Multidimensional Scaling and Relatedness

Research

Jan Graffelman^{1,2}

Iván Galván Femenía^{3,4}

Rafael de Cid⁴

Carles Barceló-i-Vidal³

¹Department of Statistics and Operations Research
Universitat Politècnica de Catalunya, Barcelona, Spain

²Department of Biostatistics
University of Washington, Seattle, USA

³Department of Computer Science, Applied Mathematics & Statistics
Universitat de Girona, Girona, Spain

⁴Genomes For Life - GCAT lab
Program of Predictive and Personalized Medicine of Cancer (PMPPC)
Institute for Health Science Research Germans Trias i Pujol (IGTP)
Can Ruti Campus, Badalona, Barcelona, Spain

- Running title: Multidimensional scaling and relatedness
- Corresponding author: Jan Graffelman, Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Avinguda Diagonal 647, 08028 Barcelona, Spain. email: jan.graffelman@upc.edu, tel: 00-34-934011739, fax: 00-34-934016575

Abstract

Multidimensional scaling is a well-known multivariate technique, that is often used in genetics for studying population substructure. In this paper we show that multidimensional scaling of marker data is of relevance for relatedness research. Relatedness is usually investigated by estimating and plotting identity-by-state and identity-by-descent allele sharing statistics. We show that outlying individuals in a map obtained by multidimensional scaling of genetic variables do not necessarily stem from a different human population, but can be the consequence of relatedness. We propose a method for classifying pairs of individuals into the standard relationship categories that combines genetic bootstrapping, multidimensional scaling and discriminant analysis. We validate our method with simulation studies. Given the variant filtering procedures, our method classifies relationships up to and including the fourth degree with high accuracy (96-97%), using only identity by state. The usefulness of the method is illustrated with data from the 1,000 genomes and the GCAT projects.

Key words: allele sharing, composition, genetic bootstrapping, identity by state, identity by descent, log-ratio transformation.

1 Introduction

Multidimensional scaling (MDS) is a versatile multivariate statistical technique that has found applications in many fields of science [Cox and Cox, 2001, Borg and Groenen, 2005]. In genetics, MDS is often used to investigate the existence of population substructure. If individuals do not come from a single homogeneous human population, but from two or more populations

with different allele frequencies, then an MDS of the genetic marker data typically separates the individuals of the different populations. An example is given in Figure 1 where we show the MDS map for a sample of 368 individuals, 165 from the CEU population, and 203 from the YRI population of the HapMap project [The International HapMap Consortium, 2007], now included in the 1,000 genomes project [The 1000 Genomes Project Consortium, 2015], using the allele-sharing distance [Mountain and Cavalli-Sforza, 1997]. Many examples of the use of MDS for detecting population substructure can be found in the genetics literature. Pemberton *et al.* (2010) used MDS to investigate the population affiliation of the HapMap Phase III individuals. Other examples of population substructure research with MDS are given by Jakobsen *et al.* (2008), Sabatti (2009), Foulkes (2009), Wang (2010) and Pemberton *et al.* (2013).

The examples in the literature and in Figure 1 suggest that observing outlying individuals in an MDS map implies these individuals are different, and therefore potentially stem from a different human populations that may be characterized by different allele frequencies for the variants under study. In this paper we show that this conclusion is not always warranted. Outlying individuals, and even groups of outlying individuals, may very well stem from the *same* human population and appear separated from their population for a very different reason: they correspond to individuals that are closely related. Figure 1 illustrates this point, where some interrelated individuals of the YRI population appear separated from their alike. In particular, the map distance of the most outlying individuals with respect to the center of the YRI population is actually *larger* than the distance between the centers of the YRI and the

Figure 1: MDS map for a sample of 165 individual of the CEU and 203 individuals of the YRI population of the 1,000 genomes project, using 48,072 LD-pruned, autosomal SNPs with $MAF > 0.40$ and non-significant in a HWE exact test.

CEU population. The four most outlying individuals are all involved in two parent-offspring (PO) relationships, whereas the two intermediate outliers are involved in one PO relationship (Pemberton *et al.* (2010), supplemental data).

Relatedness investigations in genetics are mostly based on allele-sharing statistics such as the average number of IBS alleles shared by a pair of individuals over a set of loci, or by estimating the probabilities of sharing 0, 1 or 2 alleles IBD [Cotterman, 1941, Thompson, 1991], known as Cotterman's coefficients. Plots of these sharing statistics typically show clusters that correspond to unrelated individuals (UN), parent-offspring pairs (PO), full sibs (FS), half sibs (HS), monozygotic twins (MZ), avuncular pairs (AV), first cousins (FC), grandparent-grandchild (GG) or more remote relationships. MDS is not used in such relatedness investigation, but is linked to the allele-sharing approach in the sense that a distance matrix can easily be computed from the allele sharing statistics. Moreover, recent studies [Weir and Goudet, 2017, Conomos et al., 2016] have pointed out that relatedness and population substructure are tightly connected topics in the sense that both induce allele sharing.

The remainder of this paper is organized as follows. In Section 2 we provide background on relatedness research and MDS, and we develop theory for using MDS in relatedness research by calculating genetic distances between pairs of individuals. Section 3 explores and validates our methodology with some simulation studies, that assess the effects of sample size, number of variants and minor allele frequency (MAF) on the classification error rate obtained by a discriminant analysis based on MDS results. Section 4 shows applications of the proposed methodology with data from the 1,000 genomes and GCAT projects. A Discussion section (5) finishes the paper.

2 Theory

Relatedness research

We briefly review some fairly standard procedures that are currently used in relatedness research. Relatedness investigations are focused on the extent to which alleles are shared between individuals. Two individuals can share 0, 1 or 2 alleles for any autosomal variant. Alleles can be identical by state (IBS) or identical by descent (IBD). A pair of individuals share IBS alleles if they match irrespective of their provenance; whereas they share IBD alleles only if they come from a common ancestor. In IBS studies, the means and standard deviations of the IBS allele counts over variants [Abecasis et al., 2001], or the proportions of variants sharing 0, 1 and 2 IBS alleles [Rosenberg, 2006] can be plotted. These plots reveal characteristic clusters corresponding to MZ, PO, FS, HS/AV/GG, FC or UN pairs. Alternatively, in an IBD based approach, the probability of sharing 0, 1 or 2 IBD alleles (usually denoted by $\hat{k}_0$, $\hat{k}_1$ and $\hat{k}_2$) can be represented in a scatterplot [Nembot-Simo and Graham, 2013]. The Cotterman coefficients can be estimated by the method of moments [Purcell et al., 2007] or by maximum likelihood [Thompson, 1975, Milligan, 2003, Weir et al., 2006]. A good account of the state-of-the-art of IBD relatedness research is given by Weir *et al.* (2006). In IBD studies, reference values for the standard relationships are available (see Table 1). Related pairs can also be distinguished, albeit at lower resolution, by using the coancestry coefficient defined as $\theta = k_1/2 + k_2$ or the kinship coefficient defined as $\phi = \theta/2$. An example of the use of ϕ is shown by Conomos *et al.* (2016) where plots of the kinship coefficient and k_0 also uncover clusters with the different family relationships. Galván *et al.* (2017) give an overview of graphics for related-

ness research. Figure 2 shows a panel plot of some standard graphics used in IBS and IBD studies for all the pairs of individuals from the CEU population of the HapMap project. These plots distinguish UN, PO, FS and second degree pairs. First and second degree relationships for the CEU population were documented by Pemberton *et al.* (2010) using IBS methods, and confirmed by Kyriazopoulou *et al.* (2011), who used hidden Markov models and suggested additional third and fourth degree relationships. Stevens *et al.* (2012) used IBD methods confirming the results of Pemberton *et al.* (2010).

Type of Relative	R	ϕ	IBD Probabilities		
			k_0	k_1	k_2
Monozygotic twins (MZ)	0	1/2	0	0	1
Full-siblings (FS)	1	1/4	1/4	1/2	1/4
Parent-offspring (PO)	1	1/4	0	1	0
Half-siblings/ grandchild-grandparent/ niece/nephew-uncle/aunt (HS,GG,AV)	2	1/8	1/2	1/2	0
First cousins (FC)	3	1/16	3/4	1/4	0
Unrelated (UN)	∞	0	1	0	0

Table 1: Degree of relationship (R), kinship coefficient (ϕ), and probability of sharing zero, one or two alleles identical by descent (k_0 , k_1 , k_2).

Alternatively, a Markov-chain approach with the calculation of likelihood ratios for putative and alternative relationship has been developed by Epstein *et al.* (2000) (the Relpair program) and by McPeck and Sun (2000) (the Prest-plus program).

Multidimensional scaling and relatedness research

Multidimensional scaling (MDS) is a well-known multivariate method with many applications in genetics, though its potential in relatedness research seems not to have been explored. Several variants of multidimensional scal-

Figure 2: Some standard graphics for relatedness research for the CEU sample of the HapMap project. IBS/IBD statistics were calculated over a set of 26,081 complete, LD-pruned autosomal SNPs with MAF above 0.4, and HWE exact test p-value above 0.05. A: scatterplot of the mean and standard deviation of the number of IBS alleles. B: scatterplot of the fraction of variants sharing 2 (p_2) against the fraction sharing 0 (p_0) IBS alleles. C: scatterplot of k_1 against k_0 . D: scatterplot of ϕ against k_0 .

ing exist and a distinction is usually drawn between metric and non-metric multidimensional scaling. Metric MDS is also known as classical scaling or principal coordinate analysis (PCO, Gower [1966]). In metric MDS a map of the individuals is obtained by extracting eigenvalues and eigenvectors from a conveniently transformed distance matrix. The distances in the map approximate the (genetic) distances between the individuals. The method is not invariant under monotone transformations of the original distances. Non-metric multidimensional scaling obtains the map in an iterative manner by minimizing a stress function, and is invariant under monotone transformations of the distances, as it uses only the rank order of the distances. In this paper we confine ourselves mostly to metric MDS. The theory of MDS is described in most textbooks on multivariate analysis [Mardia et al., 1979] and some books are entirely dedicated to the topic [Cox and Cox, 2001, Borg and Groenen, 2005].

Figure 3 shows metric MDS maps for simulated populations with some related individuals. Bi-allelic marker data was generated by sampling independent genotypes from a multinomial distribution with Hardy-Weinberg frequencies, using uniformly distributed allele frequencies. Some related individuals were then generated from the base set of independent individuals. Figure 3A shows that a single monozygotic twin in the database stands out as an outlier in the map (the two twin-sibs show up as two coinciding points in the first dimension). The existence of two identical points in MDS can also be inferred from the eigenvalues, as two identical points create an additional zero eigenvalue in the MDS output (one zero eigenvalue is always present and is a characteristic of the method as it results from the double-centring of the transformed distance matrix). At first sight the appearance of related individ-

Figure 3: MDS maps for simulated populations with some related individuals. (F=Father, M=Mother, C=Child, MZ=Monozygotic twin, FS=Full Sib, HS=Half Sibs)

uals as outliers seems contradictory. The MZ pair in panel (A) is shown at a large distance from the rest of the population suggesting that the genetic distance between a twin-sib and the rest of the individuals is much larger than any of the distances between any other pair of unrelated non-twin-sib individuals. This is actually not true. The genetic distances between a twin-sib and the other individuals are in fact of similar order of magnitude as the distances between any pair of unrelated individuals. The reason that the MZ pair separates out in the map at a larger distance is due to the fact that the MZ pair constitutes an outlier in the analysis. The genetic distance between the twin-sibs is zero, and this is an extreme outlier in the off-diagonal entries of the distance matrix, where all distances between unrelated pairs are much larger. MDS is sensitive to the existence of outliers and fits the outlying point much better than the other points. In regression, it is well-known that a single outlying observation can exert leverage on the slope of the regression line and can highly influence the slope [Weisberg, 2005, Chapter 9]. This phenomenon also arises in MDS, where a single outlying observation exerts leverage on the fitted plane. Consequently, the goodness-of-fit of all distances involving the twins are all high, whereas the goodness-of-fit of all other distances is very low, as is revealed in a diagnostic plot of fitted versus observed distances (see supplementary Figure S1). Similarly, most of the maps shown in Figure 3 containing some related individuals are pretty much uninformative on most of the distances, except for those distances that involve the outlying points.

MDS with pairs of individuals

An important difference between the standard allele-sharing methods outlined in the first paragraph of this Section and MDS as discussed above is that the allele-sharing methods plot *pairs* of individuals, whereas MDS is usually employed to plot the *individuals themselves*. To bring MDS maps to the same realm as the allele sharing techniques, we propose to compute a distance matrix between all possible *pairs* of individuals rather than the distance matrix between the individuals. In order to do so, we need to define a distance (or similarity) measure between two pairs of individuals. This can be done in many different ways. The distance measure needs to indicate to what extent two pairs of individuals are similar or not. We use n for the number of individuals, and k for the number of genetic variants. For bi-allelic variants there exist six possible pairs of genotypes whose counts over k variants can be laid out in a triangular array shown in Table 2, where k_{ij} refers to the number of variants that have i B alleles for one individual, and j B alleles for the other individual.

1st indiv.	AA	k_{00}		
	AB	k_{10}	k_{11}	
	BB	k_{20}	k_{21}	k_{22}
		AA	AB	BB
		2nd indiv.		

Table 2: Lower triangular matrix layout with counts for all possible genotype pairs.

Consequently, each pair can be represented by a vector of six counts which can be expressed as a composition by division by its total

$$\mathbf{x} = (k_{00}, k_{10}, k_{20}, k_{11}, k_{21}, k_{22})/k. \quad (1)$$

The total number of variants is given by $k = \sum_{i \geq j} k_{ij}$. For PO pairs this vector has, in theory, a structural zero, $k_{20} = 0$, because PO pairs share at least one IBS allele. However, for empirical data $k_{20} = 0$ is, with large k , never observed due to the existence of some mutations and genotyping error. In order to measure the distance between two pairs, $\mathbf{x}$ and $\mathbf{y}$, we consider the ordinary Euclidean distance (d_e), the Manhattan distance (d_m), the Canberra distance (d_c) and the Aitchison distance [Aitchison, 1986], d_a , where the latter is appealing for being specifically designed for compositional data. These distances are given by

$$d_e(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^D (x_i - y_i)^2}, \quad (2)$$

$$d_m(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^D |x_i - y_i|, \quad (3)$$

$$d_c(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^D \frac{|x_i - y_i|}{|x_i| + |y_i|}, \quad (4)$$

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^D \left[\ln \left(\frac{x_i}{g(\mathbf{x})} \right) - \ln \left(\frac{y_i}{g(\mathbf{y})} \right) \right]^2}, \quad (5)$$

where D represents the number of parts of the composition (here $D = 6$), and $g(\cdot)$ the geometric mean of the composition. The Aitchison distance requires strictly positive elements, which is theoretically a problem for PO pairs. Methodology for dealing with zeros in compositions has been developed (Martín-Fernández *et al.*, 2011). Additional metrics could be explored,

but are not considered here.

A drawback of the representation of pairs of individuals in an MDS map is that the type of relationship cannot be inferred. Outlying points observed in pairwise MDS maps are likely to correspond to related pairs of individuals, but without additional analysis one does not know whether the outlier is an MZ, PO, FS, HS or other pair. We resolve this by first identifying a subset of approximately unrelated individuals in the database, having a co-ancestry coefficient with other individuals that is below 0.05. We next simulate pairs of related individuals of known relationships from this subset, applying the Mendelian inheritance rules. E.g., PO pairs are simulated by first drawing two parents at random from the unrelated subset. A child is then simulated by drawing one allele at random from both these parents. The process is repeated in order to generate many random PO pairs. FS, HS and pairs of other relationships are simulated in an analogous manner. This process is based on re-sampling the alleles of the individuals, and we will refer to this principle as *genetic bootstrapping*. The artificially generated data set forms a *reference set* or *training set* against which the empirically observed data can be compared. This reference set is conditional on the allele frequencies of the observed sample. The distance matrix of all simulated pairs is calculated and used to construct an MDS reference map. The empirically observed pairs are projected onto this MDS map and their relationship is inferred, according to which simulated type of relationship is most close to the empirical pair. This can be done in a quantitative way by classifying all empirical pairs with linear discriminant analysis, using the simulated pairs as a training set. We briefly outline the calculations below.

MDS [Mardia et al., 1979, Chapter 14] is performed by the spectral de-

composition in Equation (6) of the $m \times m$ scalar product matrix $\mathbf{B}$, which is obtained from the double centred matrix with the squared distances ($\mathbf{D}$) as $\mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{D}\mathbf{H}$, where $\mathbf{H}$ is an idempotent centring matrix ($\mathbf{H} = \mathbf{I} - (1/m)\mathbf{1}\mathbf{1}'$),

$$\mathbf{B} = \mathbf{V}\mathbf{D}_\lambda\mathbf{V}' = \mathbf{X}\mathbf{X}', \quad \text{with } \mathbf{X} = \mathbf{V}\mathbf{D}_\lambda^{\frac{1}{2}}. \quad (6)$$

Matrix $\mathbf{V}$ contains the eigenvectors of $\mathbf{B}$, and $\mathbf{D}_\lambda$ is a diagonal matrix containing the eigenvalues in non-increasing order of magnitude. Matrix $\mathbf{X}$ contains the coordinates of the pairs in an MDS map of dimension r , the dimensionality chosen for the representation of the results.

Let $\mathbf{d}$ be the $m \times 1$ vector of supplementary distances, containing the squared distances between an empirical pair and all m pairs in the reference set. The squared transformed distances $\tilde{\mathbf{d}}$ are obtained as

$$\tilde{\mathbf{d}} = \mathbf{d} - \frac{1}{m}\mathbf{D}\mathbf{1} + \mathbf{1} \left(\frac{1}{2m^2}\mathbf{1}'\mathbf{D}\mathbf{1} \right). \quad (7)$$

This transformation amounts to subtracting the average of the squared distance vector and adding half the mean of the squared entries of the original distance matrix calculated for the reference set. The projection of the empirical pairs is carried out by using results that plot supplementary points in an MDS map. In particular, the coordinates of a projected pair, $\mathbf{z}$, are found by [Gower, 1968]

$$\mathbf{z} = -\frac{1}{2}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{d}}, \quad (8)$$

where the resulting coordinate vector $\mathbf{z}$ is $r \times 1$, and has as many elements as the dimensionality chosen for representation of the results. The validity of (7) and (8) is easily verified by taking a copy of one row of the original distance matrix, and applying the projection equations. The map coordinates of the projected point will then coincide exactly with the coordinates of that point in the initial MDS solution. Equations (7) and (8) exemplify that a supplementary point is found by first centring (and scaling) the observation in the same way the original observations are centered (and scaled), followed by a regression [Graffelman and Aluja-Banet, 2003].

3 Simulations

We first validate the proposed methodology with some simulations. We simulated 5,000 independent genetic bi-allelic variants by sampling from a multinomial distribution under the Hardy-Weinberg assumption, using a minor allele frequency of 0.5 for all variants. Using Mendelian inheritance rules, 100 independent pairs of each type of relationship were simulated. We used two scenarios, one with perfect application of Mendelian inheritance rules, yielding simulated data sets that are free of Mendelian inconsistencies, and one with 1% genotyping error for all three genotypes. Two-dimensional metric MDS maps using four distance measures (Euclidean, Manhattan, Canberra and Aitchison) for the six-part compositions of the pairs are shown in Figure 4. An MDS with the Euclidean distance is seen to separate the main types of relationships until the fourth degree (UN, fourth, third, and second degree, FS and PO), with overlapping clusters for UN, fifth and fourth degree. The FS group is most outlying. The Manhattan metric gives a picture that is simi-

Figure 4: MDS maps for simulated samples, with 1% genotyping error. 100 pairs of each type of relationship (UN, fifth, fourth, third (FC), second (HS), FS and PO) were generated using an allele frequencies of 0.5, assuming Hardy-Weinberg equilibrium and 5,000 independent bi-allelic variants.

lar to the Euclidean distance, but with apparently less discrimination between UN, fifth, fourth, third and second degree relationships. The Canberra metric, which is related to the Manhattan metric, places PO and FS pairs at similar positions in the first dimension, and separates them in the second. Results obtained with the Aitchison metric differ, as this metric strongly sorts out the PO pairs in the first dimension, inverting the order of the PO and FS clusters in this dimension. Scatterplots of the MDS coordinates beyond the second dimension show, in general, little separation of the different relationships, confirming that the solution is essentially two-dimensional. Exceptions are the Canberra and Aitchison distance, which do show clear clustering of relationships in the third dimension (see supplementary figure S2). Classification rates for a linear discriminant analysis, using the first three MDS dimensions, are given for the four metrics in Table 3.

		Metric			
	Genotyping error	Euclidean	Manhattan	Canberra	Aitchison
Overall	0%	0.910	0.899	0.914	0.915
	1%	0.910	0.884	0.906	0.851
Up to 4th	0%	0.966	0.952	0.960	0.964
	1%	0.952	0.942	0.946	0.934

Table 3: Overall and up to and including fourth degree relationships classification rates for linear discriminant analysis using four different metrics, and applying genetic bootstrapping without genotyping error, or with 1% genotyping error.

The non-perfect overall classification can mainly be ascribed to the fifth degree pairs, who are partly misclassified as fourth degree or unrelated. The classification rates up to and including the fourth degree pairs are about five percent higher. The differences between the metrics are small, though the Euclidean distance seems to work best, in particular in the presence of genotyping error. Somewhat better classification rates are obtained if genotyping

error is absent. When there is no genotyping error, there are structural zeros in the compositions of PO pairs, which were dealt with using methods outlined by Martín-Fernández *et al.* (2011). Similar results were obtained by using quadratic discriminant analysis (not shown).

This simulation concerns a relatively ideal dataset with independent variants and maximally polymorphic variants. For empirical data sets, the independence of the variants can be approximately achieved by LD pruning variants. In practice, many variants have a low MAF. We investigated the effect of the MAF on the discriminatory power of the different metrics, by simulating variants with different MAF. Figure 5 shows how the classification rate varies as a function of the MAF and the number of bootstrap samples for the considered metrics.

Figure 5 shows a worse classification rate for the Aitchison metric, whereas Euclidean, Manhattan and Canberra metrics achieve the best classification with high MAF variants. The observed classification rates are similar for a training set with 20 and 100 simulated pairs. The classification rate depends on the number of maximally polymorphic variants, and suggests that 25,000 variants are sufficient to almost perfectly classify PO, FS, second, third and fourth degree relationships. In the light of the simulation results, we will use one hundred bootstrap samples and the Euclidean distance for the examples presented in Section 4. In the remainder of the paper, we will refer to our method, which combines genetic bootstrapping, IBS allele sharing, MDS and LDA, as IBS-MDS.

Figure 5: Classification rate up to and including the fourth degree as a function of the minor allele frequency (MAF) and the number of maximally polymorphic variants obtained by linear discriminant analysis using four different metrics, and allowing for 1% genotyping error.

4 Applications

In this Section we apply the proposed MDS of pairwise distances to two genomic data sets. We use the CEU population of the HapMap project, now available as part of the 1,000 genomes project [The 1000 Genomes Project Consortium, 2015] (www.internationalgenome.org), whose family relationships have been analysed in detail by Pemberton *et al.* (2010), Kyriazopoulou *et al.* (2011), Huff *et al.* (2011) and Stevens *et al.* (2011, 2012). We also present a relatedness study of the population based GCAT Genomes for Life project (a cohort study of the genomes of Catalonia, www.genomesforlife.com, Obón-Santacana *et al.* [2018]). For both projects, we used Plink 1.90 [Purcell *et al.*, 2007] for data manipulation and filtering, and R [R Core Team, 2014] for MDS and discriminant analysis.

4.1 The CEU sample

We detail the analysis of the CEU panel. HapMap phase II and III variants of the CEU panel were filtered according to missingness (only variants genotyped for all individuals were used), MAF (> 0.40) and Hardy-Weinberg equilibrium test result (using the exact mid p-value > 0.05 [Graffelman and Moreno, 2013]). Variants were LD-pruned with Plink's independent pairwise option (indep-pairwise 50 5 0.2). The final data set contained 31,370 autosomal variants. The CEU panel consists of 165 individuals, mainly PO trios, giving 13,530 possible pairs of individuals. The ordinary MDS map of the individuals of this panel is shown in Figure 6. Documented related individuals are connected by lines. Figure 6 shows an outlying FS pair in the first dimension, that pulls the related families along this axis.

Figure 6: MDS map for a sample of 165 individuals of the CEU population, with outlying FS pair in the first dimension

Figure 7 shows the MDS map of the pairwise approach, using the Euclidean distance measure. The convex hulls delimit the cloud of the positions of the simulated UN, fifth, fourth, third, second, FS and PO pairs (using 100 pairs of each). The classification rate of the simulated data was as high as 99.6%. Points with open circles represent the empirical pairs which were projected onto the map using Equations (7) and (8), and are coloured according to their predicted relationship. The map suggests the CEU sample has 96 PO pairs, one FS pair, two second degree pairs, one third degree pair, five fourth degree pairs and many fifth degree pairs that merge with UN pairs. The classification of the empirical pairs by linear discriminant analysis confirmed the 96 PO and the single FS pair relationships described by Pemberton *et al.* (2010) (results not shown), as well as the additional FC pair reported by Kyriazopoulou *et al.* (2011), with a posterior probability of one for all these pairs. Third and fourth degree relationships uncovered by Kyriazopoulou *et al.* (2011) are reported in Table 4, together with the posterior probabilities obtained in our IBS-MDS analysis. 193 pairs were classified as fifth degree relationship pairs, of which 66 had a posterior probability above 0.95. We do not report all these pairs because given the overlap with the UN cluster and the poorer classification rate of the fifth degree observed in the simulations.

Figure 7: MDS map of 13,530 pairs of individuals of the CEU population. Convex hulls delimit the region of the pairs obtained by genetic bootstrapping.

Pair	ID1	sex	ID2	sex	Pem.	Kyr.	Ste.	Huf.	IBS-MDS	Posterior probabilities						
										UN	5th	4th	3rd	2nd	FS	PO
1	NA06997	2	NA12801	1	-	FC	FC	-	3rd	0.000	0.000	0.000	1.000	0.000	0.000	0.000
2	NA06993	1	NA07056	2	-	4th	-	4th	4th	0.000	0.000	1.000	0.000	0.000	0.000	0.000
3	NA06993	1	NA07022	1	-	4th	-	4th	4th	0.000	0.000	1.000	0.000	0.000	0.000	0.000
4	NA07031	2	NA12043	1	-	4th	-	-	4th	0.000	0.000	1.000	0.000	0.000	0.000	0.000
5	NA12155	1	NA12264	1	-	4th	-	4th	4th	0.000	0.000	1.000	0.000	0.000	0.000	0.000
6	NA12760	1	NA12830	2	-	FC	-	-	4th	0.000	0.000	1.000	0.000	0.000	0.000	0.000
7	NA06989	2	NA12155	1	-	4th	-	-	5th	0.000	0.999	0.001	0.000	0.000	0.000	0.000
8	NA06991	2	NA07022	1	-	4th	-	-	5th	0.000	0.997	0.003	0.000	0.000	0.000	0.000
9	NA06994	1	NA12892	2	-	4th	-	5th	5th	0.000	1.000	0.000	0.000	0.000	0.000	0.000
10	NA07014	2	NA12043	1	-	4th	-	-	5th	0.000	1.000	0.000	0.000	0.000	0.000	0.000
11	NA07031	2	NA12761	2	-	4th	-	-	5th	0.000	0.994	0.006	0.000	0.000	0.000	0.000
12	NA10831	2	NA12264	1	-	FCOR	-	-	5th	0.003	0.997	0.000	0.000	0.000	0.000	0.000
13	NA10863	2	NA12155	1	-	FCOR	-	-	5th	0.000	1.000	0.000	0.000	0.000	0.000	0.000
14	NA11931	2	NA12748	1	-	4th	-	-	5th	0.000	1.000	0.000	0.000	0.000	0.000	0.000
15	NA12752	1	NA12830	2	-	FCOR	-	-	5th	0.000	0.989	0.011	0.000	0.000	0.000	0.000
16	NA12752	1	NA12818	2	-	4th	-	-	5th	0.406	0.594	0.000	0.000	0.000	0.000	0.000
17	NA12760	1	NA12818	2	-	FCOR	-	-	5th	0.000	1.000	0.000	0.000	0.000	0.000	0.000

Table 4: Predicted relationships of third (3rd) and fourth (4th) degree pairs of the CEU sample of the HapMap project reported by Kyriazopoulou *et al.* (2011) and their posterior probabilities according to IBS-MDS. Coding and abbreviations used: sex 1 = male, 2 = female; a hyphen (-) indicates the corresponding pair is not annotated or regarded unknown by the corresponding authors; FCOR: first cousin once removed; Pem.: Pemberton *et al.* (2010); Kyr.: Kyriazopoulou *et al.* (2011); Ste.: Stevens *et al.* (2012); Huf.: Huff *et al.* (2011).

Our results confirm a third degree pair (pair 1 in Table 4) reported by Kyr-iazopoulou *et al.* (2011). We also confirm four of the fourth degree pairs reported by the latter authors (pairs 2-5 in Table 4). However, we also observed considerably incongruence of our results with those of the latter authors. We found an FC pair to be classified as fourth degree (pair 6) by our method and 11 reported fourth degree pairs were classified as fifth degree. One of these (pair 16) had a high probability of being even more remotely related. We also compared results with those published by Huff *et al.* (2011), who estimate recent shared ancestry (ERSA) by using IBD segments. Our work confirms three fourth degree pairs and one fifth degree pair report by the latter authors, though we found two additional fourth degree pairs which are not confirmed by Huff *et al.* (2011).

4.2 GCAT samples

We use samples from the GCAT Genomes for life project, a cohort study of the genomes of Catalonia (www.genomesforlife.com). GCAT is a prospective cohort study that includes 17,924 participants (40-65 years, release August 2017) recruited from the general population of Catalonia, a Mediterranean region in the northeast of Spain. Participants are mainly part of the Blood and Tissue Bank (BST), a public agency of the Catalan Department of Health. Detailed information regarding the GCAT project is described in Obón-Santacana *et al.* (2018). We study relatedness on 5,075 GCAT Spanish participants from Caucasian origin and 736,223 SNPs that passed quality control [Sumoy *et al.*, 2017]. Inferred relatives of first and second degree were confirmed by the BST public agency, for pairs sharing one surname (PO, sec-

Figure 8: MDS maps of 5,075 individuals of the GCAT study.

ond degree pairs) or two surnames (FS pairs), respecting the privacy of the participants. According to the same filtering procedures used in the CEU samples, 26,006 SNPs ($MAF > 0.40$, LD-pruned, HWE exact mid p-value > 0.05 , and missing call rate 0) were considered for relatedness analysis. A MDS map of the individuals, using the allele sharing distance, is shown in Figure 8A.

The first dimension shows an outlying triple of sibs (pairs 48, 49 and 50 in Table S1, Figure 8A). When considering second and third dimensions, this trio is masked, and these axes reveal other FS pairs (Figure 8B). Figure 9 shows the IBS-MDS approach for pairs of individuals using the Euclidean distance. Our IBS-MDS method detects four PO, 70 FS and 12 second degree pairs, and reveals 68 and 54 third and fourth degree pairs (See supplementary Table S1). The IBS-MDS-plot shows a large variability in the FS group, where many pairs fall outside their respective simulated convex hull. We suggest this excess variability is, at least in part, due to the existence of some three-quarter

siblings in the database (e.g. half-sibs where two of the three parents have an FS or PO relationship). The Cotterman coefficients for a three-quarter sibling relationship are readily deduced to be $k_0 = 3/8$, $k_1 = 1/2$ and $k_2 = 1/8$, and their co-ancestry coefficient is $\theta = 3/8$, or, equivalently, their kinship coefficient is $3/16$. They can therefore be expected to fall in-between the FS and 2nd degree clusters, as is shown for some labeled pairs in Figure 9. We therefore repeated IBS-MDS with a new training set where this type of relationship had been included. The resulting map is shown in Figure 10 and shows that one outlying second degree pair (77) and several initially outlying FS pairs (73, 63, 53, 43) are now indeed classified as three-quarter siblings. The posterior probabilities for these classifications are given in supplementary Table S2.

5 Conclusions and Discussion

Multidimensional scaling is a widely used tool in genetic analysis, and is often used to check data for homogeneity and for substructure investigation. If there exists population substructure, MDS maps typically reveal clusters corresponding to the different populations. However, outlying observations or groups of outlying observations do not necessarily *prove* the existence of population substructure. Individuals may separate out in MDS because they are related but actually stem from the *same* population. We conclude that population substructure cannot be ruled out or confirmed only by looking at MDS maps. We consider it necessary to look for additional zero eigenvalues and to inspect diagnostic plots like Figure S1 to identify influential observations, and to perform a relatedness study before inferring the existence of any population

Figure 9: MDS map of pairs of individuals of the GCAT study. Convex hulls delimit the region of the pairs obtained by genetic bootstrapping.

Figure 10: MDS map of pairs of individuals of the GCAT study, trained with an additional category for three-quarter siblings (3/4S). Convex hulls delimit the region of the pairs obtained by genetic bootstrapping.

substructure.

The proposed method for classifying pairs using MDS and discriminant analysis is seen to perform well with both simulated and empirical data. The sampling of artificially related pairs from the observed data requires a considerable number of approximately unrelated individuals to be present. We therefore suggest the method to be used for large samples, where such a substantial subset of unrelated individuals can be identified. This is probably not an obstacle for the use of our method, as increasingly large samples are being used in epidemiological genomics. The sampling of artificial pairs from the observed data respects the allele frequency distribution of the original data, and provide reference areas for the different relationships given the allele frequencies of the observed data.

If MDS is used with the primary purpose of detecting substructure, we suggest first to remove one individual from each related pair, and to apply MDS to the filtered dataset. If not, related individuals can upset the proper detection of substructure. Incidental outliers in MDS maps are likely to correspond to individuals that have a close family relationship. Figure 2 shows a set of standard patterns that are relatively easy to recognize in genetic data analysis. However, if more family relationships are present in the database, the patterns obtained by MDS are more difficult to interpret.

The proposed IBS-MDS method can be used in an exploratory and iterative way. For some data sets, a first analysis may reveal that certain relationship categories are obviously absent in the data. In such cases, IBS-MDS can

be re-run with a training set without these categories, in order to focus the analysis on those relationships that are of relevance for the data under study. Indeed, we tacitly omitted monozygotic twins (MZ) from our training set, as MZ pairs (or duplicates) are easily recognized, and this way the analysis is focused on more remote relationships. The GCAT example illustrates this iterative use, as three-quarter siblings were added to the training set in second instance to improve the analysis.

In standard applications of MDS to bi-allelic genetic variants, genetic data is often coded in (0,1,2) format and Euclidean distances between individuals are calculated in this coding, and represented by MDS in low-dimensional space. The units on the axes of such maps are not very well interpretable, because the Euclidean distances are calculated over a large number of genetic variants. We propose to scale the Euclidean distance matrix that is entered into MDS by a factor of $1/\sqrt{k}$, where k is the number of variants. This scaling will not affect the configuration of the points in the map, and neither its goodness-of-fit, but it will render the axes and distances of the map more interpretable. Two individuals that are a unit distance apart in the map now differ on average by one allele per locus. This will typically scale all the coordinates in the MDS map within the (-1,1) interval, as the maximum difference in number of alleles between two individuals is two. We note that the Euclidean distance is not the same as the allele-sharing distance we used in this paper. The allele-sharing distance is the sum of the absolute value of the difference in numbers of alleles over all loci. In multivariate analysis this metric is better known as the Manhattan distance, the city-block distance, or the taxicab metric. Division of the Manhattan distance by the number of polymorphisms gives the allele-

sharing distance. For the sake of interpretability, it is convenient to scale the Manhattan distance matrix by $1/k$ prior to using it in MDS. A pair of individuals two units apart will then differ by two alleles at all loci. The suggested scalings facilitate interpretation of MDS results, but are hampered by the fact that the map is a low-dimensional approximation to the original distance matrix, and therefore the property will not hold exactly, but only approximately so. For distance matrices that have the Euclidean property, and that therefore only have non-negative eigenvalues [Mardia et al., 1979, Chapter 14], the map distances will approximate the true distances from below, and one can say that two individuals that are, for instance, 0.5 units apart, will differ on average by at least 0.5 minor alleles per locus, or can be expected to have at least one different allele for every pair of loci. We have used Manhattan distances scaled by $1/k$ in all MDS maps of individuals used this paper (Figures 1, 3, 6 and 8).

We note that the family relationships detected in Figure 3 would have gone largely unnoticed if the data is analyzed by non-metric multidimensional scaling of the between-individual distance matrix. In non-metric MDS only the rank order of the distances is used, not the value of the distances itself. In non-metric MDS related individuals do therefore generally not appear as outliers. This makes the non-metric approach less attractive at the individual level, though it does not rule out the non-metric approach for analysing pairs.

We recommend the use of discriminant analysis in allele-sharing studies as employed in this paper. The posterior probabilities of the different relationships give a quantitative criterion for deciding upon which relationship is most likely for a given pair of individuals. In allele sharing studies this deci-

sion is mostly made graphically by inspecting a (p_0, p_2) plot in IBS studies, or a (k_0, k_2) plot in IBD studies. We note that these posterior probabilities differ from those used in a standard discriminant analysis, in the sense that they are affected by additional uncertainty generated by using a training set obtained by genetic bootstrapping of the observed data.

The GCAT sample shows, for almost all relationship categories, larger variability in the MDS coordinates than would be expected under strict Mendelian sampling of alleles from unrelated individuals. This excess variability can, at least in part, be explained by the presence of additional relatedness between (unobserved) close relatives of the individuals in the database. This leads to increased autozygosity, which is a characteristic of more endogamous populations. The occurrence of three-quarter siblings is just a particular instance of this phenomenon. Consequently, the degree of relatedness of two individuals tends to become a continuous variable, which is increasingly hard to discretize into the standard relationship categories.

Several scholars have recently distinguished between remote or background relatedness (population substructure) and more recent relatedness (family relationships), describing methods that try to address both [Conomos et al., 2016, Weir and Goudet, 2017]. We have shown in this paper that MDS can be used to detect both recent relatedness and substructure.

6 Acknowledgments

This work was partially supported by grants SGR551 and SGR1269 from the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) of the Generalitat de Catalunya, by grant MTM2015-65016-C2-2-R and ADE 10/00026 (MINECO/FEDER) of the Spanish Ministry of Economy and Competitiveness and European Regional Development Fund, and by grant R01 GM075091 from the United States National Institutes of Health. R. de Cid was supported by the Ramon y Cajal action (RYC-2011-07822).

References

- G. R. Abecasis, S. S. Cherny, C. W. O. C., and L. R. Cardon. GRR: Graphical representation of relationship errors. Bioinformatics, 17(8):742–743, 2001.
- J. Aitchison. The Statistical Analysis of Compositional Data. The Blackburn press, Caldwell, NJ, 1986. 2003 printing.
- I. Borg and P. J. F. Groenen. Modern multidimensional scaling. Springer, New York, second edition, 2005.
- M. P. Conomos, A. P. Reiner, B. S. Weir, and T. A. Thornton. Model-free estimation of recent genetic relatedness. American Journal of Human Genetics, 98(1):127–148, 2016. doi: 10.1016/j.ajhg.2015.11.022.
- C. W. Cotterman. Relatives and human genetic analysis. The Scientific Monthly, 53:227–234, 1941.
- T. F. Cox and M. A. A. Cox. Multidimensional Scaling. Chapman & Hall, second edition, 2001.

- M. Epstein, W. Duren, and B. M. Improved inference of relationship for pairs of individuals. American Journal of Human Genetics, 67:1219–1231, 2000.
- A. S. Foulkes. Applied statistical genetics with R. Springer, New York, 2009.
- I. Galván-Femenía, J. Graffelman, and C. Barceló-i Vidal. Graphics for relatedness research. Molecular Ecology Resources, 17(6):1271–1282, 2017. doi: <https://doi.org/10.1111/1755-0998.12674>.
- J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53(3):325–338, 1966.
- J. C. Gower. Adding a point to vector diagrams in multivariate analysis. Biometrika, 55:582–585, 1968.
- J. Graffelman and T. Aluja-Banet. Optimal representation of supplementary variables in biplots from principal component analysis and correspondence analysis. Biometrical Journal, 45(4):491–509, 2003.
- J. Graffelman and V. Moreno. The mid p -value in exact tests for Hardy-Weinberg equilibrium. Statistical Applications in Genetics and Molecular Biology, 12(4):433–448, 2013.
- C. Huff, D. Witherspoon, T. Simonson, J. Xing, W. Watkins, Y. Zhang, T. Tuohy, D. Neklason, R. Burt, S. Guthery, S. Woodward, and L. Jorde. Maximum-likelihood estimation of recent shared ancestry (ersa). Genome research, 21(5):768–774, 2011. doi: 10.1101/gr.115972.110.
- M. Jakobsson, S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere, H. Fung, Z. Szpiech, J. Degnan, K. Wang, R. Guerreiro, J. Bras, J. Schymick, D. Hernandez, B. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton,

- J. van de Leemput, I. Rafferty, M. Bucan, H. Cann, J. Hardy, N. Rosenberg, and A. Singleton. Genotype, haplotype and copy-number variation in worldwide human populations. Nature, 451:998–1003, 2008. doi: <http://dx.doi.org/10.1038/nature06742>.
- S. Kyriazopoulou-Panagiotopoulou, D. Kashef Haghighi, S. J. Aerni, A. Sundquist, S. Bercovici, and S. Batzoglou. Reconstruction of genealogical relationships with applications to Phase III of HapMap. Bioinformatics, 27(13):i333–41, 2011. doi: 10.1093/bioinformatics/btr243.
- K. V. Mardia, J. T. Kent, and J. M. Bibby. Multivariate analysis. Academic press London, 1979.
- J. A. Martín-Fernández, J. Palarea-Albaladejo, and R. A. Olea. Dealing with zeros. In V. Pawlowsky-Glahn and A. Buccianti, editors, Compositional Data Analysis: Theory and Applications, pages 43–58, Chichester (UK), 2011. John Wiley & Sons. ISBN 978-0-470-71135-4.
- M. S. McPeck and L. Sun. Statistical tests for detection of misspecified relationships by use of genome-screen data. American Journal of Human Genetics, 66(3):1076–1094, 2000.
- B. G. Milligan. Maximum-likelihood estimation of relatedness. Genetics, 163(3):1153–1167, 2003.
- J. L. Mountain and L. L. Cavalli-Sforza. Multilocus genotypes, a tree of individuals, and human evolutionary history. American Journal of Human Genetics, 61:705–718, 1997.
- A. Nembot-Simo and B. Graham, J. amd McNeney. CrypticIBDcheck: an R

- package for checking cryptic relatedness in nominally unrelated individuals. Source Code for Biology and Medicine, 8(5), 2013.
- M. Obón-Santacana, M. Vilardell, A. Carreras, X. Duran, J. Velasco, I. Galván-Femenía, T. Alonso, L. Puig, L. Sumoy, E. J. Duell, M. Peruchó, V. Moreno, and de Cid R. Cohort profile: Gcat genomes for life. a prospective cohort study of the genomes of catalonia. BMJ Open, 2018. To appear.
- T. J. Pemberton, C. Wang, J. Z. Li, and N. A. Rosenberg. Inference of unexpected genetic relatedness among individuals in HapMap Phase III. American Journal of Human Genetics, 87:457–464, 2010.
- T. J. Pemberton, M. DeGiorgio, and N. A. Rosenberg. Population structure in a comprehensive genomic data set on human microsatellite variation. G3: Genes, Genomes, Genetics, 3(5):891–907, 2013. doi: 10.1534/g3.113.005728. URL <http://www.g3journal.org/content/3/5/891>.
- S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. Plink: A toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81(3):559–575, 2007. URL <http://pngu.mgh.harvard.edu/purcell/plink/>.
- R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL <http://www.R-project.org/>.

- N. A. Rosenberg. Standardized subsets of the HGDP-CEPH Human Genome Diversity cell line Panel, accounting for atypical and duplicated samples and pairs of close relatives. American Journal of Human Genetics, 70(6): 841–847, 2006. doi: 10.1111/j.1469-1809.2006.00285.x.
- C. Sabatti, S. K. Service, A. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, C. G. Jones, N. A. Zaitlen, T. Varilo, M. Kaakinen, U. Sovio, A. Ruokonen, J. Laitinen, E. Jakkula, L. Coin, C. Hoggart, A. Collins, H. Turunen, S. Gabriel, P. Elliot, M. I. McCarthy, M. J. Daly, M. Järvelin, N. B. Freimer, and L. Peltonen. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nature genetics, 41(1):35–46, 2009. doi: 10.1038/ng.271.
- E. L. Stevens, G. Heckenberg, E. D. O. Roberson, J. D. Baugher, T. J. Downey, and J. Pevsner. Inference of relationships in population data using indentity-by-descent and identity-by-state. PLoS Genetics, 7(9):1–15, 2011.
- E. L. Stevens, J. D. Baugher, M. D. Shirley, L. P. Frelín, and J. Pevsner. Unexpected relationships and inbreeding in HapMap Phase III populations. PLoS ONE, 7(11):e49575, 2012. doi: 10.1371/journal.pone.0049575.
- L. Sumoy, D. Piñeyro, I. Galván-Femenía, R. Pluvionet, L. Ramos, and R. de Cid. Joint calling description and qc summary. Technical report, GCAT Project, IMPPC, Badalona (Barcelona, Spain), 2017. URL <http://www.genomesforlife.com/investigadors>.
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature, 526:68–74, 2015.

- The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature, 449:851–861, 2007.
- E. A. Thompson. Estimation of pairwise relationships. Annals of Human Genetics, 39:173–188, 1975.
- E. A. Thompson. Estimation of relationships from genetic data. In C. R. Rao and R. Chakraborty, editors, Handbook of Statistics, volume 8, pages 255–269, Amsterdam, 1991. Elsevier Science.
- C. Wang, Z. Szpiech, J. Degnan, M. Jakobsson, T. Pemberton, J. Hardy, A. Singleton, and N. Rosenberg. Comparing spatial maps of human population-genetic variation using procrustes analysis. Statistical applications in genetics and molecular biology, 9, 2010. doi: 10.2202/1544-6115.1493. Article 13.
- B. S. Weir and J. Goudet. A unified characterization of population structure and relatedness. Genetics, 2017. ISSN 0016-6731. doi: 10.1534/genetics.116.198424. URL <http://www.genetics.org/content/early/2017/05/26/genetics.116.198424>.
- B. S. Weir, A. D. Anderson, and A. B. Hepler. Genetic relatedness analysis: modern data and new challenges. Nature Reviews Genetics, 7(10):771–780, 2006.
- S. Weisberg. Applied Linear Regression. John Wiley & Sons, New Jersey, third edition, 2005.

Supplementary material

Figure S1: Fitted versus observed distance for an MDS map of a homogeneous population with one MZ pair. The single pair involving the two individuals of the MZ pair is red, pairs involving one individual of the MZ pair are blue and pairs not involving any individual of the MZ pair are black.

Figure S2: Scatterplot matrices of the coordinates of the first four dimensions of an MDS solution using Euclidean, Manhattan, Canberra and Aitchison distance. Data obtained by simulation of 5,000 independent variants with a MAF of 0.5 under Hardy-Weinberg equilibrium.

Table S1: Predicted relationships of first, second, third and fourth degree pairs of the GCAT cohort and their probabilities according to IBS-MDS. k_0 , k_1 , k_2 : Cotterman's coefficients, θ : coancestry coefficient.

	Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities						k_0	k_1	k_2	θ	
							UN	5th	4th	3rd	2nd	FS					PO
43	1	AAA3246716	F	AAA4366649	F	PO	0	0.000	0.000	0.000	0	0	1	0.000	0.992	0.008	0.504
	2	AAA4366853	M	AAA4622161	M	PO	0	0.000	0.000	0.000	0	0	1	0.000	0.998	0.002	0.501
	3	AAA3233746	F	AAA4277419	F	PO	0	0.000	0.000	0.000	0	0	1	0.000	0.995	0.005	0.502
	4	AAA3231204	M	AAA1604417	F	PO	0	0.000	0.000	0.000	0	0	1	0.000	1.000	0.000	0.500
	5	AAA3256970	F	AAA4628527	F	FS	0	0.000	0.000	0.000	0	1	0	0.342	0.457	0.201	0.429
	6	AAA4226942	F	AAA4227518	F	FS	0	0.000	0.000	0.000	0	1	0	0.254	0.479	0.266	0.506
	7	AAA4626832	F	AAA4834036	F	FS	0	0.000	0.000	0.000	0	1	0	0.187	0.518	0.295	0.554
	8	AAA4225757	M	AAA4031162	F	FS	0	0.000	0.000	0.000	0	1	0	0.190	0.508	0.303	0.556
	9	AAA4225075	M	AAA4624355	M	FS	0	0.000	0.000	0.000	0	1	0	0.267	0.442	0.290	0.511
	10	AAA4227506	M	AAA4031164	M	FS	0	0.000	0.000	0.000	0	1	0	0.222	0.538	0.240	0.509
	11	AAA4224640	F	AAA4217071	F	FS	0	0.000	0.000	0.000	0	1	0	0.311	0.460	0.229	0.459
	12	AAA4018041	F	AAA4217227	M	FS	0	0.000	0.000	0.000	0	1	0	0.248	0.514	0.237	0.494
	13	AAA3240903	F	AAA4365463	F	FS	0	0.000	0.000	0.000	0	1	0	0.258	0.490	0.253	0.498
	14	AAA4217714	M	AAA3231530	F	FS	0	0.000	0.000	0.000	0	1	0	0.237	0.496	0.267	0.515

43

	Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities						k_0	k_1	k_2	θ	
							UN	5th	4th	3rd	2nd	FS					PO
44	15	AAA4338734	F	AAA1604854	F	FS	0	0.000	0.000	0.000	0	1	0	0.310	0.471	0.220	0.455
	16	AAA4277435	F	AAA4367506	M	FS	0	0.000	0.000	0.000	0	1	0	0.254	0.530	0.216	0.481
	17	AAA4227515	M	AAA3237815	F	FS	0	0.000	0.000	0.000	0	1	0	0.275	0.526	0.199	0.462
	18	AAA3263087	F	AAA4365219	F	FS	0	0.000	0.000	0.000	0	1	0	0.247	0.440	0.313	0.533
	19	AAA4223928	F	AAA1604429	F	FS	0	0.000	0.000	0.000	0	1	0	0.279	0.471	0.250	0.485
	20	AAA4628557	F	AAA4833416	M	FS	0	0.000	0.000	0.000	0	1	0	0.262	0.509	0.229	0.484
	21	AAA4023463	F	AAA4018431	F	FS	0	0.000	0.000	0.000	0	1	0	0.148	0.549	0.302	0.577
	22	AAA3883248	F	AAA3883688	F	FS	0	0.000	0.000	0.000	0	1	0	0.252	0.528	0.221	0.485
	23	AAA4256408	F	AAA1604078	F	FS	0	0.000	0.000	0.000	0	1	0	0.216	0.512	0.272	0.528
	24	AAA4628172	F	AAA4626816	M	FS	0	0.000	0.000	0.000	0	1	0	0.236	0.449	0.315	0.540
	25	AAA1862514	F	AAA4366204	M	FS	0	0.000	0.000	0.000	0	1	0	0.168	0.523	0.308	0.570
	26	AAA4031813	M	AAA4229044	M	FS	0	0.000	0.000	0.000	0	1	0	0.280	0.492	0.228	0.474
	27	AAA3263034	F	AAA3258976	F	FS	0	0.000	0.000	0.000	0	1	0	0.264	0.430	0.306	0.521
	28	AAA4365437	M	AAA4365449	F	FS	0	0.000	0.000	0.000	0	1	0	0.243	0.505	0.252	0.505
	29	AAA3238247	F	AAA4027580	F	FS	0	0.000	0.000	0.000	0	1	0	0.191	0.508	0.301	0.555
	30	AAA3238213	F	AAA3238291	F	FS	0	0.000	0.000	0.000	0	1	0	0.232	0.511	0.257	0.512

44

	Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities						k_0	k_1	k_2	θ	
							UN	5th	4th	3rd	2nd	FS					PO
45	31	AAA3253254	M	AAA4027522	F	FS	0	0.000	0.000	0.000	0	1	0	0.225	0.566	0.208	0.492
	32	AAA3246741	M	AAA4228843	M	FS	0	0.000	0.000	0.000	0	1	0	0.282	0.432	0.285	0.501
	33	AAA3257361	F	AAA3231510	F	FS	0	0.000	0.000	0.000	0	1	0	0.316	0.427	0.258	0.471
	34	AAA4223896	M	AAA4033132	F	FS	0	0.000	0.000	0.000	0	1	0	0.231	0.542	0.227	0.498
	35	AAA1862528	M	AAA3231518	F	FS	0	0.000	0.000	0.000	0	1	0	0.361	0.439	0.200	0.419
	36	AAA3256375	M	AAA3257334	M	FS	0	0.000	0.000	0.000	0	1	0	0.247	0.543	0.210	0.481
	37	AAA4217017	F	AAA4018434	F	FS	0	0.000	0.000	0.000	0	1	0	0.221	0.446	0.332	0.555
	38	AAA3253233	F	AAA3246732	F	FS	0	0.000	0.000	0.000	0	1	0	0.310	0.446	0.245	0.468
	39	AAA4277695	M	AAA4626181	F	FS	0	0.000	0.000	0.000	0	1	0	0.286	0.494	0.220	0.467
	40	AAA4031881	F	AAA4031893	F	FS	0	0.000	0.000	0.000	0	1	0	0.305	0.485	0.211	0.453
	41	AAA4217761	M	AAA4217056	F	FS	0	0.000	0.000	0.000	0	1	0	0.278	0.511	0.211	0.466
	42	AAA4229806	F	AAA4224637	F	FS	0	0.000	0.000	0.000	0	1	0	0.224	0.522	0.254	0.515
	43	AAA4277430	M	AAA4224670	M	FS	0	0.000	0.000	0.000	0	1	0	0.326	0.520	0.154	0.414
	44	AAA4216857	M	AAA4632452	M	FS	0	0.000	0.000	0.000	0	1	0	0.218	0.479	0.303	0.543
	45	AAA4267559	M	AAA4366821	F	FS	0	0.000	0.000	0.000	0	1	0	0.257	0.451	0.292	0.517
	46	AAA4223878	M	AAA4223888	M	FS	0	0.000	0.000	0.000	0	1	0	0.312	0.477	0.211	0.450

	Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities						k_0	k_1	k_2	θ	
							UN	5th	4th	3rd	2nd	FS					PO
46	47	AAA4268343	F	AAA4216826	F	FS	0	0.000	0.000	0.000	0	1	0	0.250	0.502	0.248	0.499
	48	AAA3238629	F	AAA3238662	F	FS	0	0.000	0.000	0.000	0	1	0	0.231	0.460	0.310	0.539
	49	AAA3238629	F	AAA3238641	F	FS	0	0.000	0.000	0.000	0	1	0	0.177	0.516	0.307	0.565
	50	AAA3238662	F	AAA3238641	F	FS	0	0.000	0.000	0.000	0	1	0	0.170	0.502	0.327	0.578
	51	AAA3239884	F	AAA4031172	F	FS	0	0.000	0.000	0.000	0	1	0	0.210	0.525	0.265	0.527
	52	AAA3231280	M	AAA1594735	F	FS	0	0.000	0.000	0.000	0	1	0	0.285	0.454	0.261	0.488
	53	AAA3239877	F	AAA1862815	M	FS	0	0.000	0.000	0.000	0	1	0	0.349	0.510	0.140	0.395
	54	AAA3256915	F	AAA4023844	F	FS	0	0.000	0.000	0.000	0	1	0	0.230	0.481	0.289	0.529
	55	AAA4223962	M	AAA3258917	M	FS	0	0.000	0.000	0.000	0	1	0	0.203	0.508	0.289	0.543
	56	AAA4226441	F	AAA3231593	F	FS	0	0.000	0.000	0.000	0	1	0	0.252	0.517	0.231	0.489
	57	AAA3240904	M	AAA4229019	M	FS	0	0.000	0.000	0.000	0	1	0	0.243	0.496	0.260	0.508
	58	AAA1604858	M	AAA4223844	F	FS	0	0.000	0.000	0.000	0	1	0	0.235	0.503	0.263	0.514
	59	AAA4257889	F	AAA4031119	F	FS	0	0.000	0.000	0.000	0	1	0	0.313	0.427	0.260	0.473
	60	AAA4228456	M	AAA4277649	F	FS	0	0.000	0.000	0.000	0	1	0	0.299	0.491	0.210	0.456
	61	AAA3256773	F	AAA4226980	F	FS	0	0.000	0.000	0.000	0	1	0	0.182	0.567	0.251	0.535
	62	AAA4223868	F	AAA1604838	F	FS	0	0.000	0.000	0.000	0	1	0	0.264	0.545	0.191	0.463

	Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities							k_0	k_1	k_2	θ
							UN	5th	4th	3rd	2nd	FS	PO				
47	63	AAA4223967	M	AAA4632402	F	FS	0	0.000	0.000	0.000	0	1	0	0.366	0.464	0.170	0.402
	64	AAA3256774	M	AAA3239839	F	FS	0	0.000	0.000	0.000	0	1	0	0.212	0.525	0.263	0.525
	65	AAA1862853	M	AAA1604065	M	FS	0	0.000	0.000	0.000	0	1	0	0.259	0.514	0.227	0.484
	66	AAA4226277	M	AAA4216896	F	FS	0	0.000	0.000	0.000	0	1	0	0.259	0.559	0.182	0.462
	67	AAA4031136	M	AAA4833606	M	FS	0	0.000	0.000	0.000	0	1	0	0.367	0.427	0.206	0.419
	68	AAA1862562	M	AAA1862579	F	FS	0	0.000	0.000	0.000	0	1	0	0.271	0.503	0.226	0.478
	69	AAA4023847	M	AAA4338272	F	FS	0	0.000	0.000	0.000	0	1	0	0.183	0.518	0.299	0.558
	70	AAA4216291	F	AAA4366272	F	FS	0	0.000	0.000	0.000	0	1	0	0.279	0.498	0.223	0.472
	71	AAA4338789	F	AAA4217204	F	FS	0	0.000	0.000	0.000	0	1	0	0.216	0.497	0.287	0.535
	72	AAA4833422	F	AAA4833410	F	FS	0	0.000	0.000	0.000	0	1	0	0.233	0.548	0.218	0.492
	73	AAA3235002	F	AAA1862534	F	FS	0	0.000	0.000	0.000	0	1	0	0.323	0.505	0.172	0.424
	74	AAA4023806	M	AAA4217080	M	FS	0	0.000	0.000	0.000	0	1	0	0.262	0.523	0.215	0.476
	75	AAA3240987	F	AAA4367565	F	2nd	0	0.000	0.000	0.000	1	0	0	0.498	0.502	0.000	0.251
	76	AAA3238273	F	AAA3240980	F	2nd	0	0.000	0.000	0.000	1	0	0	0.607	0.358	0.035	0.214
	77	AAA4026850	M	AAA4834042	M	2nd	0	0.000	0.000	0.000	1	0	0	0.342	0.560	0.098	0.378
	78	AAA1862883	F	AAA3263093	F	2nd	0	0.000	0.000	0.000	1	0	0	0.446	0.554	0.000	0.277

	Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities						k_0	k_1	k_2	θ	
							UN	5th	4th	3rd	2nd	FS					PO
48	79	AAA1604057	F	AAA4622121	M	2nd	0	0.000	0.000	0.000	1	0	0	0.491	0.501	0.007	0.258
	80	AAA4275946	F	AAA4365406	M	2nd	0	0.000	0.000	0.000	1	0	0	0.450	0.550	0.000	0.275
	81	AAA3233721	M	AAA4367573	F	2nd	0	0.000	0.000	0.000	1	0	0	0.484	0.516	0.000	0.258
	82	AAA3255761	F	AAA4277424	M	2nd	0	0.000	0.000	0.000	1	0	0	0.567	0.430	0.003	0.218
	83	AAA3246729	F	AAA4366264	M	2nd	0	0.000	0.000	0.000	1	0	0	0.514	0.486	0.000	0.243
	84	AAA4245776	F	AAA3257385	F	2nd	0	0.000	0.000	0.000	1	0	0	0.465	0.535	0.000	0.267
	85	AAA1862529	M	AAA4217201	F	2nd	0	0.000	0.000	0.000	1	0	0	0.471	0.529	0.000	0.265
	86	AAA4018468	M	AAA4632461	M	2nd	0	0.000	0.000	0.000	1	0	0	0.466	0.531	0.003	0.268
	87	AAA4032204	F	AAA4032271	F	2nd	0	0.000	0.000	0.000	1	0	0	0.443	0.557	0.000	0.278
	88	AAA4225718	M	AAA4626825	F	3rd	0	0.000	0.000	1.000	0	0	0	0.649	0.351	0.000	0.176
	89	AAA1604855	M	AAA4267565	F	3rd	0	0.000	0.000	1.000	0	0	0	0.733	0.267	0.000	0.134
	90	AAA3238236	M	AAA4245710	M	3rd	0	0.000	0.000	1.000	0	0	0	0.722	0.276	0.002	0.140
	91	AAA3256382	F	AAA4217039	F	3rd	0	0.000	0.000	1.000	0	0	0	0.749	0.251	0.000	0.126
	92	AAA4245765	M	AAA4833462	F	3rd	0	0.000	0.000	1.000	0	0	0	0.762	0.234	0.004	0.121
	93	AAA4023867	M	AAA4339504	M	3rd	0	0.000	0.000	1.000	0	0	0	0.761	0.239	0.000	0.119
	94	AAA4018003	F	AAA3231592	F	3rd	0	0.000	0.000	1.000	0	0	0	0.754	0.246	0.000	0.123

	Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities						k_0	k_1	k_2	θ	
							UN	5th	4th	3rd	2nd	FS					PO
49	95	AAA4229869	M	AAA4032237	F	3rd	0	0.000	0.000	1.000	0	0	0	0.801	0.196	0.003	0.101
	96	AAA1831722	M	AAA4367079	F	3rd	0	0.000	0.000	1.000	0	0	0	0.797	0.189	0.014	0.108
	97	AAA1862295	F	AAA4018028	F	3rd	0	0.000	0.000	1.000	0	0	0	0.719	0.281	0.000	0.141
	98	AAA4026834	F	AAA4026822	F	3rd	0	0.000	0.000	1.000	0	0	0	0.743	0.257	0.000	0.129
	99	AAA4338758	F	AAA4256468	M	3rd	0	0.000	0.000	1.000	0	0	0	0.753	0.242	0.005	0.126
	100	AAA4366279	M	AAA3238678	F	3rd	0	0.000	0.000	1.000	0	0	0	0.762	0.238	0.000	0.119
	101	AAA4026851	M	AAA4629602	M	3rd	0	0.000	0.000	1.000	0	0	0	0.783	0.217	0.000	0.108
	102	AAA4026851	M	AAA4018473	F	3rd	0	0.000	0.000	1.000	0	0	0	0.788	0.209	0.003	0.108
	103	AAA4272088	F	AAA1862589	M	3rd	0	0.000	0.000	1.000	0	0	0	0.742	0.255	0.003	0.130
	104	AAA3883274	M	AAA4629667	M	3rd	0	0.000	0.000	1.000	0	0	0	0.691	0.309	0.000	0.154
	105	AAA4256408	F	AAA4624339	M	3rd	0	0.000	0.000	1.000	0	0	0	0.712	0.288	0.000	0.144
	106	AAA4277495	F	AAA4267521	F	3rd	0	0.000	0.000	1.000	0	0	0	0.743	0.254	0.003	0.130
	107	AAA4365882	M	AAA3258985	F	3rd	0	0.000	0.000	1.000	0	0	0	0.705	0.295	0.000	0.148
	108	AAA4628580	F	AAA4365476	F	3rd	0	0.000	0.000	1.000	0	0	0	0.770	0.229	0.001	0.116
	109	AAA4228878	F	AAA4622107	F	3rd	0	0.000	0.000	1.000	0	0	0	0.708	0.290	0.002	0.147
110	AAA4018096	F	AAA4216134	M	3rd	0	0.000	0.000	1.000	0	0	0	0.765	0.234	0.000	0.117	

49

	Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities						k_0	k_1	k_2	θ	
							UN	5th	4th	3rd	2nd	FS					PO
50	111	AAA4273571	F	AAA3256980	F	3rd	0	0.000	0.000	1.000	0	0	0	0.671	0.323	0.006	0.167
	112	AAA3239817	F	AAA3231221	F	3rd	0	0.000	0.000	1.000	0	0	0	0.663	0.337	0.000	0.168
	113	AAA4018432	M	AAA4272026	F	3rd	0	0.000	0.000	1.000	0	0	0	0.750	0.249	0.001	0.125
	114	AAA4217720	F	AAA3237860	F	3rd	0	0.000	0.000	1.000	0	0	0	0.768	0.232	0.000	0.116
	115	AAA4226232	M	AAA4365890	F	3rd	0	0.000	0.000	1.000	0	0	0	0.717	0.278	0.005	0.144
	116	AAA4217711	M	AAA1604063	M	3rd	0	0.000	0.000	1.000	0	0	0	0.729	0.271	0.000	0.136
	117	AAA4228452	M	AAA4366604	M	3rd	0	0.000	0.000	1.000	0	0	0	0.683	0.317	0.000	0.159
	118	AAA3233766	F	AAA4032573	M	3rd	0	0.000	0.000	1.000	0	0	0	0.782	0.218	0.000	0.109
	119	AAA4257865	F	AAA1862872	F	3rd	0	0.000	0.000	1.000	0	0	0	0.823	0.177	0.000	0.088
	120	AAA4226904	M	AAA4625993	F	3rd	0	0.000	0.000	1.000	0	0	0	0.770	0.230	0.000	0.115
	121	AAA4273539	F	AAA4626510	F	3rd	0	0.000	0.000	1.000	0	0	0	0.736	0.262	0.001	0.133
	122	AAA3256934	M	AAA4032565	M	3rd	0	0.000	0.000	1.000	0	0	0	0.791	0.209	0.000	0.105
	123	AAA4031110	M	AAA3237825	F	3rd	0	0.000	0.000	1.000	0	0	0	0.800	0.194	0.006	0.103
	124	AAA3238659	M	AAA4624324	F	3rd	0	0.000	0.000	1.000	0	0	0	0.780	0.214	0.006	0.113
	125	AAA4031824	M	AAA4626192	M	3rd	0	0.000	0.000	1.000	0	0	0	0.719	0.274	0.007	0.144
	126	AAA4629674	M	AAA4227571	F	3rd	0	0.000	0.000	1.000	0	0	0	0.759	0.228	0.012	0.127

50

	Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities						k_0	k_1	k_2	θ	
							UN	5th	4th	3rd	2nd	FS					PO
51	127	AAA4629691	M	AAA4626181	F	3rd	0	0.000	0.294	0.706	0	0	0	0.813	0.179	0.008	0.098
	128	AAA1604078	F	AAA4624339	M	3rd	0	0.000	0.000	1.000	0	0	0	0.751	0.248	0.000	0.124
	129	AAA3257107	F	AAA3260152	F	3rd	0	0.000	0.000	1.000	0	0	0	0.687	0.313	0.000	0.156
	130	AAA4256436	M	AAA1862814	F	3rd	0	0.000	0.320	0.680	0	0	0	0.822	0.175	0.003	0.090
	131	AAA3240982	M	AAA3239854	F	3rd	0	0.000	0.000	1.000	0	0	0	0.774	0.226	0.000	0.113
	132	AAA4275971	F	AAA4365222	M	3rd	0	0.000	0.000	1.000	0	0	0	0.745	0.255	0.000	0.128
	133	AAA4018054	F	AAA1604430	F	3rd	0	0.000	0.000	1.000	0	0	0	0.731	0.260	0.009	0.139
	134	AAA4018032	F	AAA1594743	F	3rd	0	0.000	0.000	1.000	0	0	0	0.772	0.226	0.002	0.115
	135	AAA3256127	F	AAA1862545	F	3rd	0	0.000	0.000	1.000	0	0	0	0.812	0.183	0.005	0.096
	136	AAA3256143	F	AAA4018005	F	3rd	0	0.000	0.000	1.000	0	0	0	0.758	0.242	0.000	0.121
	137	AAA4366285	F	AAA4023815	F	3rd	0	0.000	0.000	1.000	0	0	0	0.761	0.236	0.003	0.121
	138	AAA3256191	F	AAA4032558	M	3rd	0	0.000	0.000	1.000	0	0	0	0.672	0.328	0.000	0.164
	139	AAA4217706	F	AAA3237827	F	3rd	0	0.000	0.000	1.000	0	0	0	0.643	0.357	0.000	0.178
	140	AAA4257861	M	AAA4627650	M	3rd	0	0.000	0.000	1.000	0	0	0	0.686	0.314	0.000	0.157
	141	AAA3233770	F	AAA4625357	M	3rd	0	0.000	0.000	1.000	0	0	0	0.682	0.318	0.000	0.159
142	AAA4226907	F	AAA1604040	M	3rd	0	0.000	0.000	1.000	0	0	0	0.707	0.293	0.000	0.147	

	Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities						k_0	k_1	k_2	θ	
							UN	5th	4th	3rd	2nd	FS					PO
52	143	AAA1604851	M	AAA4277623	M	3rd	0	0.000	0.000	1.000	0	0	0	0.734	0.266	0.000	0.133
	144	AAA4216252	F	AAA4018068	F	3rd	0	0.000	0.000	1.000	0	0	0	0.748	0.251	0.001	0.127
	145	AAA4365487	F	AAA4018476	F	3rd	0	0.000	0.000	1.000	0	0	0	0.796	0.204	0.000	0.102
	146	AAA4026883	F	AAA4229078	F	3rd	0	0.000	0.000	1.000	0	0	0	0.741	0.257	0.002	0.130
	147	AAA4216247	M	AAA4023441	F	3rd	0	0.000	0.000	1.000	0	0	0	0.728	0.262	0.010	0.141
	148	AAA4227543	M	AAA4226927	F	3rd	0	0.000	0.134	0.866	0	0	0	0.802	0.198	0.000	0.099
	149	AAA4216230	M	AAA4277635	M	3rd	0	0.000	0.000	1.000	0	0	0	0.767	0.233	0.000	0.116
	150	AAA4367063	M	AAA3884118	F	3rd	0	0.000	0.000	1.000	0	0	0	0.727	0.273	0.000	0.137
	151	AAA4245758	F	AAA4031894	F	3rd	0	0.000	0.000	1.000	0	0	0	0.739	0.261	0.000	0.131
	152	AAA3233122	F	AAA4229857	F	3rd	0	0.000	0.000	1.000	0	0	0	0.723	0.277	0.000	0.139
	153	AAA3231268	M	AAA3258918	M	3rd	0	0.000	0.000	1.000	0	0	0	0.730	0.259	0.012	0.141
	154	AAA4256444	F	AAA3257181	M	3rd	0	0.000	0.000	1.000	0	0	0	0.724	0.273	0.003	0.139
	155	AAA1604890	F	AAA4026837	F	3rd	0	0.000	0.000	1.000	0	0	0	0.732	0.268	0.000	0.134
	156	AAA4223950	M	AAA4031889	F	4th	0	0.000	1.000	0.000	0	0	0	0.834	0.161	0.005	0.085
	157	AAA4225024	F	AAA1605238	M	4th	0	0.089	0.911	0.000	0	0	0	0.905	0.095	0.000	0.047
158	AAA4227551	F	AAA1604892	F	4th	0	0.000	1.000	0.000	0	0	0	0.841	0.159	0.000	0.080	

	Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities						k_0	k_1	k_2	θ	
							UN	5th	4th	3rd	2nd	FS					PO
53	159	AAA3257155	F	AAA4277414	M	4th	0	0.000	1.000	0.000	0	0	0	0.850	0.150	0.000	0.075
	160	AAA4216228	M	AAA4268395	F	4th	0	0.000	0.990	0.010	0	0	0	0.811	0.189	0.000	0.094
	161	AAA4027518	F	AAA4625901	M	4th	0	0.000	1.000	0.000	0	0	0	0.845	0.151	0.003	0.079
	162	AAA3246732	F	AAA4833889	F	4th	0	0.002	0.998	0.000	0	0	0	0.902	0.087	0.010	0.054
	163	AAA4365468	M	AAA4227575	M	4th	0	0.000	1.000	0.000	0	0	0	0.843	0.146	0.011	0.084
	164	AAA4337708	F	AAA3258927	M	4th	0	0.007	0.993	0.000	0	0	0	0.898	0.097	0.005	0.053
	165	AAA3253233	F	AAA4833889	F	4th	0	0.001	0.999	0.000	0	0	0	0.896	0.091	0.013	0.059
	166	AAA3256981	F	AAA3237805	F	4th	0	0.000	0.982	0.018	0	0	0	0.799	0.201	0.000	0.100
	167	AAA4627624	M	AAA4032287	M	4th	0	0.002	0.998	0.000	0	0	0	0.893	0.107	0.000	0.053
	168	AAA4338775	M	AAA4365809	M	4th	0	0.000	1.000	0.000	0	0	0	0.860	0.140	0.000	0.070
	169	AAA3256141	M	AAA4245715	M	4th	0	0.390	0.610	0.000	0	0	0	0.915	0.083	0.002	0.044
	170	AAA4338238	M	AAA4216874	F	4th	0	0.000	1.000	0.000	0	0	0	0.880	0.119	0.001	0.060
	171	AAA4338238	M	AAA4833007	F	4th	0	0.000	1.000	0.000	0	0	0	0.834	0.166	0.000	0.083
	172	AAA4365412	M	AAA3884116	F	4th	0	0.002	0.998	0.000	0	0	0	0.890	0.110	0.000	0.055
	173	AAA4228402	M	AAA4834032	F	4th	0	0.001	0.999	0.000	0	0	0	0.885	0.113	0.001	0.058
174	AAA4217778	M	AAA4367001	M	4th	0	0.017	0.983	0.000	0	0	0	0.893	0.106	0.001	0.054	

	Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities						k_0	k_1	k_2	θ	
							UN	5th	4th	3rd	2nd	FS					PO
54	175	AAA3233746	F	AAA3257374	F	4th	0	0.000	1.000	0.000	0	0	0	0.820	0.180	0.000	0.090
	176	AAA4277639	M	AAA4031170	M	4th	0	0.000	1.000	0.000	0	0	0	0.898	0.088	0.014	0.058
	177	AAA4277695	M	AAA4629691	M	4th	0	0.000	0.986	0.014	0	0	0	0.817	0.178	0.004	0.093
	178	AAA1862530	M	AAA1604808	F	4th	0	0.007	0.993	0.000	0	0	0	0.900	0.100	0.000	0.050
	179	AAA4277676	F	AAA4833624	F	4th	0	0.012	0.988	0.000	0	0	0	0.897	0.098	0.006	0.054
	180	AAA3257331	F	AAA4366626	F	4th	0	0.000	1.000	0.000	0	0	0	0.868	0.132	0.000	0.066
	181	AAA4226436	M	AAA4627661	M	4th	0	0.000	1.000	0.000	0	0	0	0.877	0.123	0.000	0.061
	182	AAA1862263	M	AAA3239850	F	4th	0	0.000	1.000	0.000	0	0	0	0.879	0.113	0.008	0.065
	183	AAA3256980	F	AAA4626522	F	4th	0	0.000	1.000	0.000	0	0	0	0.817	0.183	0.000	0.091
	184	AAA4257880	M	AAA1831745	M	4th	0	0.000	1.000	0.000	0	0	0	0.837	0.157	0.006	0.085
	185	AAA3260107	M	AAA1594774	F	4th	0	0.269	0.731	0.000	0	0	0	0.897	0.103	0.000	0.052
	186	AAA4217268	M	AAA3257396	M	4th	0	0.109	0.891	0.000	0	0	0	0.907	0.079	0.014	0.053
	187	AAA4367043	M	AAA4833682	F	4th	0	0.000	1.000	0.000	0	0	0	0.884	0.111	0.005	0.061
	188	AAA4018244	F	AAA3256903	M	4th	0	0.000	1.000	0.000	0	0	0	0.838	0.159	0.003	0.083
	189	AAA4339110	F	AAA4833674	F	4th	0	0.000	1.000	0.000	0	0	0	0.850	0.147	0.003	0.077
190	AAA3253275	M	AAA3258930	F	4th	0	0.000	0.974	0.026	0	0	0	0.820	0.180	0.000	0.090	

54

Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities							k_0	k_1	k_2	θ
						UN	5th	4th	3rd	2nd	FS	PO				
191	AAA3257149	F	AAA4229811	F	4th	0	0.000	1.000	0.000	0	0	0	0.903	0.089	0.008	0.053
192	AAA3233747	F	AAA4626132	M	4th	0	0.385	0.615	0.000	0	0	0	0.903	0.097	0.000	0.048
193	AAA4338708	M	AAA4229055	M	4th	0	0.000	0.998	0.002	0	0	0	0.812	0.188	0.000	0.094
194	AAA4023829	F	AAA4626556	F	4th	0	0.000	1.000	0.000	0	0	0	0.891	0.105	0.004	0.057
195	AAA3255749	F	AAA1862240	M	4th	0	0.036	0.964	0.000	0	0	0	0.891	0.109	0.000	0.055
196	AAA4216278	M	AAA3256388	M	4th	0	0.000	1.000	0.000	0	0	0	0.892	0.108	0.000	0.054
197	AAA1604812	M	AAA1862568	M	4th	0	0.085	0.915	0.000	0	0	0	0.916	0.081	0.003	0.044
198	AAA4223868	F	AAA4366818	F	4th	0	0.044	0.956	0.000	0	0	0	0.864	0.136	0.000	0.068
199	AAA4018086	F	AAA4226950	M	4th	0	0.294	0.706	0.000	0	0	0	0.892	0.108	0.000	0.054
200	AAA3253294	F	AAA4216874	F	4th	0	0.000	1.000	0.000	0	0	0	0.847	0.151	0.002	0.078
201	AAA3256909	F	AAA4026889	F	4th	0	0.007	0.993	0.000	0	0	0	0.900	0.089	0.011	0.056
202	AAA4033322	F	AAA4023803	F	4th	0	0.000	1.000	0.000	0	0	0	0.858	0.138	0.004	0.073
203	AAA4256468	M	AAA4339554	F	4th	0	0.161	0.839	0.000	0	0	0	0.908	0.082	0.009	0.051
204	AAA1862559	F	AAA4229050	M	4th	0	0.000	1.000	0.000	0	0	0	0.830	0.170	0.000	0.085
205	AAA4216176	M	AAA4267560	F	4th	0	0.072	0.928	0.000	0	0	0	0.882	0.118	0.000	0.059
206	AAA4277424	M	AAA3263022	M	4th	0	0.000	1.000	0.000	0	0	0	0.881	0.116	0.003	0.061

Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities							k_0	k_1	k_2	θ
						UN	5th	4th	3rd	2nd	FS	PO				
207	AAA4216291	F	AAA4275990	M	4th	0	0.000	1.000	0.000	0	0	0	0.869	0.129	0.002	0.067
208	AAA3238644	F	AAA3256126	F	4th	0	0.000	1.000	0.000	0	0	0	0.809	0.191	0.000	0.096
209	AAA4224487	F	AAA3233710	F	4th	0	0.002	0.998	0.000	0	0	0	0.895	0.099	0.005	0.055

Table S2: Potential three-quarter siblings of the GCAT cohort and their probabilities according to IBS-MDS. k_0 , k_1 , k_2 : Cotterman's coefficients, θ : coancestry coefficient.

Pair	ID1	sex	ID2	sex	IBS-MDS	Posterior probabilities								k_0	k_1	k_2	θ
						PO	FS	3/4S	2nd	3rd	4th	5th	UN				
35	AAA1862528	M	AAA3231518	F	3/4S	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.361	0.439	0.200	0.419
67	AAA4031136	M	AAA4833606	M	3/4S	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.367	0.427	0.206	0.419
43	AAA4277430	M	AAA4224670	M	3/4S	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.326	0.520	0.154	0.414
53	AAA3239877	F	AAA1862815	M	3/4S	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.349	0.510	0.140	0.395
63	AAA4223967	M	AAA4632402	F	3/4S	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.366	0.464	0.170	0.402
73	AAA3235002	F	AAA1862534	F	3/4S	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.323	0.505	0.172	0.424
76	AAA3238273	F	AAA3240980	F	2nd	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.607	0.358	0.035	0.214
77	AAA4026850	M	AAA4834042	M	3/4S	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.342	0.560	0.098	0.378

57