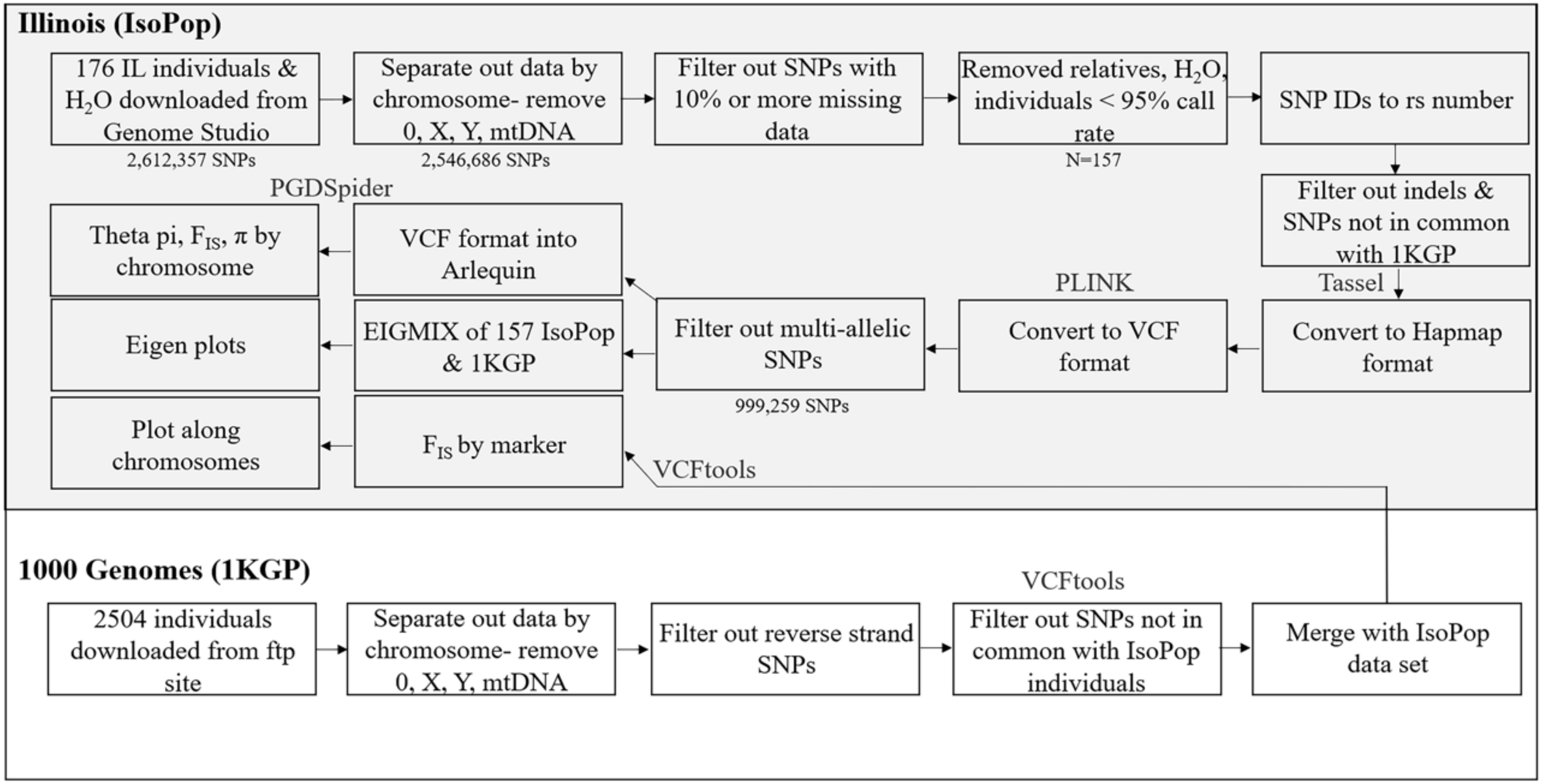

Illinois (IsoPop)
176 IL individuals & H_2O downloaded from Genome Studio
2,612,357 SNPs
Separate out data by chromosome- remove 0, X, Y, mtDNA
2,546,686 SNPs
Filter out SNPs with 10% or more missing data
Removed relatives, H_2O, individuals < 95% call rate
N=157
SNP IDs to rs number
Filter out indels & SNPs not in common with 1KGP
Tassel
Convert to Hapmap format
PLINK
Convert to VCF format
Filter out multi-allelic SNPs
999,259 SNPs
PGDSpider
VCF format into Arlequin
Theta pi, F_{IS}, π by chromosome
EIGMIX of 157 IsoPop & 1KGP
Eigen plots
VCFtools
F_{IS} by marker
Plot along chromosomes
1000 Genomes (1KGP)
2504 individuals downloaded from ftp site
Separate out data by chromosome- remove 0, X, Y, mtDNA
Filter out reverse strand SNPs
VCFtools
Filter out SNPs not in common with IsoPop individuals
Merge with IsoPop data set

Figure 2.

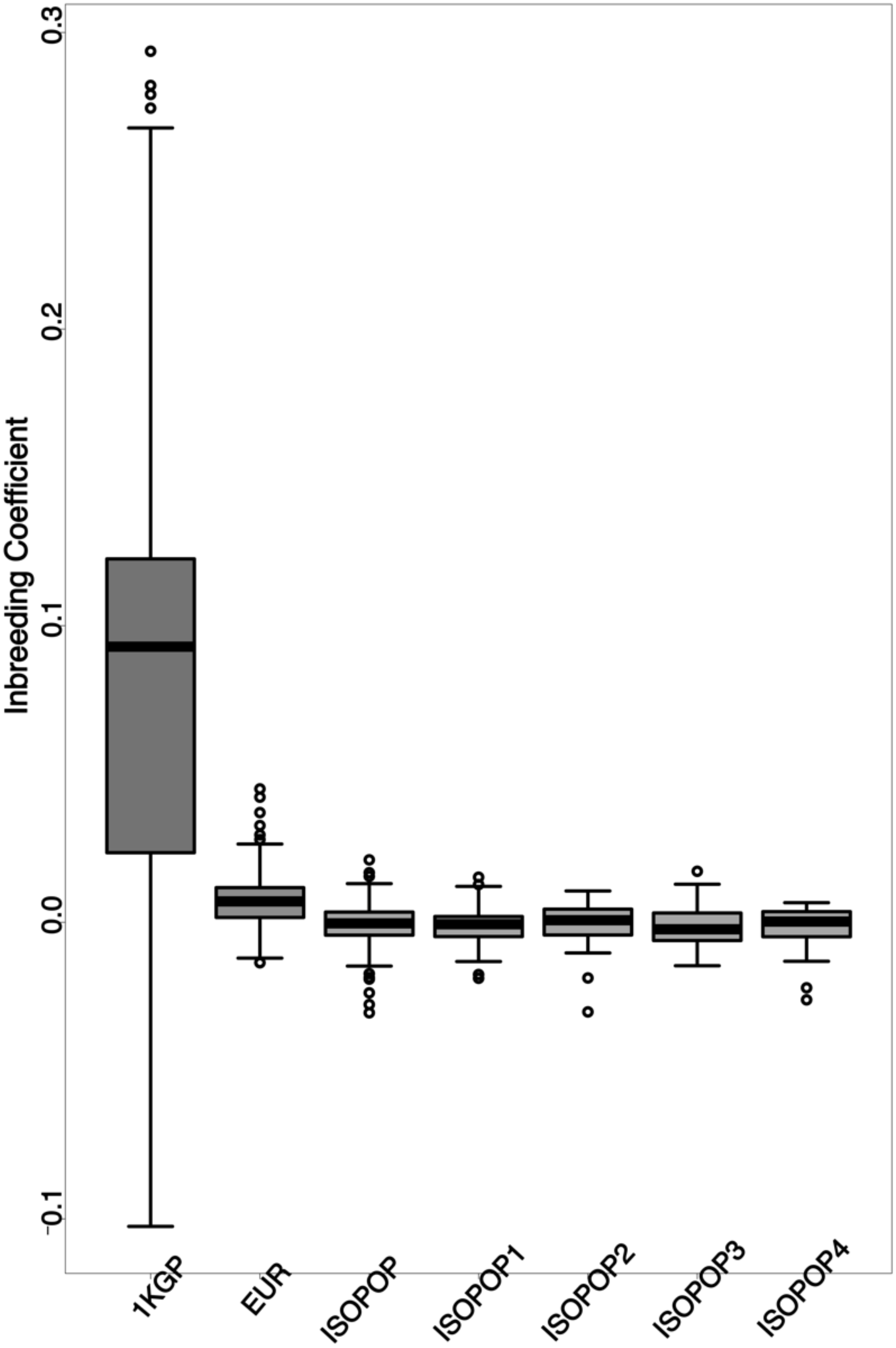

Inbreeding Coefficient
0.3
0.2
0.1
0.0
−0.1
1KGP
EUR
ISOPOP
ISOPOP1
ISOPOP2
ISOPOP3
ISOPOP4

Figure 3.

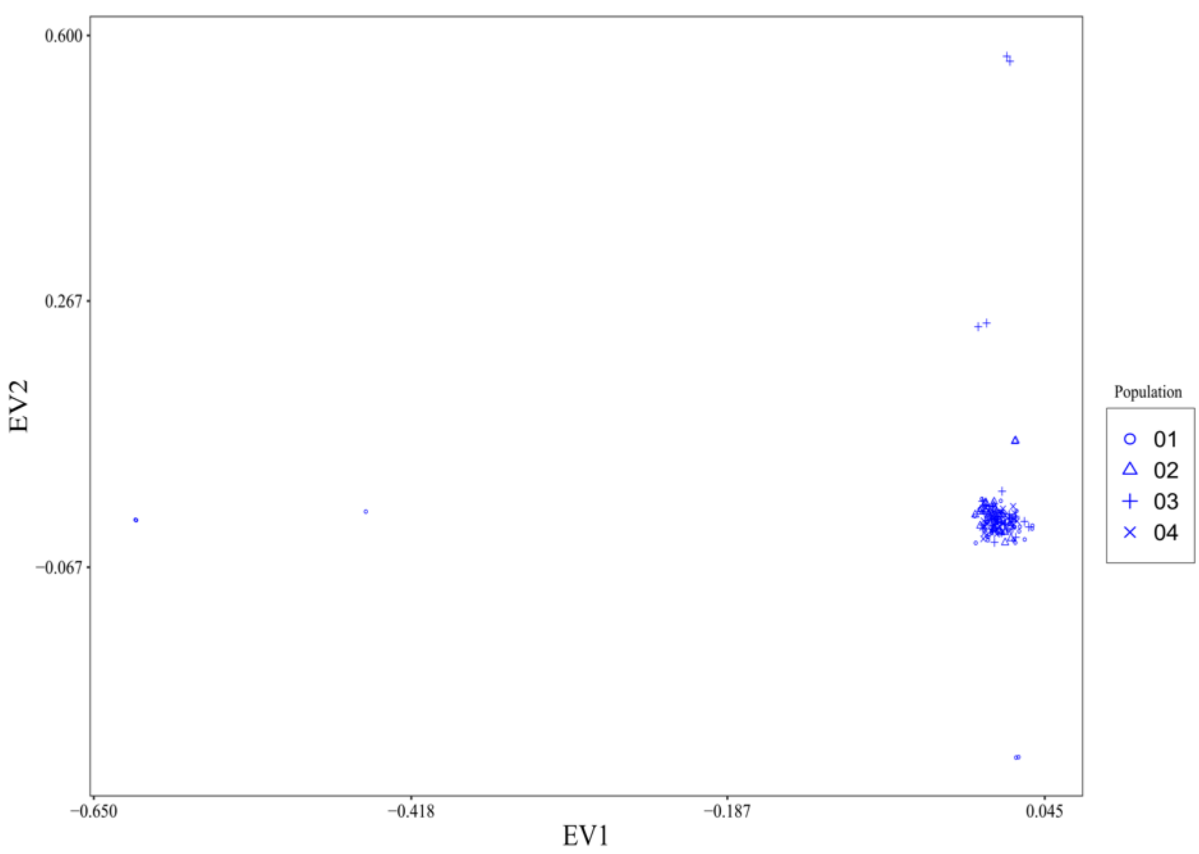

0.600
0.267
−0.067
EV2
−0.650
−0.418
−0.187
0.045
EV1
Population
○ 01
△ 02
+ 03
× 04

Figure 4.

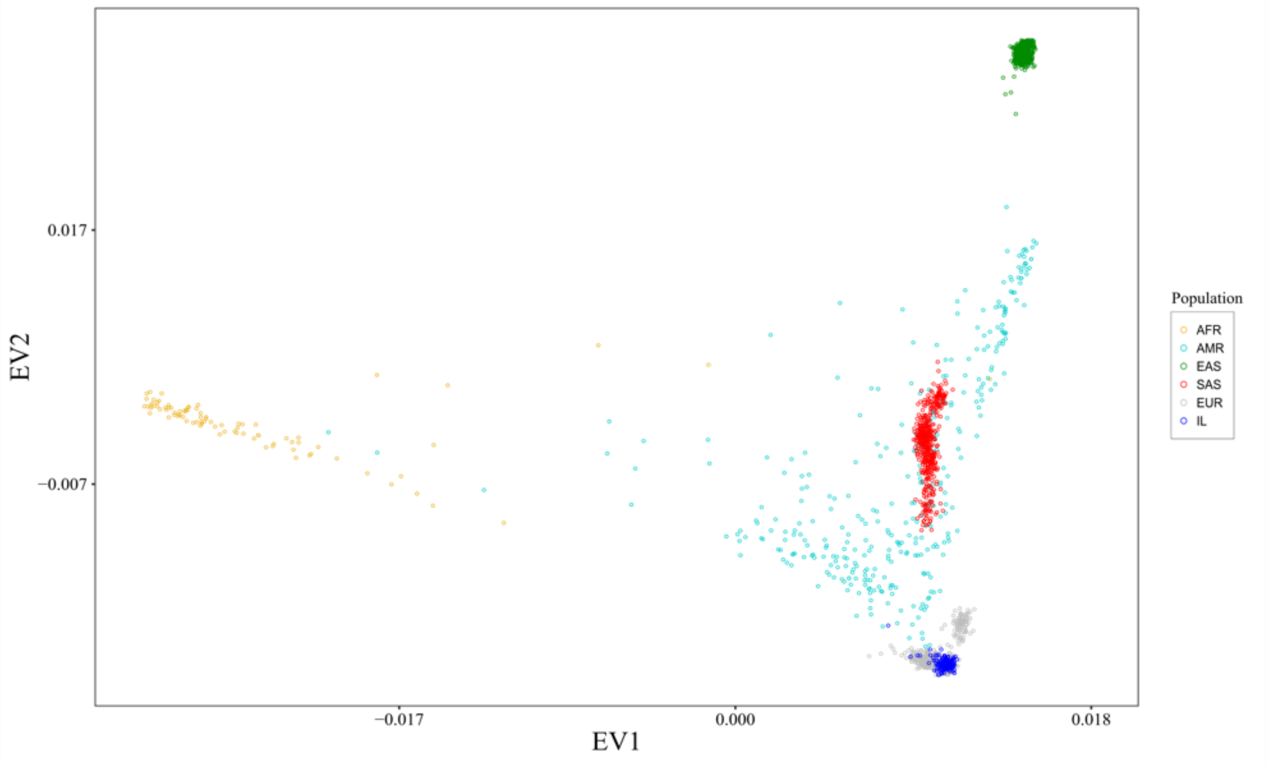
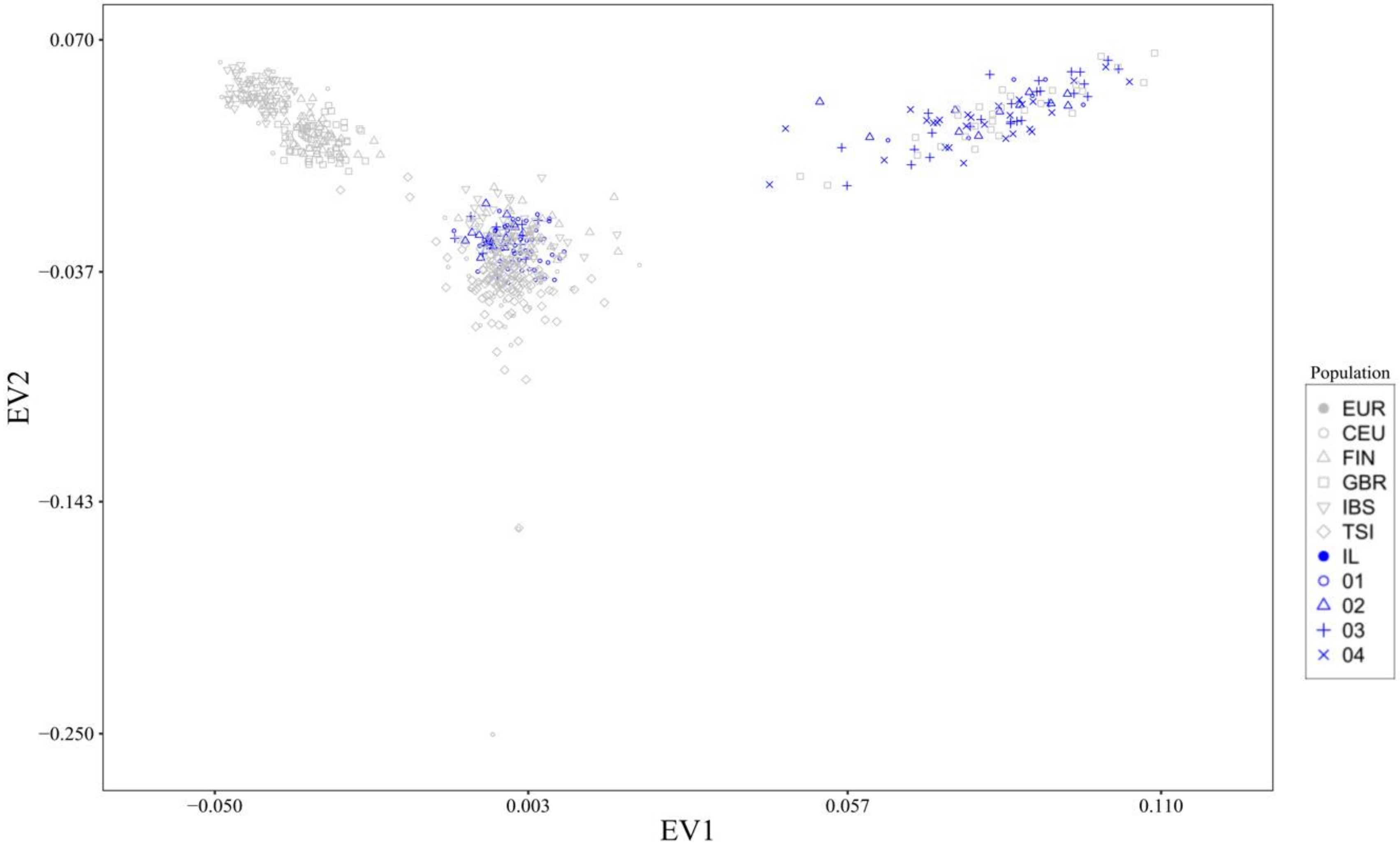
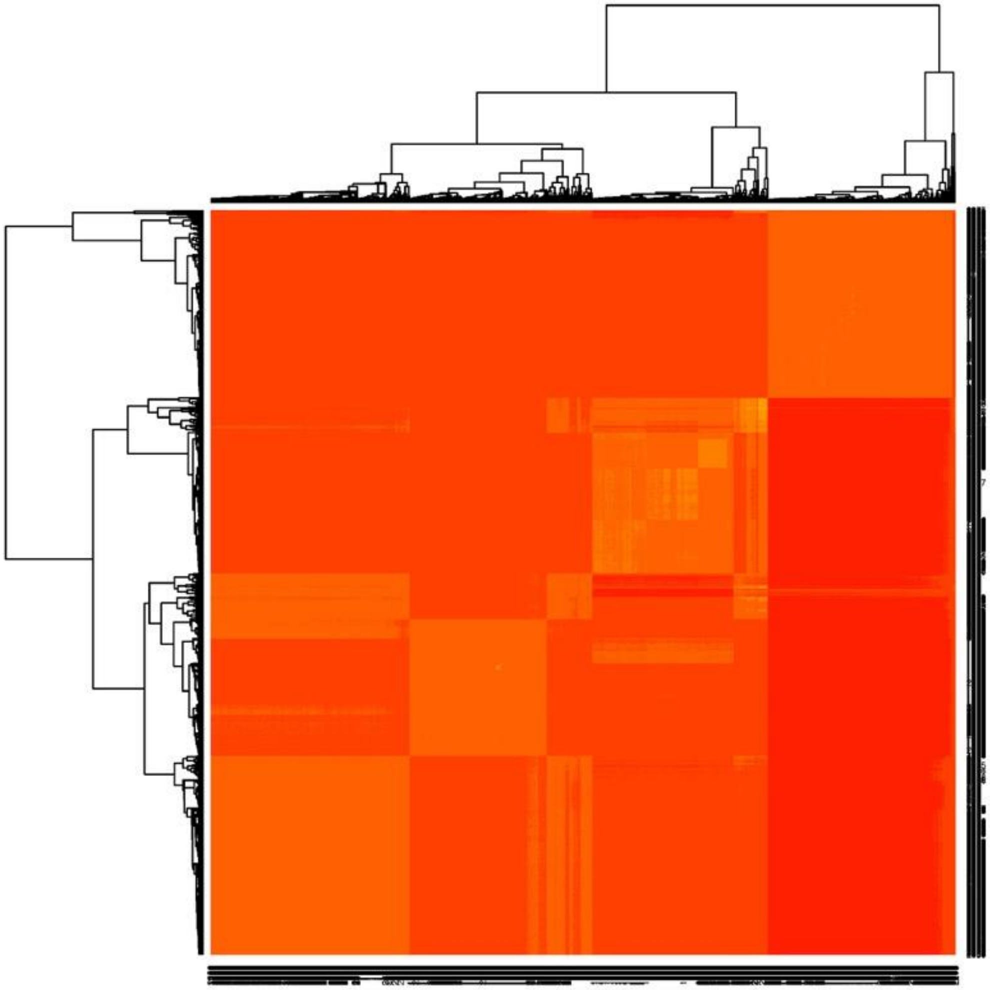

0.017
−0.007
EV2
−0.017
0.000
0.018
EV1
Population
AFR
AMR
EAS
SAS
EUR
IL

Figure 5.

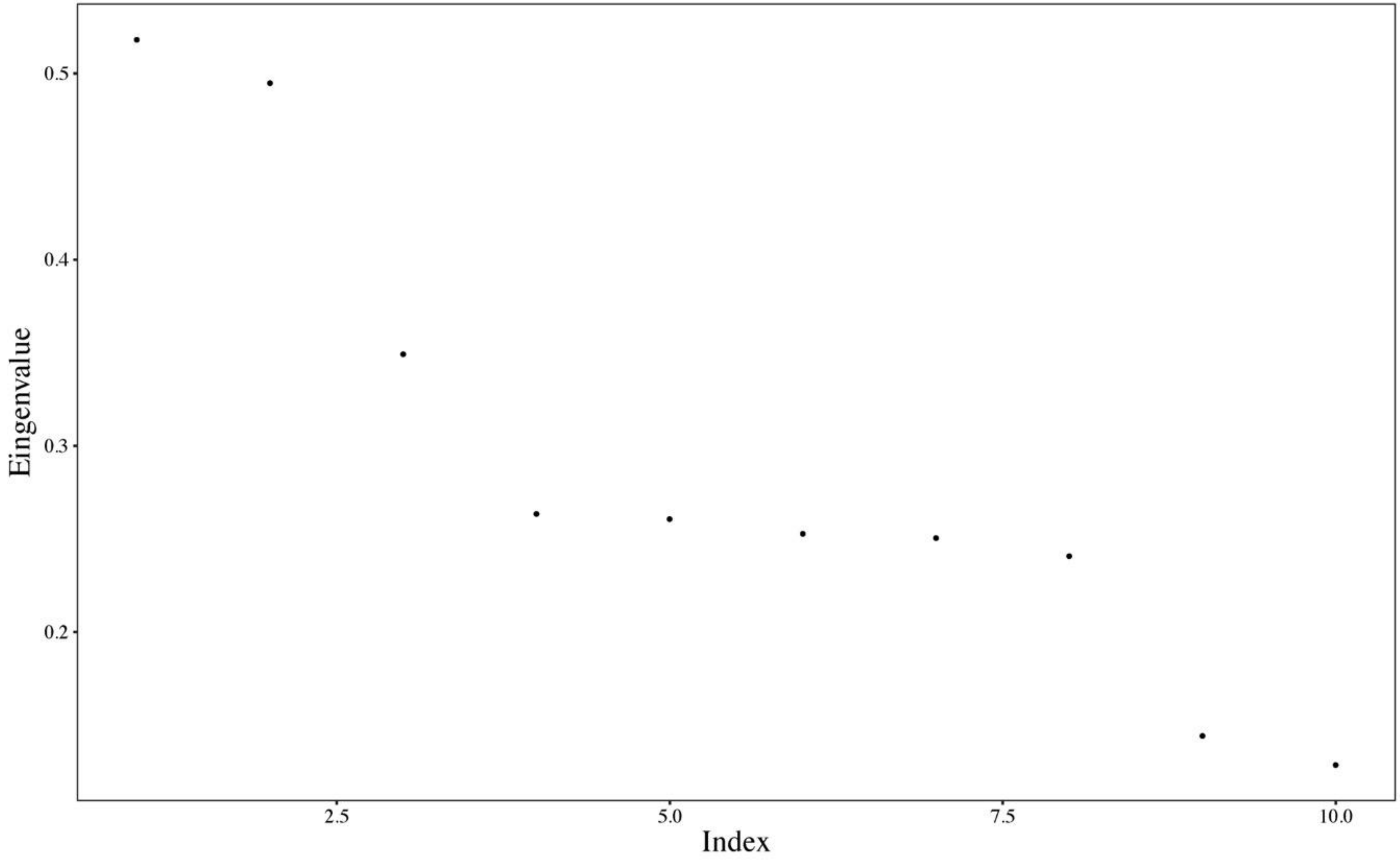


0.070
−0.037
−0.143
−0.250
EV2
−0.050
0.003
0.057
0.110
EV1
Population
EUR
CEU
FIN
GBR
IBS
TSI
IL
01
02
03
04

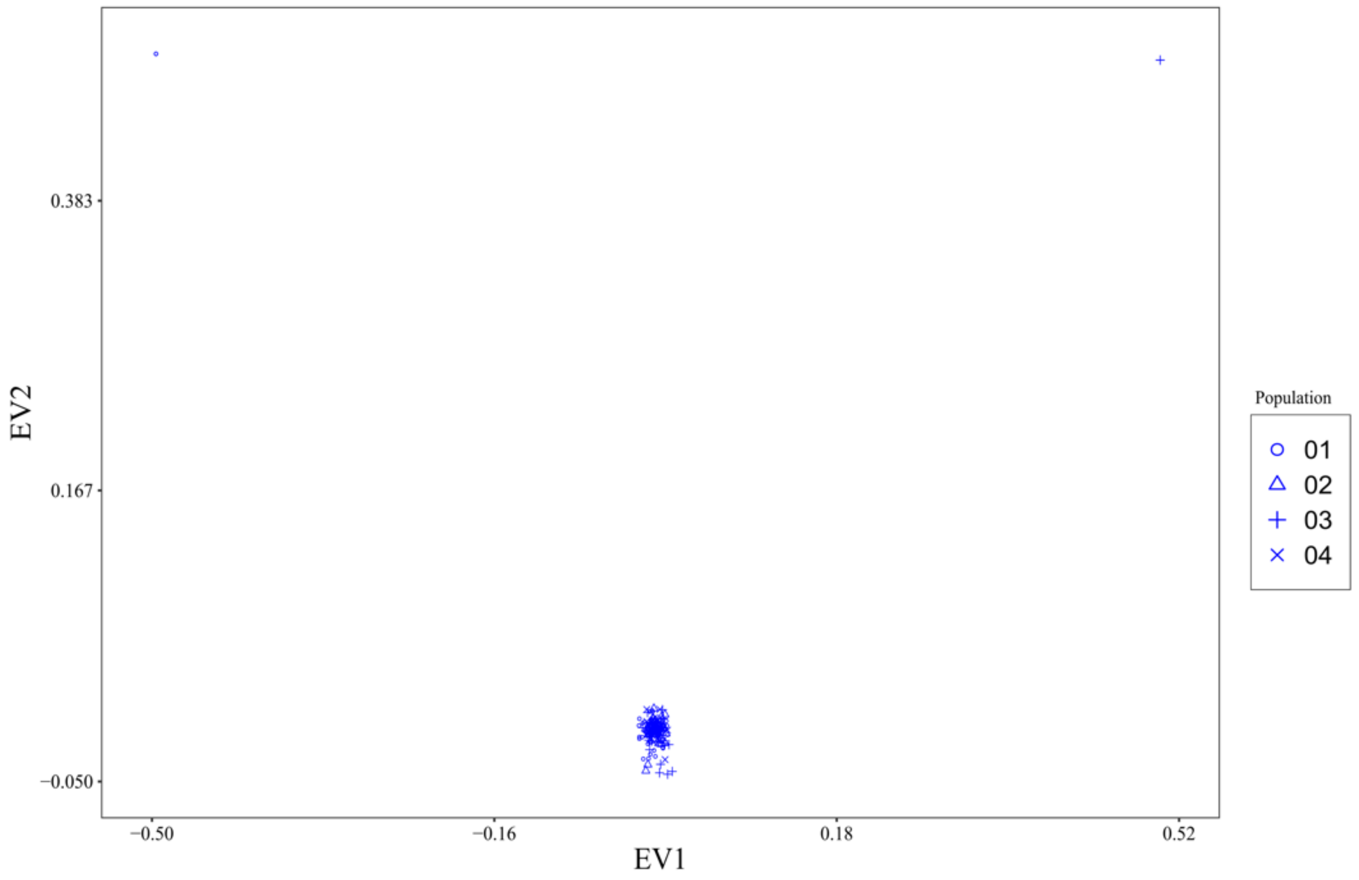
Supplementary Figure S1.



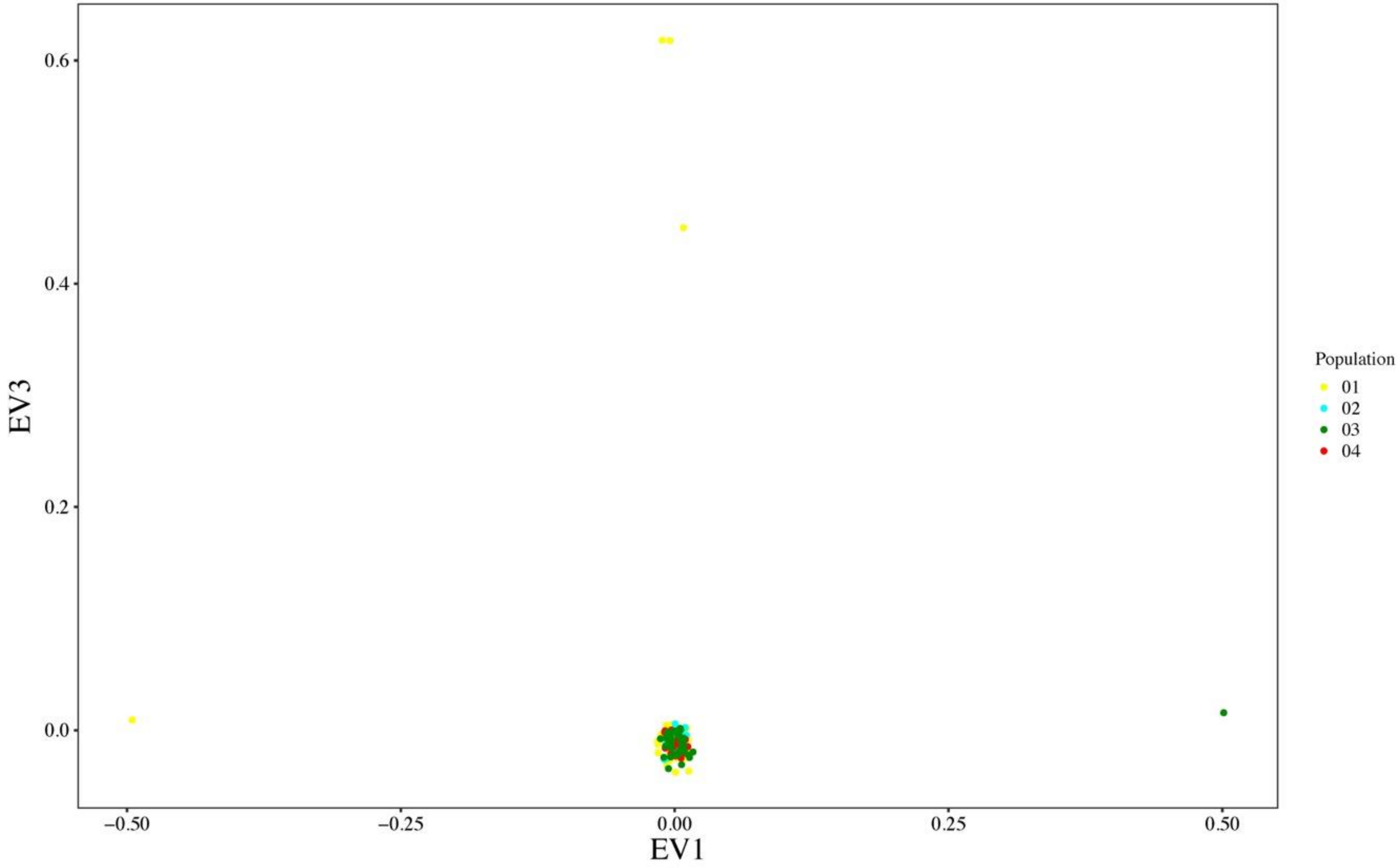
Supplementary Figure S2.



Supplementary Figure S3.

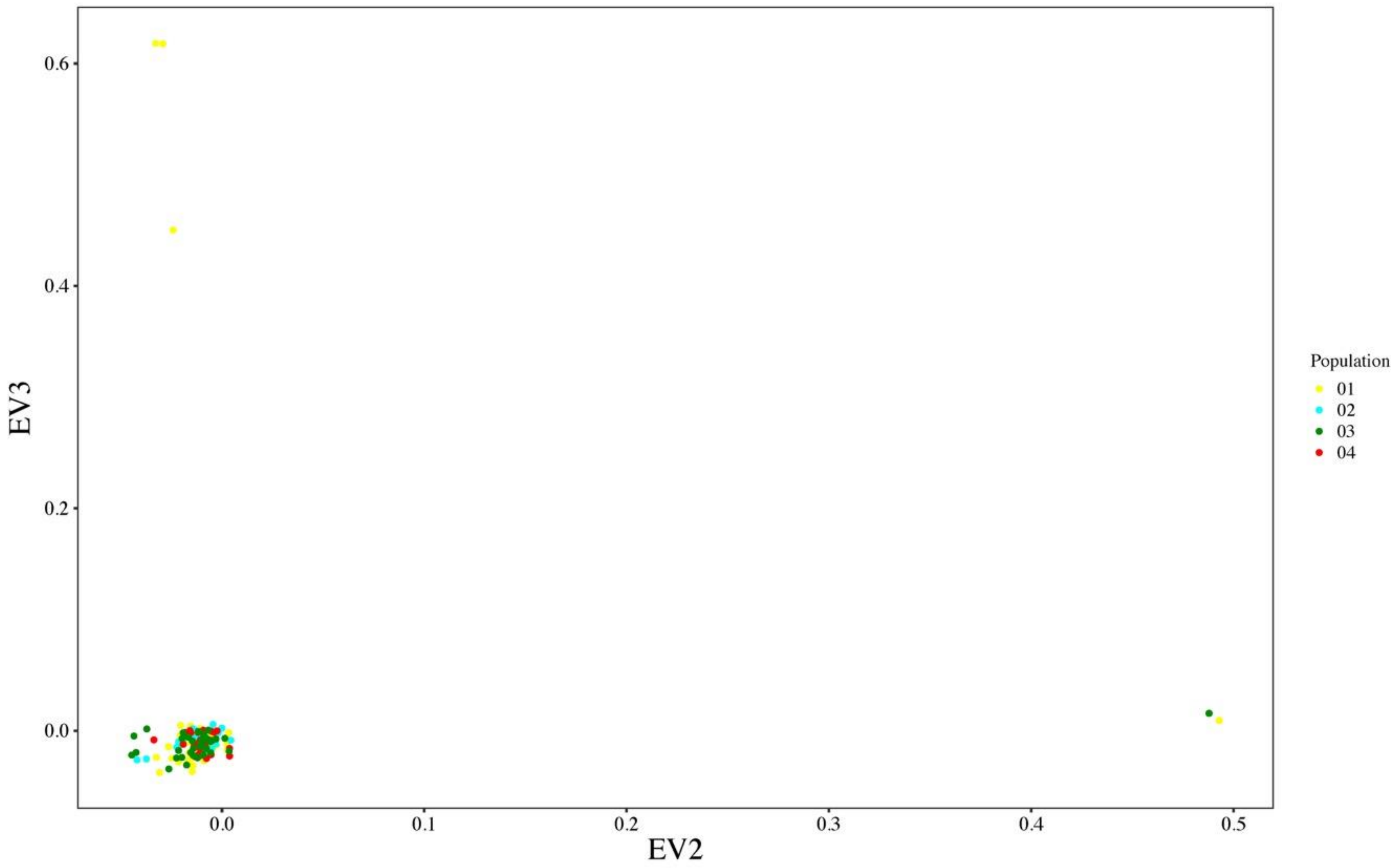


Supplementary Figure S4.

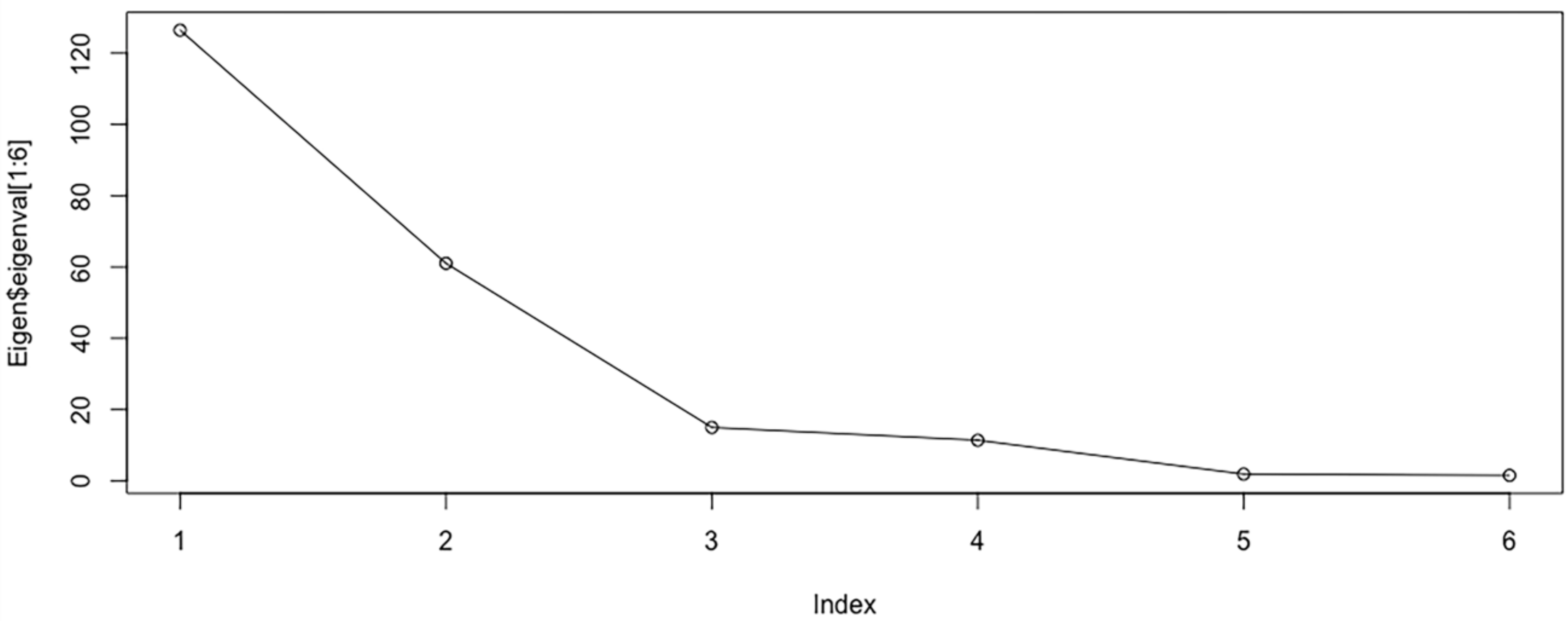


Pre-print version. Visit http://digitalcommons.wayne.edu/humbiol/ after publication to acquire the final version.

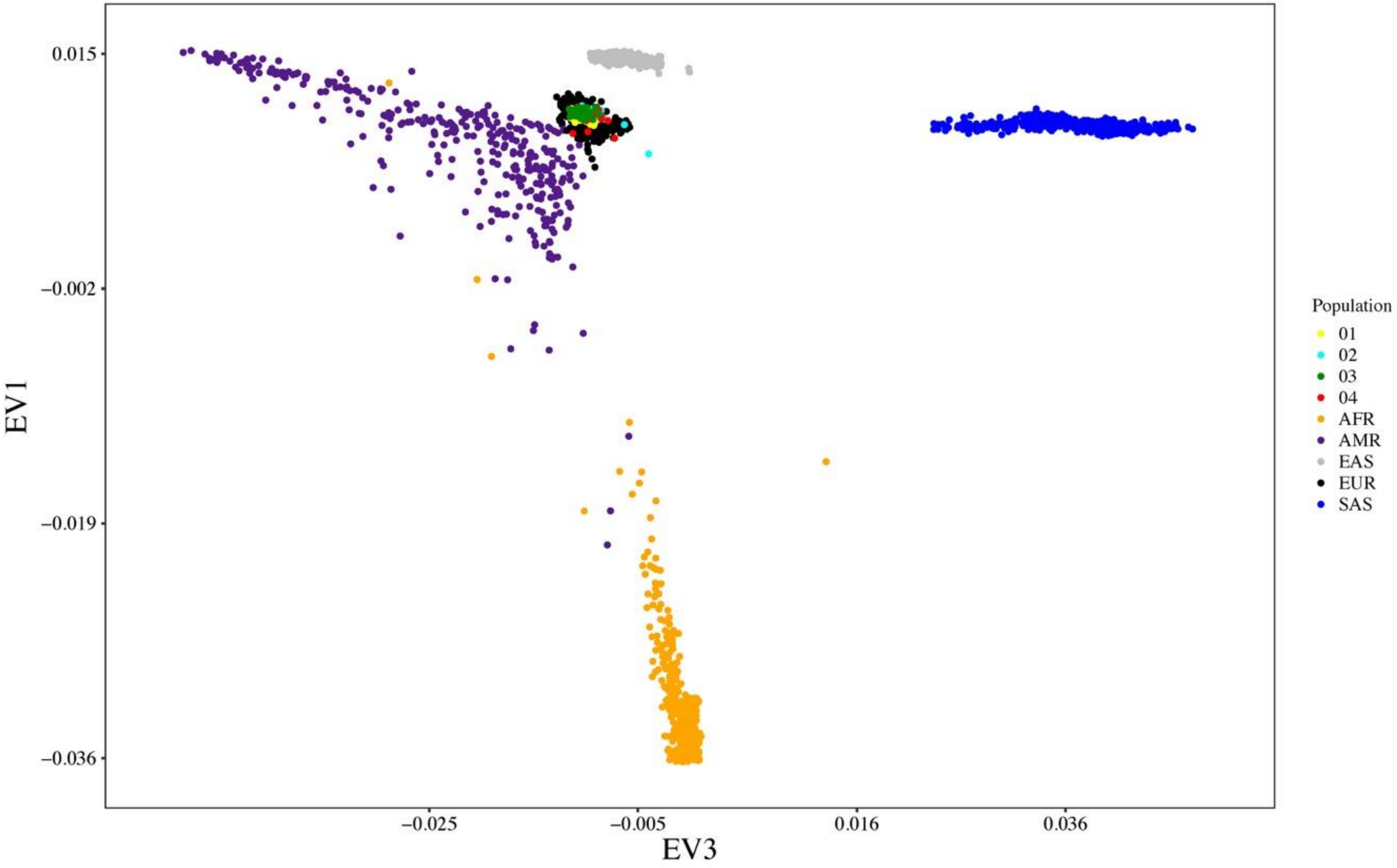
Supplementary Figure S5.



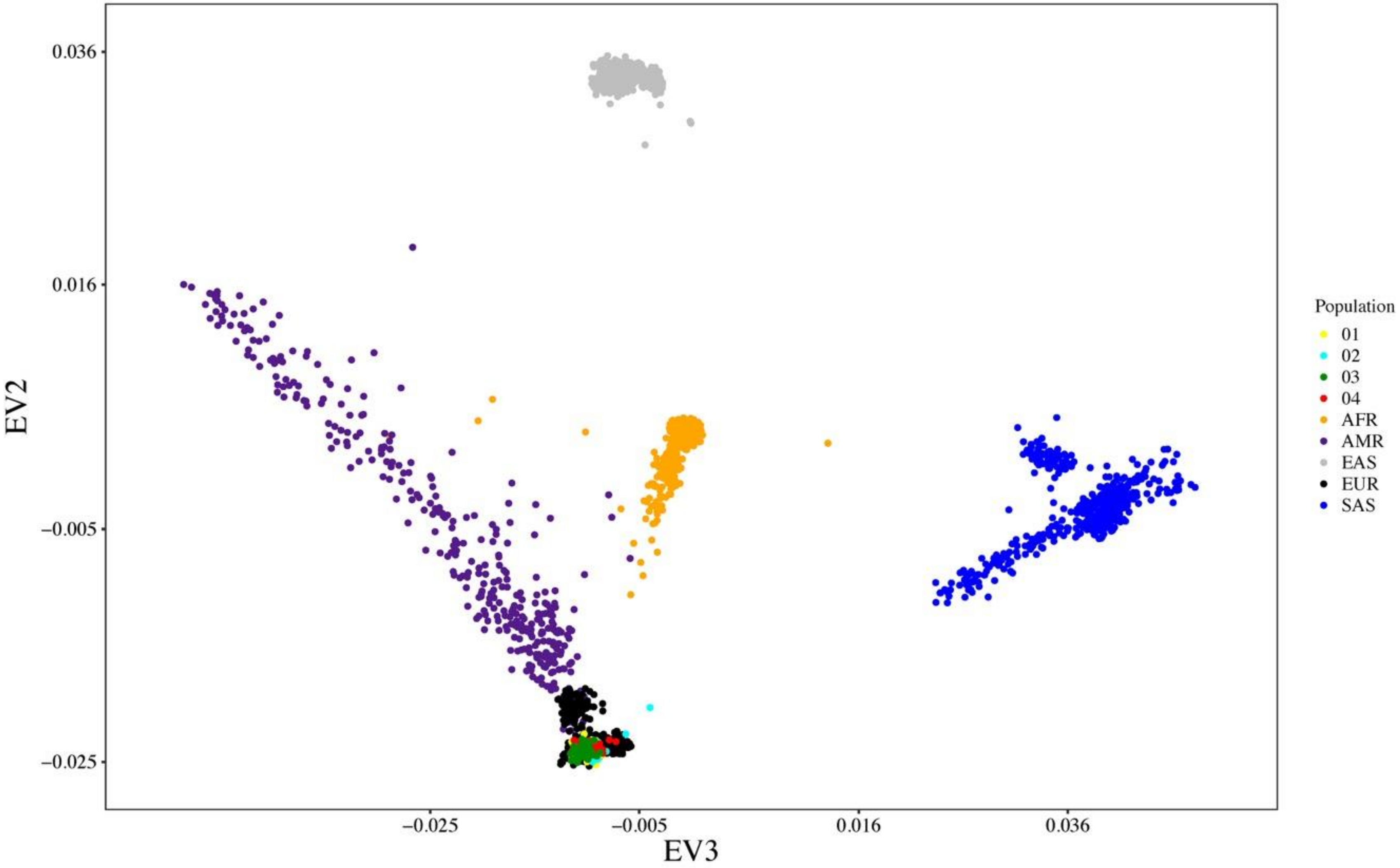
Supplementary Figure S6.



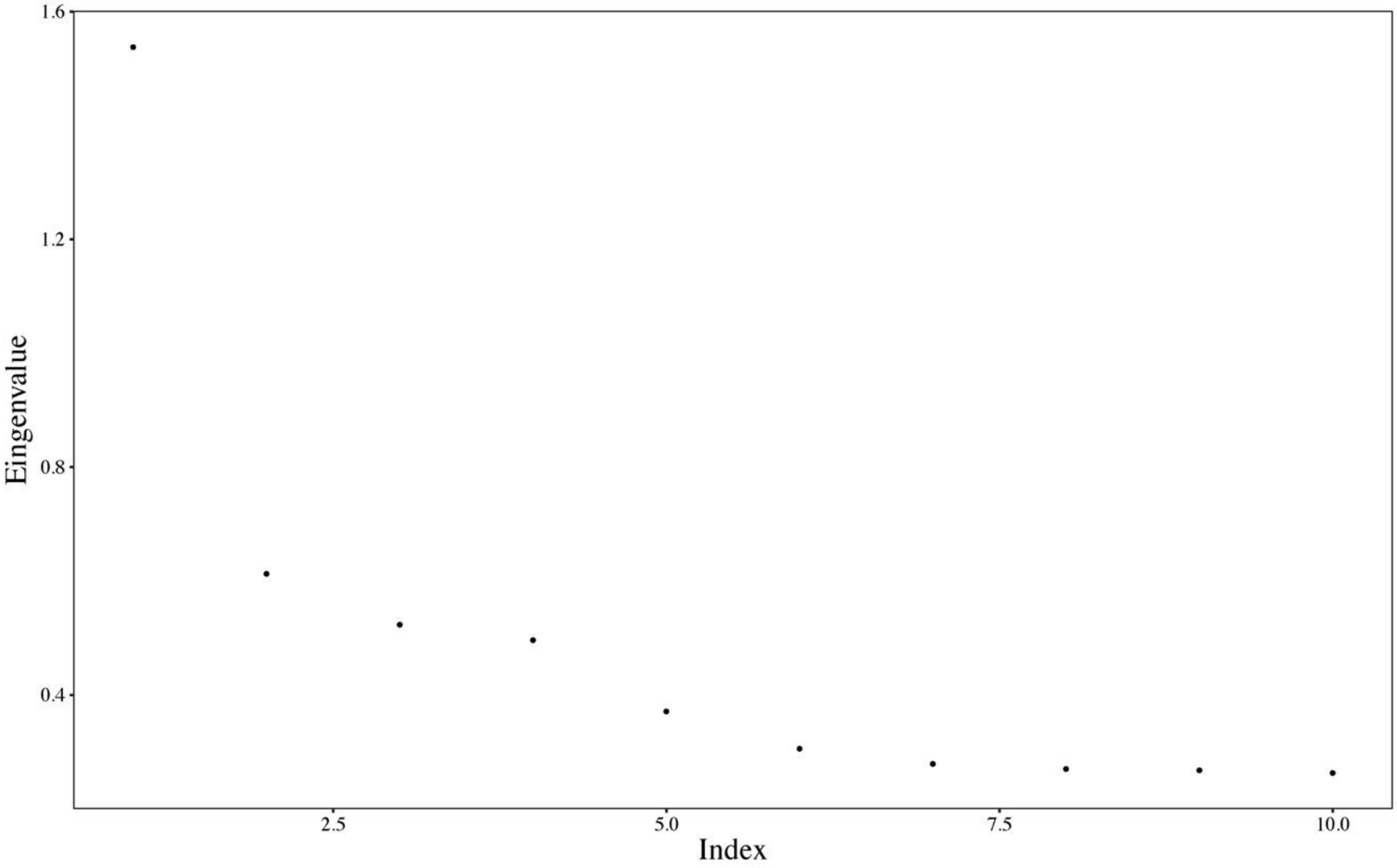
Supplementary Figure S7.



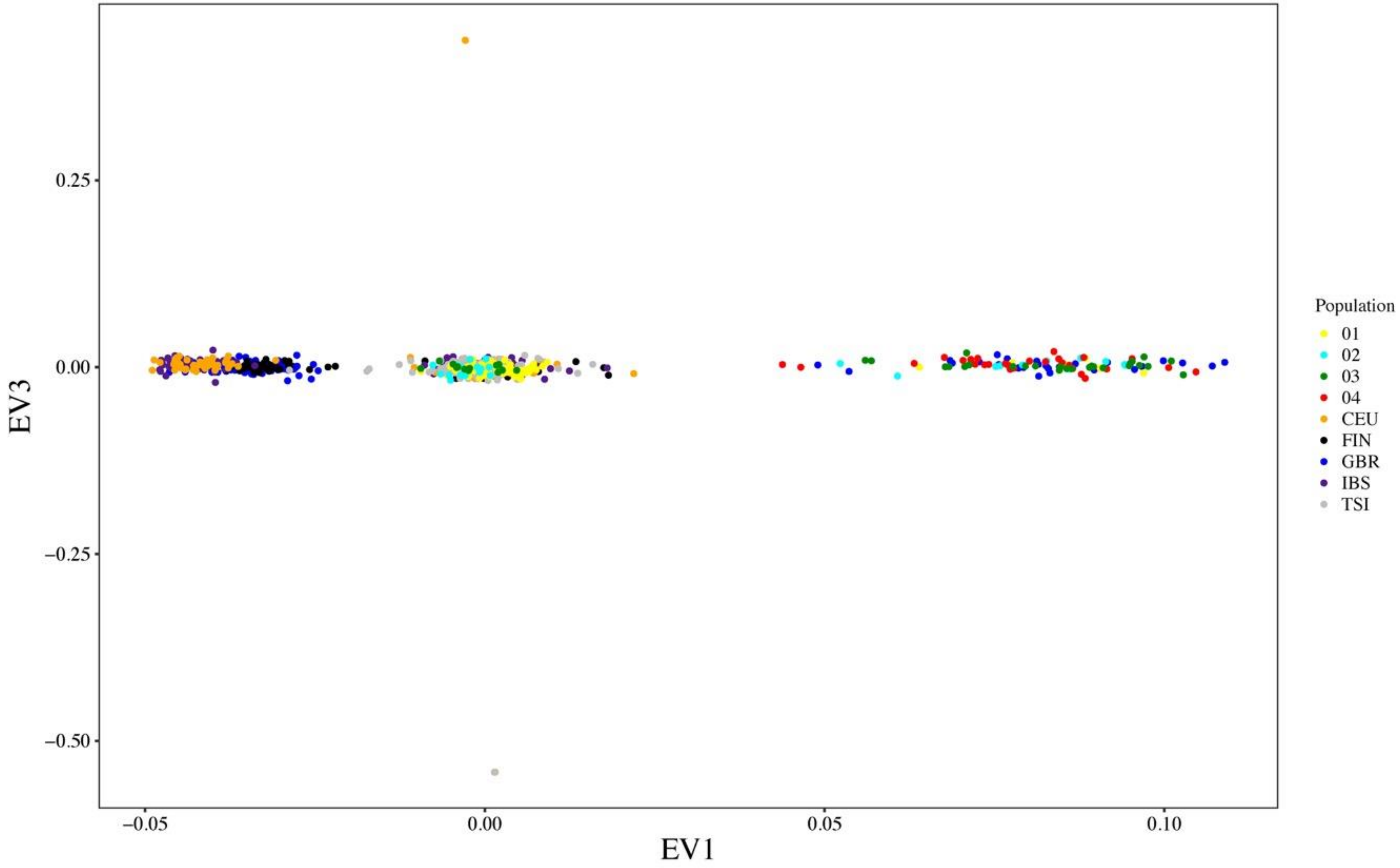
Supplementary Figure S8.



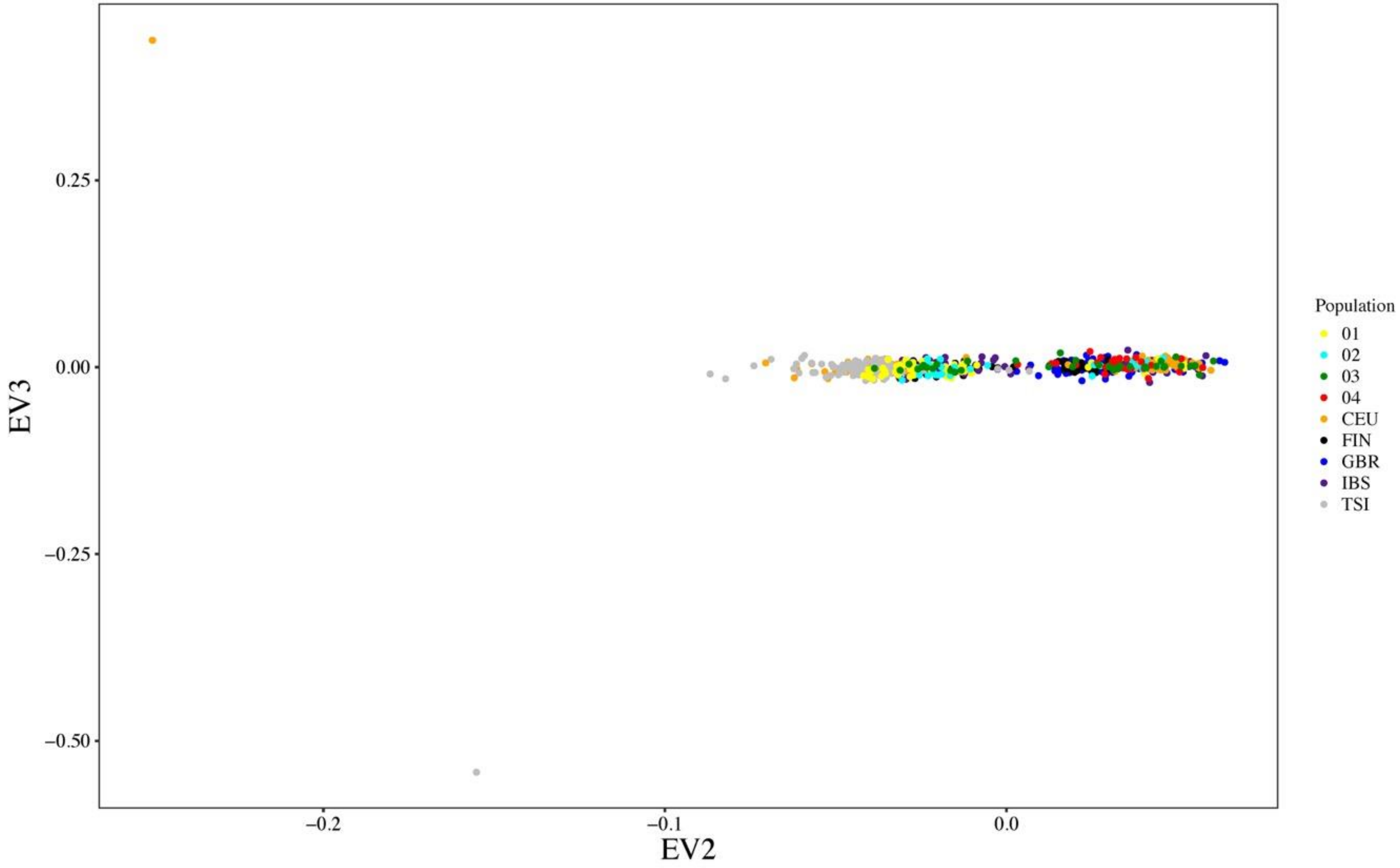
Supplementary Figure S9.



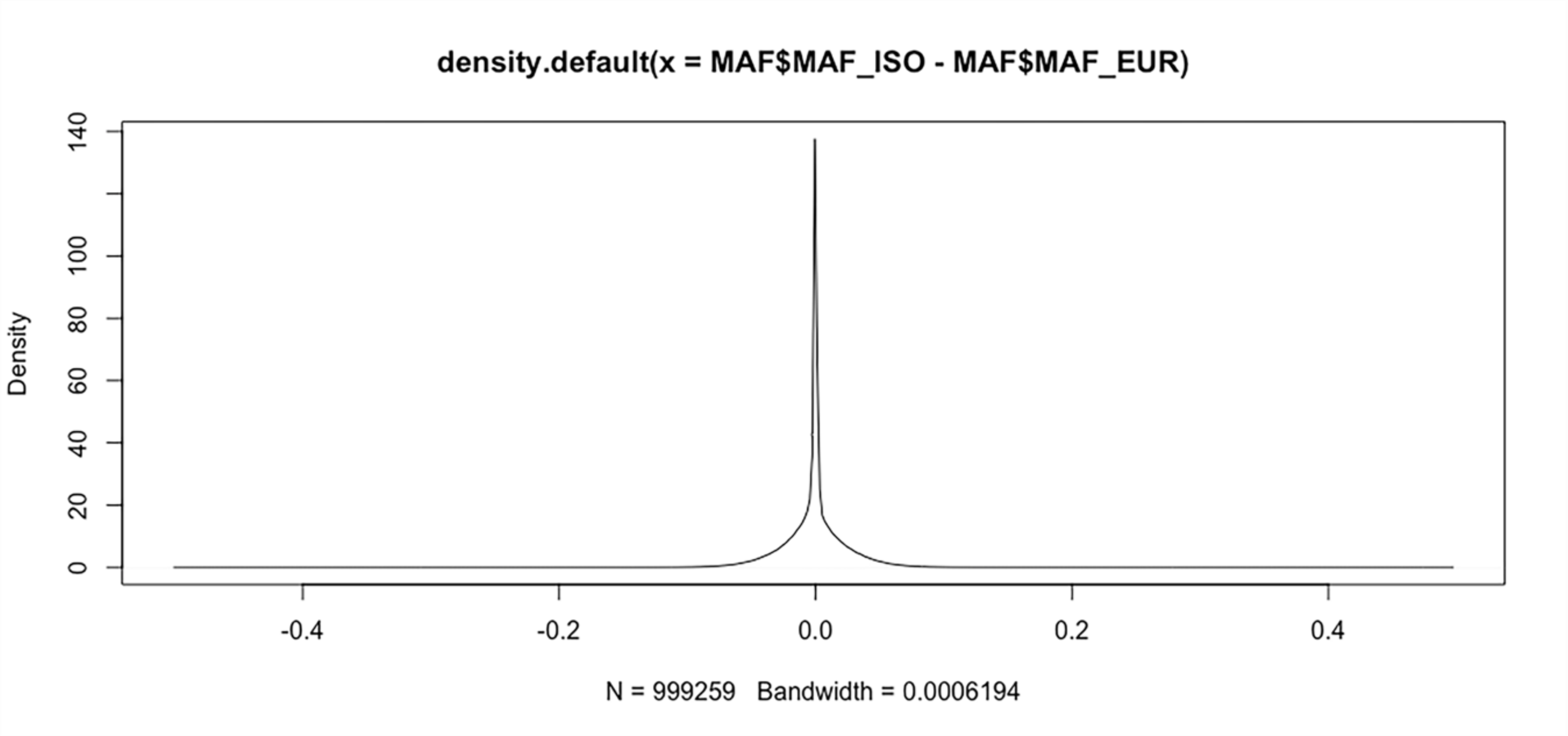
Supplementary Figure S10.



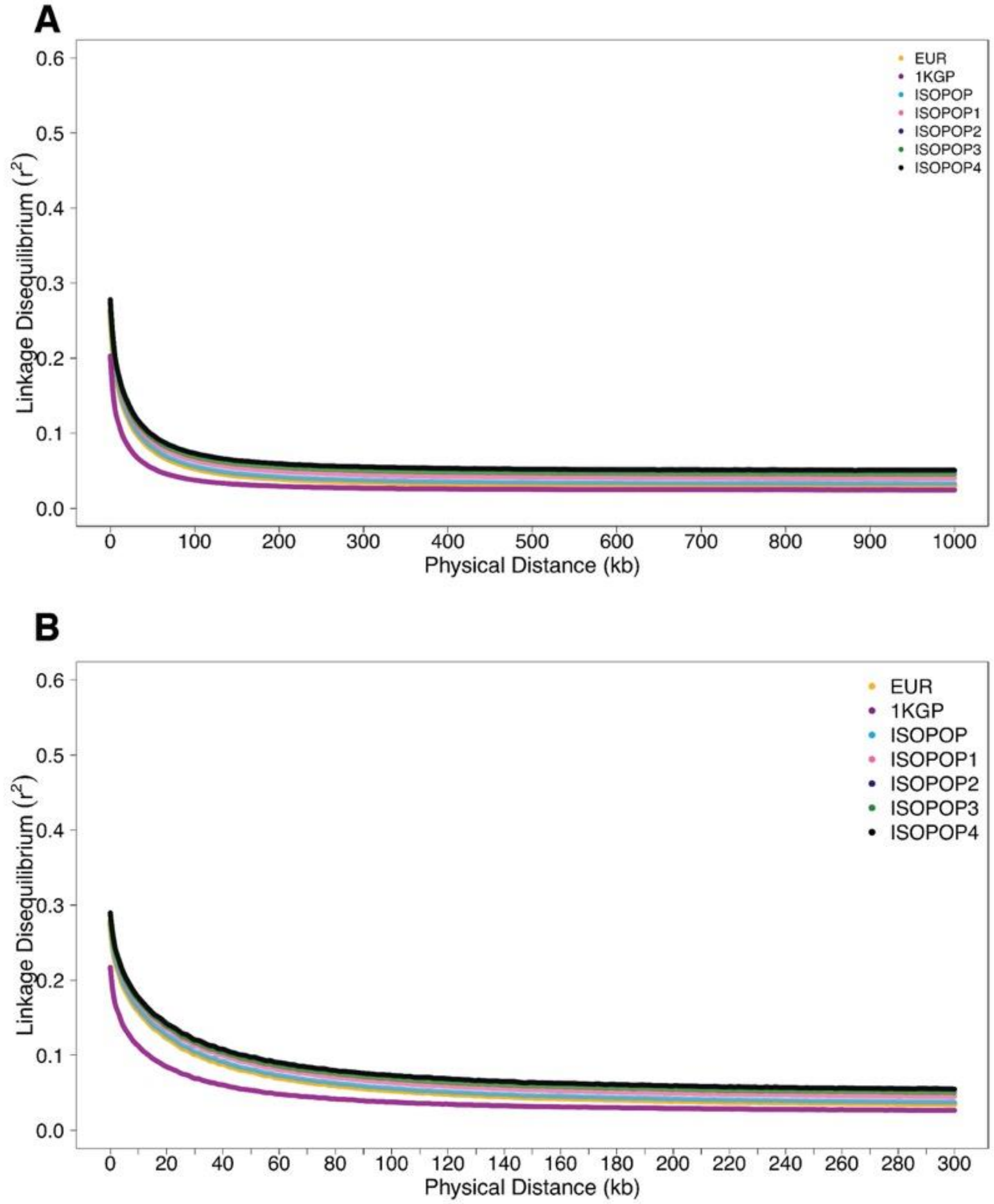
Supplementary Figure S11.



Supplementary Figure S12.



Supplementary Figure S13.


A
Linkage Disequilibrium (r^2)
Physical Distance (kb)
● EUR
● 1KGP
● ISOPOP
● ISOPOP1
● ISOPOP2
● ISOPOP3
● ISOPOP4
B
Linkage Disequilibrium (r^2)
Physical Distance (kb)
● EUR
● 1KGP
● ISOPOP
● ISOPOP1
● ISOPOP2
● ISOPOP3
● ISOPOP4