

# Archetypal Analysis for Population Genetics

Julia Gimbernat-Mayol<sup>1</sup>, Daniel Mas Montserrat<sup>2</sup>, Carlos D. Bustamante<sup>2</sup>, Alexander G. Ioannidis<sup>2,3,\*</sup>

**1** Department of Biomedical Engineering, Imperial College London, London, United Kingdom

**2** Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

**3** Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA, USA

\*correspondence: ioannidis@stanford.edu

## Abstract

The estimation of genetic clusters using genomic data has application from genome-wide association studies (GWAS) to demographic history to polygenic risk scores (PRS) and is expected to play an important role in the analyses of increasingly diverse, large-scale cohorts. However, existing methods are computationally-intensive, prohibitively so in the case of nationwide biobanks. Here we explore Archetypal Analysis as an efficient, unsupervised approach for identifying genetic clusters and for associating individuals with them. Such unsupervised approaches help avoid conflating socially constructed ethnic labels with genetic clusters by eliminating the need for exogenous training labels. We show that Archetypal Analysis yields similar cluster structure to existing unsupervised methods such as ADMIXTURE and provides interpretative advantages. More importantly, we show that since Archetypal Analysis can be used with lower-dimensional representations of genetic data, significant reductions in computational time and memory requirements are possible. When Archetypal Analysis is run in this fashion, it takes several orders of magnitude less compute time than the current standard, ADMIXTURE. Finally, we demonstrate uses ranging across datasets from humans to canids.

## Author summary

This work introduces a method that combines the singular value decomposition (SVD) with Archetypal Analysis to perform fast and accurate genetic clustering by first reducing the dimensionality of the space of genomic sequences. Each sequence is described as a convex combination (admixture) of archetypes (cluster representatives) in the reduced dimensional space. We compare this interpretable approach to the widely used genetic clustering algorithm, ADMIXTURE, and show that, without significant degradation in performance, Archetypal Analysis outperforms, offering shorter run times and representational advantages. We include theoretical, qualitative, and quantitative comparisons between both methods.

## Introduction

Estimating ancestry cluster allele frequencies and cluster membership from single nucleotide polymorphism (SNP) data is important for many applications in population

genetics and applying methods to characterize diverse human cohorts has become an essential part of large-scale genomic studies. With the growing number of samples in whole genome databases, efficient population clustering techniques that can handle such sample sizes have become increasingly important. Existing techniques for the clustering of genomes include STRUCTURE [1], FRAPPE [2] and, ADMIXTURE [3]. These compute probabilistic values referred to as *ancestry coefficients* that represent the fraction of the genome of an individual attributable to a particular population cluster. These methods can perform both supervised and unsupervised inference of *ancestry coefficients*. Supervised inference requires reference individuals from predefined ancestral populations, while unsupervised inference uses the structure of the data alone. These existing approaches perform inference via Bayesian [1] or likelihood based methods [2,3] and tend to be computationally expensive due to the high dimensionality of genomic data.

Dimensionality reduction techniques such as multidimensional scaling (MDS), principal component analysis (PCA) and uniform manifold approximation (UMAP) have been used to overcome the high dimensionality of genomic data [4,5], and have become indispensable for visualization and representation of diversity amongst genomic sequences. In PCA, samples are projected onto the axes of highest variation, each of which is a linear combination of allelic dosages across variants [6]. This method has become particularly important in genome-wide association studies and has also been used to investigate the distribution of genetic variation across geography [7]. An advantage is that no assumptions are made about ancestral populations; however, interpretation can often be misleading if sampling designs are irregular. Unsupervised clustering techniques such as ADMIXTURE or Archetypal Analysis (AA) can complement PCA to provide a detailed description of data and to augment visualization. In this work we show how AA can be coupled with PCA, specifically Single Value Decomposition (SVD), to efficiently cluster samples providing shorter run-times than STRUCTURE or ADMIXTURE. We also discuss how these techniques work, where they differ, and how they relate to well established general-purpose clustering techniques such as K-Means and K-Medoids.

## Materials and methods

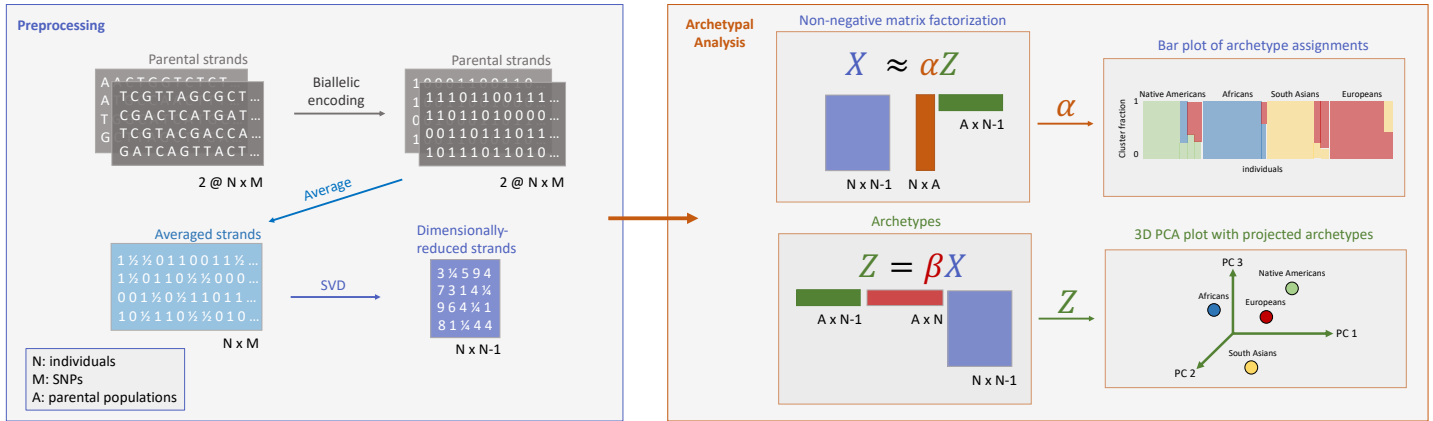
### System Overview

The complete proposed pipeline is presented in Figure 1.

### Singular Value Decomposition

Because the subspace spanned by the centered genotype vectors can have no more than  $N - 1$  dimensions with  $N$  the number of samples, there is no loss of information in projecting these centered genotype vectors onto their top  $N$  right singular vectors before applying Archetypal Analysis. This operation corresponds simply to a rotation of the coordinate system followed by a pruning of the unused dimensions and yields a space that is generally far smaller than the original, that is the number of total genotyped positions  $M$ , since typically  $N \ll M$ .

If we observe  $N$  individuals at  $M$  SNP positions, each individual  $i$  can be represented by a vector  $\mathbf{x}_i \in \{0, \frac{1}{2}, 1\}^M$ , where each position  $j$  in  $\mathbf{x}_i$  indicates the average number of alternate alleles found for each  $j$  (position) and  $i$  (individual's diploid genome). By aggregating  $\mathbf{x}_i$  the vectors for all individuals, we obtain an  $M \times N$  genotype matrix  $\mathbf{G} = [x_1 \dots x_N]$ . We center the columns of  $\mathbf{G}$  to produce data matrix  $\mathbf{X}$  and then compute the SVD:



**Fig 1. Archetypal Analysis pipeline.** The allele counts from both haplotypes of each of  $N$  individuals are averaged and then dimensionally-reduced from  $M$  SNPs to  $N - 1$  singular vectors via the SVD. Archetypal Analysis then implements an alternating non-negative matrix factorization algorithm that minimizes a constrained sum of squares to find ancestry proportions ( $\alpha$ ) and cluster centroids ( $Z$ : archetypes). Archetypal analysis models the individual genotypes as originating from the admixture of  $A$  parental populations, where  $A$  is an input parameter. For visualization we create bar plots for proportions of archetype assignments given by the matrix  $\alpha$ , and project archetypes  $Z$  into a 3D subspace using the first three principal components of the individual genotype sequences.

$$X = U\Sigma V^T \quad (1)$$

This yields  $U$  and  $V$ , the left and right-singular vectors respectively. The first  $N - 1$  scores  $U\Sigma$  can then be used as input for Archetypal Analysis. As described in [6] these vectors are made up of a linear combination (rotation) of genotypic values across the genome.

### Archetypal Analysis

This non-negative matrix factorization method was first developed by Cutler and Breiman in 1994 [8], and here it represents each individual as a convex combination of *extreme points*, or archetypes, in allele frequency space. In particular, given an  $N \times M$  multivariate data set  $X$  with  $N$  individuals and  $M$  SNPs, for a given number of archetypes or clusters  $K$ , the algorithm finds the  $M \times K$  matrix of archetypes  $Z$  according to two principles:

1. The samples are approximated as convex combinations of the archetypes such that the the residual sum of squares (RSS) between the approximation and original data is minimized:

$$RSS = \|X - \alpha Z^T\|^2 \quad (2)$$

with  $\alpha$  representing the fractional ancestry assignments, so  $\sum_{j=1}^K \alpha_{ij} = 1$ ,  $1 \geq \alpha_{ij} \geq 0$  for  $i = 1, \dots, N$ , and  $j = 1, \dots, K$ .

2. The archetypes are convex combinations of the samples:

$$Z = X^T \beta \quad (3)$$

with  $\beta$  an  $N \times K$  matrix and  $\beta_{ij}$  indicating the weight of sample  $i$  at archetype  $j$ , and  $\sum_{i=1}^N \beta_{ij} = 1$  with  $1 \geq \beta_{ij} \geq 0$ .

By combining Equation 2 and 3 we have:

$$RSS = \|X - \alpha\beta^T X\|^2 = \sum_i \|x_i - \sum_k \alpha_{ik} \sum_j \beta_{kj} x_j\|^2 \quad (4)$$

The optimization problem presented in Equation 2 and 3 consists of finding the weight matrices  $\alpha$  and  $\beta$  for a given data matrix  $X$  and a particular number of archetypes  $K$ . This is commonly solved through an iterative process of optimizing  $\alpha$  and  $\beta$  in an alternating fashion. For a fixed set of values for  $\alpha$ , finding the optimal values for  $\beta$  is reduced to a constrained least squares problems, and vice versa [8]. The iterative process is typically repeated until the quality of the decomposition reaches a pre-defined threshold, or up to a fixed maximum number of steps. The constrained least square optimization problem can be solved through a variety of techniques. Here we make use of the implementation of [9], which utilizes a non-negative least squares solver obtaining  $\alpha_{ij} \geq 0$  and  $\beta_{ij} \geq 0$ , where it adds an extra dimension to enforce  $\sum_{i=1}^n \alpha_{ij} = 1$  and  $\sum_{i=1}^n \beta_{ij} = 1$ . There are multiple open-source packages available in R [10], Python [9] and MATLAB [11] that implement Archetypal Analysis.

Unlike ADMIXTURE, Archetypal Analysis permits the use of dimensionally-rotated representations of SNP data, such as the singular value decomposition. If all singular vectors are used the residual sum of squares of the decomposition ( $RSS'$ ) using projected data  $X'$  is equivalent to the  $RSS$  of the original decomposition:

$$\begin{aligned} RSS' &= \sum_i \|x'_i - \sum_k \alpha_{ik} \sum_j \beta_{kj} x'_j\|^2 = \sum_i \|Px_i - \sum_k \alpha_{ik} \sum_j \beta_{kj} Px_j\|^2 \\ &= \sum_i \|Px_i - P(\sum_k \alpha_{ik} \sum_j \beta_{kj} x_j)\|^2 \\ &= \sum_i \|x_i - \sum_k \alpha_{ik} \sum_j \beta_{kj} x_j\|^2 \propto RSS \end{aligned} \quad (5)$$

Since the projection matrix  $P = V$ , the orthonormal rotation matrix of  $X$  onto its singular vector axes.

This permits us to perform AA clustering on a matrix of dimensions only  $N \times N - 1$  instead of  $N \times M$ . Note that although the learnt parameters of AA,  $\alpha$  and  $\beta$ , do not depend on  $M$ , the computation times for  $Z$  and the  $RSS$  do, therefore, working in lower dimensions reduces the computational load.

**Constrained Optimization.** Non-negative least squares (NNLS) is a constrained least squares problem in which coefficients are always non-negative (Eq. 8). Archetypal Analysis includes an additional constraint coefficient  $C$  and adds a row of ones to matrices involved in optimization after every NNLS iteration (Eq. 9 and 10) to ensure the coefficients also sum to one, one of the definitional properties of Archetypal Analysis.

Given an  $N \times M$  matrix  $X$  representing a multivariate data set with  $N$  observations and  $M$  attributes, for a given  $K$ , we minimize:

$$R\tilde{S}S = \|\tilde{X} - \tilde{\alpha}Z^T\|_2 \quad (6)$$

where  $\tilde{\alpha}$  is defined as:

$$\tilde{\alpha} = \begin{pmatrix} \alpha_{11}C & \alpha_{12}C & \alpha_{13}C & \dots & \alpha_{1K}C \\ \alpha_{21}C & \alpha_{22}C & \alpha_{23}C & \dots & \alpha_{2K}C \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \alpha_{N1}C & \alpha_{N2}C & \alpha_{N3}C & \dots & \alpha_{NK}C \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix} \quad (7)$$

and  $\alpha$  and archetypes are defined in the previous section.  $\tilde{X}$  is defined as:

$$\tilde{X} = \begin{pmatrix} x_{11}C & x_{12}C & x_{13}C & \dots & x_{1M}C \\ x_{21}C & x_{22}C & x_{23}C & \dots & x_{2M}C \\ \vdots & \vdots & \vdots & \dots & \vdots \\ x_{N1}C & x_{N2}C & x_{N3}C & \dots & x_{NM}C \\ 1 & 1 & 1 & \dots & 1 \end{pmatrix} \quad (8)$$

where  $C$  is a constraint coefficient for  $C > 0$  and rows of 1's are added after every NNLS iteration. This ensures the constraint  $\sum_{j=1}^k \alpha_{ij} = 1$  where the value of  $C$  represents a weighting between the importance of the constraint and NNLS minimization, with lower  $C$ 's giving a stronger importance to the constraint. The same method is applied to  $\beta$  coefficients to ensure  $\sum_{i=1}^n \beta_{ij} = 1$ .

**Archetype Initialization.** We make use of the implementation in [9] which supports three different archetype initialization strategies: (1) random initialization of the archetypes where each dimension of the archetype is sampled from a uniform distribution scaled to have the same range as the input data, (2) random selection of a sample from the input data as the archetype, and (3) the FurthestSum introduced in [11]. By default we make use of FurthestSum initialization as it efficiently generates initial archetype candidates by, after selecting the first archetype randomly, selecting each subsequent archetype as the sample that has the largest aggregate distance from the previously selected archetypes.

**Implementation Details.** Archetypal analysis was run with the following parameters (with code adapted from [9]).

- Tolerance: defines when to stop optimization when alternating between finding the best  $\alpha$ 's for given archetypes  $Z$  and finding the best  $Z$  for given  $\alpha$ 's. Specifically, the threshold applied is,

$$\frac{||RSS_c - RSS_p||}{RSS_p} > T \quad (9)$$

where  $RSS$  is the residual sum of squares defined in (Eq. 2) for the current iteration  $RSS_c$  and the previous iteration  $RSS_p$ , and  $T$  is the desired tolerance. We use a value of  $T = 0.001$ .

- Maximum number of iterations for the residual sum of squares ( $RSS$ ) minimization: 50.
- Constraint coefficient  $C$ : coefficient that ensures the summation of  $\alpha$ 's and  $\beta$ 's equals to 1. See Appendix B for further details on the constrained optimization method. We use a value of  $C = 0.001$ .
- Initialization method: we use FurthestSum [11] as the initialization method.

## Datasets

### HUMAN

Whole genomes from the Human Genome Diversity Project [12], the Simons Genome Diversity Project [13] and the 1000 Genomes Project [14] have been included in this study. The Human Genome Diversity Project whole genome cohort includes 929 individuals from 54 human populations. The Simons Genome Diversity Project contains 300 genomes from 142 diverse populations, and the 1000 Genomes Project includes 2504 individuals from 26 populations. The three datasets were merged, removing duplicated individuals between the studies and retaining only SNPs present in all three datasets, yielding an intersection of 1,411,471 SNPs for analysis. Rare variants with minor allele frequencies  $< 0.1$  were removed. In total, 3558 individuals were included in the study from 7 different continents: 683 from Europe, 805 from Africa, 34 from Oceania, 695 from South Asia, 772 from East Asia, 150 from West Asia, and 419 indigenous individuals from the Americas.

### Dogs

The heterogeneous data set of dog breeds from [15] consists of 1355 groups representing 166 dog breeds. Each sequence has a total of 150,131 SNPs. Populations with vastly different histories are included, originating from all continents except Antarctica [15].

## Results

### Human datasets

#### Principal Components and Archetypal Analysis

We first compute the principal components of the human data set and display the first two components in a plot coloured by continental population (Fig. 2, **a**). The African population displays the highest genetic variability extending across the first principal component axis (11% explained variance). We then use all principal components, that is the projection onto all the left singular vectors of the SVD, as input to the Archetypal Analysis and plot the proportional membership of each cluster for each individual in a compositional plot (Fig. 2, **b**). The African population is represented by three archetypes (A1, A2 and A8), while the East Asian and South Asian populations have one archetype each (A3 and A5 respectively). Note that Archetypal Analysis addresses the high variation within African groups by using multiple archetypes. The European and West Asian populations share a single archetype (A4), while the Oceanian populations are found on the gradient between the East Asian and South Asian archetypes. Finally, the Native American populations are represented by two archetypes (A6 and A7) and have a gradient running to the European/West Asian archetype due to colonial admixture. Example populations found along this gradient are the Puerto Ricans and Colombians.

#### Comparison of ancestry estimates

To compare the ancestry estimates derived from ADMIXTURE and Archetypal Analysis, we display the proportional ancestry cluster assignments, the  $Q$  and  $\alpha$  matrices respectively, in a bar plot for  $K = 8$  cluster (Fig. 3, **b**). Each vertical bar represents an individual and the shaded colors denote the cluster proportions. We also display individuals on a three-dimensional PCA plot with projected archetypes ( $Z$ ) and

ADMIXTURE cluster centers ( $F$ ) (Fig. 3, **a**). A theoretical comparison of both methods can be found in the Discussion section.

*Archetypal Analysis*: European (red), South Asian (turquoise), and East Asian populations (yellow) are predominantly represented by a single archetype. Native American populations are a combination of three archetypes, two of which are mostly specific to this population (light green and dark green) and a third, representing colonial admixture, which is European (red). Individuals from Puerto Rico and Colombia mostly share the third archetype with Europeans. The African population is represented by three archetypes. One archetype encompasses West African populations such as Mandeka, Gambian Mandika and Mende (ocean blue). Another includes eastern and southern groups such as Luhya and San (navy blue). A third archetype represents a few individuals from all African populations (light blue).

*ADMIXTURE*: Oceanian (purple) and East Asian populations (yellow) are predominantly represented by a single cluster center. Europeans and West Asians are a combination of two centers (red and pink) that are located outside the point cloud of individuals, differing from AA which captures both with a unique cluster. Native Americans show traces of the European and West Asian cluster components, but are mostly represented by their own, here single, cluster (light green). African populations are predominantly represented by two clusters (ocean blue and light blue), while a few populations, such as the North African Mozabites, show traces of European and West Asian components. Finally, South Asians predominantly cluster around a single cluster (turquoise), but also show traces of the European and West Asian clusters.

Overall, Archetypal Analysis provide estimates which qualitatively often match ethnolinguistic and geographical labels. AA properly captures the wide variation within African populations; however, it fails to identify a unique cluster for Oceanians. Additionally, due to its stronger constraints than ADMIXTURE, AA obtains cluster centroids that lie near actual sampled genotypes.

## Domestic dog breed dataset

### Principal Components and Archetypal Analysis

We compute the principal components of the dog breed data sets and display the first two components in a plot coloured by dog clades (Fig. 4, **a**). The Asian Spitz clade shows the highest genetic variability extending across the first principal component axis, including breeds such as Chow Chow, Greenland Sledge Dog and Siberian Husky. The latter is found close to the wolf, while the European Mastiff clade represented by breeds such as Bull Terrier, Boxer and Bulldog extends across the second principal component axis. Archetypal Analysis is then computed for  $K = 5$  and  $K = 15$  with principal components as input (Fig 4, **b** and **c**). For  $K = 5$ , dog archetypes were found to be the Asian Spitz dogs (A1), the Bulldog-derived dogs (A2), the Terriers (A3), hunting water dogs (A4) and herding dogs (A5). The remaining breeds are displayed as a combination of these main archetypes, mostly represented by A5 and A4. This matches the structure shown in PCA, where most of the breeds are clustered in the origin, except the dogs in the Bull Terrier and Husky groups. When increasing the number of archetypes to  $K = 15$ , individual dog breeds begin clustering around single archetypes, showing the growing population structure. New archetypes appear for the Boxer (A3), Irish Wolfhound (A4), Otter Hound (A5), Bullmastiff (A6), Bernese Mountain Dog (A10), Glen of Imaal Terrier (A11), French Bulldog (A12), Boston Terrier (A13), Shetland Sheepdog (A14) and Tibetan Spaniel (A15). The rest of the breeds are mostly found near A14 and A15.

## Performance metric analysis

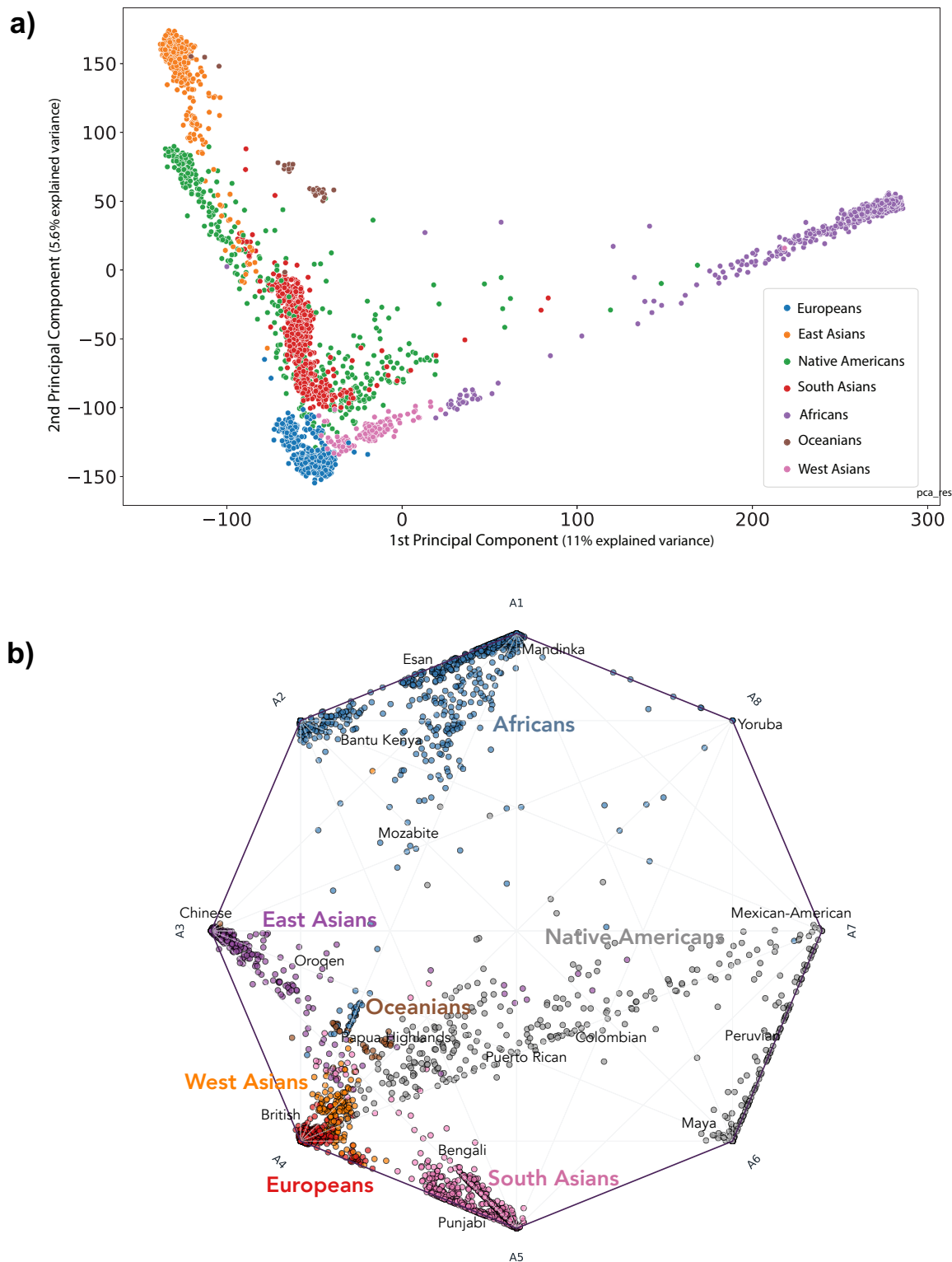
The dog breed dataset was used to benchmark the computation times and clustering quality of both ADMIXTURE and Archetypal Analysis. Running times and explained variances of ADMIXTURE and Archetypal Analysis are measured for an increasing number of archetypes/clusters  $K = 1, \dots, 22$  and  $K = 1, \dots, 30$  respectively. The initialization was set to *random* for both methods to achieve uniform comparison and results were averaged over 5 runs. Accumulated run-times increased exponentially with  $K$  for ADMIXTURE whereas they increased linearly for Archetypal Analysis (Fig. 5). An accumulated runtime of 34 minutes was taken by Archetypal Analysis to compute ancestry estimates for  $K = 2$  to  $K = 30$  clusters. For ADMIXTURE, the accumulated runtime from  $K = 2$  to  $K = 30$  was 78 hours. Thus, Archetypal Analysis ran 137 times faster than ADMIXTURE on the domestic dog breed dataset. A similar increase in relative speed was maintained, on average, for non-cumulative times (Table 1).

Explained variances increased linearly in the number of clusters for both algorithms (Fig. 5). The explained variance for Archetypal Analysis was on average 2% lower than for ADMIXTURE. For the values of  $K$  included in this analysis, the mean standard deviation for five averaged runs with random initialization was 0.007 for Archetypal Analysis and 0.0004 for ADMIXTURE. As described in the following Discussion section, the difference in explained variance is due, at least in part, to the stronger restrictions that Archetypal Analysis imposes when estimating the cluster centroids. However, as shown with human sequences in figure 3, the stronger restrictions of AA lead to centroids that are always a linear combination of actual samples, guaranteeing that they represent theoretically observable population samples.

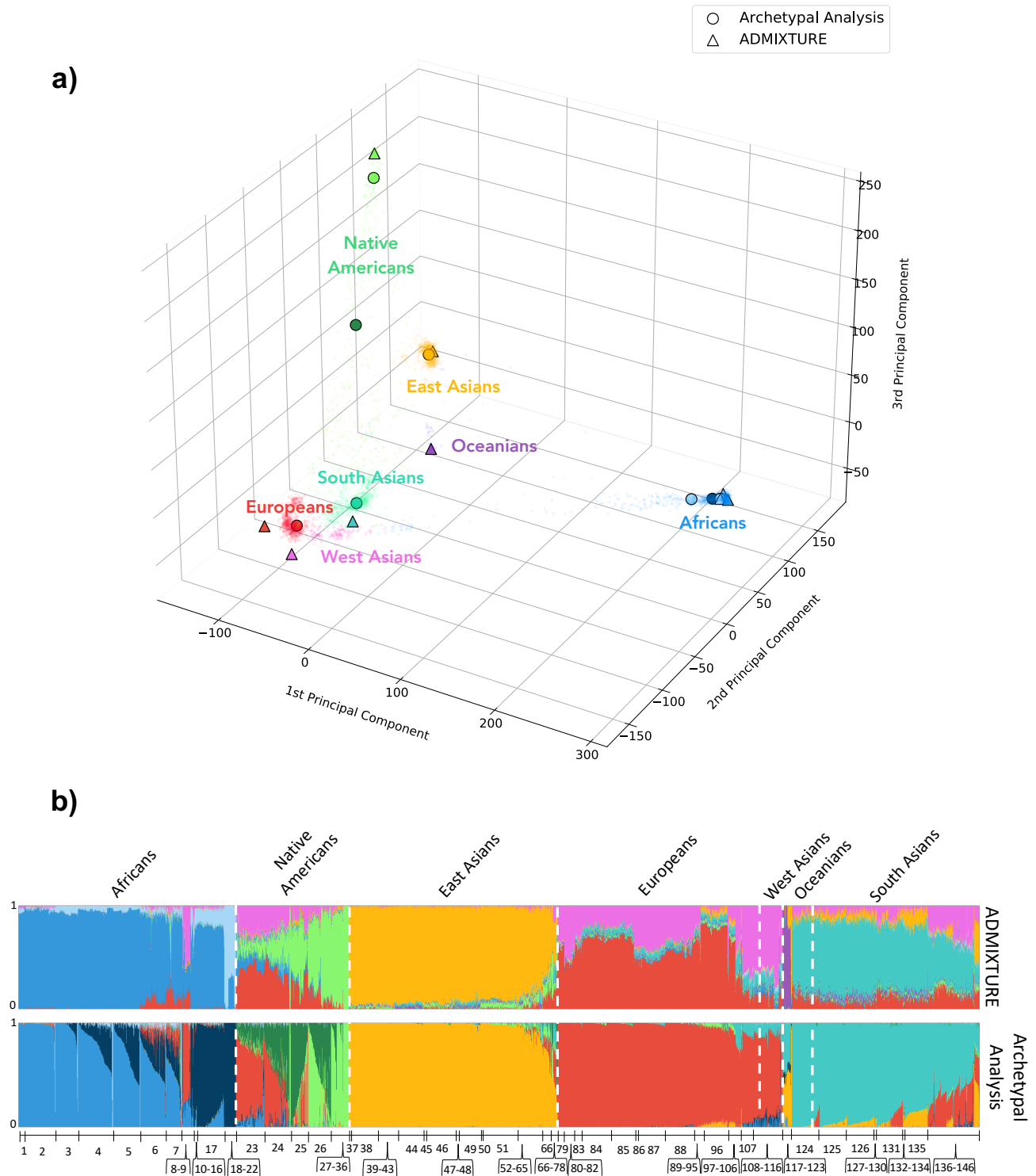
**Table 1. Runtime (in minutes) for ADMIXTURE-AA comparison**

Algorithm	K (number of clusters / archetypes)						
	[2-6)	[6-10)	[10-14)	[14-18)	[18-22)	[22-26)	[26-30)
ADMIXTURE	43	64	97	150	247	250	319
AA	0.5	0.48	0.7	1	1.4	1.9	2.4
<b>Relative speed</b>	86×	133×	139×	150×	176×	132×	132×

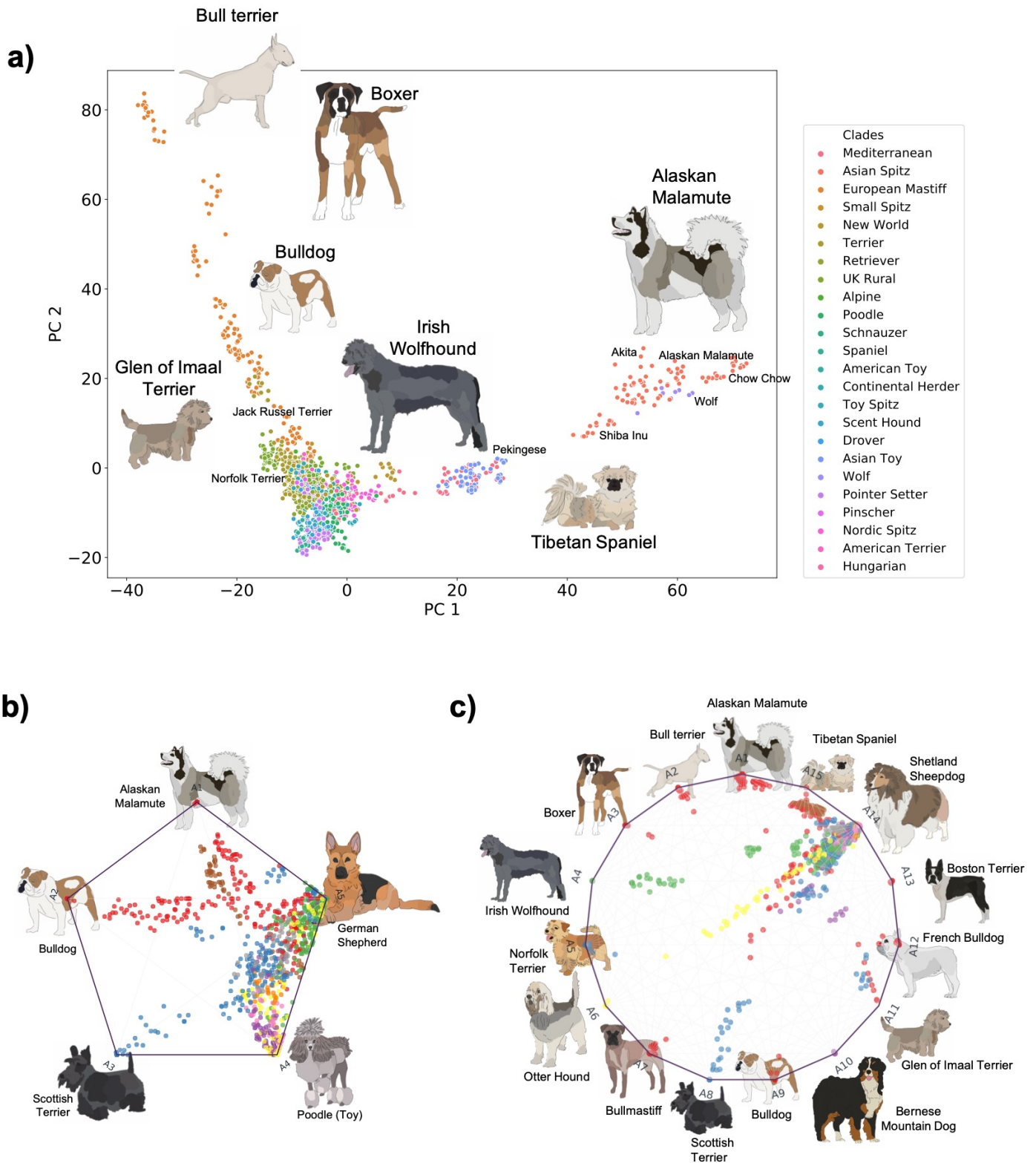




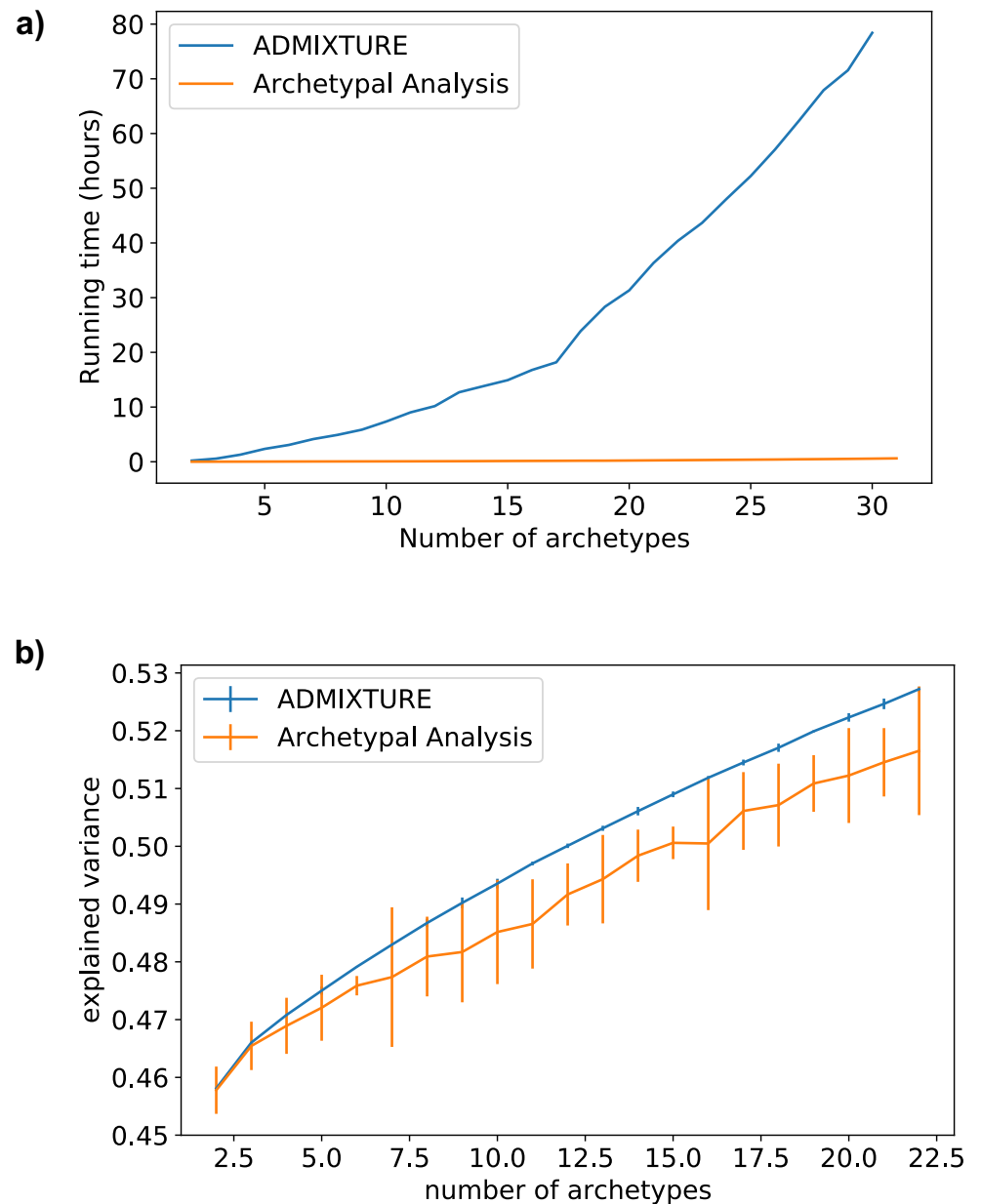
**Fig 2. Principal Component Analysis and Archetypal Analysis compositional plots for human populations (K=8).** **a)**, 2-dimensional PCA plot of human continental populations, where groups of individuals are colored by the unique regional genetic components they possess (see legend) **b)**, Compositional plot giving proportional archetype assignment for each individual (points). Points are coloured by the presence of regional genetic components and a few example sub-populations are shown given in text. Clusters of individuals from the same population are observed on the edges of the polygon and gradients between edges indicate admixed individuals.



**Fig 3. Comparison of ancestry estimates for human populations (K=8).** **a)**, three-dimensional PCA plot of individuals with projected archetypes (circles) and ADMIXTURE cluster centers (triangles). **b)**, bar plot where individuals are represented along the horizontal axis as narrow columns and ordered by population group. Colour bars along the vertical axis show the proportional cluster assignment for each individual. We compare the cluster assignments of ADMIXTURE (top) and Archetypal Analysis (bottom).



**Fig 4. Principal Component Analysis and Archetypal Analysis compositional plots for domestic dog breeds.** **a)**, two-dimensional PCA plot of domestic dog breeds where groups of dogs are colored by clade. **b)** and **c)**, proportional composition of each cluster for each individual in coordinate space for  $K=5$  and  $K=15$  respectively. Data points are coloured by clade and archetype representatives are shown as drawings. Gradients between edges indicate combinations between breeds.



**Fig 5. Performance metrics analysis.** a), runtime analysis for ADMIXTURE and Archetypal Analysis for  $K = 2$  to  $K = 30$ . Time is expressed in units of accumulated hours. b), explained variance analysis comparison for ADMIXTURE and Archetypal analysis for  $K = 2$  to  $K = 22$ . Results are averaged over five distinct random seed values for each value of  $K$  and the ranges observed are shown as vertical bars.

## Discussion

### Population structure overview

Archetypal Analysis proved to be an interpretable alternative to ADMIXTURE. It assigned separate regional archetypes that associated predominantly with Europeans, with South Asians, and with East Asians, and it recognized the high genetic variability of African populations. Differences within regions were also detectable (Fig. 3). For example, indigenous peoples across the Americas were separated from the remainder of the modern American communities as the light green archetype. Peruvians were also included in this group, most likely because indigenous groups make up 45% of the Peruvian population. Similarities in peoples that are geographically spread were also detected. For example, the Bantu peoples (Bantu Herero, Bantu Tswana, Bantu Kenya, Bantu South Africa and Luhya) comprise several hundred indigenous ethnic groups in Africa spread over a vast area from Central Africa to Southern Africa, but those present in our dataset were grouped together forming the dark blue archetype.

European-like archetype components seen in African peoples due to geographic proximity and migration were also found in the Saharawi and Mozabites from the northwestern part of Africa. As also observed in previous studies, American populations, such as Puerto Rico and Colombia, showed European representation due to Spanish colonization. The suggested effects of this historical event can also be observed in (Fig. 2, b), which shows a gradient of relatedness to Europeans that runs through Puerto Ricans, Colombians, Peruvians through the Mexican-Americans. Archetypal Analysis also identified South Asian communities having a shared component with Europeans that ADMIXTURE did not detect (Fig. 3, b). For example, the Brahui, Kalash and Baloch were identified with a European-like archetype by Archetypal Analysis and not by ADMIXTURE. These might reflect the influence of Indo-European migrations and the Ancestral North Indians [16], an ancestral genetic grouping in India that shares some ancestry with other Indo-European speakers from India to Iran to Europe.

### Relationship between Archetypal Analysis and ADMIXTURE

The popular algorithm ADMIXTURE estimates individual ancestries by computing maximum likelihood estimates in a parametric model. Specifically, it maximizes the biconcave log-likelihood of the model using block relaxation:

$$\mathcal{L}(Q, F) = \sum_{i,j} (n_{ij} \ln p_{ij} + (2 - n_{ij}) \ln (1 - p_{ij})) \quad (10)$$

where genotype  $n_{ij}$  for individual  $i$  at SNP  $j$  represents the number of type '1' alleles observed. Given  $K$  populations, the success probability  $p_{ij} = \sum_{k=1}^K q_{ik} f_{kj}$  in the binomial distribution  $n_{ij} \sim \text{Bin}(2, p_{ij})$  depends on the fraction  $q_{ik}$  of  $i$ 's ancestry attributable to population  $k$  and on the frequency  $f_{kj}$  of the allele 1 in population  $k$ , where  $q_{ik}$  and  $f_{kj}$  are the entries of  $Q$  and  $F$  respectively [3].

ADMIXTURE and Archetypal Analysis share similar modeling assumptions. Both  $q_{kj}$  ADMIXTURE and  $\alpha$  archetype fractions can be interpreted as partial cluster assignments while ADMIXTURE frequency coefficients  $f_{kj}$  and archetype coordinates  $Z$  encode cluster center locations in SNP space. A key difference is that ADMIXTURE cluster centroids have  $M$  (# of SNPs) free parameters, in other words, the frequency at each SNP for each cluster ( $f_{kj}$ ) is a parameter that needs to be learnt. Instead, in AA, cluster centroids have  $N$  (number of samples) free parameters, that is, a coefficient ( $\beta$ ) for each training sample needs to be learnt for each cluster center. When  $M \gg N$  (the typical scenario when working with genomic data), AA has far fewer free-parameters than ADMIXTURE. This can lead to lower explained variance values (or higher

reconstruction errors), but guarantees centers that exist within the convex hull of real samples (and thus could represent a real descendant individual), while ADMIXTURE can over-fit, yielding centers outside the hull of the observed data (see Results section) that may represent no population that has ever existed. Furthermore, because AA does not optimize each of the  $M$  free-parameters, it can work with rotated data (the left singular vectors of the SVD) without any loss of information, or with dimensionally-reduced data, allowing for a much more efficient computation.

The likelihood function of ADMIXTURE can be understood as an error or distance metric between the input sequences  $X$  (where both haplotypes have been averaged) and a decomposed product  $QF$ . In fact, when  $X \approx QF$ :

$$\frac{1}{2}\mathcal{L}(Q, F) = \sum_{i,j} (x_{ij} \ln q_{ij} f_{ij} + (1 - x_{ij}) \ln (1 - q_{ij} f_{ij})) \approx \sum_{i,j} \|x_{ij} - q_{ij} f_{ij}\|^2 \quad (11)$$

Therefore, the likelihood function resembles the  $RSS$  problem of AA. In fact, ADMIXTURE can be understood as a type of likelihood-based relaxed archetypal analysis, where the constraints imposed on the cluster centroids are loosened.

Another shared aspect of both methods is the alternating nature of the optimization procedure. In both methods, cluster centers and cluster assignments are optimized in an iterative manner. Once the cluster assignments are fixed, optimizing centers becomes a convex problem, and vice versa, allowing for fast convergences.

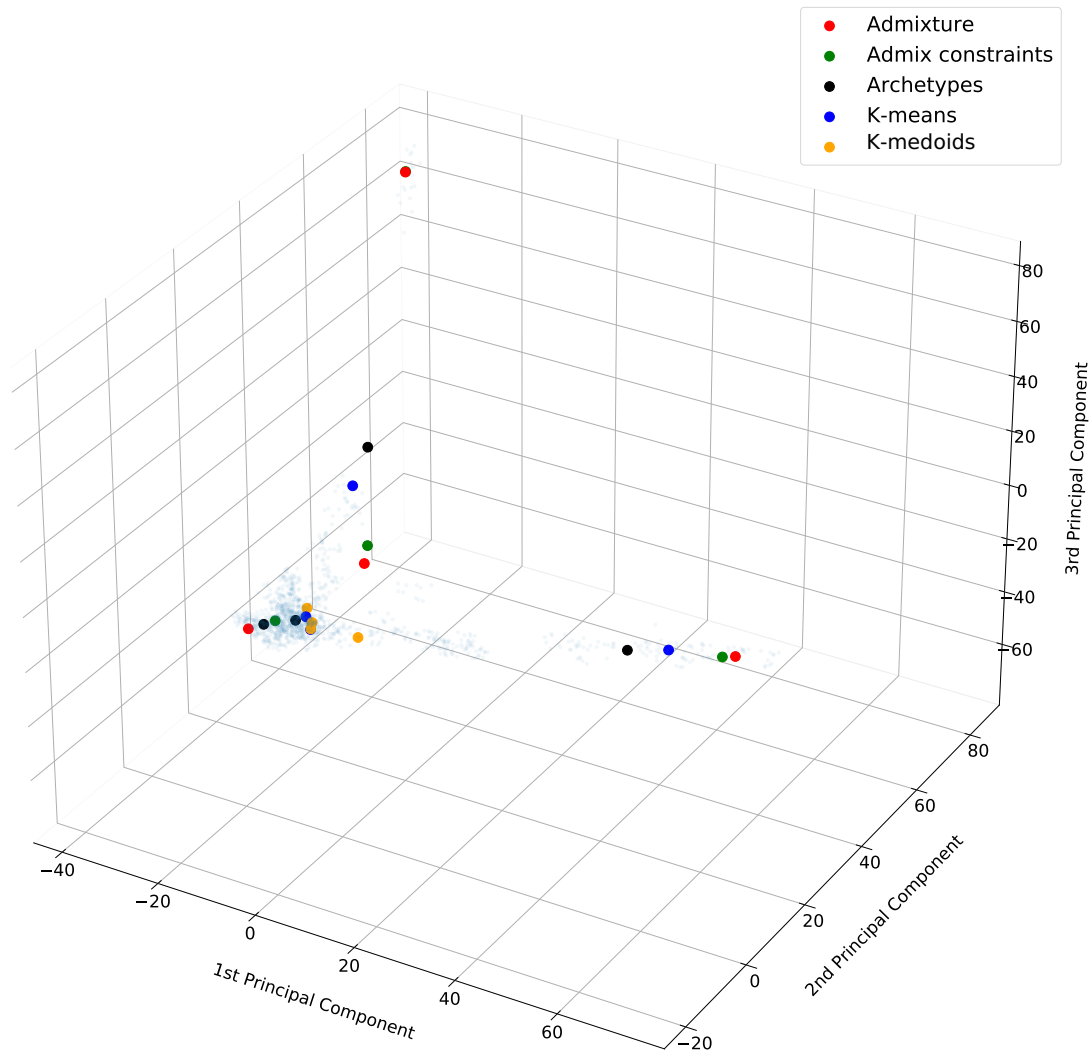
**Table 2. ADMIXTURE and Archetypal Analysis comparison**

	ADMIXTURE	Archetypal Analysis
<b>Model</b>	$X \approx QF$	$X \approx \alpha Z^T$
Loss Function	log-likelihood	RSS
Free-parameters	$(N + M)K - N$	$2NK - N - K$
<b>Cluster Assignments (CA)</b>	$Q$	$\alpha$
CA Dimensions	$N \times K$	$N \times K$
CA Free-parameters	$N(K - 1)$	$N(K - 1)$
CA Constraints	$\sum_{j=1}^K Q_{ij} = 1$ and $Q_{ij} \geq 0$	$\sum_{j=1}^K \alpha_{ij} = 1$ and $\alpha_{ij} \geq 0$
<b>Cluster Centroids (CC)</b>	$F$	$Z = \bar{X}^T \beta$
CC Dimensions	$K \times M$	$K \times M$
CC Free-parameters	$KM$	$K(N - 1)$
CC Constraints	$0 \leq F_{ij} \leq 1$	$\sum_{j=1}^N \beta_{ij} = 1$ and $\beta_{ij} \geq 0$

### Relationship between Archetypal Analysis, ADMIXTURE, K-Means, and K-Mediod Clustering

Archetypal Analysis and ADMIXTURE hold a strong relationship with K-Means and K-Mediods. As already stated in [11], if the constraints on the archetypes  $Z$  are relaxed, and cluster assignments are limited to binary values  $\alpha \in \{0, 1\}$  and  $\sum_{j=1}^k \alpha_{ij} = 1$ , then archetypal analysis becomes equivalent to K-Means. Similarly, if the sparsity regularization used in ADMIXTURE [3] is strongly applied, the clusters assignments  $Q$  become binary and the technique becomes similar to K-Means. In a similar fashion, if both  $\alpha$  and  $\beta$  are restricted to be binary,  $\alpha, \beta \in \{0, 1\}$ , Archetypal Analysis becomes equivalent to K-Mediods. Therefore, AA can be understood as a smooth or fuzzy version of K-Mediods. Note that both K-Means and K-Mediods are also typically optimized in an iterative alternating nature, similar to AA and ADMIXTURE.

Figure 6 shows a qualitative comparison of all four of these methods when  $K = 4$ . Examples with  $K = 3$  and  $K = 5$  can be found in the supplement. We can observe that ADMIXTURE with sparsity constraints (green) obtains cluster centroids less extremal than ADMIXTURE without these constraints, showing a behaviour that tends to K-Means. Note that the differences between cluster centers will not depend only on differences in modelling assumptions for each technique, but also in differences in implementation details and initialization approaches of each method.



**Fig 6. Comparison of cluster centroids from different methods.** Cluster centers learned by ADMIXTURE, ADMIXTURE with sparsity regularization, Archetypal Analysis, K-Means, and K-Medoids

## Conclusion

In this paper we show how Archetypal Analysis can be used as a fast alternative to ADMIXTURE for population clustering. We also show how the Archetypal Analysis

model has fewer degrees of freedom, restricting the centroids of clusters within convex hull combinations of the training samples, which leads to lower explained variance than ADMIXTURE, but provide more interpretable cluster centroids. We apply our proposed system to human and dog genotypes, showing that AA can perform more than two orders of magnitude faster than ADMIXTURE while still properly capturing the population structure of the data.

## Acknowledgments

This work was supported in part by the Chan Zuckerberg Biohub (through CDB) and by the Royal Academy of Engineering Leaders Scholarship (awarded to JGM). We would like to thank Santiago De Vilallonga for his illustrations of dogs.

## References

1. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000;155(2):945–59.
2. Tang H, Peng J, Wang P, Risch NJ. Estimation of individual admixture: analytical and study design considerations. *Genetic epidemiology*. 2005;28(4):289–301.
3. Alexander DH, Lange K. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*. 2011;12. doi:10.1186/1471-2105-12-246.
4. Reich D, Price AL, Patterson N. Principal component analysis of genetic data. *Nature Genetics*. 2008;40. doi:10.1038/ng0508-491.
5. Diaz-Papkovich A, Anderson-Trocmé L, Ben-Eghan C, Gravel S. UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS genetics*. 2019;15(11):e1008432.
6. McVean G. A genealogical interpretation of principal components analysis. *PLoS Genetics*. 2009;5. doi:10.1371/journal.pgen.1000686.
7. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. *Nature*. 2008;456(7218):98–101.
8. Cutler A, Breiman L. Archetypal analysis. *Technometrics*. 1994;36. doi:10.1080/00401706.1994.10485840.
9. Motevalli Soumehsaraei A, Benyamin; Barnard. Archetypal Analysis Package. v1. CSIRO. Software Collection.; 2019.
10. Eugster MJA, Leisch F. From Spider-Man to Hero - Archetypal Analysis in R. *Journal of Statistical Software*. 2009;30. doi:10.18637/jss.v030.i08.
11. Mørup M, Hansen LK. Archetypal analysis for machine learning and data mining. *Neurocomputing*. 2012;80:54–63.
12. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*. 2020;367. doi:10.1126/science.aay5012.



13. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*. 2016;538(7624):201–206.
14. Consortium GP, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68.
15. Parker HG, Dreger DL, Rimbault M, Davis BW, Mullen AB, Carpintero-Ramirez G, et al. Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development. *Cell Reports*. 2017;19. doi:10.1016/j.celrep.2017.03.079.
16. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. *Nature*. 2009;461. doi:10.1038/nature08365.

## Supporting information

**Human bar plot labels** Table 3 displays all the subpopulation labels used in Figure 3b. The details of the dataset can be found in the Datasets subsection in the Methods section.

**Table 3. Bar plot labels**

ID	Population	ID	Population	ID	Population	ID	Population
1	Mandenka	2	Gambian Mandinka	3	Mende	4	Yoruba
5	Esan	6	African-Caribbean	7	African-American SW	8	Bantu Herero
9	Mozabite	10	Saharawi	11	Luo	12	Bantu South Africa
13	Dinka	14	Somali	15	Masai	16	Bantu Kenya
17	Luhya	18	Bantu Tswana	19	Khomani San	20	Mbuti
21	Biaka	22	San	23	Puerto Rican	24	Colombian
25	Mexican-American	26	Peruvian	27	Maya	28	Pima
29	Zapotec	30	Quechua	31	Mixe	32	Chane
33	Mixtec	34	Piapoco	35	Surui	36	Karitiana
37	Ami	38	Southern Han Chinese	39	Miao	40	Tujia
41	Han	42	She	43	Dai	44	Dai Chinese
45	Atayal	46	Han Chinese	47	Igorot	48	Korean
49	Kinh Vietnamese	50	Lahu	51	Japanese	52	Northern Han
53	Yi	54	Naxi	55	Hezhen	56	Daur
57	Tu	58	Thai	59	Xibo	60	Mongolian
61	Oroqen	62	Cambodian	63	Ulchi	64	Burmese
65	Even	66	Yakut	67	Altaiian	68	Itelman
69	Kyrgyz	70	Tubalar	71	Eskimo Chaplin	72	Eskimo Sireniki
73	Eskimo Naukan	74	Uyгур	75	Mansi	76	Chukchi
77	Aleut	78	Tlingit	79	Basque	80	Czech
81	Sardinian	82	Bergamo Italian	83	French	84	British
85	CEPH	86	Albanian	87	Tuscan	88	Spanish
89	Orcadian	90	Norwegian	91	Icelandic	92	Hungarian
93	Polish	94	Estonian	95	Bulgarian	96	Finnish
97	Crete	98	Greek	99	Russian	100	Samaritan
101	Adygei	102	Lezgin	103	Abkhasian	104	Saami
105	Chechen	106	North Ossetian	107	Druze	108	Yemenite Jew
109	Palestinian	110	Jordanian	111	Armenian	112	Bedouin
113	Georgian Mingrelian	114	Iraqi Jew	115	Turkish Cappadocia	116	Iranian
117	Papuan Highlands	118	Australian	119	Papuan Sepik	120	Bougainville
121	Maori	122	Hawaiian	123	Dusun	124	Gujarati
125	Indian Telugu	126	Sri Lankan	127	Brahmin	128	Yadava
129	Mala	130	Kapu	131	Punjabi	132	Madiga
133	Relli	134	Irula	135	Bengali	136	Sindhi
137	Pathan	138	Kalash	139	Burusho	140	Balochi
141	Brahui	142	Khonda Dora	143	Makrani	144	Tajik
145	Kusunda	146	Hazara				

**Domestic dog breeds details.** Tables 4 and 5 show the breeds of all the dogs included in our study. Details about the dataset can be found above in the Datasets subsection within the Methods section.

**Archetypal Analysis compositional plots.** Figures 7 and 8 show further examples of Archetypal Analysis predictions with the dataset of dog genotypes. The figures show that as the number of archetypes increases, more breeds are clustered in their individual archetype (e.g. A6 to A15 in the 15 archetypes plot), while the rest of breeds (the majority of the breeds) are represented as a combination of a few number of archetypes (e.g. A1 to A5 in the 15 archetypes plot).

**Genome-Wide Association Studies.** In order to depict how Archetypal Analysis can be included in GWAS, we display a Manhattan plot (Figure 9) and a Q-Q plot (Figure 10) of an association study of the height of dog breeds with and without archetype proportions as covariates.

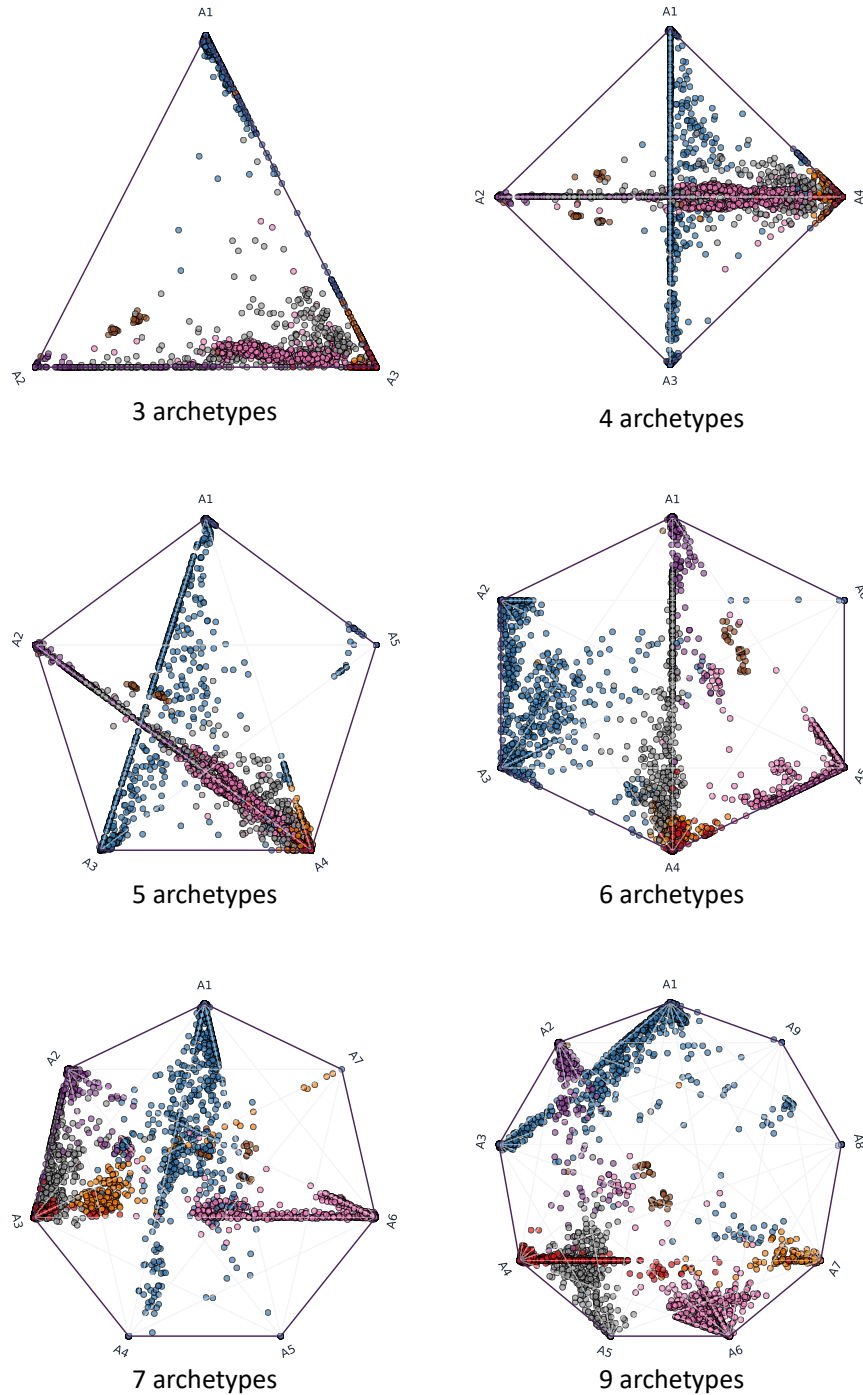
**Relationship between AA, K-Means, and K-Medioids.** We include additional plots comparing the  $K = 3$  cluster centers of AA, ADMIXTURE, K-Means, and K-Medioids in Figure 11 and  $K = 5$  in Figure 12.

**Table 4. Domestic dog breeds details (1)**

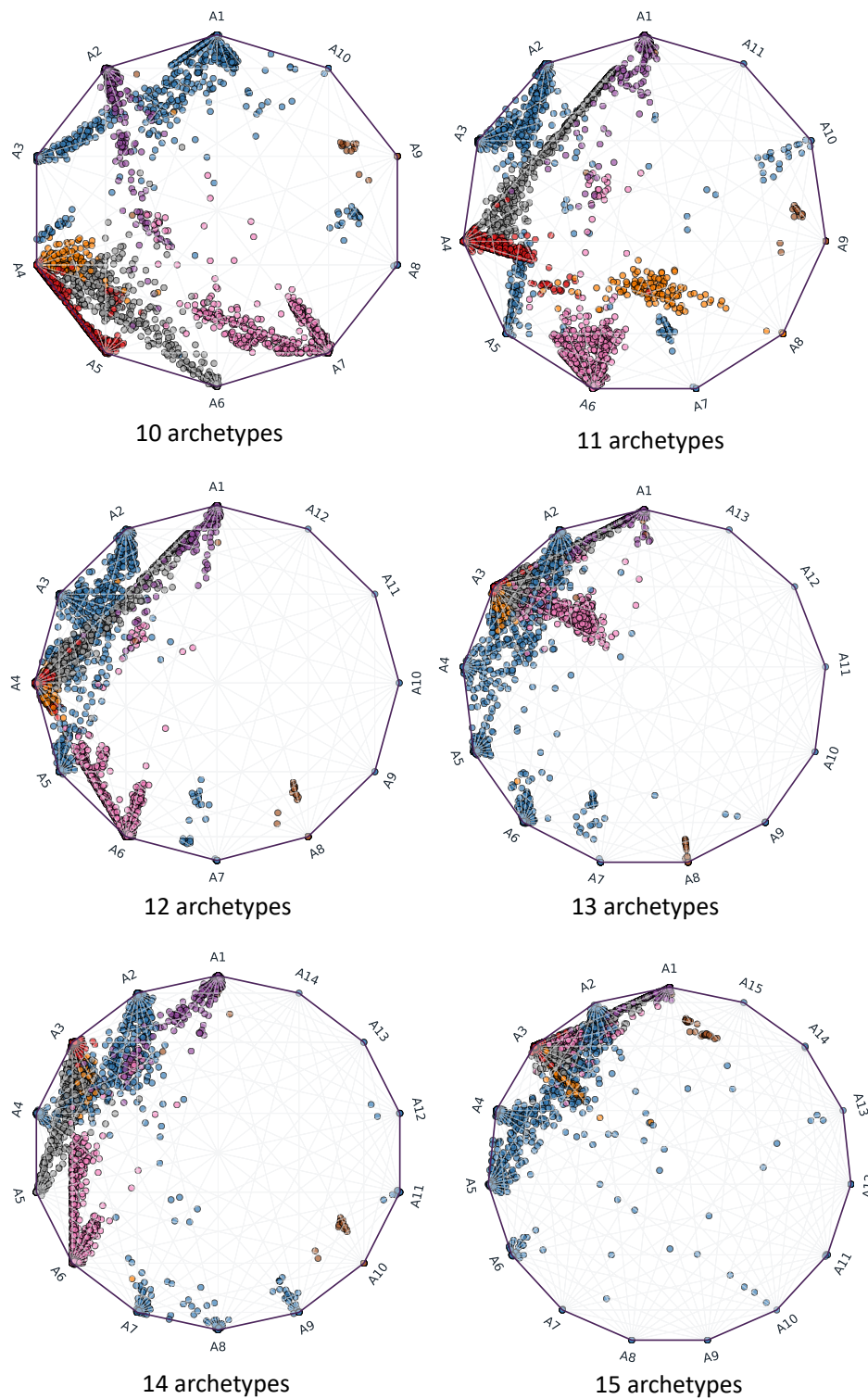
Breed	Clade	N	Breed	Clade	N
American Cocker Spaniel	Spaniel	10	American Eskimo Dog	Nordic Spitz	6
Afghan Hound	Mediterranean	10	American Hairless Terrier	American Terrier	10
Airedale Terrier	Terrier	3	Akita	Asian Spitz	10
Alaskan Malamute	Asian Spitz	10	American Staffordshire Terrier	European Mastiff	6
Anatolian Shepherd	Mediterranean	6	Australian Cattle Dog	UK Rural	10
Australian Shepherd	UK Rural	10	Australian Terrier	Terrier	10
Azawakh	Mediterranean	5	Basset Hound	Scent Hound	10
Beagle	Scent Hound	10	Bedlington Terrier	Terrier	7
Belgian Sheepdog	Continental Herder	10	Bearded Collie	UK Rural	3
Bichon Frise	Poodle	10	Bloodhound	Scent Hound	10
Belgian Malinois	Continental Herder	6	Bernese Mountain Dog	Alpine	10
Boerboel	European Mastiff	3	Border Collie	UK Rural	10
Border Terrier	Terrier	10	Borzoi	UK Rural	10
Boston Terrier	European Mastiff	10	Bouvier des Flandres	Continental Herder	8
Boxer	European Mastiff	10	Berger Picard	New World	3
Briard	Continental Herder	10	Brittany	Pointer Setter	10
Black Russian Terrier	Drover	4	Brussels Griffon	Toy Spitz	2
Basenji	-	10	Bulldog	European Mastiff	10
Bulmastiff	European Mastiff	10	Bull Terrier	European Mastiff	10
Cairn Terrier	Terrier	10	Cane Corso	European Mastiff	9
Cardigan Welsh Corgi	UK Rural	10	Curly Coated Retriever	Retriever	6
Chihuahua	American Toy	10	Japanese Chin	Asian Toy	4
Chow Chow	Asian Spitz	10	Cirneco dell'Etna	Mediterranean	5
Cavalier King Charles Spaniel	Spaniel	10	Collie	UK Rural	10
Chinook	New World	10	Coton du Tulear	Poodle	2
Cane Paratore	New World	2	Chinese Crested	American Toy	10
Dachshund	Scent Hound	10	Dalmatian	Pointer Setter	9
Great Dane	European Mastiff	10	Dogue de Bordeaux	European Mastiff	6
Scottish Deerhound	UK Rural	10	Doberman Pinscher	Drover	10
English Cocker Spaniel	Spaniel	10	English Setter	Pointer Setter	10
English Springer Spaniel	Spaniel	10	Eurasier	-	10
French Bulldog	European Mastiff	10	Flat-coated Retriever	Retriever	10
Field Spaniel	Spaniel	4	Finish Spitz	-	10
Foxhound	Scent Hound	10	Glen of Imaal Terrier	Terrier	9
Golden Retriever	Retriever	10	Gordon Setter	Pointer Setter	10
Great Pyrenees	Mediterranean	10	Greenland Sledge Dog	Asian Spitz	10
Greyhound	UK Rural	10	German Shepherd Dog	New World	10
German Shorthaired Pointer	Pointer Setter	10	Greater Swiss Mountain Dog	Alpine	6
Giant Schnauzer	Drover	10	German Wirehaired Pointer	Pointer Setter	2
Havanese	Poodle	10	Siberian Husky	Asian Spitz	10
Ibizan Hound	Mediterranean	10	Icelandic Sheepdog	Nordic Spitz	2
Peruvian Hairless dog	New World	10	Irish Terrier	Terrier	7
Irish Setter	Pointer Setter	9	Italian Greyhound	UK Rural	10
Irish Wolfhound UK	Rural	10	Irish Water Spaniel	Retriever	10
Jack Russell Terrier	Terrier	10	Keeshond	NordicSpitz	10
Kelpie	UK Rural	2	Kerry Blue Terrier	Terrier	4
Komondor	Mediterranean	2	Kuvasz	Mediterranean	10
Labrador Retriever	Retriever	10	Large Munsterlander	Pointer Setter 3	
Leonberger	Mediterranean	10	Lhasa Apso	Asian Toy	10
Levriero Meridionale	Mediterranean	2	Mastino Abruzzese	Mediterranean	2
Maltese	Poodle	10	English Mastiff	European Mastiff	10
Miniature Bull Terrier	European Mastiff	10	Toy Mnachester Terrier	Pinscher	2
Miniature Pinscher	Pinscher	10	Miniature Schnauzer	Schnauzer	10

**Table 5. Domestic dog breeds details (2)**

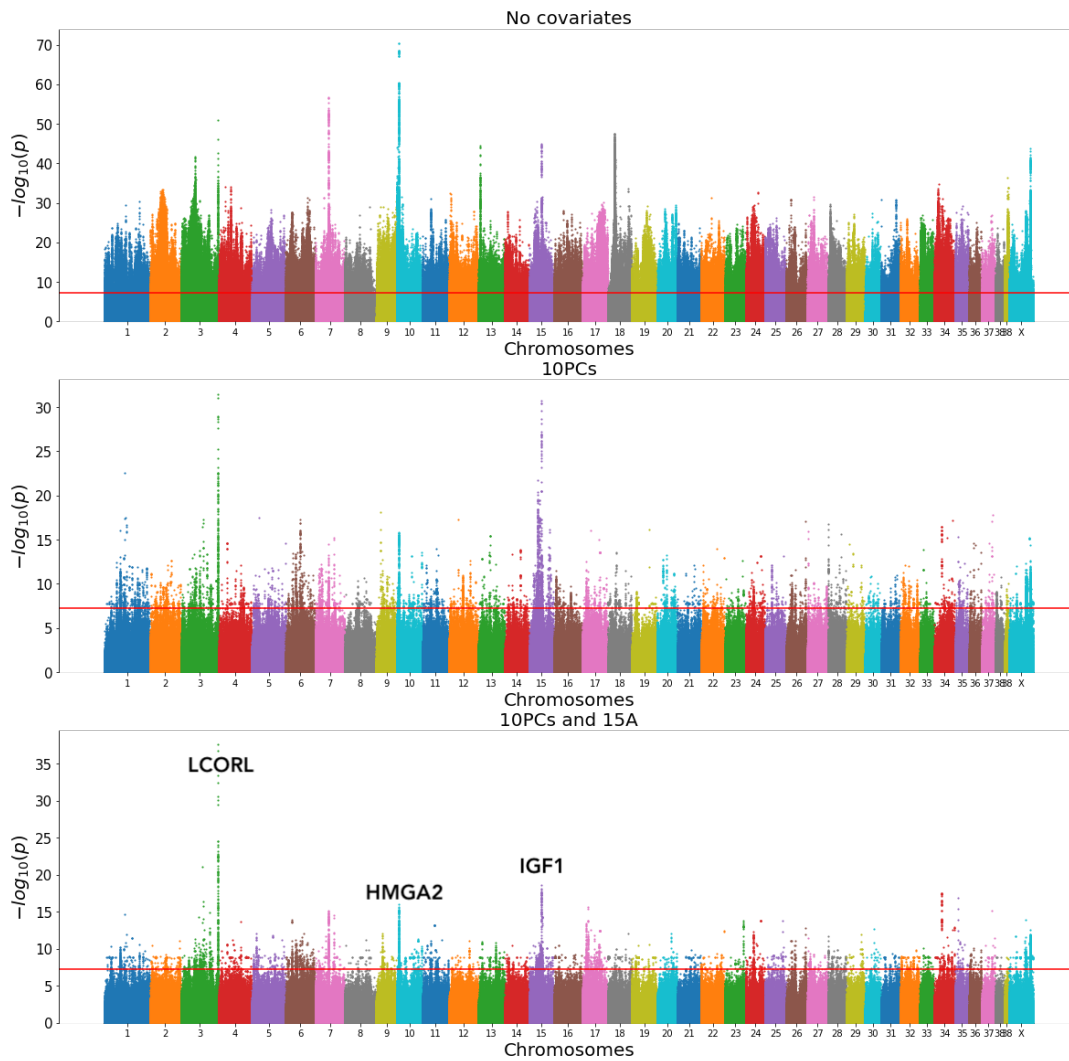
Breed	Clade	N	Breed	Clade	N
Neapolitan Mastiff	European Mastiff	6	Chinese Shar-pei	Asian Spitz	10
Norwegian Elkhound	Nordic Spitz	10	Shiba Inu	Asian Spitz	8
Newfoundland	Retriever	10	Shih Tzu	Asian Toy	10
Norfolk Terrier	Terrier	10	Silky Terrier	Terrier	4
Norwich Terrier	Terrier	10	Schipperke	Toy Spitz	10
Nova Scotia Duck Tolling Retriever	Retriever	10	Sloughi	Mediterranean	5
Old English Sheepdog	UK Rural	10	Spinone Italiano	Pointer Setter	2
Otter Hound	Scent Hound	9	Shetland Sheepdog	UK Rural	10
Papillon	Toy Spitz	10	Standard Schnauzer	Schnauzer	10
Parsons Russell Terrier	Terrier	2	Staffordshire Bull Terrier	European Mastiff	10
Petit Basset Griffon Vendéen	Scent Hound	10	Saint Bernard	Alpine	10
Pekingese	Asian Toy	10	Swedish Valhund	Nordic Spitz	6
Pembroke Welsh Corgi	UK Rural	10	Tibetan Mastiff	Asian Spitz	10
Pharaoh Hound	Mediterranean	2	Tibetan Spaniel	Asian Toy	10
Pomeranian	Small Spitz	10	Tibetan Terrier	-	10
Poodle - Miniature	Poodle	10	Belgian Tervuren	Continental Herder	10
Poodle - Standard	Poodle	10	Toy Fox Terrier	American Terrier	4
Poodle - Toy	Poodle	10	Vizsla	Pointer Setter	7
Portuguese Water Dog	Poodle	10	Volpino Italiano	Small Spitz	4
Pug Dog	Toy Spitz	10	Weimaraner	Pointer Setter	10
Puli	Hungarian	4	Wire Fox Terrier	Terrier	10
Pumi	Hungarian	5	Whippet	UK Rural	10
Rat Terrier	American Terrier	2	Wirehaired Pointing Griffon	Pointer Setter	6
Redbone Coonhound	Scent Hound	2	West Highland White Terrier	Terrier	10
Rhodesian Ridgeback	European Mastiff	9	Xigou	Asian Spitz	5
Rottweiler	Drover	10	Xoloitzcuintle	New World	5
Saluki	Mediterranean	19	Xoloitzcuintle - Miniature	New World	5
Samoyed	-	10	Yorkshire Terrier	Terrier	10
Scottish Terrier	Terrier	10	Grey Wolf	-	7
Soft Coated Wheaten Terrier	Terrier	4	Golden Jackal	-	2



**Fig 7. Archetypal Analysis compositional plots for human continental populations.** Archetypal Analysis polygon compositions of human data (3-9 archetypes, excepting the 8-archetype polygon which can be found as Fig. 2 in Section 4.1.1). The colours represent the continental origins: EUR - European (red), AFR - African (blue), EAS - East Asian (purple), WAS - West Asian (orange), OCE - Oceanian (brown), SAS - South Asian (pink), AMR - American (gray).

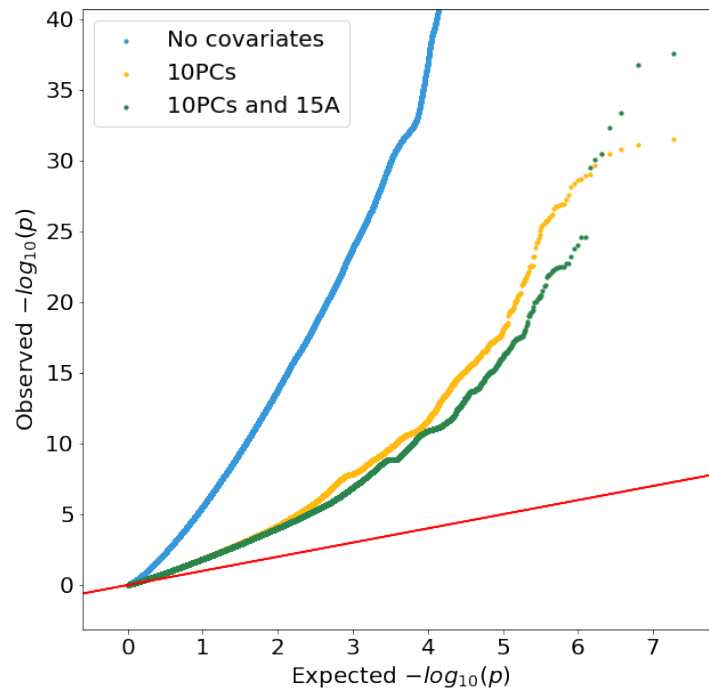


**Fig 8. Archetypal Analysis compositional plots for human continental populations.** Archetypal Analysis polygon compositions of human data (10-15 archetypes). The colours represent the continental origins: EUR - European (red), AFR - African (blue), EAS - East Asian (purple), WAS - West Asian (orange), OCE - Oceanian (brown), SAS - South Asian (pink), AMR - American (gray).

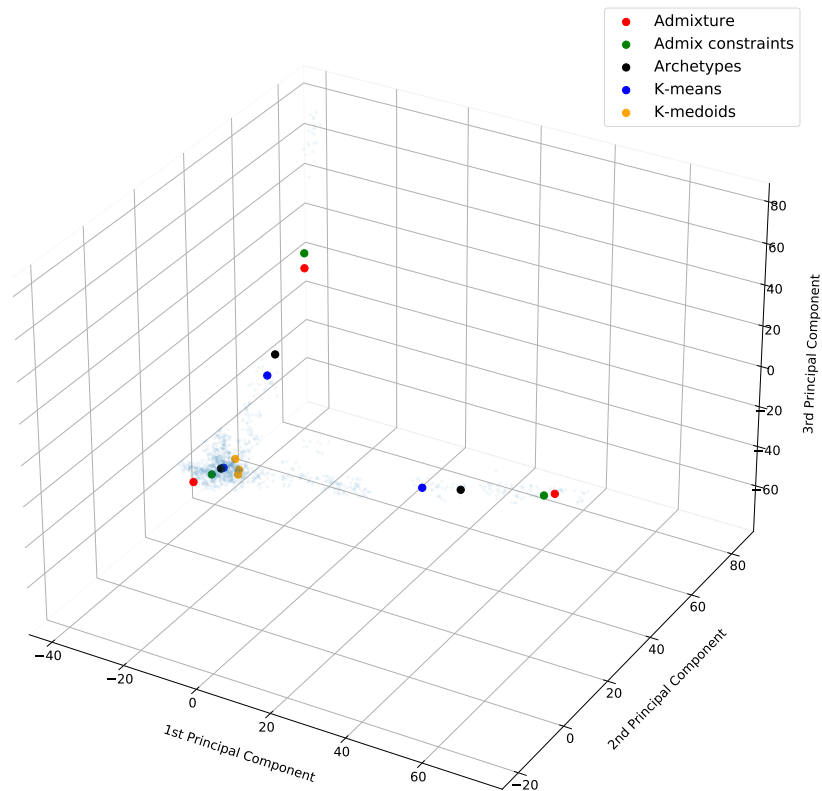


**Fig 9. Manhattan plot** Manhattan plots for no covariates (top), when adding 10 PCs (middle) and when adding 10 PCs and 15 Archetype coefficients (bottom).

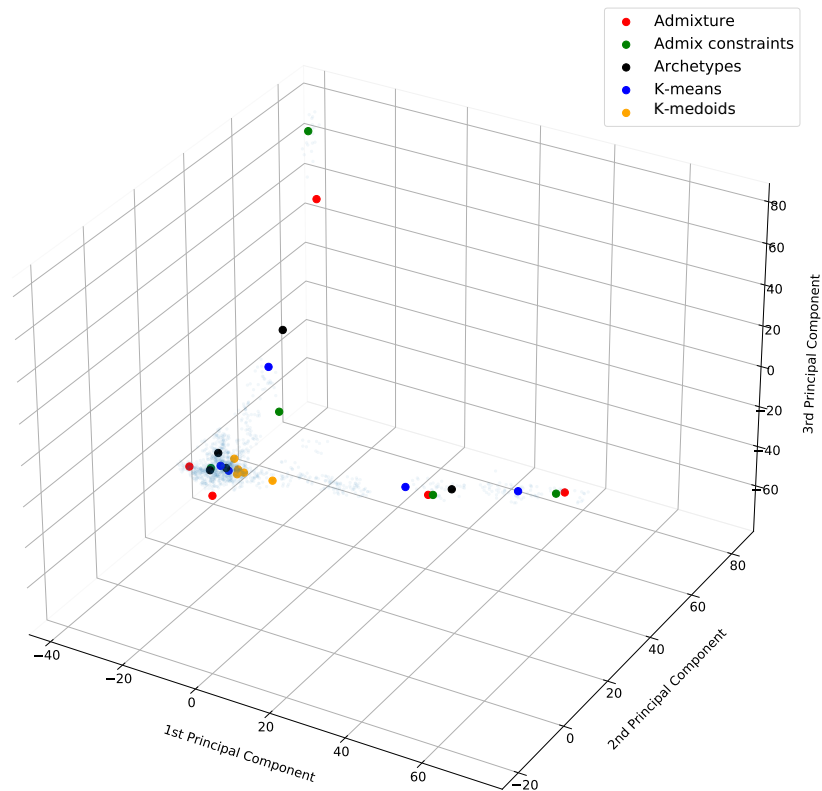




**Fig 10. Quantile-Quantile plot (Q-Q)** Q-Q plots for no covariates (blue), when adding 10 PCs (yellow) and when adding 10 PCs and 15 Archetype coefficients (green).



**Fig 11. Comparison of cluster centroids from different methods.** Cluster centroids learned by ADMIXTURE, ADMIXTURE with sparsity regularization, Archetypal Analysis, K-Means, and K-Medoids



**Fig 12. Comparison of cluster centroids from different methods.** Cluster centroids learned by ADMIXTURE, ADMIXTURE with sparsity regularization, Archetypal Analysis, K-Means, and K-Medoids