Determining population structure from k-mer frequencies.

Yana Hrytsenko (✉ yana_hrytsenko@uri.edu)  
University of Rhode Island

Noah M. Daniels  
University of Rhode Island

Rachel S. Schwartz  
University of Rhode Island

Research Article

Keywords: population structure, population differentiation, k-mers, k-mer frequencies

Posted Date: May 27th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1689838/v1

License: ☇️ This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

Background:

Determining population structure helps us understand connections among different populations and how they evolve over time. This knowledge is important for studies ranging from evolutionary biology to large-scale variant-trait association studies. Current approaches to determining population structure include model-based approaches, statistical approaches, and distance-based ancestry inference approaches. In this work, we outline an approach that identifies population structure from k-mer frequencies using principal component analysis (PCA). This approach can be classified as statistical; however, prior work employing PCA has used multilocus genotype data (SNPs, microsatellites, or haplotypes), while here we analyze k-mer frequencies. K-mer frequencies can be viewed as a summary statistic of a genome and have the advantage of being easily derived from a genome by counting the number of times a k-mer occurred in a sequence. No genetic assumptions must be met to generate k-mers, whereas current population structure approaches often depend on several genetic assumptions and require careful selection of ancestry informative markers to identify populations.

Results:

In this work, we show that PCA is able to determine population structure just from the frequency of k-mers found in the genome. The application of PCA and a clustering algorithm to k-mer profiles of genomes provides an easy approach to detecting the number and composition of populations (clusters) present in the dataset. We describe this approach and show that the results are comparable to those found by a model-based approach using genetic markers. We validate our method using 48 human genomes from populations identified by the 1000 Human Genomes Project. We also compare our results to those from 
\textit{mash}, which determines relationships among individuals using the number of matched k-mers between sequences.

Conclusions:

This study shows that PCA, together with a clustering algorithm, is able to detect population structure from k-mer frequencies and can identify samples of admixed and non-admixed origin. In contrast, \textit{mash} (based on the number of k-mer matches) was highly sensitive to the parameters of k-mer length and sketch size. Using k-mer frequencies to determine population structure has the potential to avoid some challenges of existing methods.

Background

Population structure is a complex phenomenon influenced by the combined effects of various processes such as geographic or demographic barriers. Samples that are located in close geographic proximity tend to look more genetically similar than those that are more remote\textsuperscript{1}. When populations are subdivided, they can evolve as separate lineages experiencing differences in recombination, mutation, genetic drift,
demographic history, and natural selection. Population structure is thus observed as a systematic difference in allele frequencies among populations due to non-random mating among individuals. Genetic differences within and among populations are examined by studies that investigate changes in frequencies of alleles and genotypes over time.

Identification of population structure and gene flow among populations is informative for genetic ancestry and provides information about both demographic history and geographic origins. For example, gene flow, or a gene transfer from one population to another, is indicative of migration processes. When individuals of a single population possess recent ancestry from two or more separate sources, this population is considered admixed. Admixed populations contain high levels of genetic diversity that reflect contributions of the intermixture of source populations with different genetic variants.

Understanding gene flow among populations informs a diversity of studies of species. For example, studies in population structure across marine species have analyzed connectivity among populations, leading to the establishment of networks of marine protected areas. Understanding this connectivity among populations, which entails evaluation of population structure across taxa, is a key factor for the effective design of these networks and preserving biodiversity on a large scale. This knowledge is fundamental for ensuring the long-term survival of ecosystems inhabited by these species.

Population structure is also an important confounding variable in GWAS due to the possibility of inaccurate associations between genotype and the trait of interest in a genetic study. The presence of population structure may cause false positive or negative associations between genotype and trait due to differences in local ancestry that are not related to disease risk or trait variance. Thus identification of population structure, and controlling for it, removes the confounding factors. This step enables GWAS to find new genetic associations and improve the detection, treatment, and prevention of certain diseases.

There are three types of population inference approaches: model-based, distance-based, and statistical (sometimes referred to as algorithmic). An example of a model-based population inference approach is Structured Association, which assigns samples to subpopulation clusters (possibly allowing fractional cluster membership) using a model-based clustering program such as structure. However, the applicability of this approach to large genome-wide data sets is limited by its high computational cost when allowing fractional cluster membership. Faster model-based approaches, such as admixture, fastStructure, and frappe adopt the likelihood model embedded in structure but incorporate relaxation methods for improving computational efficiency.

However, because these approaches are based on genetic assumptions about the data, including Hardy-Weinberg equilibrium (HWE) within populations and linkage equilibrium (LE) between loci, violating these assumptions may lead to misleading results. Marker deviation from equilibrium can also signify a
possible sequencing \cite{28-31} or genotyping error and thus such markers should be excluded from further analysis \cite{32}. Incorrect inference of genotypes are known to occur due to low coverage of DNA sequencing \cite{33}. The assumptions of Hardy-Weinberg proportions, which must be met for the marker to be included in the analysis, sometimes are not met due to genotyping errors \cite{34}. Markers that do not meet this requirement should not be included in the analysis \cite{32}.

Thus, marker genotypes of SNPs or microsatellites, or haplotype frequencies generated from the sequence data, require careful data preprocessing steps \cite{32}. Additionally, before running these methods, the number of populations (K) must be set but may not be known in advance. While model-based approaches are very powerful in population structure identification, they are thus limited by computational cost, operate on genetic assumptions that must be held, and are sensitive to sample size \cite{35}.

Alternative distance-based population inference approaches adopt a pairwise distance matrix computed among each pair of individuals. Some examples of implemented distance-based approaches are Genetic Similarity Score Matching (GSM) \cite{36}, Spectral-GEM \cite{37}, and FastPop \cite{38}. GSM and Spectral-GEM require high computational intensity when the sample size is large. FastPop results in complex computation and has not been established when inferring genetic ancestry among more than 4 populations \cite{39}.

Statistical approaches, such as PCA, can be applied to genotyped data (individual allele frequencies, single-nucleotide polymorphism (SNPs)) to extract linear combinations of individuals that share the greatest similarities. A graphical overview (scatter plot) of the population structure can be shown using principal components as axes of variations. PCA is efficient and has been implemented for ancestry inference in eigenstrat \cite{23,62} and smartpca \cite{62}. Statistical approaches are able to handle large-scale genomic datasets and are not restricted by genetic assumptions \cite{32}.

PCA has advantages over the model-based approaches in that it is a non-parametric method (it does not require a predefined number of populations) and does not rely on modeling assumptions (HWE, LE). PCA is also computationally efficient. Albeit, current PCA-based approaches, just as model-based approaches, operate on genomic markers, which require careful identification to be useful for population structure analysis.

Overall, existing methods to identify population structure are computationally intensive. The number of available markers grows as the number of samples included in the analysis increases, thus reducing the efficiency of computation. Identification of these genotypes requires rigorous steps, and reducing the number of informative markers is often desirable for efficient population structure determination \cite{32,40}. Ancestry informative markers are usually determined as a set of minimum markers needed to determine the population structure and lower the genotyping cost. Selection of informative markers using the supervised method relies on self-reported ancestry information from individuals, while the unsupervised
approach applies PCA to determine markers that are associated with the significant principal components and then score each marker 41.

In this work, we investigate our ability to determine population structure with PCA using frequencies of k-mers present in a genome. Information on the presence and absence of k-mers has shown promising results in population differentiation whole-genome sequencing reads from two distinct superpopulations 42. K-mers are shorter substrings that can be “overlapped” to reconstruct the full sequence and deliver equivalent genomic information as a whole sequence 43. K-mers or k-mer profiles of a sequence (k-mer and its frequency in the genome) can be generated efficiently 44. Then the structure of the genome can be investigated directly from the k-mer profiles of the genome. The problem of sequencing or genotyping errors may be reduced because PCA aggregates the k-mer frequency information; therefore extra counts of frequencies that could potentially accrue due to sequencing errors should not substantially influence the PCA projection. Additionally, sequencer errors can be identified and removed by filtering out k-mers of frequency one (singletons), which are generally considered a result of sequencer errors 45,46, and not to be included in the final PCA computation.

In this study, we examined the ability to determine population structure based on k-mer frequencies present in a genome, with the goal of formulating a quick and accurate approach for population structure identification. We used samples from across human populations from the 1000 Human Genome Project 47. The structure of these populations has been established previously and includes both the separation of the five superpopulations, independent populations within these, and populations with mixed ancestry.

We were able to apply PCA to frequencies of k-mers present in genomes to accurately separate superpopulations and populations, and identify admixed individuals. For comparison, we investigated population stratification based on the number of k-mer matches between pairs of genomes using a popular alignment-free sequence comparison tool, mash 48. We were able to build accurate population trees using this approach; however, the results depended on the parameter selection and it was difficult to identify a priori the k value (k-mer length) and sketch (reduced representation of a sequence) size needed for accurate results. Thus, the practicality of this approach is limited compared to our PCAs of k-mer frequencies.

Results

Population structure of five human superpopulations shows two major groups

To differentiate samples from the five human superpopulations identified by the 1000 Genomes Project 35 we used principal component analysis with a K-Means clustering algorithm. 80% of the variance was captured by 21 PCs. The first two PCs contained 14.5% of the variance.
Using 21 PCs, which hold 80% of the variance, caused undecidability in the K-Means clustering algorithm where no obvious inflection (elbow) point was observed (Fig. 1a). However, using the first 2 PCs, which hold 14.5% of the variance, showed deterministic results in K-Means clustering (Fig. 1b). Thus, we applied K-Means clustering to the first 2 PCs throughout this study. The African superpopulation was strongly differentiated from all other samples along PC1. The “elbow point” of the K-Means scree plot (when the change in the value of inertia is no longer significant) indicated the samples grouped into three clusters. There was a strong differentiation of samples of African Ancestry (AFR_LWK) from all other populations on PC1, and differentiation within this population along PC2 (Fig. 2).

Population structure from four human superpopulations shows four groups

To determine whether we could differentiate the non-African populations we repeated the PCA analysis excluding AFR_LWK samples. 80% of the variance was captured by 18 PCs. The first two PCs contained 13% of the variance. Plotting the first two PCs, we observe four distinct groups corresponding to the four superpopulations (Fig. 3). The elbow point of the K-Means scree plot indicated the samples grouped into four clusters (Supplementary material Appendix 1, Figure A1).

Population structure from four human superpopulations including samples of admixed origin

To differentiate the samples of admixed origin we repeated the PCA analysis with additional samples of admixed origin from EAS and EUR superpopulations. 80% of the variance was captured by 36 PCs. The first two PCs contained 8.4% of the variance. Plotting the first two PCs, we observe four distinct groups corresponding to the four superpopulations; individual populations were not clearly differentiated (Fig. 4). The “elbow point” of the K-Means scree plot indicated the samples grouped into four clusters (Supplementary material Appendix 1, Figure A2).

Population structure from three populations of admixed and non-admixed origin

To differentiate the samples of admixed origin we repeated the PCA analysis with samples from the East Asian superpopulation including one population of admixed origin. 80% of the variance was captured by 14 PCs. The first two PCs contained 13% of the variance. Plotting the first two PCs, we observe three distinct groups corresponding to the three populations CDX, CHB, and JPT (superpopulation EAS), with the CHB population placed in between CDX and JPT populations. The “elbow point” of the K-Means scree plot was not prominent and indicated the samples of admixed origin formed a continuous grouping (Supplementary material Appendix 1, Figure A3). However, we observed a strong differentiation of samples by population on PC1 (Fig. 5).
Additionally, we repeated the PCA analysis to differentiate samples of admixed origin with samples from the European superpopulation including one population of admixed origin. 80% of the variance was captured by 14 PCs. The first two PCs contained 13% of the variance. Plotting the first two PCs, the three populations CEU, FIN, and TSI (superpopulation EUR) appear visually differentiated along PC 1, with CEU (admixed) placed in between FIN and TSI (Fig. 6). However, the “elbow point” of the K-Means scree plot was not prominent, indicating a lack of strong grouping of samples (Supplementary material Appendix 1, Figure A4). While the scatter plot of the PCA provides a graphical overview of the population structure along the axes of variations of the first 2 PCs, K-Means clusters individuals into genetically homogeneous subpopulations by placing each observation to the cluster with the nearest mean. Consequently, there is a visual discordance between sample populations and K-Means assignments.

Accuracy of K-Means clustering

The AMI score for K-Means clusters and the expected clusters from the analysis of the five superpopulations of non-admixed origin was 0.35 but improved to 1.0 when excluding AFR. The AMI for K-Means clusters and the expected clusters of the four superpopulations including samples of admixed and non-admixed origin was 1.0. When analyzing individual superpopulations including samples of admixed and non-admixed origin to differentiate populations, the AMI was 0.83 for the populations in EAS, and 0.43 for the populations in EUR. For the explained variance plots and for the cumulative explained variance plots see Supplementary material (Appendix 1, Figure A5 - A14).

Memory usage when analyzing population structure with k-mer frequencies

These analyses (using k-mer length of 21) stored a dictionary data structure with k-mer content for each sequence. The dictionary of 48 vectors of k-mers and their frequencies occupied 41Gb of space in a pickled (compressed) format. Reduction of space to 38Gb was possible by calculating the intersection of k-mers across all vectors. Generating these data structures required an HPC node with at least 250Gb of RAM for the dictionary compression step. Additionally, calculating PCA on this dataset required ~ 239Gb of RAM, and took 2 hrs 36 minutes and 51 seconds of job wall-clock time on a 36-core HPC node.

Population structure from k-mer presence alone

For comparison with our PCA approach, we used mash to estimate population structure from the same samples. Monophyly of the superpopulation groups was observed on the unrooted tree for various parameters of k-mer length (k) and sketch size (s) (Supplementary material Appendix 1, Figure A15–18). Specifically, the trend of accurate grouping by population was observed with shorter k-mer length and higher sketch size, and conversely lower sketch size and longer k-mer length produced trees with monophyletic groupings of samples by population (Fig. 7).

While we saw a general trend of improved accuracy with low k-mer length and higher sketch size, and conversely longer k-mers and smaller sketch size, we saw exceptions in the accuracy trend for various
parameters. For example, k-mer length 24 and sketch size 3000 produced accurate results, however, the accuracy dropped with setting $k = 24$ and $s = 5000$ and 7000, and the accuracy picked up again when $k = 24$ and $s = 8000–30000$. While this test case allowed us to compare results to the ground truth, it is difficult to select \textit{a priori} the parameters that produce accurate results.

**Discussion**

In this work, we showed that population structure can be accurately detected from k-mer frequencies using PCA, including accurate representation of samples from admixed origin. Application of PCA to vectors of k-mer frequencies reduces the dimensionality of the data, which makes the dataset manageable for the following step of applying a clustering algorithm to detect structure in the dataset. We suggest that k-mer frequencies may be easier to calculate compared to accurate genotypes and allow the estimation of population structure hierarchically.

**Population identification**

We hypothesize that our initial observation of two larger clusters in the data from the five human superpopulations (Africa and all others) is likely due to individuals from African populations possessing the greatest genetic variation, as predicted by the out-of-Africa model of human origins\textsuperscript{35}. Thus, the initial dominating force in the signal differentiates this superpopulation in a way that is accurate although not described by the 1000 Genomes Project. Excluding this superpopulation, PCA groups the samples by superpopulation. Again the majority of variation is among superpopulations, leading us to take a hierarchical approach to examine clusters of samples individually for substructure (i.e. separation of individual populations).

PCA analysis of individual superpopulations successfully differentiates populations. For admixed populations, PCA places samples in between the two populations of non-admixed origin. The continuum in the placement of samples of admixed origin likely corresponds to the fractional membership of samples as determined by the 1000 Human Genome Project\textsuperscript{35}. More specifically, if there are two established populations, then samples of admixed origin are placed on the coordinates along edges joining the centers of the established populations as described by\textsuperscript{50}. While it is not surprising that these samples are not accurately assigned to a third population, as this same result was found by the 1000 Human Genome Project, when samples were grouped into three populations, there did appear to be a third intermediate population, which was largely comprised of the expected admixed samples.

**K-mers versus marker genotypes**

Our analysis of k-mer frequencies using PCA produced comparable results to those produced by the model-based approach using marker genotypes\textsuperscript{35}. However, k-mer frequencies can be simpler to count and errors may be more readily identified. When analyzing marker genotype data, a number of preprocessing steps to evaluate data quality are necessary\textsuperscript{21,27}, as population structure estimation using
genetic markers, is susceptible to genotyping errors. This evaluation includes an assessment of SNP call rates, MAFs, verification of the HWE assumptions, and relatedness between individuals. Additionally, the identification of ancestry informative markers, which constitute a minimal number of markers needed to obtain population structure, is necessary to ensure the accuracy of the results. In contrast, k-mer frequencies can be viewed as summary statistics of a genome resulting from SNPs. While errors in k-mer counts occur due to sequencing or genotyping errors, these can easily be identified as k-mers of low frequency. The overwhelming majority of k-mers of frequency 1 are not found in a genome and thus are most likely due to sequencing errors and therefore can easily be discarded.

Model-based population analysis methods can also be limited by inadequate sample sizes and the number of markers analyzed. These methods depend on an estimation of allele frequency that is sensitive to small samples. In contrast, our PCA-based k-mer frequency approach does not depend on allele frequency estimation and thus is not affected by sample size. We were able to produce accurate results with a number of samples as low as six per population.

The PCA-based approach also has the advantage of operating without a preset number of populations and no modeling assumption requirements. This makes the analysis of population structure using a non-model-based approach an appealing choice. The application of PCA and K-Means to k-mer profiles of genomes makes it easy to detect a number of populations (clusters) present in the dataset, which is a major parameter in the model-based method that is required to be set in advance.

However, the two approaches deliver different types of information. While a more complex but sophisticated method such as structure describes the population structure by probabilistic assignment to classes, the PCA combined with K-Means, provides a graphical and quantitative representation of population structure along axes of variation in the dataset. Thus, the goal of the investigation of population structure should be taken into consideration when deciding which approach is more applicable for the analysis.

Consideration of computational efficiency

The PCA-based approach is computationally efficient and can handle genomic marker data for thousands of individuals. PCA was also shown to be efficient when applied to k-mer frequencies in our work. The major computational cost comes with the data preprocessing step when analyzing either marker genotypes or k-mer frequencies. The preprocessing of genetic data can consume computational time. On the other hand, when analyzing k-mer frequencies, the memory requirements needed to store k-mer profiles grow with k-mer length, as does the computational power needed to calculate distances between sequences based on k-mer frequencies. However, shorter k-mers are less informative while longer k-mers are more unique. Thus, when analyzing k-mers, there is a tradeoff between specificity, computational efficiency, and memory requirements.

Identification of population structure based on the number of shared k-mers
mash’s identification of all five superpopulations, suggests that mash has an adequate amount of sensitivity to differentiate samples by superpopulation even with the presence of samples with greater variation (AFR). However, grouping individuals into populations based on k-mer presence using mash distances was more difficult than using PCA and k-mer frequencies. mash produced accurate groupings (samples were placed on the tree according to the superpopulation) only for particular combinations of k-mer length and sketch size, and accuracy was not necessarily predictable. We were able to identify the k-mer length that produced accurate results by checking with the results produced by the 1000 Genome Project; however, finding the right parameter for k-mer length would be hard without knowing the correct population structure in advance. Thus, while mash is a robust tool for identifying genome relationships on a species level \cite{51}, on a population level, in comparison to the PCA approach, mash showed less viable performance for differentiating populations due to its high sensitivity to parameter selection which is unknown in advance.

Conclusion

In sum, principal component analysis together with K-Means clustering appears to successfully identify population structure based on the k-mer frequencies present in genomes. This approach is robust in differentiating samples at the superpopulation and population levels. Notably this approach differentiated samples hierarchically, and PCA was able to discern the population signal in samples of admixed and non-admixed origin within a single superpopulation. These results are comparable to model-based approaches that identify populations using genotypes, and which provide information on fractional membership of a sample to a population. However, using k-mer frequencies does not depend on genetic assumptions or the process of marker selection curation. In contrast, the method using k-mer presence to group samples lacked sensitivity to consistently identify populations. With the increasing availability of whole-genome data, we anticipate that the use of k-mer frequencies combined with PCA and k-means clustering will be a strong addition to population structure investigations.

Methods

Differentiating human superpopulations

To examine the use of k-mer-based methods in identifying population structure we first obtained genome data for humans from the International Genome Sample Resource (IGSR) \cite{62}. These data have been used to identify human population structure \cite{35}, thus providing a comparison for the population structure we identify with our alternative approaches. This dataset contains genome data for individuals classified by superpopulation as Africa (AFR), East Asia (EAS), Europe (EUR), South Asia (SAS), and the Americas (AMR). Each superpopulation contains multiple populations. For our initial analysis, we analyzed one population per superpopulation. We downloaded sequenced reads that were aligned to the GRCh38 human reference sequence from 6 samples from a single population per superpopulation identified as having the least admixed origin (referred to here as “non-admixed”) \cite{62}: 

Page 10/22
1. Luhya population in Webuye, Kenya (LWK) of African Ancestry
2. Peruvian population in Lima, Peru (PEL) of American Ancestry
3. Toscani population in Italy (TSI) of European Ancestry
4. Indian Telugu population in the UK (ITU) of South Asian Ancestry
5. The Japanese population in Tokyo, Japan (JPT) of East Asian Ancestry

All samples were sequenced using PCR-free high coverage technology and listed under the “1000 Genomes 30x on GRCh38” data collection. Data were accessed as cram files. Each cram file was converted into a bam file using samtools. We then used the bcftools mpileup command to filter out regions with low-quality scores, call the variants, and perform pileups. Finally, the fasta files were built using the bcftools consensus command.

**Population structure from k-mer frequencies using PCA**

For each sample, we built profiles of canonical k-mers (k-mer or its reverse complement, whichever comes first lexicographically) from the fasta file using Jellyfish, a tool for fast, memory-efficient counting of k-mers in the DNA sequence, for k-mer length 21. When choosing the length of k-mers we were guided by a general rule used by the alignment-free methods for sequence comparisons - shorter k-mers are more likely to be present in a sequence (e.g. 1-mers); thus they are less informative in analyzing closely related genomes; however, longer k-mers are more unique to particular species and are therefore more useful for similarity identification across species. Moreover, widely used alignment-free sequence comparison tools such as mash found a k-mer length of 21 to give accurate estimates of sequence similarities. We filtered out singletons to account for possible errors produced by the sequencer. We then sorted each k-mer profile in alphabetical order and calculated the intersection of k-mer profiles across all samples. Additionally, sequencer errors were identified and removed by filtering out k-mers of frequency one. For this purpose, the flag -L 2 is used when analyzing the k-mer content of the sequence with Jellyfish (applicable to our method) to only include k-mer frequencies above 1.

We performed a Principal Component Analysis (PCA) on vectors of k-mer frequencies of each sample in python using the scikit-learn (sklearn) library. We normalized the data (frequencies of k-mers) by scaling all the values to be between 0 and 1 using the StandardScaler function from the scikit-learn library. Normalization of the dataset is a necessary step due to PCA calculating a new projection of the dataset with new axes based on the standard deviation of the variables. Variables with different standard deviations (high versus low) will have different weights for axes calculation. Normalization of data allows for uniform standard deviation across all variables, thus PCA calculates axes with all variables having equal weight. We visualized the projection of the two PCs with the most variance using a scatter plot from the matplotlib package.

To identify populations, we used K-Means clustering based on the PCs. We found the optimal number of principal components that capture the greatest amount of variance in the data by plotting the explained variances in a scree plot. We used explained variance ratio as a metric to evaluate the
usefulness of the principal components and to choose how many components to use in the model. The explained variance ratio is the percentage of variance that is attributed by each of the selected components. To avoid overfitting the model we chose the number of components to include in the model by adding the explained variance ratio of each component until we reach a total of around 80%, which is considered an adequate amount of variance to derive informative results. We used these PCs to cluster samples into different numbers of groups (k from 1–10). We determined the optimal number of clusters (K) by using the “elbow method” heuristic approach. The sum of the squared distances to the nearest cluster center (aka inertia) was measured using the K-Means model for each k. From the scree plot, we determined the “elbow point”, i.e. the point after which the inertia starts decreasing in a linear fashion. We clustered the dataset by fitting the numbers of PCs with 80% of the variance into the K-Means model with k number of clusters.

Because the PCs with 80% of the variance caused undecidability in the K-Means clustering algorithm, we hypothesized that the dataset contains noise that suppresses the true biological signal, and only a small fraction of k-mers “drive” (dominate) the data. From the scree plots of PC variance and K-Means inertia plots, we saw that the variance is spread out, mostly equally, through all the PCs. Commonly, the first three PCs contain most of the variance of the data (around 80%), however, our dataset shows that the first three PCs have only a slightly higher variance than the rest of the PCs. Nevertheless, when using the first two PCs we saw deterministic results by K-Means and they matched the expected results. Thus we used the first two PCs throughout this work.

**Human population structure and substructure**

The analysis of the dataset including all five superpopulations showed strong differentiation of the AFR superpopulation from the rest of the superpopulations. Thus, we repeated the analysis for four superpopulations excluding AFR, with 24 samples of non-admixed origin (AMR_PEL, EAS_JPT, EUR_TSI, and SAS_ITU).

**Differentiating human populations with k-mer frequencies using PCA**

Because we were able to differentiate human superpopulations, we examined whether this approach could differentiate populations within those superpopulations. We obtained an additional 12 samples (6 per population) of European Ancestry and East Asian Ancestry of non-admixed origin:

1. Finnish population in Finland (FIN) of European Ancestry
2. Dai Chinese population in Xishuangbanna, China (CDX) of East Asian Ancestry

Genomes for each individual were obtained as described above. We repeated our analysis with these six populations from four superpopulations (excluding AFR).

**Admixed populations**
To determine whether we could identify populations of admixed origin, we selected samples from populations that were previously identified as comprising roughly equal parts of fractional membership of two non-admixed populations from the same superpopulation (we refer to these as “admixed” populations). We obtained six samples from the Han Chinese population in Beijing China (CHB) of East Asian ancestry that appears to be a mixture of JPT and CDX ancestry, and another six samples from Utah residents (CEU) of Northern and Western European ancestry that appear to be a mixture of FIN and TSI ancestry. Genomes for each individual were obtained as described above. We repeated our analysis with these eight populations from four superpopulations (AMR_PEL, EAS_CDX, EAS_CHB, EAS_JPT, EUR_CEU, EUR_FIN, EUR_TSI, and SAS_ITU; excluding AFR).

Because we continued to differentiate the four superpopulations, we focused a subsequent analysis on the single EAS superpopulation with 18 samples (12 of non-admixed and 6 of admixed origin; EAS_CDX, EAS_CHB, and EAS_JPT) to determine our ability to identify admixture. We repeated our PCA analysis on these samples alone. Additionally, we repeated the analysis on the single EUR superpopulation with 18 samples (12 of non-admixed and 6 of admixed origin; EUR_CEU, EUR_FIN, EUR_TSI).

**Population structure from a number of shared k-mers between sequence pairs**

For comparison with our kmer frequency-based PCA approach, we examined population identifiability using *mash*. *mash* can be used to build phylogenies for family-level data and shows promise for population genetic analyses of polyploid sequences. The principle behind *mash* is that each sequence is converted into a MinHash sketch, a vastly reduced representation of a sequence, then two sketches are compared by calculating the fraction of shared k-mers between a pair of sequences (Jaccard index). Finally, the *mash distance* is calculated, which estimates the rate of sequence mutation under a simple evolutionary model. *mash* has been investigated for basic population genetic analyses of polyploid and diploid species and showed some promising results in the population stratification of plants.

For each sample we built *mash* sketches using a *mash* sketch command with -m 2 flag to filter out single k-mers, -k N flag to analyze the k-mer length of N, and -s M flag to build sketches of size M. We repeated the process for parameters of -k N = 21, 24, 27, 29, 32 and -s M = 1000, 3000, 5000, 7000, 8000, 9000, 10000, 12000, 15000, 18000, 20000, 23000, 25000, 28000, 30000 to compare the results for a different set of parameters. We calculated pairwise distances between each pair of samples using *mash* and used the distances to build a Neighbor-Joining tree (NJ) for each set of parameters - k-mer length and sketch size. *mash* does not assign samples by population thus to verify that grouping by superpopulation we checked for monophyly of each of the groups in the NJ tree by superpopulation label. We used the *is.monophyletic* function in the R package to check whether each population was monophyletic in the resulting tree and *ggplot* to plot the Boolean values for whether the tree with all the populations contained clades that are all monophyletic.

**Declarations**
Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All the code used in this work is available in the GitHub repository https://github.com/YanaHrytsenko/determining_population_structure_from_k-mer_frequencies and can be used freely for reproducibility purposes.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was funded by a grant to R. Schwartz from the National Science Foundation (DBI-1942273) and by the USDA National Institute of Food and Agriculture, Hatch project accession no. 1017848.

Authors' contributions

YH implemented the code, performed all the experiments, and wrote the manuscript. ND and RS guided the study. All authors contributed to conceiving the study, provided contributions to the written manuscript, and read and approved the final manuscript.

Acknowledgments

We thank members of the Schwartz lab for valuable comments and suggestions. The High-Performance Research Computing (HPC) facility at the University of Rhode Island is acknowledged for providing the computational resources for the analyses in this manuscript. Additionally, we thank Kevin Bryan, the HPC systems manager, for invaluable help with software installation on the cluster and GPU operations management.

References
1. Bradburd, G. S., Ralph, P. L. & Coop, G. M. A Spatial Framework for Understanding Population Structure and Admixture. PLoS Genet. 12, e1005703 (2016).

2. Andam, C. P., Challagundla, L., Azarian, T., Hanage, W. P. & Robinson, D. A. Population Structure of Pathogenic Bacteria. Genetics and Evolution of Infectious Diseases 51–70 (2017) doi:10.1016/b978-0-12-799942-5.00003-2.

3. Clark, A. G. Population Genetics. Encyclopedia of Genetics 1513–1519 (2001) doi:10.1006/rwgn.2001.1016.

4. Novembre, J. et al. Genes mirror geography within Europe. Nature 456, 98–101 (2008).

5. Smock, P. J. & Schwartz, C. R. The Demography of Families: A Review of Patterns and Change. J. Marriage Fam. 82, 9–34 (2020).

6. Elliot A. Schulman, F., Levin, M., Alvin E. Lake, I. I. I. & Elizabeth Loder, M. P. H. Refractory Migraine: Mechanisms and Management. (Oxford University Press, 2010).

7. Mills, M. C. & Rahal, C. at 75 years: An empirical review. Popul. Stud. 75, 7–25 (2021).

8. Choudhuri, S. Bioinformatics for Beginners: Genes, Genomes, Molecular Evolution, Databases and Analytical Tools. (Elsevier, 2014).

9. Boca, S. M., Huang, L. & Rosenberg, N. A. On the heterozygosity of an admixed population. J. Math. Biol. 81, 1217–1250 (2020).

10. Ellstrand, N. C. & Rieseberg, L. H. When gene flow really matters: gene flow in applied evolutionary biology. Evol. Appl. 9, 833–836 (2016).

11. Pascual, M. Population structure in marine organisms: from genetics to genomics. Frontiers in Marine Science vol. 6 (2019).

12. Shen, Y. et al. Population structure, demographic history and local adaptation of the grass carp. BMC Genomics 20, 467 (2019).

13. Zhao, L. et al. Population genomics provides insights into the population structure and temperature-driven adaptation of Collichthys lucidus. BMC Genomics 22, 729 (2021).

14. Bayless, T. M., Brown, E. & Paige, D. M. Lactase Non-persistence and Lactose Intolerance. Curr. Gastroenterol. Rep. 19, 23 (2017).

15. Hellwege, J. N. et al. Population Stratification in Genetic Association Studies. Curr. Protoc. Hum. Genet. 95, 1.22.1–1.22.23 (2017).

16. Li, M., Reilly, M. P., Rader, D. J. & Wang, L.-S. Correcting population stratification in genetic association studies using a phylogenetic approach. Bioinformatics 26, 798–806 (2010).

17. Friedrich, D. C., Santos, S. E. B., Ribeiro-dos-Santos, Â. K. C. & Hutz, M. H. Several different lactase persistence associated alleles and high diversity of the lactase gene in the admixed Brazilian population. PLoS One 7, e46520 (2012).

18. Ingram, C. J. E. et al. A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? Hum. Genet. 120, 779–788 (2007).
19. Mulcare, C. A. et al. The T Allele of a Single-Nucleotide Polymorphism 13.9 kb Upstream of the Lactase Gene (LCT) (C − 13.9kbT) Does Not Predict or Cause the Lactase-Persistence Phenotype in Africans. The American Journal of Human Genetics vol. 74 1102–1110 (2004).

20. Tishkoff, S. A. et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. 39, 31–40 (2007).

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

22. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. Genetics 155, 945–959 (2000).

23. Rosenberg, N. A. et al. Genetic structure of human populations. Science 298, 2381–2385 (2002).

24. Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet. 11, 459–463 (2010).

25. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197, 573–589 (2014).

26. Tang, H., Peng, J., Wang, P. & Risch, N. J. Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. 28, 289–301 (2005).

27. Liu, N. & Zhao, H. A non-parametric approach to population structure inference using multilocus genotypes. Human Genomics vol. 2 353 (2006).

28. Glenn, T. C. Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. 11, 759–769 (2011).

29. Goodwin, S., McPherson, J. D. & Richard McCombie, W. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics vol. 17 333–351 (2016).

30. Mardis, E. R. Next-generation sequencing platforms. Annu. Rev. Anal. Chem. 6, 287–303 (2013).

31. Salk, J. J., Schmitt, M. W. & Loeb, L. A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat. Rev. Genet. 19, 269–285 (2018).

32. Alhusain, L. & Hafez, A. M. Nonparametric approaches for population structure analysis. Hum. Genomics 12, 25 (2018).

33. Lachance, J. Hardy–Weinberg Equilibrium and Random Mating. Encyclopedia of Evolutionary Biology 208–211 (2016) doi:10.1016/b978-0-12-800049-6.00022-6.

34. Laurie, C. C. et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genet. Epidemiol. 34, 591–602 (2010).

35. Gao, X. & Starmer, J. D. AWclust: point-and-click software for non-parametric population structure analysis. BMC Bioinformatics 9, 77 (2008).

36. Guan, W., Liang, L., Boehnke, M. & Abecasis, G. R. Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. Genet. Epidemiol. 33, 508–517 (2009).

37. Lee, A. B., Luca, D., Klei, L., Devlin, B. & Roeder, K. Discovering genetic ancestry using spectral graph theory. Genet. Epidemiol. 34, 51–59 (2010).
38. Li, Y. et al. FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data. BMC Bioinformatics vol. 17 (2016).
39. Byun, J. et al. Ancestry inference using principal component analysis and spatial analysis: a distance-based analysis to account for population substructure. BMC Genomics 18, 789 (2017).
40. Stevens, E. L. et al. Inference of relationships in population data using identity-by-descent and identity-by-state. PLoS Genet. 7, e1002287 (2011).
41. Paschou, P. et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 3, 1672–1686 (2007).
42. Rahman, A., Hallgrímsdóttir, I., Eisen, M. & Pachter, L. Association mapping from sequencing reads using k-mers. eLife vol. 7 (2018).
43. Compeau, P. E. C., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. Nature Biotechnology vol. 29 987–991 (2011).
44. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764–770 (2011).
45. Shendure, J. & Ji, H. Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135–1145 (2008).
46. Wang, T. T. et al. High efficiency error suppression for accurate detection of low-frequency variants. Nucleic Acids Res. 47, e87 (2019).
47. 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature 526, 68–74 (2015).
48. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 132 (2016).
49. Fränti, P. & Sieranoja, S. K-means properties on six clustering benchmark datasets. Applied Intelligence vol. 48 4743–4759 (2018).
50. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. PLoS Genet. 2, e190 (2006).
51. Zielezinski, A., Vinga, S., Almeida, J. & Karłowski, W. M. Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol. 18, 186 (2017).
52. Prakash, A. & Tompa, M. Measuring the accuracy of genome-size multiple alignments. Genome Biol. 8, R124 (2007).
53. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5, e11147 (2010).
54. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14, 1394–1403 (2004).
55. Lee, C., Abdool, A. & Huang, C.-H. PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics 10 Suppl 1, S73 (2009).
56. Pompanon, F., Bonin, A., Bellemain, E. & Taberlet, P. Genotyping errors: causes, consequences and solutions. Nat. Rev. Genet. 6, 847–859 (2005).
57. Erbert, M., Rechner, S. & Müller-Hannemann, M. Gerbil: a fast and memory-efficient -mer counter with GPU-support. Algorithms Mol. Biol. **12**, 9 (2017).
58. Kokot, M., Długosz, M. & Deorowicz, S. KMC 3: counting and manipulating k-mer statistics. Bioinformatics vol. 33 2759–2761 (2017).
59. Melsted, P. & Pritchard, J. K. Efficient counting of k-mers in DNA sequences using a bloom filter. BMC Bioinformatics **12**, 333 (2011).
60. Lawson, D. J., van Dorp, L. & Falush, D. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nat. Commun. **9**, 3258 (2018).
61. Porras-Hurtado, L. *et al.* An overview of STRUCTURE: applications, parameter settings, and supporting software. Front. Genet. **4**, 98 (2013).
62. Clarke, L. *et al.* The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. Nucleic Acids Res. **45**, D854–D859 (2017).
63. Bernard, G., Chan, C. X. & Ragan, M. A. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. Sci. Rep. **6**, 28970 (2016).
64. Greenfield, P. & Roehm, U. Answering biological questions by querying k-mer databases. Concurrency and Computation: Practice and Experience vol. 25 497–509 (2013).
65. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *arXiv [cs.LG]* (2012).
66. Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. Philos. Trans. A Math. Phys. Eng. Sci. **374**, 20150202 (2016).
67. Hunter. Matplotlib: A 2D Graphics Environment. **9**, 90–95 (2007).
68. Ding, C. & He, X. K-means clustering via principal component analysis. *Twenty-first international conference on Machine learning - ICML ’04* (2004) doi:10.1145/1015330.1015408.
69. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. **11**, 94 (2010).
70. Yuan, C. & Yang, H. Research on K-Value Selection Method of K-Means Clustering Algorithm. Jvol. 2 226–235 (2019).
71. VanWallendael, A. & Alvarez, M. Alignment-free methods for polyploid genomes: Quick and reliable genetic distance estimation. Mol. Ecol. Resour. (2021) doi:10.1111/1755-0998.13499.
72. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
73. Wilkinson, L. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. *Biometrics* vol. 67 678–679 (2011).

**Figures**
Figure 1

a) Scree plot showing the non-deterministic number of clusters (no “elbow point”) determined by K-Means using 21 PCs (80% of the variance) from k-mer frequencies of five superpopulations.  
b) Scree plot showing a deterministic number of clusters = 3 (“elbow point”) determined by K-Means using 2 PCs (14.5% of the variance) from k-mer frequencies of five superpopulations.

Figure 2

PCA generated using k-mer frequencies from a single population from each of five superpopulations using 2PCs. Samples are labeled by population. K-Means algorithm identified three clusters (circled) present in the data: two in Africa (AFR) and one including all other populations, Americas (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS).

Figure 3

PCA generated using k-mer frequencies from four superpopulations of non-admixed origin (America (AMR), East Asia (EAS), Europe (EUR), South Asia (SAS), but excluding Africa (AFR)) using 2PCs. Samples are labeled by population (colors). K-Means algorithm identified four clusters present in the data (circled).
Figure 4

PCA generated using k-mer frequencies from four superpopulations (America (AMR), East Asia (EAS), Europe (EUR), South Asia (SAS)) including samples of admixed and non-admixed origin in EAS and EUR using 2PCs. Samples are labeled by population. K-Means algorithm identified four clusters present in the data (circled).

Figure 5

PCA generated using k-mer frequencies from the EAS superpopulation including samples of admixed and non-admixed origin (CDX, CHB, and JPT) using 2PCs. Samples are labeled by population. K-Means algorithm identified three clusters present in the data (circled) corresponding closely to the expected populations.

Figure 6

PCA generated using k-mer frequencies from the EUR superpopulation including samples of admixed and non-admixed origin (CEU, FIN, and TSI) using 2PCs. Samples are labeled by population. K-Means algorithm identified three clusters present in the data (circles). While populations appear to separate along PC 1, K-Means clustering differentiates the two non-admixed populations (FIN and TSI) but groups finds overlap between CEU and TSI, as well as between CEU and FIN.
Figure 7

Heatmap plot for the presence of trees with all superpopulations being monophyletic built from pairwise mash distances for different k-mer length and sketch size parameters.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.
