COMPOSITIONAL METHODOLOGY AND STATISTICAL INFERENCE OF FAMILY RELATIONSHIPS USING GENETIC MARKERS

Iván Galván Femenía

DOCTORAL THESIS

Compositional methodology and statistical inference of family relationships using genetic markers

Iván Galván Femenía

2020

Universitat de Girona

DOCTORAL THESIS

Compositional methodology and statistical inference of family relationships using genetic markers

Iván Galván Femenía

2020

Doctoral Programme in Technology

Advisors

Dr. Jan Graffelman

and

Dr. Carles Barceló i Vidal

Tutor

Dr. Josep A. Martín Fernández

Dissertation submitted to obtain the academic degree of Doctor by the University of Girona

Universitat de Girona

Dr. Jan Graffelman, professor of the departament of Statistics and Operations Research of the Technical University of Catalonia, and Dr. Carles Barceló i Vidal, emeritus professor of the departament of Computer Science, Applied Mathematics and Statistics of the University of Girona,

WE STATE

The dissertation entitled *Compositional methodology and statistical inference of family relationships using genetic markers*, submitted by Mr. Iván Galván Femenía to obtain the academic degree Doctor, has been carried out under our supervision.

And, for the record, we sign this document.

Signatures,

Dr. Jan Graffelman

Dr. Carles Barceló i Vidal

Girona, December 3, 2019.

"Logic is the foundation of the certainty of all the knowledge we acquire."

Leonhard Euler

"Oh, people can come up with statistics to prove anything, Kent. Forty percent of all people know that."

Homer Simpson

To the memory of my mother, Cristina Femenía-Navarro, who passed away during my doctoral studies and to my father, José Antonio Galván-Granado.
For your unconditional love and your efforts to give me the best education you could.

Acknowledgements

This adventure started several years ago and there have been many people with whom I have shared experiences and whom I would like to thank for their support and for their contributions, directly or indirectly, to this thesis.

My first deep gratitude goes to my supervisors Jan Graffelman and Carles Barceló-i-Vidal. Jan, thank you for introducing me to the field of statistical genetics, for all your good advice and patience. I enjoyed every time spent in your office learning from you, discussing and brainstorming different future research lines. Carles, thank you for your continuous suggestions and your time despite your status of emeritus professor, your insights were always valuable. I feel very happy to have had both of you as supervisors both in academic and personal level.

Specially thanks to Martin, Pepus, Santi, Glòria, Marina, Marc, Vera and Juanjo for opening me the doors to the compositional data research group and also for giving me the opportunity of teaching. No doubt I have learnt a lot from each one of you.

My also sincere gratitude to Rafael de Cid for the opportunity of learning from you and working together. I feel very lucky to combine this PhD thesis with the research in the GCAT project. It has been a plus for my doctorate and academic training. Thank you Rafa and all the nice colleagues from the GCAT lab for the great atmosphere and all your support during the last years.

Thank you to the Data Analytics 'Beabloo' team for all the great working experience and good moments we had fun. To 'la colla amb en Txorri' for all the good dinners and calçotades in Girona. To my friends from the Master MEIO. To my ex-flatmates and friends Clara and Eva, it was a memorable year living together. To the BrokenBike Mountain group for the nice biking routes. To my brothers from other mothers from the 'Pelels' and 'OBM' families. To my friend and 'mestre' Francisco José Santonja for your boost to complete this adventure.

A special thank to my love Anna for trying to get a smile in the most difficult moments and all the love and support received. To my parents, for your understanding and your encouragement to fulfill my dream of obtaining a doctorate. Mom, I whish you were here. And finally, to my aunts and cousins for all the love from the distance in this adventure. To everyone, thank you.

Publications

This thesis is presented as a compendium of the following research articles:

- Galván-Femenía, I., Graffelman, J. & Barceló-i-Vidal, C. (2017). **Graphics for relatedness research**. *Molecular Ecology Resources*, 17(6), 1271-1282. doi:10.1111/1755-0998.12674. Ranking of the *Journal Citation Reports* (JCR) from the *Institute of Scientific Information*: 30/298 (Biochemistry & Molecular Ecology); 10/164 (Ecology); 7/50 (Evolutionary Biology). First quartile (Q1) in all three areas. Impact factor: 7.332.

- Graffelman, J., Galván-Femenía, I., de Cid, R. & Barceló-i-Vidal, C. (2019). **A log-ratio biplot approach for exploring genetic relatedness based on identity by state**. *Frontiers in Genetics*, 10, 341. doi:10.3389/fgene.2019.00341. Ranking of the *Journal Citation Reports* (JCR) from the *Institute of Scientific Information*: 56/174 (Genetics & Heredity). Second quartile (Q2). Impact factor: 3.517.

- Galván-Femenía, I., Barceló-i-Vidal, C., Sumoy, L., Moreno, V., de Cid, R. & Graffelman, J. (2020). **A likelihood ratio approach for identifying three quarter siblings in genetic databases**. *Heredity*. Submitted. Ranking of the *Journal Citation Reports* (JCR) from the *Institute of Scientific Information*: 46/171 (Genetics & Heredity); 15/49 (Evolutionary Biology); 31/158 (Ecology). Second quartile (Q2). Impact factor: 3.179.

Apart from the articles of the journals, the following contributions in conferences are part of this doctoral thesis:

- Oral contribution at 6th International Workshop on Compositional Data Analysis (CODAWORK 2015). **A compositional approach to allele sharing analysis**. Galván-Femenía, I., Graffelman, J. & Barceló-i-Vidal, C. L'Escala, Spain. ISBN: 978-84-8458-451-3. P. 97-105.

- Poster contribution at 6th International Workshop on Compositional Data Analysis (CODAWORK 2015). **An application of the isometric logratio transformation for relatedness research**. Graffelman, J. & Galván-Femenía, I. L'Escala, Spain. ISBN: 978-84-8458-451-3. P. 134-142.

- Oral contribution at XVth Spanish Biometric Conference and Vth Ibero-American Biometric Meeting 2015. **IBS.IBD.studies: an R package for relatedness research using SNPs and microsatellite data**. Graffelman, J. & Galván-Femenía, I. Bilbao, Spain.

- Poster contribution at European Human Genetics Conference 2016. **Graphical tools for estimating family relationships**. Galván-Femenía, I., Graffelman, J., de Cid, R. & Barceló-i-Vidal, C. Barcelona, Spain. Page 384 from book of abstracts. European Journal of Human Genetics Volume 24 E-Supplement 1, May 2016.

- Poster contribution at 7th International Workshop on Compositional Data Analysis (CODAWORK 2017). **Multidimensional scaling for relatedness research: an application of the Aitchison distance in the GCAT population based cohort**. Galván-Femenía, I., Graffelman, J., Barceló-i-Vidal, C., Sumoy, L., Moreno, V. & de Cid, R. Abbadia San Salvatore, Italy.

In addition, two book chapters of the Proceedings of the 6th International Workshop on Compositional Data Analysis (CODAWORK 2015) are published also as part of this doctoral thesis:

- Galván-Femenía, I., Graffelman, J. & Barceló-i-Vidal, C. (2016) **A Compositional Approach to Allele Sharing Analysis**. In: Martín-Fernández J., Thió-Henestrosa S. (eds) Compositional Data Analysis. CoDaWork 2015. Springer Proceedings in Mathematics & Statistics, vol 187. Springer, Cham.

- Graffelman, J. & Galván-Femenía, I. (2016) **An Application of the Isometric Log-Ratio Transformation in Relatedness Research**. In: Martín-Fernández J., Thió-Henestrosa S. (eds) Compositional Data Analysis. CoDaWork 2015. Springer Proceedings in Mathematics & Statistics, vol 187. Springer, Cham.

List of Abbreviations

3/4S	Three-quarter siblings
alr	Additive log-ratio
clr	Centered log-ratio
CoDa	Compositional Data
DNA	Deoxyribonucleic acid
FS	Full siblings
GCAT	Genomes of Catalonia
GWAS	Genome-wide association studies
HGDP-CEPH	Human Genome Diversity Cell Line Panel-Centre d'Etude du Polymorphisme Humain
IBD	Identical by descent
IBS	Identical by state
ilr	Isometric log-ratio
PO	Parent-offspring
SNP	Single nucleotide polymorphism
STR	Short tandem repeat

List of Figures

List of Tables

Contents

Abstract

The present thesis is a compendium of *three* research articles produced between 2015 and 2019. All these three articles have a common link: they are different contributions based on compositional statistical methodology and statistical inference of genetic relatedness. In brief, *Compositional Data* are random vectors with strictly positive components whose sum is constant. These components represent parts of a whole which only carry relative information. Therefore, Compositional Data is usually represented as proportions or percentages. Relatedness is based on the principle of allele sharing between individuals for a given set of genetic markers. The larger the proportion of alleles shared between a pair of individuals, the more likely they are to be related.

In the first work presented in this thesis, we review the classical graphical methods used to detect relatedness and introduce the analysis of Compositional Data for relatedness research. For any genetic marker, two individuals can share 0, 1 or 2 alleles. Allele sharing analysis is based on alleles identical by state (IBS) and alleles identical by descent (IBD). Two alleles are IBS if they are identical in terms of their DNA composition and do not necessarily come from a common ancestor. Otherwise, two alleles are IBD if they are derived from a common ancestor. A remarkable difference between IBS and IBD alleles is that IBD is an unobservable measure, and therefore it is necessary to estimate the probabilities of sharing 0, 1 or 2 IBD allelles by maximum likelihood procedures. The IBD probabilities are essential for relatedness research, since they have reference values for any family relationship category and it allows to classify them. Classical graphical methods based on IBS alleles depict the mean and the standard deviations of the number of shared IBS alleles over genetic variants. The scatterplot of the proportion of sharing zero and two IBS alleles has been also considered in the literature. Both representations of allele sharing data are able to detect outliers which correspond to potentially related individuals. Regarding the graphics based on IBD alleles, some authors represent data in an scatterplot of any combination of two out three IBD probabilities. Therefore, we propose the use of tools of Compositional Data analysis such as the ternary diagram and the isometric log-ratio transformation of the IBS/IBD probabilities. The ternary diagram is used to represent simultaneously all three IBS/IBD allele probabilities in contrast to the classical two-dimensional scatterplot. On the other hand, we introduce the isometric log-ratio transformation to overcome the problems of the Euclidean distance interpretation in the constrained space of the IBS/IBD allele sharing data.

In the second article, we propose the analysis of IBS genotype sharing data instead of the classical IBS allele sharing data. This allows us to analyse the genetic data in more than three dimensions. We consider genotype sharing counts as a six-part composition and explore the data using log-ratio biplots based on principal component analysis. Classification of pairs of individuals into family relationship categories is performed using linear discriminant analysis. In this context, the log-ratio biplot approach is compared with the classical plots in a simulation study. In a non-inbred homogeneous population the classification rate of the log-ratio principal component approach outperforms the classical graphics across the whole allele frequency spectrum. Furthermore, the log-ratio biplot is able to identify accurately family relationships up to and including fourth degree relationships. The log-ratio biplot methodology uncovered the

detection of three-quarter siblings, a family relationship which has received very little attention in the literature. Consequently, the third article finishes the thesis with the development of an additional statistical methodology such as the likelihood ratio approach. The likelihood ratio approach is developed to infer three-quarter siblings in genetic databases. We derive the IBD probabilities for three-quarter siblings and calculate likelihood ratios to distinguish three-quarter siblings from full-siblings and half-siblings.

To illustrate all the results of this doctoral thesis we use genetic markers from worldwide human population projects such as the Human Genome Diversity Project and the 1000 Genomes Project, as well as from a local prospective human cohort of the Genomes of Catalonia (GCAT).

Resum

Aquesta tesi doctoral és un compendi de *tres* articles de recerca produïts entre els anys 2015 i 2019. Els tres articles tenen un vincle comú: són aportacions diferents basades en la metodologia de les dades composicionals i en la inferència estadística de relacions familiars. Breument, les *dades composicionals* són vectors aleatoris amb components estrictament positius la suma dels quals és constant. Aquests components representen parts d'un tot que només aporten informació relativa. Per això, les dades composicionals acostumen a representar-se en proporcions o percentatges. L'anàlisi de relacions familiars es basa en el principi del compartiment d'al·lels entre individus per a un conjunt determinat de marcadors genètics. Com més gran és la proporció d'al·lels compartits entre un parell d'individus, més probabilitat hi ha que siguin individus de la mateixa familia.

En el primer treball d'aquesta tesi, revisem els mètodes gràfics clàssics utilitzats per detectar relacions familiars i introduïm l'anàlisi de les dades composicionals per a la investigació de relacions familiars. Per a qualsevol marcador genètic, dos individus poden compartir 0, 1 o 2 al·lels. L'anàlisi de comparticiød'al·lels es basa en al·lels idèntics per estat (identical by state, IBS) i al·lels idèntics per descendència (identical by desent, IBD). Dos al·lels són IBS si són idèntics quant a la seva composició d'ADN i no necessàriament provenen d'un avantpassat comú. En cas contrari, dos al·lels són IBD si provenen d'un avantpassat comú. Una diferència notable entre els al·lels IBS i IBD és que l'IBD és una mesura que no es pot observar, i per tant cal estimar. Les estimacions de les probabilitats de compartir 0, 1 o 2 al·lels IBD es poden realitzar per màxima versemblança. Les probabilitats IBD són essencials per a la investigació en relacions familiars, ja que tenen valors de referència per a qualsevol categoria de relació familiar i això permet classificar-les. Els mètodes gràfics clàssics basats en al·lels IBS representen la mitjana i les desviacions estàndard del número d'al·lels compartits sobre un conjunt de marcadors genètics. Altrament, el gràfic de la proporció de compartir zero i dos al·lels IBS també s'ha considerat en la literatura. Ambdues representacions permeten detectar parelles d'individus que són potencialment de la mateixa familia. Pel que fa als gràfics basats en al·lels IBD, alguns autors representen en un diagrama de dispersió qualsevol combinació de dues de les tres probabilitats IBD. Per tant, proposem l'ús d'eines pròpies de l'anàlisi de dades composicionals com ara el diagrama ternari i la transformació isométrica log-quocient de les probabilitats IBS/IBD. El diagrama ternari s'utilitza per representar simultàniament les tres probabilitats d'al·lels IBS/IBD en contrast amb el clàssic diagrama de dispersió bidimensional. D'altra banda, introduïm la transformació isométrica log-quocient per superar els problemes de la interpretació de la distància Euclidea a l'espai restringit dels al·lels IBS/IBD.

En el segon article, es proposa l'anàlisi de dades de genotips compartits IBS en lloc de les clàssiques dades d'al·lels compartits IBS. D'aquesta manera, podem interpretar les dades amb una dimensionalitat més gran. Considerem que els recomptes de genotips compartits són una composició de sis parts i explorem les dades mitjançant biplots basats en log-quocients derivats de l'anàlisi dels components principals. La classificació de parelles d'individus en les diferents categories de relacions familiars es realitza mitjançant l'anàlisi discriminant lineal. En aquest context, es compara el biplot basat en log-quocients amb els gràfics clàssics en un estudi de simulació.

En una població homogènia sense endogamia, la taxa de classificació correcta del biplot basat en log-quocients és superior als gràfics clàssics a tot l'espectre de freqüències al·leliques. A més, el biplot basat en log-quocients permet identificar amb precisió relacions familiars fins a quart grau. La metodologia del biplot basada en log-quocients va permitir la detecció de tres quarts germans, una relació familiar que ha rebut molt poca atenció en la literatura. En conseqüència, el tercer article finalitza la tesi amb l'elaboració d'una metodologia estadística addicional basada en la raó de versemblances. Aquest enfocament es desenvolupa per inferir tres quarts germans en bases de dades genètiques. Derivem les probabilitats IBD per a tres quarts germans i calculem les raons de versemblança per distingir els tres quarts germans d'entre germans i mig germans (germanastres).

Per il·lustrar tots els resultats d'aquesta tesi doctoral, s'utilitzen marcadors genètics de projectes de població humana procedent de tot el món com el Projecte de la Diversitat del Genoma Humà i el Projecte 1000 Genomes, així com d'una cohort humana prospectiva local dels genomes de Catalunya (GCAT).

Resumen

Esta tesis es un compendio de *tres* artículos de investigación producidos entre 2015 y 2019. Los tres artículos tienen un vínculo común: son diferentes contribuciones basadas en la metodología de los datos composicionales y en la inferencia estadística de relaciones familiares. Brevemente, los *datos composicionales* son vectores aleatorios con componentes estrictamente positivos cuya suma es constante. Estos componentes representan partes de un todo que solo aportan información relativa. Por ello, los datos composicionales generalmente se representan como proporciones o porcentajes. El análisis de relaciones familiares se basa en el principio de alelos compartidos entre individuos en un conjunto de datos de marcadores genéticos. Cuanto mayor sea la proporción de alelos compartidos entre una pareja de individuos, más probable es que sean de la misma familia.

En el primer trabajo presentado en esta tesis, revisamos los métodos gráficos clásicos utilizados para detectar relaciones familiares y presentamos el análisis de los datos composicionales para la investigación de relaciones familiares. Para cualquier marcador genético, dos individuos pueden compartir 0, 1 o 2 alelos. El análisis de compartimiento de alelos se basa en alelos idénticos por estado (IBS) y alelos idénticos por descendencia (IBD). Dos alelos son IBS si son idénticos en términos de su composición de ADN y no necesariamente provienen de un ancestro común. Por otro lado, dos alelos son IBD si se derivan de un antepasado común. Una diferencia notable entre los alelos IBS e IBD es que IBD es una medida no observable y, por lo tanto, es necesario estimarla. Las estimaciones de las probabilidades de compartir 0, 1 o 2 alelos IBD se pueden realizar por máxima verosimilitud. Las probabilidades IBD son esenciales para la investigación de relaciones familiares, ya que dispone de valores de referencia para cualquier categoría de relación familiar y esto permite clasificarlas. Los métodos gráficos clásicos basados en alelos IBS representan la media y las desviaciones estándar del número de alelos IBS compartidos sobre un conjunto de marcadores genticos. Así mismo, el diagrama de dispersión de las proporciones de marcadores para los cuales dos individuos comparten cero o dos alelols IBS también se ha considerado en la literatura. Ambas representaciones pueden detectar parejas de individuos que son potencialmente de la misma familia. Con respecto a los gráficos basados en alelos IBD, algunos autores representan en un diagrama de dispersión cualquier combinación de dos de las tres probabilidades IBD. En este primer trabajo proponemos el uso de las técnicas propias del análisis de datos composicionales, como son el diagrama ternario y la transformación isométrica log-cociente de las probabilidades IBS/IBD. El diagrama ternario se utiliza para representar simultáneamente las tres probabilidades de alelos IBS/IBD en contraste con el diagrama de dispersión bidimensional clásico. A su vez, presentamos la transformación isométrica log-cociente para superar los problemas de la interpretación de la distancia Euclidea en el espacio restringido de los alelos compartidos IBS/IBD.

En el segundo artículo, proponemos el análisis de datos del compartimiento de genotipos IBS en lugar del compartimiento de alelos IBS clásico. De esta forma, podemos interpretar los datos del compartamiento de genotipos con mayor dimensionalidad. Consideramos que los recuentos de compartimiento de genotipos constituyen una composición de seis partes y exploramos los datos utilizando biplots basados en log-cocientes derivados del análisis de componentes principales. La

inferencia estadística de las relaciones familiares se realiza mediante análisis discriminante lineal. En este contexto, el enfoque basado en el bibplot de log-cocientes se compara con las gráficas clásicas en un estudio de simulación. En una población homogénea no endogámica, la tasa de clasificación correcta del enfoque basado en el bibplot de log-cocientes supera a los gráficos clásicos en todo el espectro de frecuencias alélicas. Además, el biplot basado en log-cocientes es capaz de identificar con precisión las relaciones familiares hasta las relaciones de cuarto grado inclusive. La metodología biplot basada en log-cocientes permitió la detección de tres cuartos hermanos, una relación familiar que ha recibido muy poca atención en la literatura. Por ello, el tercer artículo con que finaliza la tesis presenta el desarrollo de una metodología estadística adicional, como es el enfoque basado en la razón de verosimilitud. Este enfoque se desarrolla para inferir tres cuartos hermanos presentes en bases de datos genéticas. Deducimos las probabilidades IBD para tres cuartos hermanos y calculamos razones de verosimilitud para distinguir tres cuartos hermanos de hermanos y medios hermanos (hermanastros).

Para ilustrar todos los resultados de esta tesis doctoral, utilizamos marcadores genéticos de proyectos de población humana procedente en todo el mundo, como son el Proyecto de Diversidad del Genoma Humano y el Proyecto 1000 Genomas, así como de una cohorte humana prospectiva local del Genoma de Cataluña (GCAT).

Chapter 1

Introduction

1.1 Motivation

The research work developed during this doctoral thesis is part of the coordinated project CODA-RETOS "Análisis de datos composicionales y métodos relacionados" (Ref: MTM2015-65016-C2-1-R; Ministerio de Economia y Competitividad), specifically, part of the subproject TRANS-CODA, of which one of the main research lines is "Biomarcadores y marcadores genéticos". The topic of this thesis is based on compositional methodology and statistical inference of family relationships by using genetic markers and forms part of this research line.

Statistical genetics is a branch of statistics based on the analysis of genetic variation and inherited traits. Population-based genetic association studies form an important area in statistical genetics (Foulkes, 2009). These studies search for genetic factors related to disease and assume a random sample of unrelated individuals from a homogeneuos human population. However, in practice, samples of individuals used in association studies often contain one or more individuals from the same family. To accomplish with the assumption of independent individuals, pairs of individuals from the same family are usually identified and removed prior to association analyses (Anderson *et al.*, 2010). Thus, the motivation of this thesis is to study and where possible, improve different statistical techniques that can detect and identify family relationships between individuals from the same human population.

1.2 Background

A good understanding of basic genetic principles is necessary for what follows (Chapter 1; Laird and Lange (2010)). Briefly, the human genome refers to all the genetic material that is inherited across generations. The genetic information is stored on chromosomes located in the nucleus of the cell. The human genome is composed of 23 pairs of chromosomes. From each pair, one chromosome is inherited from the mother and the other one from the father. A chromosome is a deoxyribonucleic acid (DNA) molecule formed by large nucleotide sequences, associated with proteins. The DNA sequences are constituted of four nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). Variations in the DNA sequence at a specific site in the human genome are referred to as alleles. The genotype of one individual is formed by the two alleles inherited from his or her parents. A pair of genotypes is in linkage disequilibrium if their alleles are inherited together by the descendants (Chapter 5; Laird and Lange (2010)). The group of alleles that are inherited together from a single parent is known as a haplotype. Genetic markers are known locations of the human genome that have allelic variation and contain information concerning the family relationships of the individuals that have been genotyped. The most common types of genetic markers are the single nucleotide polymorphism (SNP) and the microsatellite or short tandem repeat (STR).

Microsatellites are short DNA sequences which are repeated. The length of the repeated DNA sequences is constant for each STR and ranges from 2 to 6 nucleotides. Individuals have genetic variability because their alleles of determined regions of DNA vary. Microsatellites are very powerful to distinguish each individual from the population due to the presence of multiple alleles which lead to high genetic variability between individuals. Particularly, the number of the repeated sequences across STRs varies between unrelated individuals. For this reason, they are commonly used for forensic DNA studies and paternity testing (Vieira *et al.*, 2016). On other hand, SNPs are common throughout the human genome, they occur once in every 300 nucleotides on average. Nowadays, the economic cost of genotyping multiple SNPs in the human genome is decreasing. Consequently, SNPs are mostly used for large scale genome wide association studies (GWAS, Visscher *et al.* (2017)).

Many relatedness investigations are based on the principle of allele-sharing. A pair of individuals can share 0, 1 or 2 alleles for any genetic marker. The allele sharing can be considered either *by state* if the DNA composition of the alleles are identical but the alleles do not come from a common ancestor (identical by state, IBS), or *by descent* if the alleles originate from a common ancestor (identical by descent, IBD) (pages 195-196; Laird and Lange (2010)). The degree to which individuals share alleles indicates the extent to which they are related. Thus, individuals from the same family share on average more alleles than unrelated individuals.

Relatedness investigations are performed for mainly two reasons: verifying documented relationships and guaranteeing independence of individuals in the database. Population based genetic association studies are focused on the analysis of data derived from homogeneous populations of unrelated individuals with the aim to measure the disease status under investigation such as cancer or other complex diseases (Foulkes, 2009). Many statistical methods used in these studies (such as standard regression models, t-tests, logistic regression) assume that the observations (individuals) are independent. These techniques can fail and inflate the false positive rate in association genetic studies if independence is not satisfied. This phenomenon is also referred to as "cryptic relatedness" (Voight and Pritchard, 2005). For this reason, this thesis analyzes the degree of dependence between individuals by using allele sharing analysis. This allows the analyst to filter the database by removing one individual from each pair of the detected family relationships prior to association analyses. However, classical allele sharing analysis based on proportions or probabilities of IBS/IBD alleles does not take into account that these data occupy a constrained space. This property is characteristic of Compositional Data where standard statistical methods can lead to a misinterpretation of the data (Chapter 1, Pawlowsky-Glahn *et al.* (2015)). Hence, the novelty of the present thesis is to introduce compositional statistical techniques in IBS and IBD allele sharing studies.

Compositional Data (CoDa) are random vectors (also referred to as compositions) with strictly positive components whose sum is constant. These components represent parts of a whole which only carry relative information. Many examples of CoDa can be found in different fields: geology (geochemical elements), economy (income/expenditure distribution), demography (population percentages), ecology (abundance of different species), metabolomics (molar concentrations), microbiome (relative abundances of bacteria), genetics (genotype frequencies), etc. As aforementioned, Compositional Data have the property that the compositions occupy a constrained space whose values range from 0 to 100, or any other constant. Such a restricted space is known as simplex (Aitchison, 1986). This is a remarkable difference with the standard statistical approaches which assume that variables occupy the usual Euclidean space whose values range from $-\infty$ to $+\infty$. Another property of CoDa is that distances between compositions should satisfy the scale invariance, permutation invariance and subcompositional coherence principles (Aitchison,

1992). These principles are not satisfied in the case of using the standard Euclidean distance between compositions (Aitchison *et al.*, 2000). To deal with the restricted space of CoDa and the problems of interpretation of distances between compositions, transformations such as the additive log-ratio transformation (alr, Aitchison (1986)), the centered log-ratio transformation (clr, Aitchison (1986)) or the isometric log-ratio transformation (ilr, Egozcue *et al.* (2003)) are commonly applied (Pawlowsky-Glahn and Buccianti, 2011; Pawlowsky-Glahn *et al.*, 2015). This is also referred to as the principle of working in coordinates because operations and metrics in the simplex are equivalent to ordinary operations and metrics in coordinates (Mateu-Figueras *et al.*, 2011). In order to satisfy the former principles described of CoDa, we take into account the compositional nature of the proportions or probabilities of the allele sharing analysis of identical by state/descent alleles. In this way, the inference of family relationships based on scatterplots of log-ratio transformations using compositional distances will be properly defined.

This thesis illustrates two situations where the techniques from the analysis of Compositional Data enrich the statistical methods used in the allele sharing analysis:

- Graphics.
  There are several graphical methods for representing pairs of individuals and detecting family relationships by allele sharing analysis. It is possible to detect related individuals by plotting percentages of sharing 0, 1 or 2 IBS alleles across all the genetic markers (Rosenberg, 2006). Another option for detecting close relatives graphically is by plotting the means and variances of the number of shared IBS alleles between individuals across genetic markers (Abecasis *et al.*, 2001). Studies of relatedness can also be based on the probabilities that the alleles are shared identical by descent. These probabilities depend on the relatedness: monozygotic twins, full-siblings, parent-offspring, avuncular, first cousins, etc. Identity by descent is essential to research on relatedness. The probabilities of sharing 0, 1 and 2 alleles IBD are known as the Cotterman coefficients, denoted by k_0, k_1 and k_2 respectively (Cotterman, 1941). These can be estimated by maximum likelihood (Thompson, 1975, 1991; Milligan, 2003), the IBD probabilities (k_0, k_1, k_2) are chosen to maximize the probability of an observed pair of genotypes given by the likelihood function across genetic markers (Wagner *et al.*, 2006; Weir *et al.*, 2006). The kinship coefficient (ϕ) is also relevant for relatedness research and is defined as $\phi = k_1/4 + k_2/2$. The IBD probabilities and the kinship coefficient are commonly plotted in a scatterplot for determining the family relationships. Compositional Data Analysis can be applied to allele-sharing analysis because the fraction of sharing 0, 1 and 2 IBS alleles (denoted by p_0, p_1 and p_2 respectively) and the Cotterman coefficients can be considered as 3-part compositions. This is due to the fact that the three components sum to one ($p_0 + p_1 + p_2 = 1$ and $k_0 + k_1 + k_2 = 1$). This approach provides two graphical methods to detect family relationships by plotting the vector of (p_0, p_1, p_2) or (k_0, k_1, k_2) in a ternary diagram and by plotting the isometric log-ratio transformation of the vector of (p_0, p_1, p_2) or (k_0, k_1, k_2) in ilr-coordinates (Egozcue *et al.*, 2003).

- Log-ratio biplots.
  As aforementioned, the usual graphical approach for detecting family relationships is to plot allele sharing probabilities, either IBS or IBD, in a two-dimensional scatterplot. This approach ignores that allele sharing data across individuals has in reality a higher dimensionality, and neither regards the compositional nature of the counts of shared genotypes. The log-ratio biplot based on principal component analysis of Compositional Data overcomes these restrictions (Aitchison, 1983; Aitchison and Greenacre, 2002). This leads to entirely new graphics that are essentially useful for exploring relatedness in genetic databases from homogeneous populations.

On other hand, the log-ratio biplot methodology uncovered the existence of three-quarter siblings, a family relationship which has received very little attention in the literature. For this reason, this thesis develops additional statistical methodology such as the likelihood ratio approach in order to confirm and identify three-quarter siblings:

- Three-quarter siblings.
  Existing IBS/IBD methods are able to identify first degree family relationships (parent-offspring or full-siblings), second degree (half-siblings, avuncular or grandparent-grandchild) or more distant relationships. Three-quarter siblings (3/4S) is a family relationship whose individuals share fewer alleles than first degree relationships and more alleles than second degree relationships. A 3/4S pair has one common parent, while their unshared parents can be full-siblings or parent-offspring. In practice, the 3/4S relationship is hard to discover in the usual scatterplot of the IBD probabilities. Thus, it opens the doors to other methodologies such as the likelihood ratio approach (Thompson, 1986; Boehnke and Cox, 1997; Weir *et al.*, 2006; Katki *et al.*, 2010; Heinrich *et al.*, 2016) to infer this type of relationship.

1.3 Research articles

The situations described in the former section derive three original works focused on the estimation of family relationships by using genetic markers.

- The first article of this doctoral thesis is entitled **Graphics for relatedness research** and has been published in *Molecular Ecology Resources*. A copy of this article can be found at page 31. In this article, we review the most common graphics used in IBS/IBD allele sharing analysis for identifying family relationships. Furthermore, two additional graphical methods from the field of compositional data analysis are proposed: the ternary diagram to display all three allele sharing probabilities simultaneously and scatterplots of isometric log-ratios of IBS/IBD probabilities to overcome the problems with the Euclidean distance interpretation in the classical graphics. We illustrate all graphical tools with genetic data from the HGDP-CEPH diversity panel (Rosenberg *et al.*, 2002), using 377 microsatellites genotyped for 25 individuals from the Maya population of this panel. R functions for making the graphics of this article are available from `https://github.com/ivangalvan/graphics-relatedness-research`.

- The second article is entitled **A log-ratio biplot approach for exploring genetic relatedness based on identity by state** and has been published in *Frontiers in Genetics*. A copy of this article can be found at page 45. In this article, we propose a log-ratio biplot approach based on principal component analysis to identify family relationships by using only IBS alleles. The proposed approach takes into account the compositional nature of the 6-part composition derived from the genotype sharing counts. This leads to new graphics for detecting relatedness with higher dimensionality than the two-dimensional classical graphics. The discriminatory power of the log-ratio biplot approach outperforms the classical plots in a simulation study. Furthermore, simulations show that with 35,000 independent biallelic variants, log-ratio principal component analysis, combined with discriminant analysis, can correctly classify relationships up to and including the fourth degree. Genome-wide SNP datasets from the 1,000 Genomes Project (1000 Genomes Project Consortium *et al.*, 2015) and the GCAT Genomes For Life cohort project (Obón-Santacana *et al.*, 2018; Galván-Femenía *et al.*, 2018) are used to illustrate the proposed method. R code for reproducing the log-ratio biplot developed in this this article is available from `https://github.com/ivangalvan/LR-kinbiplot`.

- The third article is entitled **A likelihood ratio approach for identifying three quarter siblings in genetic databases** and has been submitted to *Heredity*. A copy of this article can be found at page 63. In this article, we derive the theoretical IBD probabilities and the kinship coefficient for three-quarter siblings (3/4S), a relationship whose individuals share fewer alleles than first degree relationships but more alleles than second degree relationships. We show that the detection of 3/4S in a scatterplot of the IBD probabilities is difficult. For this reason, we propose a likelihood ratio approach to distinguish 3/4S from full-siblings and second degree relatives. We use simulated data and a genome-wide array dataset from the GCAT Genomes for Life cohort project to illustrate the procedure. R code for using the likelihood ratio approach to detect 3/4S is available from `https://github.com/ivangalvan/LR-3.4S`.

1.4 Structure of the thesis

The remainder of this doctoral thesis is structured as follows. Chapter 2 describes the objectives of this thesis. Chapter 3 summarizes the statistical methods used to identify family relationships. Chapter 4 is the core of this thesis and shows a copy of the published and submitted articles. Chapter 5 synthesizes the main results and contains a discussion of the articles. The thesis finishes with the conclusions and further lines of research.

Chapter 2

Objectives

The main aim of this doctoral thesis is to study and propose statistical methods to identify family relationships by using genetic markers. The methods based on the analysis of Compositional Data provide two original contributions in the field of relatedness research. The first contribution is based on the use of ternary diagrams and scatterplots of log-ratio transformations. The second contribution is based on log-ratio biplots derived from principal components analysis. On the other side, statistical inference based on a likelihood ratio approach provides another contribution for identifying three quarter siblings relationships.

2.1 Objectives of the core of the doctoral thesis

The objectives of this thesis can be enumerated as follows:

Obj. 1. Review or study the classical statistical methods based on identity by state/descent for identifying family relationships.

Obj. 2. Use of the statistical methods from Compositional Data Analysis for relatedness research.

Obj. 3. Provide a statistical method for identifying three quarter sibling relationships.

These objectives have been addressed in three different publications that have been reviewed or are in revision by external reviewers. Table 2.1 shows the relationship between the publications and the objectives addressed in this thesis.

	Objective		
Research article	1	2	3
Graphics for relatedness research	✓	✓	
A log-ratio biplot approach for exploring genetic relatedness based on identity by state	✓	✓	
A likelihood ratio approach for identifying three quarter siblings in genetic databases	✓		✓

Table 2.1: Relationship between the objectives from the core of the thesis and the research articles published or submitted.

Chapter 3

Methodology

In this chapter the main methods used in this doctoral thesis are described. Firstly, the statistical methods from the field of Compositional Data Analysis used for relatedness research are illustrated. We outline the basic principles of the log-ratio methodology, the geometric structure of the simplex and the log-ratio principal component analysis, followed by an overview of the classical methods for identifying family relationships based on identical by state/descent alleles.

3.1 Compositional Data

This section is a summary of the basis of the analysis of Compositional Data (CoDa). The examples, notation and organization of the following five subsections have been extracted with prior consent from the doctoral thesis of Comas Cufí (2019), whose text is based on the lecture notes of Pawlowsky-Glahn *et al.* (2011) and the doctoral thesis of Martín-Fernández (2001); Mateu-Figueras (2003) and Vives-Mestres (2014). Further reading on the analysis of Compositional Data can be found in the books Pawlowsky-Glahn and Buccianti (2011) and Pawlowsky-Glahn *et al.* (2015).

Following on from the developments of Compositional Data in the last decades, a compositional vector of D parts, $\mathbf{x} = (x_1, x_2, \ldots, x_D)$, is defined as a vector in which the only relevant information is contained in the ratios between its components. All the components of the vector are assumed strictly positive whose sum is constant. Hereafter, components are called *parts* and a compositional vector is called a *composition*.

The assertion that all the relevant information is contained in the ratios implies that, if α is a real positive number, then $(x_1, x_2, \ldots, x_D)$ and $(\alpha x_1, \alpha x_2, \ldots, \alpha x_D)$ convey the same information and are indistinguishable. Therefore, a composition is a class of equivalent compositional vectors (Barceló-Vidal and Martín-Fernández, 2016).

The constant sum constraint of Compositional Data can complicate the statistical analysis and the interpretation of the data. For example, the classical correlation coefficient between two components of a composition cannot be interpretable as usual. In fact, K. Pearson stated that components with the same denominator provide a false or spurious correlation (Pearson, 1897). This fact makes the interpretation of Compositional Data from a classical statistical point of view difficult.

Consequently, Aitchison (1982, 1986) developed a specific methodology with the main idea that Compositional Data represent parts of a total and therefore the only information they have is the relative. That is, the only way to obtain information from one part is by comparing with another part. This leads to the ratios between parts and for mathematical convenience to the

analysis of log-ratios. Hence, it can be said that the analysis of Compositional Data is based on the analysis of the log-ratios between the parts of a composition.

3.1.1 Basic concepts

Definition 3.1 A D-part composition is a vector $(D \times 1)$ whose components $x_1, x_2, \ldots, x_D$ are strictly positive real numbers (i.e. $x_1 > 0, x_2 > 0, \ldots, x_D > 0$), whose sum is constant $x_1 + x_2 + \ldots + x_D = \kappa$ and have relative information.

Commonly, $\kappa = 1$ or $\kappa = 100$ if the data is transformed to proportions or percentages respectively. It is worth noting that a composition do not necessarily sum a constant, data that is measured in concentrations units such as mg/l or molar, and data that represent relative abundances such as microbiome datasets are also compositional (Gloor *et al.*, 2017).

Definition 3.2 The sample space of the compositions is the simplex $\mathcal{S}^D$, and is defined as

$$\mathcal{S}^D = \{(x_1, x_2, \ldots, x_D) | x_i > 0, i = 1, 2, \ldots, D; \sum_{i=1}^{D} x_i = \kappa\}.$$

The 3-part compositions with $\kappa = 1$ $(D = 3)$ are inscribed in an equilateral triangle in $\mathbb{R}^3$, located at the perpendicular plane to the vector $(1, 1, 1)$ (Figure 3.1, left). However, it is more usual to represent the data in the ternary diagram (Figure 3.1, right), which is an equivalent representation. A ternary diagram is an equilateral triangle whose points $\mathbf{x} = (x_1, x_2, x_3)$ are located to a distance x_1 from the opposite side of the vertex X_1, to a distance x_2 from the opposite side of the vertex X_2 and to a distance x_3 from the opposite side of the vertex X_3. In the case $D = 4$, the simplex is represented on a regular tetrahedron of height equals to one.

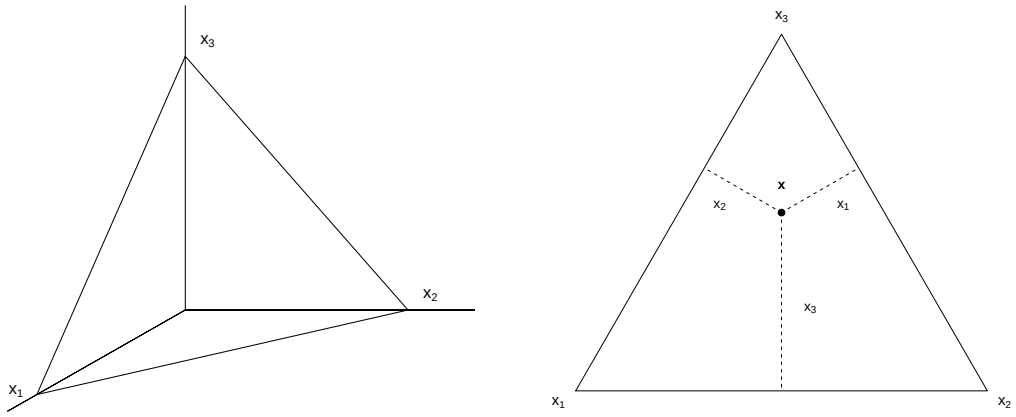

Figure 3.1: Equivalent representations of the 3-part compositions in $\mathbb{R}^3$ (left) and in the ternary diagram (right).

In order to obtain the constant sum constraint κ, the composition is divided by the total sum of the parts and multiplied by κ. This operation is known as *closure*.

Definition 3.3 For any vector of D real positive components $\mathbf{x} = (x_1, x_2, \ldots, x_D) \in \mathbb{R}_+^D$ the closure of $\mathbf{x}$ is defined as:

$$\mathcal{C}(\mathbf{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right).$$

From a compositional point of view, note that the closure operation does not modify the relative information between the parts of a composition. The closure operation provides a criterion to select one of the representatives of the equivalence class (Barceló-Vidal and Martín-Fernández, 2016). The graphical representation of the closure operation is shown in Figure 3.2 (left): the closure of $\mathbf{x}$ moves the point through the line (equivalence class) from the origin to $\mathbf{x}$ to the intersection with the plane $\sum x_i = \kappa$. Instead of $\mathbf{x}$, the new representative of the equivalence class will be $\mathcal{C}(\mathbf{x})$.

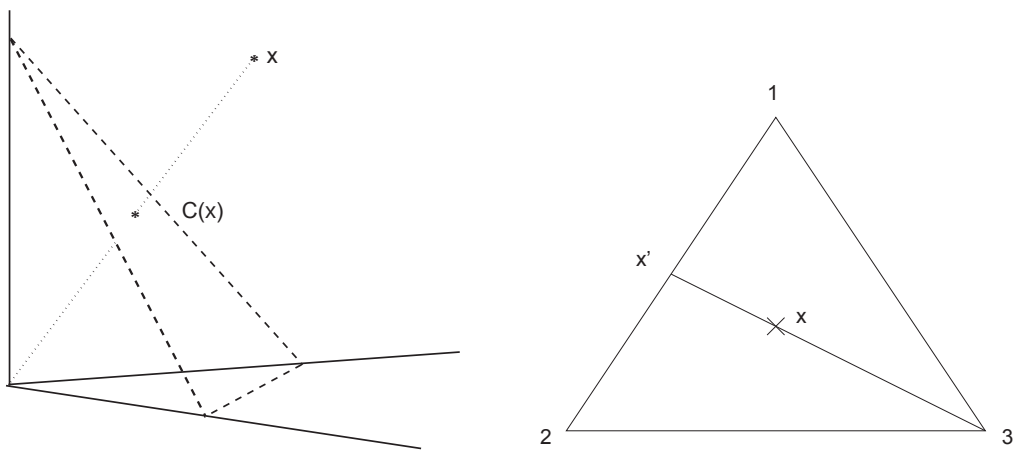

Figure 3.2: Left: Graphical representation of the closure operation. Right: Subcomposition $\mathbf{x}' \in \mathcal{S}^2$ represented as a linear projection of $\mathbf{x} \in \mathcal{S}^3$.

Because the analysis of Compositional Data relies on the relative information (ratios between parts), it can be shown that all the log-ratios of a D-part composition can be obtained by the $D-1$ ratios x_i/x_D for $i = 1, 2, \ldots, D-1$ (Aitchison, 1986). Thus, the dimension of a D-part composition is $D-1$.

Frequently, attention is focused on a group of parts of a composition. Thus, ratios of parts within the selected group are considered relevant, whereas ratios involving some part, not in the group, are ignored. This corresponds to the definition of a subcomposition including only the parts in this group. The graphical representation of the subcomposition in the simplex can be considered as a composition of lower dimension obtained by projection. Figure 3.2 (right) shows the subcomposition $\mathbf{x}' \in \mathcal{S}^2$ formed by the first two parts of $\mathbf{x} \in \mathcal{S}^3$, this is the result of projecting $\mathbf{x}$ into the side 12 from the vertex 3.

3.1.2 Principles of the analysis of Compositional Data

The statistical methods applied to Compositional Data should satisfy specific principles. These principles should be coherent with the nature of Compositional Data. All principles of Compositional Data analysis must be based on the following general principle: "Compositional Data quantitatively describe the parts of some whole and they provide only relative information between their components". This general principle is embodied in the following particular principles: scale invariance, permutation invariance and subcompositional coherence (Aitchison, 1992).

The scale invariance principle states that the results of the analysis are the same for any measured unit of the composition. The analysis of ratios satisfies this principle, because in the ratio $x_1/x_2 = (\lambda x_1)/(\lambda x_2)$ the scaling factor λ cancels out. However, the order of the ratios is relevant,

that is $x_1/x_2 \neq x_2/x_1$. Hence, the log transformation is adequate for the analysis, $\log x_1/x_2$, because the inversion of the order of the components causes only a change of the sign. This implies a symmetry with regard to the order of the parts.

The permutation invariance principle states that the conclusions of the analysis of Compositional Data do not depend on the order of the parts. The results obtained are the same if the order of the parts are changed.

The subcompositional coherence principle states that the results obtained in the analysis of a subcomposition do not change the results obtained from the analysis of the full composition. Note that the ratio of two components remains unchanged when we move from full composition to any subcomposition. Therefore, as Compositional Data analysis is based on log-ratios, it accomplishes this principle of subcompositional coherence

3.1.3 Transformations of the simplex to the real space based on log-ratios

In this section we show two common transformations of the simplex that allow to work in the coordinates of the real space. These transformations are the additive log-ratio (alr) transformation and the centered log-ratio (clr) transformation (denoted $\boldsymbol{w}$ and $\boldsymbol{z}$). By using these transformations, it is possible to work with the ordinary operations in the real space as if you were in the simplex. This is known as the principle of working in coordinates (Mateu-Figueras *et al.*, 2011), as we will see in section 3.1.5 .

Alr transformation

Aitchison (1986) defines the additive log-ratio transformation as:

Definition 3.4 Let $\mathbf{x}$ be a D-part composition, the additive log-ratio transformation of $\mathbf{x} \in \mathcal{S}^D$ to $\boldsymbol{w} \in \mathbb{R}^{D-1}$ is defined as

$$\boldsymbol{w} = \operatorname{alr}(\mathbf{x}) = \left(\log \frac{x_1}{x_D}, \log \frac{x_2}{x_D}, \ldots, \log \frac{x_{D-1}}{x_D} \right).$$

The alr transformation is bijective and the inverse transformation is alr^{-1} that is defined as

$$x_i = \frac{\exp w_i}{\sum_{j=1}^{D-1} \exp w_j + 1} \qquad (i = 1, 2, \ldots, D-1),$$

$$x_D = 1 - \left(\sum_{i=1}^{D-1} x_i \right) = \frac{1}{\sum_{j=1}^{D-1} \exp w_j + 1}.$$

A limitation of the alr transformation is the lack of symmetry, because the component of the denominator of the log-ratio acquires a special attention in comparison with the others components. In fact, it is possible to choose another component as a common denominator.

Clr transformation

Aitchison (1986) defines the centered log-ratio transformation as:

Definition 3.5 $\mathbf{x}$ be a D-part composition, the additive log-ratio transformation of $\mathbf{x} \in \mathcal{S}^D$ to $\boldsymbol{z} \in \mathbb{R}^D$ is defined as

$$\boldsymbol{z} = \operatorname{clr}(\mathbf{x}) = \left(\log \frac{x_1}{g(\mathbf{x})}, \log \frac{x_2}{g(\mathbf{x})}, \ldots, \log \frac{x_D}{g(\mathbf{x})} \right)$$

where $g(\mathbf{x}) = (x_1 \cdot x_2 \cdots x_D)^{1/D}$ is the geometric mean of the D compononents of $\mathbf{x}$.

In this case, the transformation is symmetric. The transformed data is located at the hyperplane V of $\mathbb{R}^D$ that intersect with the origin and is orthogonal to $(1, 1, \ldots, 1)$, that is, $V = \text{clr}(\mathcal{S}^D) = \{\boldsymbol{z} \in \mathbb{R}^D; \sum_{i=1}^D z_i = 0\}$. It implies another limitation, the covariance of the clr coordinates is singular, since the sum of the components of the transformed vector is equals to zero.

The clr transformation satisfies the scale and permutation invariance principles. Furthermore, this transformation is bijective between the simplex and the hyperplane V and the inverse (clr^{-1}) is defined as

$$\mathbf{x} = \text{clr}^{-1}(\boldsymbol{z}) = \mathcal{C}(e^{z_1}, e^{z_2}, \ldots, e^{z_D}).$$

3.1.4 The geometric structure of the simplex

The methodology of Compositional Data based on the analysis of log-ratios is equivalent to defining an Euclidean metric structure in the simplex (Egozcue and Pawlowsky-Glahn, 2006; Barceló-Vidal and Martín-Fernández, 2016). In the simplex, standard operations and metrics are not the same as in the real space. However, it is possible to define two operations in order to find a way of working that is completely analogous. These operations are the *perturbation* and *powering*. The perturbation is analogous to addition in real space and powering is analogous to multiplication by a scalar in real space. Both require in their definition the closure operation.

Definition 3.6 Let $\mathbf{x}, \mathbf{x}^*$ be two D-part compositions. Then, the perturbation operation is defined as:

$$\mathbf{x} \oplus \mathbf{x}^* = \mathcal{C}\left(x_1 x_1^*, x_2 x_2^*, \ldots, x_D x_D^*\right)$$

Definition 3.7 Let $\mathbf{x}$ be a D-part composition and let α be a scalar in $\mathbb{R}$. Then, the powering operation is defined as:

$$\alpha \otimes \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)$$

Figure 3.3 shows the results of these operations in a sample of compositions in $\mathcal{S}^3$.

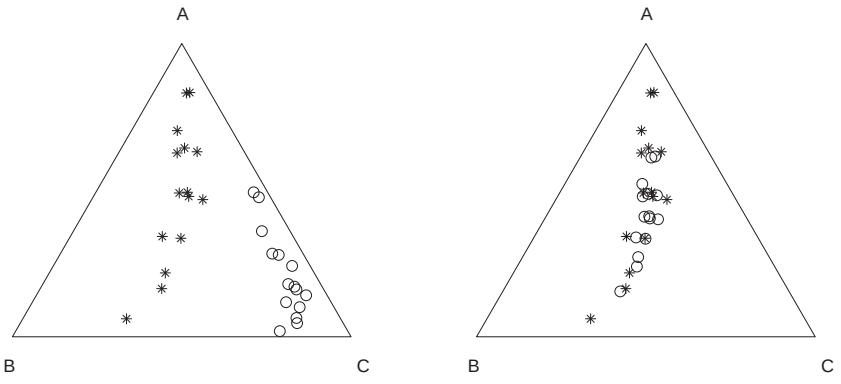

Figure 3.3: Left: Perturbation of the initial compositions $*$ by $p = (0.1, 0.1, 0.8)$ that leads to $\circ$. Right: Powering of the initial compositions $*$ by $\alpha = 0.2$ that leads to $\circ$.

The perturbation and powering operators satisfy the properties required to give a vector space structure to the simplex $\mathcal{S}^D$. The clr transformation defined in the former section is a linear map from the vector space $\mathcal{S}^D$ to the real $(D-1)$ subspace V of $\mathbb{R}^D$, since it holds

$$\text{clr}(x \oplus y) = \text{clr}(x) + \text{clr}(y) \text{ and } \text{clr}(\alpha \otimes x) = \alpha \cdot \text{clr}(x),$$

for any $\mathbf{x} \in \mathcal{S}^D$ and $\alpha \in \mathbb{R}$.

This property allows to define in the simplex an inner product, a norm and a distance in correspondence with the same standard operations defined in the subspace V of $\mathbb{R}^{D-1}$. All these operations confer to simplex an Euclidean vector space structure that allows to work with Compositional Data in the simplex in the same manner as in the real space. Further reading on the geometric structure of the simplex can be found on the books Pawlowsky-Glahn and Buccianti (2011) and Pawlowsky-Glahn *et al.* (2015).

3.1.5 Ilr transformation: working in coordinates

Egozcue *et al.* (2003) define an isometry between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$ by using the Aitchison distance. The main motivation of this transformation is to overcome the limitations of the two previous defined transformations: the non permutation invariance of the alr and the singular covariance structure of the clr coordinates.

The ilr transformation appears naturally from the clr transformation. The condition $\sum z_k = 0$ satisfied by the components of the subspace $V = \text{clr}(\mathcal{S}^D)$, indicate that the clr coordinates are located in the hyperplane with normal vector $(1, 1, \ldots, 1)$. Thus, it is possible to choose an orthonormal basis to identify any clr coordinate in the subspace V of dimension $D-1$ of $\mathbb{R}^D$.

This procedure, clr transformation followed by change to orthonormal basis and orthogonal projection to the subspace V, leads to an isometry between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$ by considering the Aitchison distance defined from the clr transformation. Egozcue *et al.* (2003) define this transformation as follows:

Definition 3.8 Given an orthonormal basis from the simplex $\mathcal{S}^D$, $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1})$, and the matrix of dimension $(D-1 \times D)$ in $\mathbb{R}^{D-1}$, $\mathbf{\Psi} = \begin{pmatrix} \text{clr}(\mathbf{e}_1) \\ \text{clr}(\mathbf{e}_2) \\ \ldots \\ \text{clr}(\mathbf{e}_{D-1}) \end{pmatrix}$, the isometric log-ratio transformation of a composition $\mathbf{x} \in \mathcal{S}^D$ to a vector $\boldsymbol{y} \in \mathbb{R}^{D-1}$ is

$$\boldsymbol{y} = \text{ilr}(\mathbf{x}) = clr(\mathbf{x}) \cdot \mathbf{\Psi}'.$$

In the same way as the alr and clr transformations, the ilr transformation has an inverse transformation (ilr^{-1}) and satisfies the three principles of the analysis of Compositional Data.

The ilr transformation is not unique, given that the orthonormal basis of $\mathcal{S}^D$ is not specified in its definition of isometry and therefore it is possible to choose the basis freely. The coordinates of the compositions with respect to the orthonormal basis are just the *ilr coordinates*. The ilr coordinates allow to work with compositions as usual (principle of working in coordinates, Mateu-Figueras *et al.* (2011)). That is, it is possible to use the ordinary operations of the real space in order to work with the Euclidean distance and to apply the ordinary inner product to the ilr transformed data. Egozcue *et al.* (2003) provide the relationships between the three alr, clr and ilr transformations.

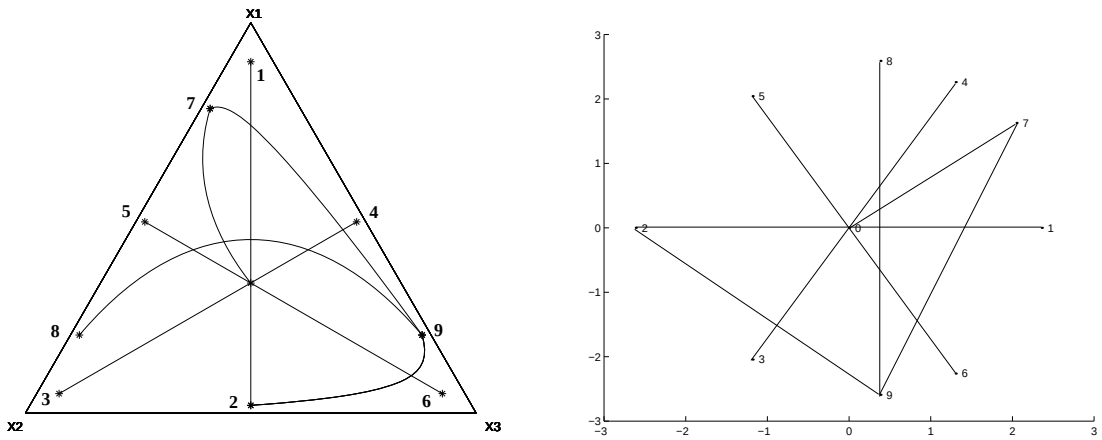

Figure 3.4: Left: Compositional lines in the simplex $\mathcal{S}^3$. Right: Equivalence of the compositional lines in the real space $\mathbb{R}^2$.

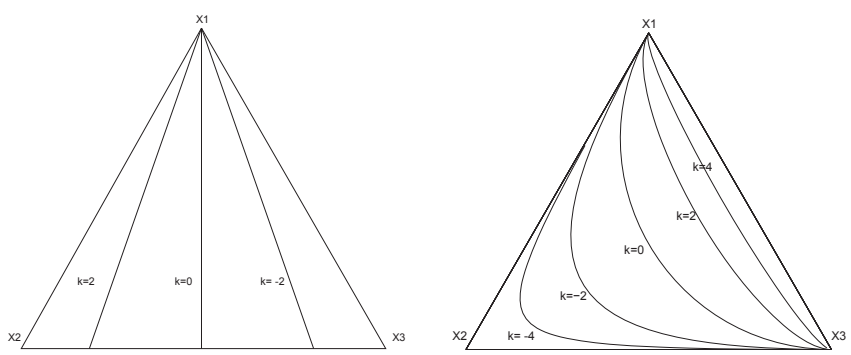

Figure 3.5: Parallel lines in the simplex. Left: $\log x_2 - \log x_3 = k$ for $k = -2, 0, 2$. Right: $\log x_1 - 2\log x_2 + \log x_3 = k$ for $k = -4, -2, 0, 2, 4$.

3.1.6 Geometry in the simplex

In this section, we present some Figures with the aim to show graphically that the geometry in the simplex is different to the usual Euclidean geometry of the real space.

Figure 3.4 shows compositional lines in the simplex $\mathcal{S}^3$ and equivalent lines in the space of orthonormal coordinates. It is shown that the perpendicular lines of the real space $\overline{12}$ and $\overline{89}$ are deformed in the constrained space. This phenomenon also occurs with angles. The right angle between the segments $\overline{50}$ and $\overline{07}$ of the real space (Figure 3.4 right) is deformed in the simplex (Figure 3.4 left).

Figures 3.5 and 3.6 show examples of parallel and orthogonal lines in the simplex $\mathcal{S}^3$ respectively. From these Figures, it can be shown that the visualization of line, parallelism and orthogonality in the real space is not valid in the space of Compositional Data, despite the fact that both are Euclidean metric spaces.

Finally, Figure 3.7 shows some circumferences in the $\mathcal{S}^3$. In the same manner that occurs with lines, the profiles of these compositional circumferences are not similar with the standard profiles of circumferences in the real space. From an Euclidean point of view, the closeness of the circumferences to the border of the simplex draws distortions in the profiles. This is due to the fact that distance between two nearby points located close to the border of the simplex is much larger

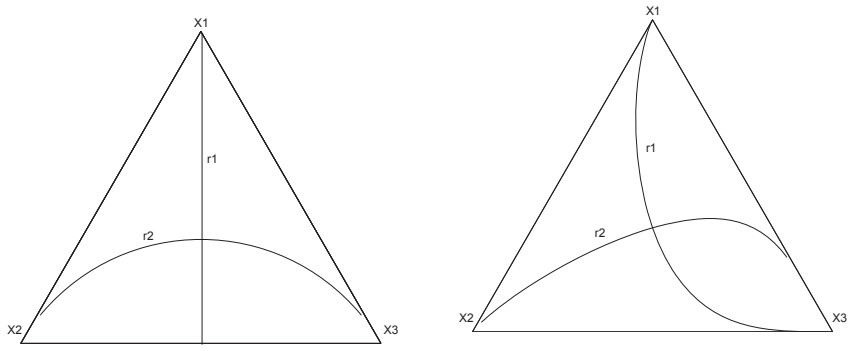

Figure 3.6: Orthogonal lines in $\mathcal{S}^3$. Left: $r_1 : x_2 = x_3$ i $r_2 : 2\log x_1 - \log x_2 - \log x_3 = 0$. Right: $r_1 : \log x_1 - 3\log x_2 + 2\log x_3 = 0$ i $r_2 : 5\log x_1 - \log x_2 - 4\log x_3 = 0$.

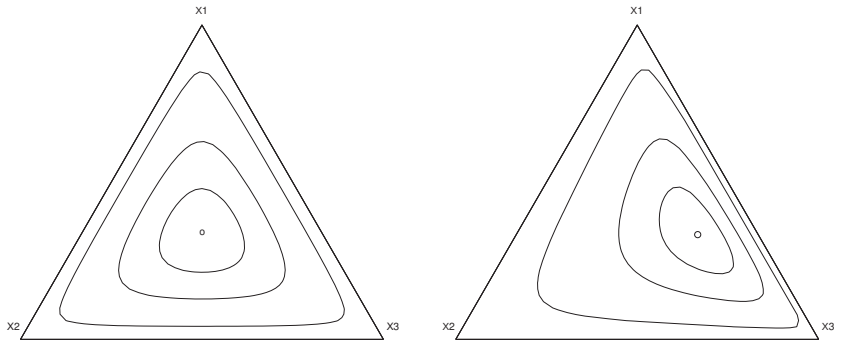

Figure 3.7: Circumferences in $\mathcal{S}^3$ of radius $r = 0.5, 1, 2$. Left: Center ($\circ$) at $(1/3, 1/3, 1/3)$ which is the barycenter of the triangle. Right: Center ($\circ$) at $(2/6, 1/6, 3/6)$.

than the distance between two points with the same closeness located at the center of the simplex.

3.1.7 Log-ratio principal component analysis

Aitchison (1983) proposed log-ratio principal component analysis (PCA) for the exploration of Compositional Data. The log-ratio PCA is usually performed by applying the centered log-ratio transformation to the Compositional Data. A detailed derivation of this approach can be found in Aitchison and Greenacre (2002) and Pawlowsky-Glahn *et al.* (2015). We briefly summarize log-ratio PCA and biplot construction.

Let X be a matrix with n compositions in its rows, and having D parts (columns). Let X_l be the log transformed compositions, that is $X_l = \ln(X)$. The clr transformed data can be obtained by just centering the rows of this matrix, using the centering matrix $H_r = I - \frac{1}{D}11'$, where I is the identity matrix and 1 is a vector of ones. Then

$$X_{clr} = X_l H_r.$$

The rows of X_{clr} are subject to a zero sum constraint because $H_r 1 = 0$. Now we column-center the clr transformed data, producing a double-centered data matrix that has zero column and row means:

$$X_{cclr} = H_c X_{clr} = H_c X_l H_r,$$

where H_c is the centering matrix $H_r = I - \frac{1}{n}11'$. Matrix X_{cclr} is used as the input for a classical principal component analysis. We perform PCA by the singular value decomposition:

$$X_{cclr} = UDV' = F_p G_s',$$

with $F_p = UD$ and $G_s = V$. Matrix F_p contains the principal components, and its first two columns contain the biplot coordinates of the compositions. The columns of G_s are the eigenvectors of the covariance matrix of X_{cclr}, its first two columns contain the biplot coordinates of the parts of the compositions.

The projection of supplementary compositions onto a given biplot can be accomplished by regression (Graffelman and Aluja-Banet, 2003). The biplot coordinates, $\tilde{F}_p$, of a matrix of supplementary compositions, Y, can be found as:

$$\tilde{F}_p = (G_s' G_s)^{-1} G_s' Y_{cclr},$$

where Y_{cclr} contains the clr-transformed supplementary compositions, but centered with respect to the compositions in X, that is

$$Y_{cclr} = Y_{clr} - \frac{1}{n}11' X_{clr}.$$

3.2 Relatedness research

This section presents an overview of the statistical methods for relatedness research based on the principle of allele sharing analysis. A pair of individuals can share 0, 1 or 2 alleles for any autosomal genetic marker. This sharing can be assessed either by state or by descent (pages 195-196; Laird and Lange (2010)):

- Two alleles are *identical by state* (IBS) if they are identical in terms of their DNA composition and do not necessarily come from a common ancestor.
- Two alleles are *identical by descent* (IBD) if they are derived from a common ancestor.

Then, by ignoring if the alleles of any pair of individuals are derived from a common ancestor, the match of a pair of alleles can be considered as identity by state. Table 3.1 shows all the possible combinations of the IBS alleles shared for a pair of individuals at a biallelic variant.

	AA	AB	BB
AA	2	1	0
AB	1	2	1
BB	0	1	2

Table 3.1: Number of IBS alleles for possible combinations of genotypes.

In terms of relatedness, the larger the number of IBS alleles shared between a pair of individuals, the more likely they are to be close related. For instance, a pair with the proportion of sharing 2 IBS alleles equals to 1 implies they are monozygotic twins or duplicated individuals in the dataset. Otherwise, a pair with the proportion of sharing 1 IBS allele larger or equal to 1 suggests they may have a parent-offspring relationship. A detailed approach to detect graphically

these types of relationships is shown in section 3.2.1.

Whereas identity by state alleles can be quantified directly from genotypes, identity by descent alleles are unobservable, since genetic data disregards which allele is inherited from the father and which from the mother. Furthermore, identity by state does not imply identity by descent. For example, Figure 3.8 shows a family tree in a particular genotype site, where □ is the nomenclature of a male and ◯ represents a female. Imagine that the genotypes from the parents are $\alpha\beta$ and AB, where $\alpha = A$ and $\beta = B$ are identical by state. If we observe the genotype of their children, a daughter has received the allele B from the mother and the allele α from the father, whereas the other daughter has received the allele A from the mother and the allele β from the father, and so, the full-siblings share 0 IBD alleles but share 2 IBS alleles. At this point, the challenge is to infer identity by descent from only the genotype status. For this purpose, section 3.2.2 shows the most common statistical methods to infer identity by descent alleles.

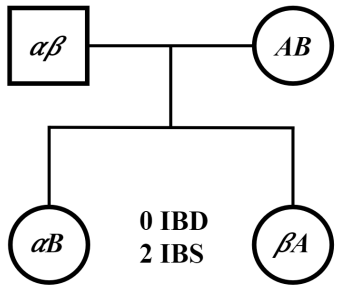

Figure 3.8: A family tree where the sharing IBD alleles is not equal to the sharing IBS alleles.

3.2.1 Identical by state analysis

Scatterplot of means and standard deviations of IBS alleles

Let n, m be the number of individuals and the number of genetic markers of the population under analysis respectively. Abecasis *et al.* (2001) propose to display the means and standard deviations of the IBS alleles of all the pairs of individuals in a scatterplot. Figure 3.9 (left) shows an example of this approach for 165 individuals from the CEU population of the 1000 Genomes Project (1000 Genomes Project Consortium *et al.* (2015), Section 3.2.3). The related pairs are colored according to the relationships reported elsewhere (Pemberton *et al.*, 2010). It can be shown that the related individuals have the larger means of the IBS alleles. In fact, full-siblings and parent-offspring relationships are the most outlying pairs, followed by second degree relationships that are located between unrelated pairs and full-siblings.

Scatterplot of the proportion of sharing of IBS alleles

Rosenberg (2006) represents IBS allele sharing data of all the pairs of individuals in a scatterplot of the proportion of sharing 0 IBS alleles (p_0) against the proportion of sharing 2 IBS alleles (p_2). Figure 3.9 (right) shows this representation for all the pairs of individuals from the CEU population. Note that parent-offspring pairs are located in the Y-axis of the plot with p_0 values close to zero and the full-sibling pair is the most outlying pair.

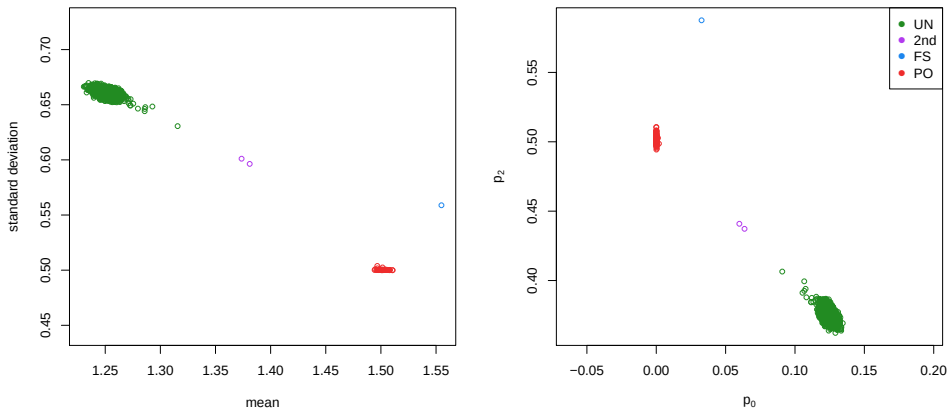

Figure 3.9: Left: Scatterplot of means and standard deviations of the IBS alleles for 13,530 pairs of individuals from the CEU population. Right: Scatterplot of the proportion of sharing 0 IBS alleles (p_0) against the proportion of sharing 2 IBS alleles (p_2) for 13,530 pairs of individuals from the CEU population. PO: parent-offspring, FS: full-siblings, 2nd: second degree relationships, UN: unrelated.

3.2.2 Identical by descent analysis

The graphical tools presented in Figure 3.9 are useful to identify the most outlying pairs of individuals as candidate familial relationships. However, to infer these relationships, the estimation of the probabilities of sharing 0, 1 and 2 IBD alleles (k_0, k_1, k_2, also referred to as Cotterman coefficients Cotterman (1941)) are of great help since each family relationship has different known theoretical values (Table 3.2). For example, parent-offspring pairs have $(k_0, k_1, k_2) = (0, 1, 0)$ and full-siblings have $(k_0, k_1, k_2) = (0.25, 0.50, 0.25)$. From these probabilities, the kinship coefficient ($\phi = k_1/4 + k_2/2$) is also widely used for relatedness research.

			Probability of IBD Sharing		
Type of Relative	R	ϕ	k_0	k_1	k_2
Monozygotic twins (MZ)	0	1/2	0	0	1
Parent-offspring (PO)	1	1/4	0	1	0
Full-siblings (FS)	1	1/4	1/4	1/2	1/4
Three-quarter siblings (3/4S)	-	3/16	3/8	1/2	1/8
Half-siblings/ grandchild-grandparent/ niece/nephew-uncle/aunt (2nd)	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 3.2: 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).

The theoretical IBD probabilities for each relationship can be deduced by using the inheritance patterns in a family tree. For instance, the IBD probabilities of the full-siblings relationship can be deduced as follows. Suppose that we label the parental genotypes as a/b and c/d. Then, there are 4 possible children (a/c, a/d, b/c and b/d). The possible combinations of the IBD status for them are shown in Table 3.3. Since each of these 16 possible genotype pairs is equally likely, it can be deduced that the probability of sharing 0, 1 and 2 IBD alleles for full-siblings is

$(k_0, k_1, k_2) = (4/16, 8/16, 4/16) = (0.25, 0.50, 0.25)$, as aforementioned.

	a/c	a/d	b/c	b/d
a/c	2	1	1	0
a/d	1	2	0	1
b/c	1	0	2	1
b/d	0	1	1	2

Table 3.3: The possible combinations of the IBD status of a pair of full-siblings given that the parental genotypes are a/b and c/d.

Maximum likelihood estimation of the identity by descent probabilities

A detailed derivation of the maximum likelihood estimation of the Cotterman coefficients (k_0, k_1, k_2) is given by Thompson (1975, 1991) and Milligan (2003). Briefly, consider bi-allelic genetic variants with alleles A and B having allele frequencies p and q respectively. Let G_1/G_2 be the genotypes for a pair of individuals, let k_t with $t = 0, 1, 2$ be their IBD probabilities (shown in Table 3.2) and let R be their family relationship. Then, the probability of observing G_1/G_2, given R is:

$$\begin{aligned} P(G_1/G_2|R) = & P(G_1/G_2|t=0)k_0 \\ & + P(G_1/G_2|t=1)k_1 \\ & + P(G_1/G_2|t=2)k_2. \end{aligned} \tag{3.1}$$

The terms $P(G_1/G_2|t=0), P(G_1/G_2|t=1)$ and $P(G_1/G_2|t=2)$ are the probabilities of observing each pair of genotypes given the number of IBD alleles (Table 3.4).

G_1/G_2	$t=0$	$t=1$	$t=2$
AA/AA	p^4	p^3	p^2
AA/AB	$2p^3q$	p^2q	0
AA/BB	p^2q^2	0	0
AB/AB	$4p^2q^2$	pq	$2pq$

Table 3.4: Possible pairs of biallelic genotypes and the probability of each pair given the number of alleles identical by descent (t). We assume that the order of the genotypes is irrelevant, i.e. the probabilities for G_1/G_2 and G_2/G_1 are the same.

Thus, regarding the Equation (3.1), the likelihood function of the Cotterman coefficients can be defined as follows:

$$L(k_0, k_1, k_2) = \prod_m P(G_1/G_2|R)$$

where m is the total number of genetic markers in the dataset. The maximum likelihood estimate is found by searching the maximum over the parameter space. This maximum is not trivial since the parameter space have the following constraints: $k_0 + k_1 + k_2 = 1$, $0 \leq k_i \leq 1$ and $k_1^2 \geq 4k_0k_2$. The last inequality follows from the assumption of absence of inbreeding (Thompson, 1975). The function *solnp* from the R-package Rsolnp (Ghalanos and Theussl, 2015) solves general nonlinear programming problems and allows for inequalities and nonlinear equalities, and can handle the

maximization problem. Otherwise, it is possible to simplify the problem by transforming the likelihood function in log-ratio coordinates, since the parameter space is located in the simplex (Graffelman and Galván-Femenía, 2016). In this way, the problem can be solved using R's general-purpose optimization routines such as *optim* and *nlminb*.

Other methods to estimate identity by descent probabilities

Purcell *et al.* (2007) use a method-of-moments approach to estimate the probabilities of sharing 0, 1 and 2 IBD alleles for all the pairs of individuals from the same homogeneous population. The algorithm is implemented in the PLINK software. Alternatively, the approach of the KING software (Manichaikul *et al.*, 2010) is focused on modeling genetic distances between pairs of individuals as a function of their allele frequencies and kinship coefficient. Another option is the PC-relate algorithm (Conomos *et al.*, 2016) based on residuals from linear regression models that include the top principal components as predictors. Both KING and PC-relate algorithms are robust estimators of the IBD probabilities in the presence of population structure (individuals from different ethnicities). On the other hand, the RELPAIR program infers relationships based on a Markov chain on underlying states of IBD status with the calculation of likelihood ratios for putative and alternative relationship (Boehnke and Cox, 1997; Epstein *et al.*, 2000).

Scatterplot of the estimated identity by descent probabilities

Once the identity by descent probabilities are estimated for each pair of individuals, the scatterplot of $\hat{k}_0$ and $\hat{k}_1$ reveals characteristic clusters for each family relationships. Figure 3.10 shows the scatterplot of $\hat{k}_0$ and $\hat{k}_1$ for all the pairs of individuals from the CEU population. The IBD probabilities are estimated with the PLINK software. Note that the related pairs of individuals are located close to their theoretical values.

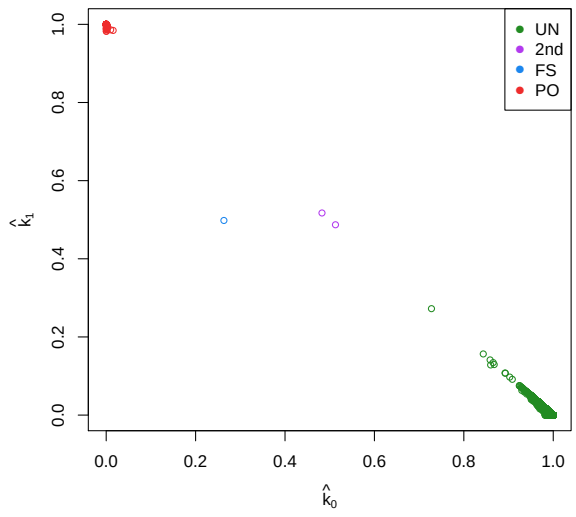

Figure 3.10: Scatterplot of $\hat{k}_0$ and $\hat{k}_1$ for 13,530 pairs of individuals from the CEU population. PO: parent-offspring, FS: full-siblings, 2nd: second degree relationships, UN: unrelated.

3.2.3 Genome wide databases

The following genetic databases are used in this doctoral thesis:

- HGDP-CEPH diversity panel (Rosenberg *et al.*, 2002). This dataset contains 377 microsatellites genotyped in 1056 individuals from 52 world-wide populations. The dataset is available from `https://rosenberglab.stanford.edu/diversity.html`.

- 1000 Genomes Project (1000 Genomes Project Consortium *et al.*, 2015). The Phase III of this dataset contains 81.7 million genetic variants (mainly SNPs) in 2,504 individuals from 26 world-wide populations. The dataset is available from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/`.

- GCAT Genomes For Life cohort project (Obón-Santacana *et al.*, 2018; Galván-Femenía *et al.*, 2018). This dataset contains almost 2 million SNPs in 5,000 individuals from European ancestry. The dataset is available from the European Genome-Phenome archive (EGA) `https://ega-archive.org/studies/EGAS00001003018`.

3.2.4 Simulations

Simulations of genetic markers and related individuals have formed an important part of this doctoral thesis. The simulations have been useful in order to evaluate the developed statistical methods for identifying family relationships.

To simulate related individuals from an empirical dataset, we identify a subset of approximately unrelated individuals with kinship coefficient below 0.05. From these individuals, we construct artificial pedigrees by sampling alleles across markers according to Mendelian laws. For example, parent-offspring pairs are simulated by first drawing two parents at random from the unrelated subset. Then, from each parent, we draw one allele at random from each genetic marker and join the alleles to generate a child. The process is repeated in order to generate many random parent-offspring pairs. To generate full-siblings or other pairs of relationships, the pedigree is simulated in an analogous manner.

On the other hand, to simulate biallelic genetic markers with allele frequencies p and q, we sample each marker from a multinomial distribution under the Hardy-Weinberg assumption ($p^2 + 2pq + q^2 = 1$). We consider a minor allele frequency (MAF) of 0.5 for all markers in order to obtain maximally polymorphic variants. In this way, the set of simulated variants are all independent and contains only unrelated individuals. From these variants, we follow the procedure of constructing artificial pedigrees as previously described to obtain individuals with known relationships.

3.3 Software

The statistical methods, data mining, data visualization and estimation of IBS/IBD alleles based on microsatellites were carried out with the R statistical software (R Core Team, 2019). The genetic data manipulation, filtering and estimation of IBS/IBD alleles based on SNPs were carried out with PLINK 1.90 (Chang *et al.*, 2015). The source R and PLINK codes used in this doctoral thesis are available from github: `https://github.com/ivangalvan/`.

Chapter 4

Research articles

4.1 Molecular Ecology Resources

This first article accomplishes with the objectives Obj. 1 and Obj. 2 described in 2.1. In summary, the classical graphical methods for relatedness research based on identity by state/descent are reviewed. Furthermore, ternary diagrams and scatterplots of isometric log-ratio transformations are proposed to identify family relationships.

This article has been published in *Molecular Ecology Resources* journal.
Volume: 17, Issue: 6, Pages: 1271-1282, Submitted: May 2016, Accepted: March 2017.
DOI: 10.1111/1755-0998.12674.
Impact factor: 7.332 (Q1). Journal Citation Reports Ranking: 30/298 (Biochemistry & Molecular Ecology); 10/164 (Ecology); 7/50 (Evolutionary Biology).

Received: 27 May 2016 | Revised: 15 March 2017 | Accepted: 21 March 2017

DOI: 10.1111/1755-0998.12674

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Graphics for relatedness research

Iván Galván-Femenía[1,2] | Jan Graffelman[3,4] | Carles Barceló-i-Vidal[1]

[1]Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona, Girona, Spain

[2]Disease Genomics-GCAT Group, Germans Trias Health Research Institute (IGTP)-Program of Predictive and Personalized Medicine of Cancer (PMPPC), Can Ruti Campus, Badalona, Barcelona, Spain

[3]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain

[4]Department of Biostatistics, University of Washington, Seattle, WA, USA

Correspondence
Iván Galván-Femenía, Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona, Girona, Spain.
Email: ivan.galvan@udg.edu

Funding information
United States National Institutes of Health, Grant/Award Number: R01 GM075091; Spanish Ministry of Economy and Competitiveness, Grant/Award Number: MTM2015-65016-C2-2-R, MTM2015-65016-C2-1-R

Abstract

Studies of relatedness have been crucial in molecular ecology over the last decades. Good evidence of this is the fact that studies of population structure, evolution of social behaviours, genetic diversity and quantitative genetics all involve relatedness research. The main aim of this article was to review the most common graphical methods used in allele sharing studies for detecting and identifying family relationships. Both IBS- and IBD-based allele sharing studies are considered. Furthermore, we propose two additional graphical methods from the field of compositional data analysis: the ternary diagram and scatterplots of isometric log-ratios of IBS and IBD probabilities. We illustrate all graphical tools with genetic data from the HGDP-CEPH diversity panel, using mainly 377 microsatellites genotyped for 25 individuals from the Maya population of this panel. We enhance all graphics with convex hulls obtained by simulation and use these to confirm the documented relationships. The proposed compositional graphics are shown to be useful in relatedness research, as they also single out the most prominent related pairs. The ternary diagram is advocated for its ability to display all three allele sharing probabilities simultaneously. The log-ratio plots are advocated as an attempt to overcome the problems with the Euclidean distance interpretation in the classical graphics.

KEYWORDS
compositional data analysis, identical by state/descent, isometric log-ratio, microsatellite, relatedness, ternary diagram

1 | INTRODUCTION

Statistical methods for the analysis of the genetic relationships between individuals of a population are of great relevance for molecular ecology (Blouin, 2003). Studies of relatedness are crucial for studying population structure, evolution of social behaviour, genetic diversity, quantitative genetics, etc. It is known that the estimation of quantitative genetic parameters in wild populations is less biased and more precise if we dispose of pedigree information (Bérénos, Ellis, Pilkington, & Pemberton, 2014). The role of relatedness for selective breeding is also recognized. Loughnan, Smith-Keune, Jerry, Beheregaray, and Robinson (2016) recommend low levels of relatedness and high levels of neutral genetic diversity to form a base population for selective breeding. The exclusion of duplicated individuals and close relatives is a previous quality control filter used in studies of population structure (Gonder et al., 2015). Relatedness estimation is also important for conservation programmes, and the performance of several estimators has been compared in that context (Oliehoek, Windig, van Arendonk, & Bijma, 2006). It plays an important role in structuring societies with fusion–fission dynamics (Croft et al., 2012; Snyder-Mackler, Alberts, & Bergman, 2014; Spencer et al., 2015), can bias estimates of allele frequencies (Hansen, Nielsen, & Mensberg, 1997) and violates the assumption of independent individuals in trait-gene association studies (Foulkes, 2009). Thus, statistical methods that can verify documented or uncover undocumented family relationships in the database are important tools in molecular ecology.

Relatedness investigations can be carried out in an entirely numerical manner by inspecting estimated IBS (identity by state) and IBD (identity by descent) probabilities, likelihood ratios or confusion matrices (Boehnke & Cox, 1997; Epstein, Duren, & Boehnke, 2000). Graphics greatly facilitate the interpretation of the results of relatedness studies and are increasingly being used (Abecasis, Chemy, Cookson, & Cardon, 2001; Pemberton, Wang, Li, & Rosenberg, 2010; Rosenberg, 2006). The main aim of this article was to summarize the state of the art of the graphical methods used in relatedness research. Relatedness investigations are based on allele sharing, and we will consider techniques that use IBS alleles as well as those using IBD alleles. A plot of the means against the standard deviations of the IBS counts is a powerful tool to detect relatedness (Abecasis et al., 2001). We explore this tool in detail and establish the domain of this graphic from a mathematical point of view. Plots of the proportions of markers with 0, 1 or 2 IBS counts (p_0, p_1 or p_2) are often used to assess the existence of family relationships (Rosenberg, 2006). Nevertheless, if the researcher is interested in identifying the degree of relatedness, plotting the probabilities of sharing 0, 1 or 2 IBD alleles (k_0, k_1 or k_2) is the best strategy. The IBD probabilities depend directly on relatedness and enable us to accurately infer the type of relationship. In addition to the former graphical methods, we propose to use graphics from compositional data analysis (CoDA) for both IBS and IBD allele sharing studies. Due to the fact that the proportions (p_0, p_1, p_2) and the probabilities (k_0, k_1, k_2) are constrained to sum to one, it is possible to apply all the graphical and analytical CoDA techniques introduced by Aitchison (1986) and developed posteriorly by Pawlowsky-Glahn and Buccianti (2011). Two graphics, commonly used in CoDA, are of particular relevance for relatedness studies: the ternary diagram (also known as a *de Finetti* diagram in genetics) and a scatterplot of log-ratios. We show the ternary diagram to be useful for plotting the proportions of the IBS counts and for plotting the estimated Cotterman coefficients (IBD probabilities). Moreover, the theoretical IBD sharing probabilities for the standard family relationships can be used as reference points in the ternary diagram (Thompson, 2000). Furthermore, the CoDA techniques allow us to introduce the isometric log-ratio coordinates (ilr-coordinates) of the vectors $\boldsymbol{p} = (p_0, p_1, p_2)$ and $\boldsymbol{k} = (k_0, k_1, k_2)$, which we can represent in a scatterplot. These ilr-coordinates allow us to measure the degree of similarity between two vectors of IBS proportions or IBD probabilities. The graphics we propose are of universal value and can be used in any relatedness study that concerns diploid individuals.

The remainder of this article is organized as follows. Section 2 gives an overview of the IBS allele sharing analysis and the graphical methods used to detect family relationships. Section 3 presents the basic principles of IBD estimation and the most common graphics used for relatedness estimation in the IBD context. The former sections also detail the graphical methods from the field of CoDA used in IBS-IBD approaches: the ternary diagram and the scatterplot of log-ratios. Section 4 presents a way to enhance IBS and IBD graphics with convex hulls that express the degree of uncertainty about a relationship. Section 5 presents a case study with individuals from the Maya population. Finally, Section 6 summarizes the principal conclusions of this article and the pros and cons of each graphical method are discussed.

2 | IBS STUDIES

IBS studies disregard if the alleles for any diploid individual are derived from a common ancestor. IBS allele sharing concerns the number of matches between the alleles of the genotypes of two individuals. Two diploid individuals can share 0 (e.g., A1/A1 and A2/A2 or A1/A2 and A3/A3), 1 (e.g., A1/A1 and A1/A2 or A1/A2 and A1/A3) or 2 (e.g., A1/A1 and A1/A1) IBS alleles for a specific genetic marker, and we will refer to these as IBS counts. To detect family relationships in a given population of n individuals and m genetic markers, the number of matches between IBS alleles (the IBS counts) is considered for each pair of individuals across genetic markers. That is, we move from a data set of n individuals and m genetic markers to a data set of $\binom{n}{2}$ pairs of individuals with the information of the IBS counts for m genetic markers. There are different ways to deal with this type of data as described below. First, we focus on the plot of means and standard deviations of the IBS counts (Abecasis et al., 2001). Second, we detail the plot of the proportions of the IBS counts (Rosenberg, 2006). To conclude this section, graphics from CoDA (Aitchison, 1986; Pawlowsky-Glahn & Buccianti, 2011) are presented.

To illustrate the different IBS graphics that are introduced in this Section, we use five pairs of individuals with the information of IBS counts and IBS proportions for 377 microsatellites (see Table 1). The individuals are from the Maya population which we will analyse in Section 5. We consider a parent-offspring (PO) pair, a full-sib (FS) pair, a half-sib (HS), avuncular (AV) or grandparent-grandchild (GG) pair, a pair of first cousins (FC) and a pair of unrelated individuals (UN). We discuss the different graphics in the sections below.

2.1 | $(\bar{\boldsymbol{x}}, \boldsymbol{s})$-plot

Let x_{ijk} be the number (0, 1 or 2) of shared IBS alleles between individual i and j for the genetic marker k. Abecasis et al. (2001) proposed to compute the mean ($\bar{x}_{ij}$) and variance (s_{ij}^2) over K genetic markers. The plot $\bar{x}_{ij}$ versus s_{ij} reveals characteristic clusters that correspond to the different family relationships for a given population.

The statistics $\bar{x}_{ij}$ and s_{ij}^2 are constrained due to the limited number of outcomes (0, 1 or 2), and we proceed to derive their range of variation (Figure 1a). As an example, we consider a table with all possible outcomes of the allele sharing counts (0, 1 or 2) for a set of 100 markers. The rows of this table represent possible pairs of individuals. There are 3^{100} combinations (rows), if the order of the outcomes is considered relevant. However, in terms of means or standard deviations, the order of the IBS counts (0, 1 or 2) over the different markers is irrelevant but their multiplicity is important. For example, a pair of individuals sharing 1 IBS allele for the first marker and 0 for all other markers will have the same mean and variance as

TABLE 1 Computations for five pairs of individuals from the Maya population. Mean and standard deviation of IBS counts, proportion of sharing 0, 1 and 2 IBS alleles (p_0, p_1, p_2) and estimated Cotterman coefficients ($\hat{k}_0, \hat{k}_1, \hat{k}_2$) are shown

	IBS studies					IBD studies		
Type of relative	Mean	Standard deviation	p_0	p_1	p_2	$\hat{k}_0$	$\hat{k}_1$	$\hat{k}_2$
PO	1.34	0.48	0.002	0.650	0.348	0.009	0.991	0.000
FS	1.32	0.60	0.073	0.532	0.395	0.214	0.617	0.169
HS, AV or GG	1.09	0.64	0.160	0.581	0.259	0.447	0.553	0.000
FC	1.00	0.67	0.225	0.546	0.229	0.657	0.343	0.000
UN	0.86	0.67	0.308	0.526	0.166	0.731	0.269	0.000

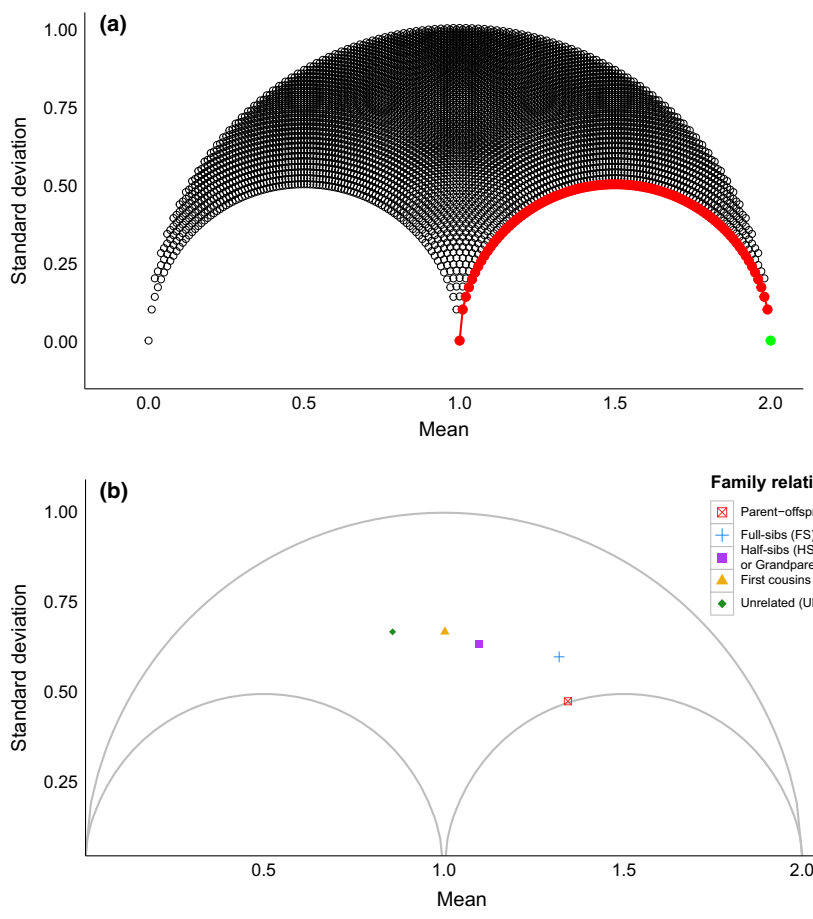

FIGURE 1 a. Plot of means and standard deviations of all possible combinations of IBS counts for a table of 100 genetic markers. The red curve shows the pairs of individuals that are parent-offspring. The green point represents a monozygotic twin pair or a pair of duplicated individuals. b. Plot of means versus standard deviations of the IBS counts for five pairs from the Maya population

a pair of individuals sharing 1 IBS allele for the k-th marker and 0 for all others. Mathematically, the combinations of the IBS counts for a pair of individuals form a multiset (Stanley, 1997, Section 1.2) of cardinality m (the number of markers) made of a basic set of cardinality $k = 3$ (the outcomes 0, 1 and 2). The possible number of $(\bar{x}, s)$ pairs in the plot can be no larger than the number of multisets of cardinality k, where the latter is given by the multiset coefficient

$$\left(\!\binom{k}{m}\!\right) = \binom{k+m-1}{k}, \tag{1}$$

Thus, for 100 genetic markers there will be at most $\left(\!\binom{3}{100}\!\right) = \binom{3+100-1}{100} = \binom{102}{100} = 5151$ different $(\bar{x}, s)$ pairs. Figure 1a shows the means and standard deviations of the 5151 combinations of IBS counts for 100 genetic markers. The figure has the shape of an umbrella and represent the domain of the $(\bar{x}, s)$-plot. For empirical data, it will be impossible to observe a $(\bar{x}, s)$ point outside the umbrella region. It is clear that the mean of the IBS counts ranges from zero to two. The maximum variance equals one and is reached when the array of IBS counts has fifty 0 IBS alleles and fifty 2 IBS alleles, whereas the minimum variance equals zero and is reached when the array of IBS counts has either one hundred 0 IBS alleles, one hundred 1 IBS allele or one hundred 2 IBS alleles.

The red points on the right hand curve of the "umbrella" correspond presumably to parent–offspring relationships for having a mean larger than 1 and low variance. The first point of the curve

with mean equal to 1 IBS allele and standard deviation equal to 0 IBS alleles corresponds to an array of one hundred ones. The second point of the curve corresponds to an array of 99 markers with 1 IBS alleles and one marker with 2 IBS alleles, and so on. In other words, this red curve represents the pairs of individuals who have a mean larger than or equal to 1 and the smallest standard deviation of all possible IBS counts. This can be related with the fact that the probability of sharing 1 IBD allele between a parent-offspring equals 1, as we will see in the next Section (Table 2). For parent-offspring pairs, we have that $\bar{x}_{ij} \geq 1$ because children inherit at least 1 IBS allele from their parents. And for monozygotic twins (MZ) or duplicated individuals, we have $\bar{x}_{ij} = 2$ and $s_{ij} = 0$ (green point in Figure 1a).

Figure 1b shows the $(\bar{x}, s)$ plot for the five Maya pairs in Table 1. The larger the mean of the IBS counts for any pair of individuals, the more likely they are to be closely related. The PO pair (red point) is located on the right hand curve of the umbrella, the FS pair (blue point) with mean larger than 1 is separated from second- and third-degree family relationships (violet and gold points respectively), whereas, the unrelated individuals have the smallest mean (green point).

2.2 | (p_i, p_j)-plots

Let x_{ij} be the vector of the IBS counts between individual i and j as large as the number of the genetic markers in the data set. Let p_0, p_1 and p_2 be the proportions of 0, 1 and 2 IBS alleles, respectively, for each pair of individuals. Rosenberg (2006) proposed a graphical method for relatedness research by plotting the proportion of sharing 2 IBS alleles (p_2) versus the proportion of sharing 0 IBS alleles (p_0) for all pairs of individuals from a given population. Similarly, Sun (2012) uses IBS proportions for relatedness research by plotting p_1 versus p_0. In fact, any combination of the three proportions could be plotted for relatedness research. We refer to these graphics as (p_i, p_j)-plots (for i, j = 0, 1, 2 and $i < j$) were p_i corresponds to the X-axis of the plot and p_j to the Y-axis.

Monozygotic twins (MZ) or duplicated individuals are easy to identify in the (p_i, p_j)-plots because they have p_2 close to 1. PO pairs have low values of p_0 and are also easy to detect visually because they are on the p_1 or p_2-axis. FS usually have large values of p_2 and are separated from unrelated individuals. Second degree and third degree are more difficult to detect because positions in the plot depend on the allele frequencies of the population under study. Figures 2a, b and c show the (p_0, p_2)-, (p_0, p_1)- and (p_1, p_2)-plots for the five Maya pairs (Table 1). Notice that the distance between pairs of individuals is not the same in the three plots. For instance, the FS pair (blue point) is most close to the PO pair (red point) in the (p_0, p_2)-plot, but closer to the HS pair (violet point) in the (p_0, p_1)-plot. If the distances between pairs of individuals are different depending on the plotted proportions, then it is not appropriate to draw conclusions about the family relationship between individuals from the (p_i, p_j)-plots.

TABLE 2 Cotterman coefficients for the different type of family relationship and degree of relatedness

Type of relative	Degree	k_0	k_1	k_2
Monozygotic twins (MZ)	0	0	0	1
Parent-offspring (PO)	1	0	1	0
Full-siblings (FS)	1	1/4	1/2	1/4
Half-siblings (HS)/avuncular (AV)/grandchild-grandparent (GG)	2	1/2	1/2	0
First cousins (FC)	3	3/4	1/4	0
Unrelated (UN)	∞	1	0	0

2.3 | Ternary diagrams

Let $\boldsymbol{p}$ be the vector (p_0, p_1, p_2) of proportions of the IBS counts. Because the three components of $\boldsymbol{p}$ sum to one ($p_0 + p_1 + p_2 = 1$), we can plot the vector $\boldsymbol{p}$ in a ternary diagram. Mathematically, the set of the vectors of proportions $\boldsymbol{p} = (p_0, p_1, p_2)$ forms the simplex, S^3. Figure 3 shows the ternary diagram for the vectors of proportions for the five Maya pairs (Table 1). The PO pair (red point) is located on the opposite side of the vertex p_0; the FS pair (blue point) has the largest value for p_2 and is the closest to the p_2 vertex. The UN pair (green point), FC pair (gold point) and the HS, AV or GG pair

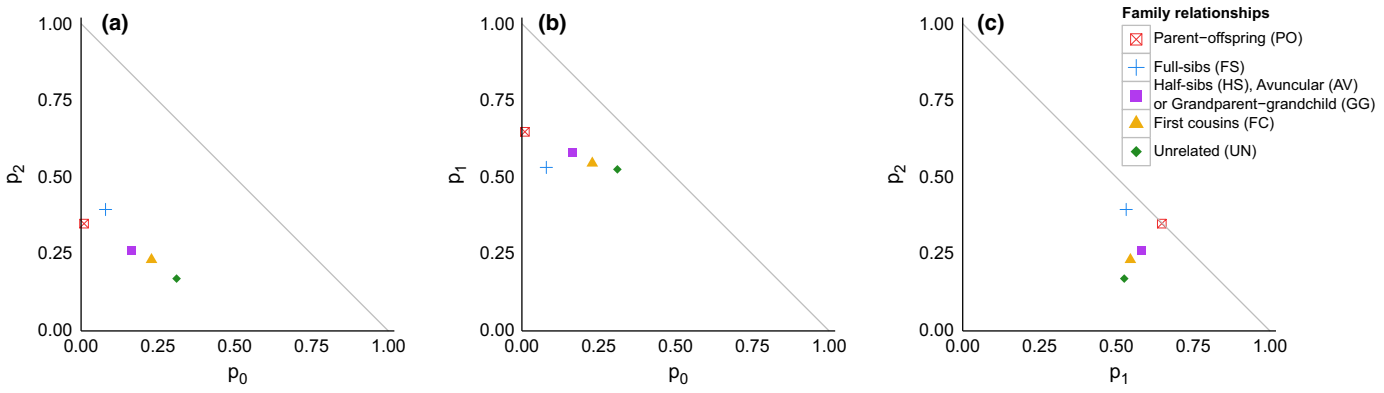

FIGURE 2 (p_i, p_j)-plots for five individuals from the Maya population. a. Plot of the proportion of sharing 0 IBS alleles (p_0) versus the proportion of sharing 2 IBS alleles (p_2): (p_0, p_2)-plot. b. Plot of the proportion of sharing 0 IBS alleles (p_0) versus the proportion of sharing 1 IBS allele (p_1): (p_0, p_1)-plot. c. Plot of the proportion of sharing 1 IBS allele (p_1) versus the proportion of sharing 2 IBS alleles (p_2): (p_1, p_2)-plot. [Colour figure can be viewed at wileyonlinelibrary.com]

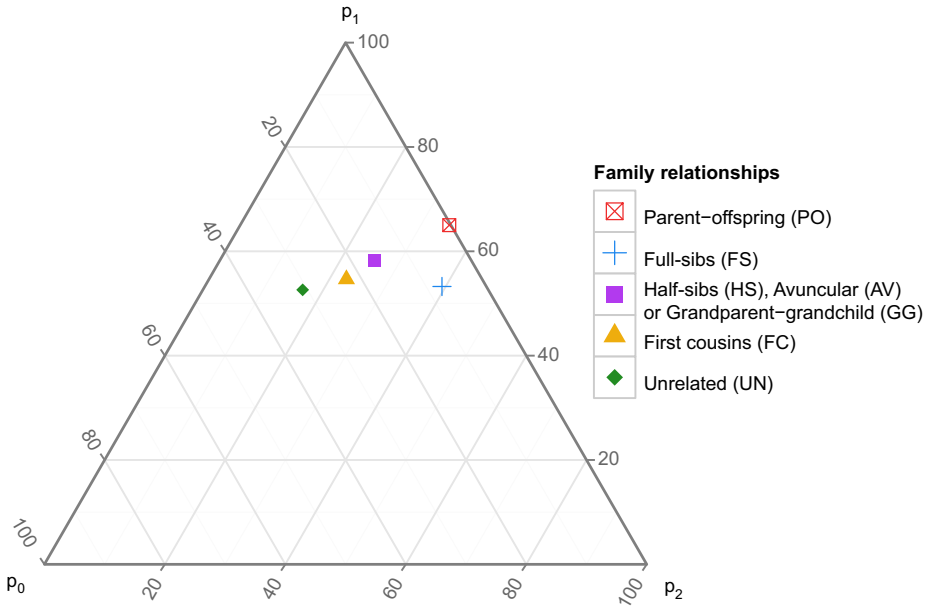

FIGURE 3 Ternary diagram of the IBS proportions for five pairs from the Maya population. [Colour figure can be viewed at wileyonlinelibrary.com]

(violet point) have lower values of p_2. The UN pair has the lowest values for p_2 and p_1 and is closest to the p_0 vertex. The main advantage of this graphical tool is that it represents the three proportions p_0, p_1 and p_2 simultaneously in contrast to the (p_i, p_j)-plots that represent only two of them.

2.4 | ilr-plots

Aitchison (1986) stated that it is not meaningful to interpret the distances between two vectors of proportions in the ternary diagram as if we were in an Euclidean space. Aitchison (1986) defines a new distance based on the log-ratio of the components of the vectors of proportions. This distance, jointly with the perturbation and powering operators (analogous to translation and scalar multiplication in the real space, respectively), forms the structure of the simplex in a two-dimensional metric space (Aitchison, Barceló-Vidal, Martín-Fernández, & Pawlowsky-Glahn, 2000; Pawlowsky-Glahn & Buccianti, 2011). Thereby, the vectors of proportions $\boldsymbol{p} = (p_0, p_1, p_2)$ can be expressed in coordinates using any orthonormal basis defined in the simplex (Egozcue, Pawlowsky-Glahn, Mateu-Figueras, & Barceló-Vidal, 2003). These coordinates are called isometric log-ratio coordinates (ilr-coordinates). The distance between two vectors of proportions is calculated as the Euclidean distance between their ilr-coordinates. The ilr-coordinates of a vector of proportions depend on the orthonormal basis used in the simplex. The most commonly used ilr-coordinates $\boldsymbol{z_0}$, $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ of a vector of proportions (p_0, p_1, p_2) are given by

$$\boldsymbol{z_0} = \begin{cases} z_{01} = \frac{1}{\sqrt{2}} \ln\left(\frac{p_2}{p_1}\right) \\ z_{02} = \frac{1}{\sqrt{6}} \ln\left(\frac{p_1 p_2}{p_0^2}\right) \end{cases} \quad \boldsymbol{z_1} = \begin{cases} z_{11} = \frac{1}{\sqrt{2}} \ln\left(\frac{p_2}{p_0}\right) \\ z_{12} = \frac{1}{\sqrt{6}} \ln\left(\frac{p_0 p_2}{p_1^2}\right) \end{cases} \quad \boldsymbol{z_2} = \begin{cases} z_{21} = \frac{1}{\sqrt{2}} \ln\left(\frac{p_1}{p_0}\right) \\ z_{22} = \frac{1}{\sqrt{6}} \ln\left(\frac{p_0 p_1}{p_2^2}\right) \end{cases}, \quad (2)$$

Figures 4a, b and c plot the ilr-coordinates for the five Maya pairs (Table 1). Notice that the distance between any pair of points is exactly the same in the three graphics, irrespective of the ilr-coordinates ($\boldsymbol{z_0}$, $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$) that are plotted. The PO pair (red point) in Figures 4a–c is an outlying pair. The FS pair (blue point) is also isolated from pairs of second and third degree of relationships. The degree of relationship decreases with the z_{02}, z_{11} and z_{21} ilr-coordinates (close relatives with a first-degree relationship (PO, FS) have larger values for these coordinates than second-degree relationships (HS, AV, GG)).

3 | IBD STUDIES

Studies of relatedness based on IBD alleles are based on the probabilities that a pair of individuals shares 0, 1 or 2 IBD alleles. These probabilities are commonly referred to as Cotterman's coefficients (Cotterman, 1941) and denoted by the vector of proportions $\boldsymbol{k} = (k_0, k_1, k_2)$. Table 2 shows the values of the Cotterman coefficients for some standard relationships. Cotterman's coefficients can be estimated by the maximum-likelihood method (Milligan, 2003; Weir, Anderson, & Hepler, 2006). The maximum-likelihood estimates reveal the most likely relationship for a pair given the observed genotype data. Let R represents a possible relationship between two individuals with genotypes G_1 and G_2, respectively. The likelihood of R is defined by the probability of observing G_1 and G_2 given relationship R. This probability depends on the allele frequencies of the population under study and is conditioned by the Cotterman coefficients. This likelihood is calculated across loci to obtain the most likely values (estimates) of the Cotterman coefficients. These estimates provide a first indication of the possible relationship between a pair of individuals. A hypothesis test is recommended to confirm or refute this relationship (García-Magariños, Egeland, López-de-Ullibarri, Hjort, & Salas, 2015). More details are explained by Wagner, Creel, and Kalinowski (2006). Under the assumption of absence of inbreeding, the inequality $k_1^2 \geq 4k_0k_2$ applies and constrains the Cotterman coefficients (Thompson, 1991).

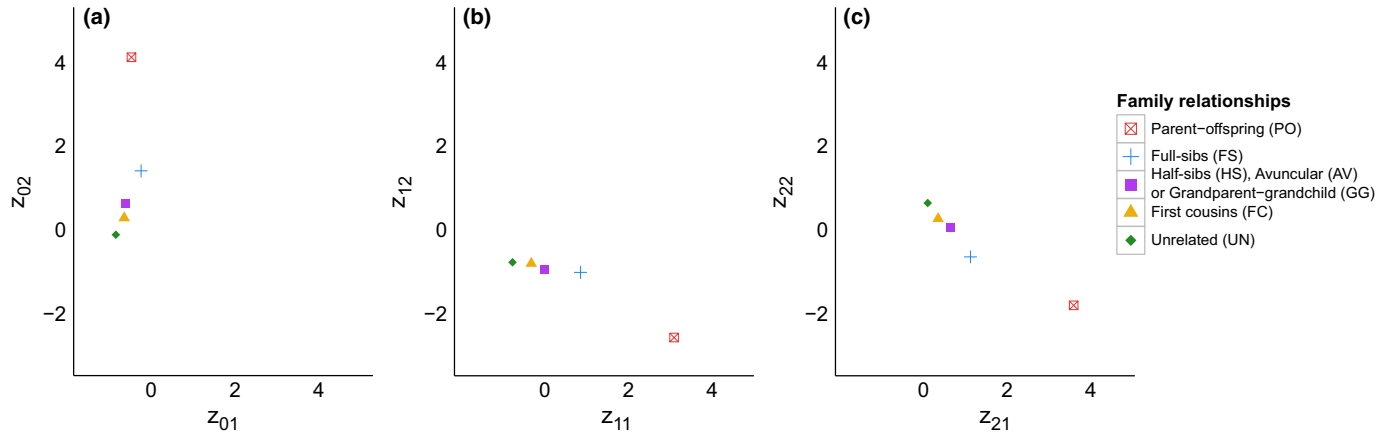

FIGURE 4 llr-coordinates of the IBS proportions for five pairs of individuals from the Maya population. a. $\boldsymbol{z}_0 = (z_{01}, z_{02})$. b. $\boldsymbol{z}_1 = (z_{11}, z_{12})$. c. $\boldsymbol{z}_2 = (z_{21}, z_{22})$. [Colour figure can be viewed at wileyonlinelibrary.com]

Analogously to the vector of proportions $\boldsymbol{p} = (p_0, p_1, p_2)$ of the IBS counts, Cotterman's coefficients also satisfy $k_0 + k_1 + k_2 = 1$. We can use the same graphical techniques described for $\boldsymbol{p} = (p_0, p_1, p_2)$ to identify relatedness from the estimated Cotterman coefficients $\hat{\boldsymbol{k}}$. The Cotterman coefficients can be represented in a $(\hat{k}_i, \hat{k}_j)$-plot, in a ternary diagram or in an ilr-plot with the ilr-coordinates $\boldsymbol{z}_0$, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, defined in the Equation (2), substituting p_i for $\hat{k}_i$. With the aim of describing each graphical method used in IBD studies, we compute maximum-likelihood estimates of the Cotterman coefficients for the five Maya pairs (Table 1).

3.1 | $(\hat{k}_i, \hat{k}_j)$-plots

In the literature, the estimated Cotterman coefficients are plotted in different ways to identify relatedness. Nembot-Simo, Graham, and McNeney (2013) use the $(\hat{k}_0, \hat{k}_1)$-plot. Similarly, Moltke and Albrechtsen (2014) use the $(\hat{k}_1, \hat{k}_2)$-plot. The remaining possibility, the $(\hat{k}_0, \hat{k}_2)$-plot, could be also considered. Figure 5a shows the plot for the five Maya pairs (Table 1). The grey curve in the $(\hat{k}_0, \hat{k}_1)$-plot corresponds to the equation $k_1^2 = 4k_0k_2$. This curve jointly with the hypotenuse and the vertical axis delimits the feasible region $k_1^2 \geq 4k_0k_2$. PO pairs are points located on the k_1-axis with values close to 1, FS pairs are located close to the centre of the grey curve according to the theoretical IBD probabilities (Table 2) and second and third degree pairs are located around the centre of the hypotenuse. UN pairs theoretically have $k_0 = 1$ and are located between the hypotenuse and the grey curve, near to the vertex $\hat{k}_0 = 1$. Finally, the origin of the $(\hat{k}_0, \hat{k}_1)$-plot is the position for any MZ pair. As previously shown for IBS studies with the (p_i, p_j)-plots, only two of the three Cotterman coefficients are plotted and the relative positions and distances between points vary depending on the $(\hat{k}_i, \hat{k}_j)$-plot used. For this reason, we propose graphics from CoDA.

3.2 | Ternary diagrams

The theoretical IBD probabilities for the standard family relationships can be represented in a ternary diagram (Thompson, 2000). These probabilities form reference points against which the empirical estimates can be compared. Figure 5b shows the ternary diagram for the estimated Cotterman coefficients for the five Maya pairs

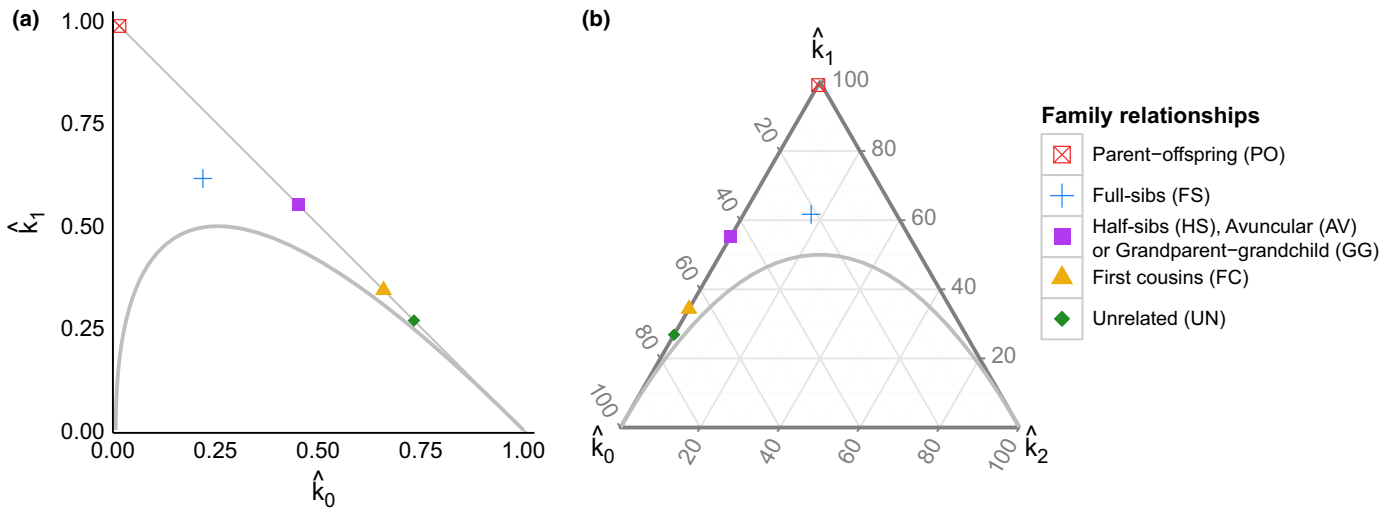

FIGURE 5 $(\hat{k}_0, \hat{k}_1)$-plot (a) and ternary diagram (b) for five pairs of individuals from the Maya population. [Colour figure can be viewed at wileyonlinelibrary.com]

(Table 1). Most pairs in Table 1 are close to their theoretical IBD probabilities given in Table 2. However, values of k_1 are larger than expected for the FS, HS, AV and notably, the UN pair (see the Discussion section). The domain has the shape of an arrowhead inside the ternary diagram. The curve delimiting the arrowhead from below corresponds to the inequality $k_1^2 \geq 4k_0k_2$.

3.3 | ilr-plots

It has been shown that the maximum-likelihood estimates of the Cotterman coefficients in the simplex are the same as the estimates obtained by maximizing the likelihood in ilr-coordinates (Graffelman & Galván-Femenía, 2016). With the aim of establishing reference zones for the standard family relationships in the ilr space, we compute the maximum-likelihood estimates of the Cotterman coefficients from the ilr-coordinates defined by the Equation (2) and we plotted the $\boldsymbol{z}_1 = (z_{11}, z_{12})$ ilr-coordinates as is shown in Figure 6. All the family relationships have values lower than $-\sqrt{(2/3)}\ln(2)$ for z_{12} which corresponds to the grey line in the graph. This line corresponds to the curve shown in the former graphs (Figure 5a and b). Due to the fact that some Cotterman coefficients equals 0, some of the (or both) ilr-coordinates tend to $+/-$ infinity. Thus, given that it is impossible to represent the point, we are limited to indicate the direction of the infinity in the ilr-plot for each type of family relationship. Regarding Figure 6, PO pairs have a large variability of values, either positive or negative for z_{11}; FS have values close to 0 for z_{11} and $-\sqrt{(2/3)}\ln(2)$ for z_{12}. HS, AV, GG and FC are located between PO, FS and UN. UN pairs have negative values of z_{11} which correspond to the green point of the left hand. If present, MZ pairs are points with positive values of z_{11} located on the right hand side of the plot.

4 | UNCERTAINTY IN IBS/IBD GRAPHICS

With the previously described graphics, one can try to infer the relationship of a pair for which the relationship is not documented, or try to confirm the documented relationships. Such graphical inference is hampered by the fact that the statistics represented in the graphs (means and standard deviations of the IBS counts, p_0, p_1, p_2, k_0, k_1, k_2) are subject to uncertainty. For a given sample, relationships are not represented by points, but by zones. Some insight into this uncertainty and the corresponding zones can be obtained by simulation. Ideally, this would require a large sample for which a subset of unrelated individuals can be identified. From these individuals, by sampling alleles across markers according to Mendelian laws, the reproductive process can be simulated allowing us to generate artificial children, leading to artificial PO pairs, FS pairs and artificial pairs of any other desired relationship. For example to simulate a PO pair we sample two UN individuals at random without replacement from the database. From each UN individual, we sample one allele at random from each marker and join the alleles to form a child. The process of sampling UN pairs and child generation is repeated many times, generating many artificial PO pairs. We can calculate the IBS/IBD statistics of the artificial pairs, and add these to the graphics of the previous sections by representing them individually or with a convex hull. A convex hull for a given set of points X is the unique convex polygon whose vertices are points from X and that contains all points of X (de Berg, van Kreveld, Overmars, & Schwarzkopf, 2000). By generating a large number of artificial pairs and representing these in the IBS/IBD graphics of interest, the zones corresponding to the different relationships can be approximated. Such simulations are conditional on the observed allele frequencies and can quantify the uncertainty in a graphical assessment of the relationship to some extent. We illustrate this with examples in the next section where all graphics are enhanced with hulls based on 80 PO, 48 FS, 120 second degree, 36 FC and 1256 UN artificially generated pairs.

5 | CASE STUDY

We applied all the graphical methods detailed in the previous sections using empirical data extracted from a world-wide data set from the Noah A. Rosenberg Research lab at Stanford University (Rosenberg et al., 2002). This world-wide database is derived from the Human Genome Diversity Cell Line Panel (HGDP, Cavalli-Sforza, 2005). The genetic information is given by 377 microsatellites genotyped for 52 human populations around the world. We used all 25 available individuals of the Maya sample to illustrate all graphical methods for relatedness research. All the family relationships present in this sample were reported by Rosenberg (2006). All the Figures

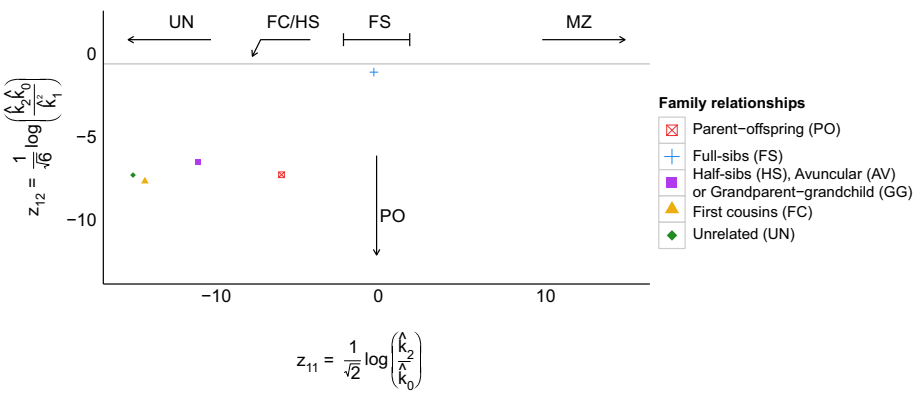

FIGURE 6 Ilr-coordinates $\boldsymbol{z}_1 = (z_{11}, z_{12})$ of the estimated Cotterman coefficients $(\hat{k}_0,\hat{k}_1,\hat{k}_2)$ for five pairs of individuals from the Maya population. [Colour figure can be viewed at wileyonlinelibrary.com]

presented throughout this article are made with the R software (R Core Team, 2015) using the R packages **ggplot2** (Wickham, 2009) and **ggtern** (Hamilton, 2015).

5.1 | IBS graphics

Figure 7 shows all IBS graphics for all pairs of the Maya population. In the $(\bar{x}, s)$-plot (Figure 7a), the points with the smallest standard deviation close to the grey curve are two PO pairs. The relationships of first and second degree are the points with a mean above 1. Note that some pairs of FC are mixed with UN pairs. Figure 7b (the (p_0, p_2)-plot) clearly separates the family relationships of first and second degree from the UN pairs. In the ternary diagram (Figure 7c), PO pairs are points on the opposite side of the vertex p_0, meaning that the p_0 is close to 0. The FS pair is the point closest to the vertex p_2, which has the largest p_2; the violet points represent the family relationships of second degree are separated from the green points representing UN pairs. In Figure 7d, the first ilr-coordinate (z_{11}) clearly discriminates first-degree relatives from UN pairs. Pairs with larger values for z_{11} are more likely to correspond to related individuals. PO pairs are extreme outliers because they have p_0 values close to 0 which increase the first coordinate of the corresponding log-ratio. The scatterplot of the log-ratios is seen to produce a larger degree of separation between FS and PO pairs, and between first-degree relationship pairs and all other pairs. The convex hulls for the simulated related pairs in Figure 7 are seen to enclose the sample estimates of the PO, FS, HS and FC pairs and so confirm the assigned relationships.

5.2 | IBD graphics

We estimated IBD probabilities for all pairs of the Maya population. All IBD graphics are shown in Figure 8. The $(\hat{k}_0, \hat{k}_1)$-plot (Figure 8a) separates the first, second and some pairs of third degree of relatedness. In the ternary diagram of $\hat{\boldsymbol{k}}$ (Figure 8b), it is easy to identify PO pairs at the vertex of $\hat{k}_1$, a FS pair close to the barycenter of the triangle and other family relationships of second degree on the opposite side of the $\hat{k}_2$ vertex. UN pairs are on the $k_0 - k_1$ edge and tend towards the k_0 vertex. Third-degree pairs are mixed with unrelated individuals. In the ilr-plot (Figure 8c), the pairs with a close family relationship tend to have larger values of z_{11}. The family relationships of the first degree (FS and PO) are located according to the directions indicated in Figure 6. The ilr-plot clearly separates out these FS and PO relationships from all other pairs. Notice that Figure 8a and b show only one pair with a second degree relationship (the violet point), whereas in Figure 8c, there are two visible violet pairs. The IBD graphics were also amplified with convex hulls of artificially generated related pairs to show the approximate expected

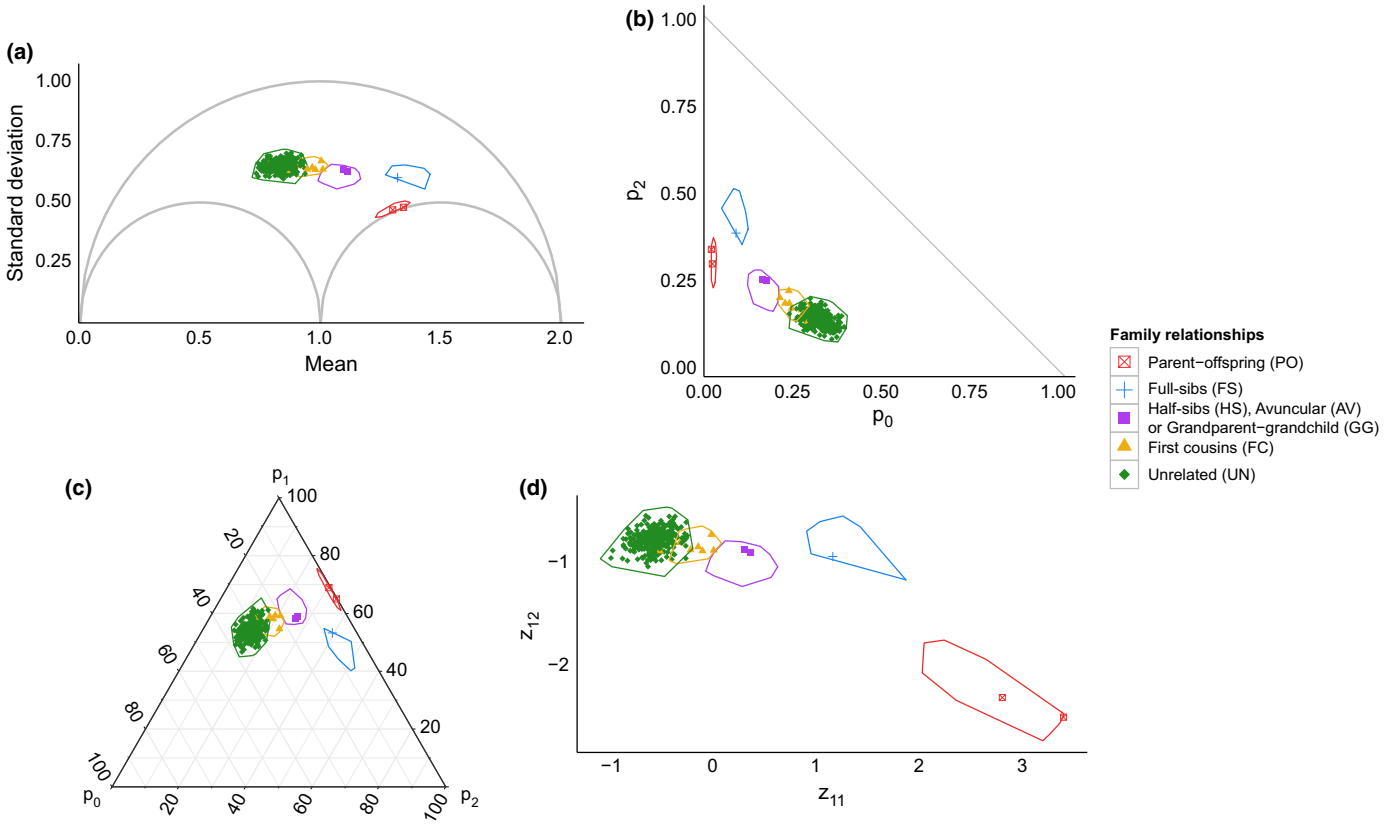

FIGURE 7 Identical by state (IBS) alleles for all the pairs of individuals from the Maya population. a. Plot of means versus standard deviations. b. (p_2, p_0)-plot. c. Ternary diagram. d. Ilr-coordinates: $\boldsymbol{z}_1 = (z_{11}, z_{12})$. The convex hulls are obtained by simulating artificial children from a subset of unrelated individuals from the Maya population and each hull is based on 80 PO, 48 FS, 120 second degree, 36 FC and 1256 UN artificial pairs

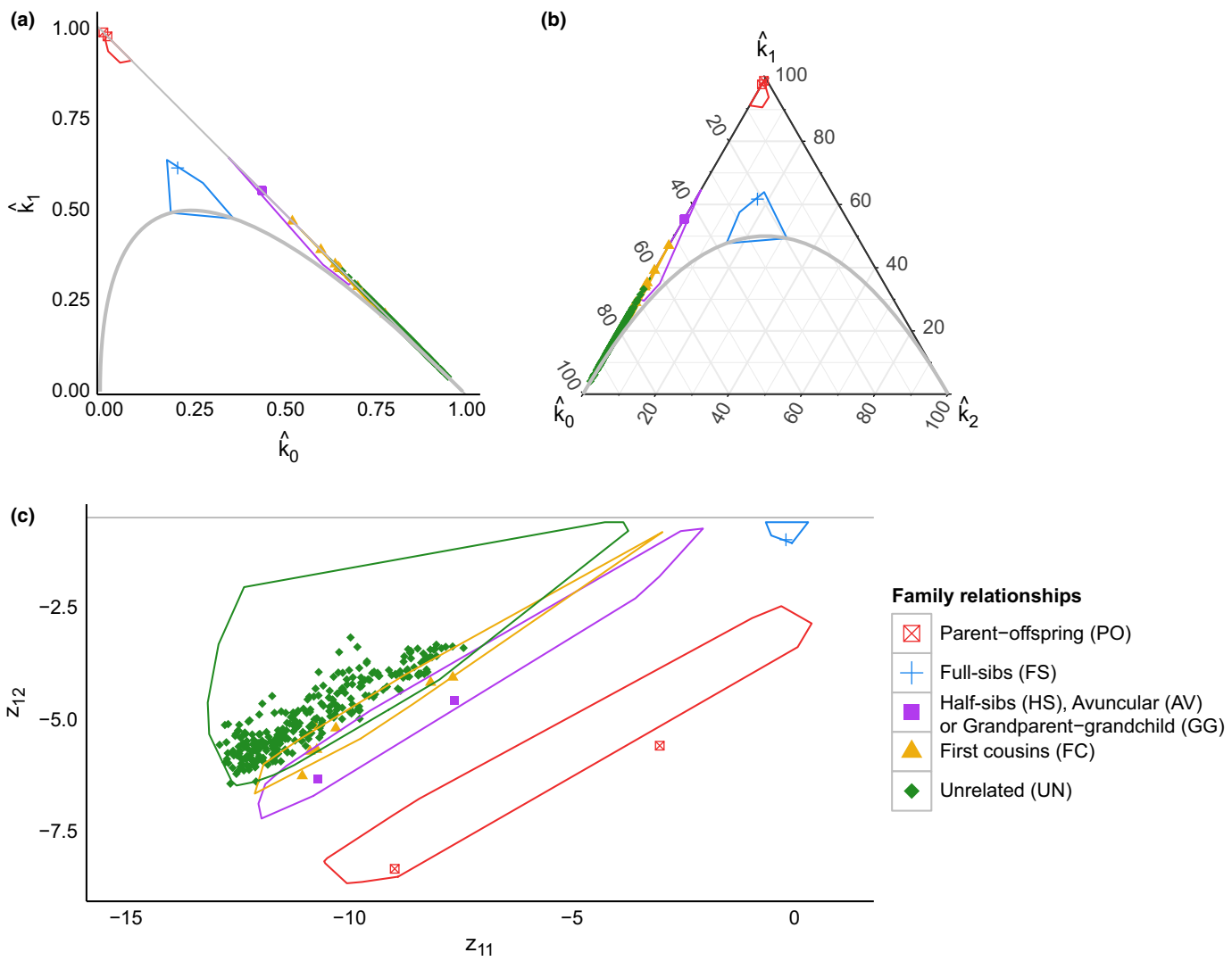

FIGURE 8 Identical by descent (IBD) alleles for all the pairs of individuals from the Maya population. a. $(\hat{k}_0, \hat{k}_1)$-plot. b. Ternary diagram. c. Ilr-coordinates: $\boldsymbol{z}_1 = (z_{11}, z_{12})$

positions for the different relationships. These hulls mainly confirm the assigned relationships. In ilr-coordinates, PO hulls do not capture all observed PO pairs (see Discussion).

6 | DISCUSSION

The main aim of this article was to review all graphical methods used in relatedness research. We have distinguished graphics based on IBS and IBD allele sharing. Plotting means versus standard deviations of the IBS counts allows us to detect monozygotic twins (MZ), parent-offspring (PO) and full-sibs (FS) pairs. However, higher degree relationships are more difficult to detect visually. The distances between unrelated and related pairs depend on the allele frequency distribution of the markers under study. The larger the heterozygosity in a population, the larger the distances between related and unrelated individuals are. A disadvantage of this mean-variance plot is that there are no fixed reference points for the standard relationships. Such reference points could eventually be found by calculating expectations of the mean and the variance of the IBS counts. These do depend on the allele frequency distribution and will therefore depend on the population that has been sampled, and on the distribution of the allele frequencies in that population. The (p_i, p_j)-plots allow easy detection of MZ pairs (or duplicated individuals) because they have p_2 values close to 1, and PO pairs have low values of p_0 and are also easy to detect. FS pairs are located between PO pairs and the pairs with large values of p_0. However, it remains hard to detect relationships of the second and third degree. The (p_i, p_j)-plots neither have a fixed reference position for the standard relationships. Moreover, as has been noted in Section 2, the Euclidean distance between two pairs in a (p_i, p_j)-plot is not invariant with respect to the chosen index (0, 1 or 2), for example, is not the same in a (p_0, p_1) and a (p_0, p_2)-plot. $(\hat{k}_i, \hat{k}_j)$-plots have, in comparison with (p_i, p_j)-plots, the advantage that fixed reference positions for the standard relationships exist, as given in Table 2. This is of great practical value when inferring relationships. Moreover, IBD plots are more reliable for classifying relationships because they show a larger degree of separation between the different relationships than their

IBS counterparts. This is clearly visible when one compares Figures 2 with 5a, 3 with 5b, 7b with 8a and 7c with 8b. However, the IBD-based $(\hat{k}_i, \hat{k}_j)$-plots suffer from the same problem as their IBS counterparts: the Euclidean distances between pairs (and reference points) depend on the index (0, 1 or 2) that is used.

We comment on some peculiarities of the HGDP-CEPH database analysed in the article. We found the high estimate of k_1 (0.27) in Table 1 for the reported UN pair to be not too unusual for Maya UN pairs, being the median of k_1 0.17 for UN pairs of this population. The relatively high k_1 estimates are probably to some extent due to inbreeding, as the South American populations had the largest medians of k_1 for UN pairs. However, for many other less inbred populations k_1 estimates of UN pairs had a large median too, in the range 0.1–0.2. We suggest the database could be affected by a certain degree of sample contamination, as this will increase the number of heterozygote calls, leading to overestimated IBD (Andoh, Sato, Sakamoto, Yoshida, & Ohtaki, 2010).

We continue with some remarks on the graphics from CoDA proposed in this article. We advocate the ternary diagram as an alternative for the (p_i, p_j)-plots because it clearly shows all three proportions simultaneously. MZ twins are close to the vertex p_2; PO pairs are easy to identify on the opposite side of the vertex p_0. FS pairs usually have large values of p_2 and are separated from unrelated pairs which have lower values of p_2. We also advocate the ternary diagram for IBD studies for the same reasons: all three estimated IBD probabilities are represented in one single graph with all three k_i axes. The theoretical IBD probabilities (Table 2) are easily added for use as reference points. The ternary diagram resolves the indeterminacy of the Euclidean distances between pairs due to the choice of axes observed above in (p_i, p_j) and (k_i, k_j) scatterplots. However, the interpretation of Euclidean distances in the ternary diagram remains a tricky issue, because the simplex is a constrained space. We note that the Euclidean distance is regarded inadequate for the comparison of compositions, and for this reason, we have considered isometric log-ratio coordinates of IBS and IBD probabilities. The Euclidean distances between the pairs in ilr-coordinates correspond to Aitchison distances between (p_0, p_1, p_2) (or (k_0, k_1, k_2)) compositions. The Aitchison distance is considered to be an adequate metric for representing compositions (Pawlowsky-Glahn, Egozcue, & Tolosana-Delgado, 2015, Chapter 3). Plotting the ilr-coordinates of the IBS proportions is useful for detecting related individuals because usually unrelated individuals are concentrated in a cloud of points and most outlying individuals correspond to related pairs. Plotting the ilr-coordinates of the estimated Cotterman coefficients gives reference zones over the ilr space for the different relationships (Figure 6). Standard family relationships can be inferred depending on the values of z_{11} and z_{12}. UN pairs are mainly represented in the scatterplot of the isometric log-ratios of IBD probabilities by a central cloud of points around $(-10, -5)$ (Figure 8c) but also by points close to the upper limit of the second ilr-coordinate $(-\sqrt{(2/3)}\ln(2))$. A small change in the tolerance or the initial point of the maximization algorithm can greatly influence the final position of an UN pair. Both IBS- and IBD-based log-ratio plots show a strong discrimination of PO and FS pairs which typically appear as outliers in these plots. We also note that all inference on relationships in all presented graphical methods relies on the judgement of the analyst, who interprets distances between points in a graph. Depending on the sample size of the study, the number of markers used for the genotyping and the distributions of their allele frequencies, those distances will be subject to some degree of uncertainty which complicates graphical inference on relationships. By simulating artificial related pairs using the genotypes of unrelated pairs of the database, convex hulls for the expectation of the standard relationships can be obtained, which are conditional on the observed sample allele frequencies. These convex hulls assess the degree of uncertainty that can be expected for the different related pairs and are helpful for confirming putative relationships. In the present work, the convex hulls are limited by the fact that they assumed independent markers. This may explain why some related pairs are outlying with respect to their corresponding convex hulls. The accuracy of the convex hulls depends on the sample size, and in particular on the number of UN individuals in the sample from which it is generated. More accurate convex hulls may be obtained if linkage disequilibrium is taken into account and artificial pairs are generated by sampling from haplotypes instead of by sampling individual markers independently. Convex hulls of PO pairs in ilr-coordinates often do not capture all observed PO pairs (Figure 8). We suggest this might be due to a small sample size combined with numerical instability. The position of a PO pair in ilr-coordinates has a high variability and depends on the tolerance and initial point used in the maximization of the likelihood (Graffelman & Galván-Femenía, 2016). If the sample size is small, or the number of simulated pairs is small, the PO hull many not cover the full area compatible with PO pairs. It is worth remarking that PO and FS convex hulls do not intersect each other and do not overlap with the rest of the hulls, having a valuable discrimination power (Figures 7 and 8). We think the current simulated convex hulls are helpful to assess uncertainty but of limited value and see a clear need for methods of formal statistical inference on relationships by means of hypothesis testing and confidence regions (García-Magariños et al., 2015).

7 | SOFTWARE

R functions for making the graphics in this manuscript are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.2532d.

ACKNOWLEDGEMENTS

This study was supported by grants MTM2015-65016-C2-2-R and R01 GM075091 from the United States National Institutes of Health (JG) and by MTM2015-65016-C2-1-R (2015-2017) of the Spanish Ministry of Economy and Competitiveness (IGF and CBV). Part of this work was presented at the 6th International Workshop on

Compositional Data Analysis in L'Escala, Spain in 2015. The authors thank Noah Rosenberg for making the HGDP-CEPH diversity panel used in this manuscript publicly available on the website of the Rosenberg Research lab at Stanford University (https://rosenberglab.stanford.edu/). We also thank the editors and all anonymous referees whose comments have helped us to improve the article.

AUTHOR CONTRIBUTIONS

All authors contributed to the writing of this article. I.G.F. analysed data.

DATA ACCESSIBILITY

The data analysed in this article are freely available on the web of the Rosenberg lab at Stanford University (https://rosenberglab.stanford.edu/diversity.html\#2002).

REFERENCES

Abecasis, G. R., Chemy, S. S., Cookson, W. O., & Cardon, L. R. (2001). GRR: graphical representation of relationship errors. *Bioinformatics*, *17*, 742–743.

Aitchison, J. (1986). *The statistical analysis of compositional data*. UK: Chapman & Hall.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., & Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Mathematical Geology*, *32*, 271–275.

Andoh, M., Sato, Y., Sakamoto, H., Yoshida, T., & Ohtaki, M. (2010). Detection of inappropriate samples in association studies by and IBS-based method considering linkage disequilibrium between genetic markers. *Journal of Human Genetics*, *55*, 436–440.

Bérénos, C., Ellis, P. A., Pilkington, J. G., & Pemberton, J. M. (2014). Estimating quantitative genetic parameters in wild populations: a comparison of pedigree and genomic approaches. *Molecular Ecology*, *23*, 3434–3451.

de Berg, M., van Kreveld, M., Overmars, M., & Schwarzkopf, O. (2000). *Computational geometry: algorithms and applications*. Berlin:Springer. 2–8.

Blouin, M. S. (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology and Evolution*, *18*, 503–511.

Boehnke, M., & Cox, N. J. (1997). Accurate inference of relationships in sib-pair linkage studies. *The American Journal of Human Genetics*, *61*, 423–429.

Cavalli-Sforza, L. L. (2005). The Human Genome Diversity Project: past, present and future. *Nature Reviews, Genetics*, *6*, 333–340.

Cotterman, C. W. (1941). Relative and human genetic analysis. *The Scientific Monthly*, *53*, 227–234.

Croft, D. P., Hamilton, P. B., Darden, S. K., Jacoby, D. M. P., James, R., Bettaney, E. M., & Tyler, C. R. (2012). The role of relatedness in structuring the social network of a wild guppy population. *Oecologia*, *170*, 955–963.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, *35*, 279–300.

Epstein, M. P., Duren, W. L., & Boehnke, M. (2000). Improved inference of relationship for pairs of individuals. *The American Journal of Human Genetics*, *67*, 1219–1231.

Foulkes, A. S. (2009). *Applied statistical genetics with R*. New York, NY: Springer.

García-Magariños, M., Egeland, T., López-de-Ullibarri, I., Hjort, N. L., & Salas, A. (2015). A parametric approach to kinship hypothesis testing using identity-by-descent parameters. *Statistical Applications in Genetics and Molecular Biology*, *14*(5), 465–479.

Gonder, M. K., Mitchell, M. W., Locatelli, S., Ghobrial, L., Pokempner, A. A., Sesink-Clee, P. R., . . . Hahn, B. H. (2015). The population genetics of wild chimpanzees in Cameroon and Nigeria suggests a positive role for selection in the evolution of chimpanzee subspecies. *BMC Evolutionary Biology*, *15*, 3.

Graffelman, J., & Galván-Femenía, I. (2016). An application of the isometric log-Ratio transformation for relatedness research. In: J. A. Martín-Fernández & S. Thió-Henestrosa (Eds.), *Compositional Data Analysis*, *Springer Proceedings in Mathematics & Statistics*, Vol *187*. (pp. 75–84). Cham: Springer International Publishing.

Hamilton, N. (2015). ggtern: An extension to `ggplot2', for the creation of ternary diagrams. R package version 2.1.1. http://CRAN.R-project.org/package=ggtern

Hansen, M. M., Nielsen, E. E., & Mensberg, K. L. D. (1997). The problem of sampling families rather than populations: relatedness among individuals in samples of juvenilebrown trout Salmo trutta L. *Molecular Ecology*, *6*, 469–474.

Loughnan, S. R., Smith-Keune, C., Jerry, D. R., Beheregaray, L. B., & Robinson, N. A. (2016). Genetic diversity and relatedness estimates for captive barramundi (Lates calcarifer, Bloch) broodstock informs efforts to form a base population for selective breeding. *Aquaculture Research*, *47*, 3570–3584.

Milligan, B. G. (2003). Maximum-likelihood estimation of relatedness. *Genetics*, *163*, 1153–1167.

Moltke, I., & Albrechtsen, A. (2014). RelateAdmix: a software tool for estimating relatedness between admixed individuals. *Bioinformatics*, *30*, 1027–1028.

Nembot-Simo, A., Graham, J., & McNeney, B. (2013). CrypticIBDcheck: an R package for checking cryptic relatedness in nominally unrelated individuals. *Source Code for Biology and Medicine*, *8*, 5.

Oliehoek, P. A., Windig, J. J., van Arendonk, J. A. M., & Bijma, P. (2006). Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics*, *173*, 483–496.

Pawlowsky-Glahn, V., & Buccianti, A. (2011). *Compositional data analysis: theory and applications*. Chichester: Wiley.

Pawlowsky-Glahn, V., Egozcue, J.J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. Chichester, United Kingdom: Wiley & Sons.

Pemberton, T. J., Wang, C., Li, J. Z., & Rosenberg, N. A. (2010). Inference of unexpected genetic relatedness among individuals in HapMap phase III. *The American Journal of Human Genetics*, *87*, 457–464.

R Core Team (2015). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org/

Rosenberg, N. A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of Human Genetics*, *70*, 841–847.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science*, *298*, 2381–2385.

Snyder-Mackler, N., Alberts, S. C., & Bergman, T. J. (2014). The sociogenetics of a complex society: female gelada relatedness patterns mirror association patterns in amultilevel society. *Molecular Ecology*, *23*, 6179–6191.

Spencer, P. B. S., Hampton, J. O., Pacioni, C., Kennedy, M. S., Saalfeld, K., Rose, K., & Woolnough, A. P. (2015). Genetic relationships within social groups influence the application of the Judas technique: a case study with wild dromedary camels. *The Journal of Wildlife Management*, *79*, 102–111.

Stanley, R. P. (1997). *Enumerative combinatorics*. Vol. *I*. New York, NY: Cambridge University Press.

Sun, L. (2012). Statistical human genetics: methods and protocols. Chapter 2, 25–46.

Thompson, E. A. (1991). Estimation of relationships from genetic data. In: C. R. Rao & R. Chakraborty (Eds.), *Handbook of Statistics*, Vol *8*. (pp. 255–269). Amsterdam: Elsevier Science.

Thompson, E. A. (2000). Statistical inference from genetic data on pedigrees. NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. *6*. Chapter 3, 29–46.

Wagner, A. P., Creel, S., & Kalinowski, S. T. (2006). Estimating relatedness and relationships using microsatellite loci with null alleles. *Heredity*, *97*(5), 336–345.

Weir, B. S., Anderson, A. D., & Hepler, A. B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews*, *Genetics*, *7*, 771–780.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. New York: Springer.

How to cite this article: Galván-Femenía I, Graffelman J, Barceló-i-Vidal C. Graphics for relatedness research. *Mol Ecol Resour*. 2017;17:1271–1282. https://doi.org/10.1111/1755-0998.12674

4.2 Frontiers in Genetics

The second article accomplishes with the objectives Obj. 1 and Obj. 2 described in 2.1. In summary, we propose a log-ratio biplot approach for identifying family relationships using only identity by state alleles. We show that the allele sharing statistics can be considered as a 6-part composition, and then, the proposed log-ratio biplots have a higher dimensionality than the two-dimensional classical graphical methods. This points to new graphics that have been shown useful for relatedness research.

This article has been published in *Frontiers in Genetics* journal.
Volume: 10, Submitted: December 2018, Accepted: March 2019.
DOI: 10.3389/fgene.2019.00341.
Impact factor: 3.517 (Q2). Journal Citation Reports Ranking: 56/174 (Genetics & Heredity).

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 April 2019
doi: 10.3389/fgene.2019.00341

A Log-Ratio Biplot Approach for Exploring Genetic Relatedness Based on Identity by State

Jan Graffelman [1,2*†], ***Iván Galván Femenía*** [3,4†], ***Rafael de Cid*** [4‡] ***and Carles Barceló Vidal*** [3]

[1] *Department of Statistics and Operations Research, Technical University of Catalonia, Barcelona, Spain,* [2] *Department of Biostatistics, University of Washington, Seattle, WA, United States,* [3] *Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Girona, Spain,* [4] *Genomes For Life - GCAT Lab, Institute for Health Science Research Germans Trias i Pujol (IGTP), Badalona, Spain*

OPEN ACCESS

Edited by:
Steven J. Schrodi,
Marshfield Clinic, United States

Reviewed by:
Himel Mallick,
Merck, United States
Wei-Min Chen,
University of Virginia, United States
William C. L. Stewart,
The Research Institute at Nationwide Children's Hospital, United States
Fabrice Larribe,
Université du Québec à Montréal, Canada

****Correspondence:***
Jan Graffelman
jan.graffelman@upc.edu

†These authors have contributed equally to this work

‡On behalf of GCAT Project Team

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *05 December 2018*
Accepted: *29 March 2019*
Published: *24 April 2019*

Citation:
Graffelman J, Galván Femenía I, de Cid R and Barceló Vidal C (2019) A Log-Ratio Biplot Approach for Exploring Genetic Relatedness Based on Identity by State. Front. Genet. 10:341. doi: 10.3389/fgene.2019.00341

The detection of cryptic relatedness in large population-based cohorts is of great importance in genome research. The usual approach for detecting closely related individuals is to plot allele sharing statistics, based on identity-by-state or identity-by-descent, in a two-dimensional scatterplot. This approach ignores that allele sharing data across individuals has in reality a higher dimensionality, and neither regards the compositional nature of the underlying counts of shared genotypes. In this paper we develop biplot methodology based on log-ratio principal component analysis that overcomes these restrictions. This leads to entirely new graphics that are essentially useful for exploring relatedness in genetic databases from homogeneous populations. The proposed method can be applied in an iterative manner, acting as a looking glass for more remote relationships that are harder to classify. Datasets from the 1,000 Genomes Project and the Genomes For Life-GCAT Project are used to illustrate the proposed method. The discriminatory power of the log-ratio biplot approach is compared with the classical plots in a simulation study. In a non-inbred homogeneous population the classification rate of the log-ratio principal component approach outperforms the classical graphics across the whole allele frequency spectrum, using only identity by state. In these circumstances, simulations show that with 35,000 independent bi-allelic variants, log-ratio principal component analysis, combined with discriminant analysis, can correctly classify relationships up to and including the fourth degree.

Keywords: allele sharing, composition, identity by state, identity by descent, log-ratio transformation

1. INTRODUCTION

The detection of pairs of related individuals in genomic databases is important in many areas of genetic research. In population-based gene-disease association studies, the assumption of independent observations which is usually made in the statistical modeling of the data, may be violated due to related individuals. Cryptic relatedness can lead to an increased false positive rate in association studies, in particular if related individuals are oversampled (Voight and Pritchard, 2005). In conservation genetics, unrelated individuals are carefully selected in breeding programs in order to maximize genetic diversity (Oliehoek et al., 2006). In quality control of genetic variants produced by high-throughput techniques, accidental duplication of samples in genetic studies can be detected by a relatedness analysis (Abecasis et al., 2001). In ecology, samples of species

often contain an excess of close relatives. This can lead to biased estimates of population-genetic parameters, lower the precision of their estimates, and inflated type 1 error rates of tests for genetic equilibria (Wang, 2018). In practice, most relatedness investigations are based on allele-sharing statistics such as the average number of identical-by-state (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 identical-by-descent (IBD; Thompson, 1975, 1991), known as Cotterman's coefficients (Cotterman, 1941). Plots of these sharing statistics typically show clusters that correspond to unrelated pairs (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 (see **Figures 1A–C**).

All these methods collapse the data to two statistics, that can summarize relatedness in two dimensions. Classical plots are the mean vs. the standard deviation of the shared number of alleles over loci [the (m, s) plot, see **Figure 1A**], the fractions of loci for which a pair of individuals shares 0 or 2 IBS alleles [the (p_0, p_2) plot, see **Figure 1B**], or the estimated probabilities of sharing 0 or 1 allele IBD [the $(\hat{k}_0, \hat{k}_1)$ plot, see **Figure 1C**]. However, all allele sharing statistics are estimated from the genotype data. For a pair of individuals with bi-allelic variants, there exist six possible pairs of genotypes, and their counts over the k variants determine the IBS allele sharing statistics. From this perspective, the observed genotype sharing data consists of vectors of six elements, that, when expressed in percentage form, occupy a five dimensional space. This suggests that the classical approaches of collapsing the data into two dimensions by plotting the summary statistics may not extract all information about relatedness that is present in the data. In this paper we propose to explore the data in five dimensions by using log-ratio principal component analysis (PCA), which is specially designed for analyzing compositional data (Aitchison, 1983). A log-ratio PCA allows us to construct comprehensive biplots that uncover the main relatedness features of the data.

Biplots are widely used in genetic research, in particular for the graphical representation of quantitative traits of genotypes in plant genetics (Anandan et al., 2016; Pandit et al., 2017; Sharma et al., 2018). In relatedness research, a PCA of bi-allelic genetic variants, coded in 0, 1, 2 format (for AA, AB, and BB respectively) is often used to investigate the existence of population substructure, that is, remote genetic relatedness. The plots obtained by this kind of PCA are, in principle, biplots, though often the genetic variants are omitted in such plots because there are too many of them. Substructure is also often investigated by multidimensional scaling (MDS) of allele sharing distances between individuals. The resulting MDS maps only represent individuals, and some authors prefer the term monoplots for such graphics (Gower et al., 2011). If MDS is based on the Euclidean distances, then a covariance-based PCA and MDS are in fact equivalent (Mardia et al., 1979). The MDS plots, PCA biplots without variable vectors for the genetic variants, are particularly popular in substructure investigations in human genetics (Jakobsson et al., 2008; Sabatti et al., 2009; Pemberton et al., 2010, 2013; Wang et al., 2010).

The biplot approach proposed in this paper differs from the classical applications described above in several ways. We propose a biplot of the genetic data of *pairs* of individuals, that represents artificially related pairs of a reference set of given familial relationships, generated by a respampling of the genetic data. The empirically observed pairs are used in a supplementary way, and are projected onto the reference biplot. The data matrix used in this biplot is not a (0, 1, 2) genetic data matrix, neither a distance matrix of allele sharing distances, but consists of vectors of counts of genotype patterns [(AA,AA), (AA,AB), etc.] which we treat as compositions, and we therefore use a log-ratio approach. More details are given in the section 2 below.

An important additional advantage of using log-ratio PCA in this context is that it allows us to explore the data iteratively with a *peel and zoom* procedure. A first log-ratio PCA may clearly reveal a cluster of FS pairs. Once identified, the corresponding pairs can be removed from the database, and log-ratio PCA can be repeated on the remaining pairs. The second analysis will focus more closely on more remote relationships that may be present in the database, and thereby act as a magnifying glass for the latter. The aforementioned classical graphics do not have this property, as they are invariant under removal of a relationship category.

The remainder of this paper is organized as follows. In the section 2 we provide background on relatedness research and log-ratio PCA, and show how to construct biplots that are useful for relatedness research. In the section 3 we study the discriminative power of log-ratio PCA and compare this with the classical plots in a simulation study. We also describe two empirical examples of our method with data from two different population-based datasets; a next generation sequencing dataset from the 1,000 Genomes Project (The 1000 Genomes Project Consortium, 2015) and a genome-wide SNP array technology dataset from the GCAT Genomes For Life Cohort Study of the Genomes of Catalonia (Galván-Femenía et al., 2018; Obón-Santacana et al., 2018). A discussion finishes the paper.

2. METHODS

We first summarize some basic methods for relatedness research (section 2.1), then give a brief account of log-ratio PCA (section 2.2), and finally show how log-ratio PCA can be used in relatedness research (section 2.2).

2.1. 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. **Table 1** shows all the possible combinations of the IBS alleles shared for a pair of individuals at a biallelic variant. Considering k biallelic variants, each pair of individuals has a vector of 0, 1, and 2 IBS counts of length k. In IBS studies, the means (m) and standard deviations (s) of the vector of the IBS allele counts

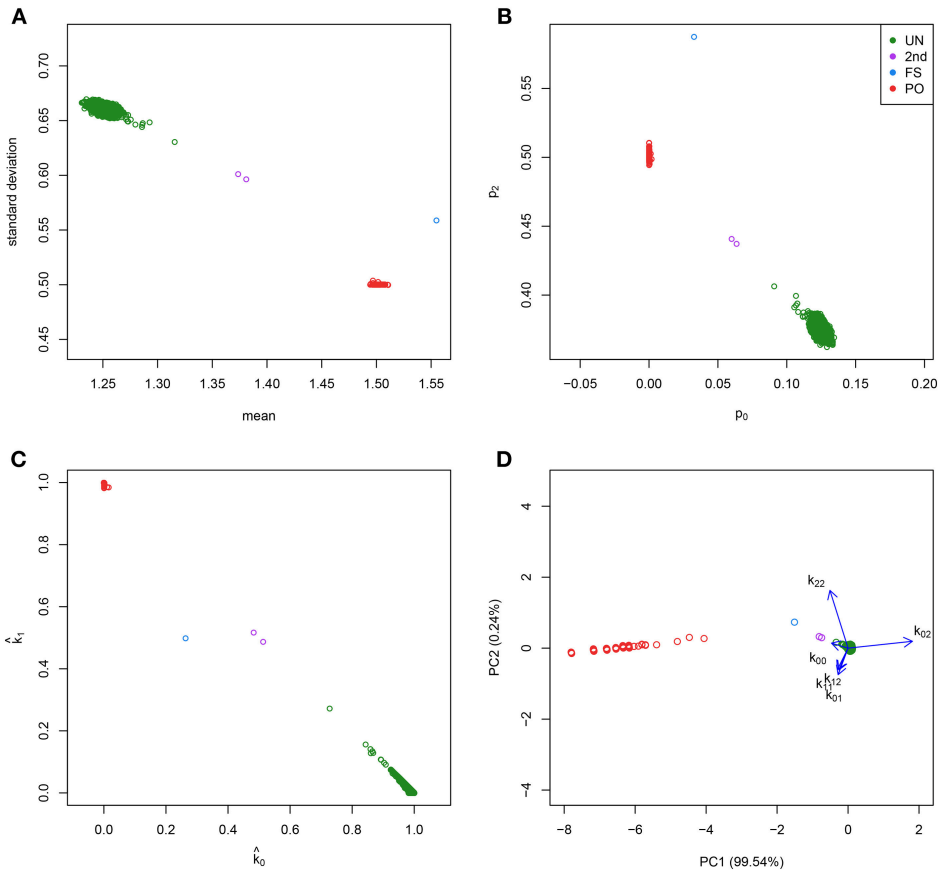

FIGURE 1 | Classical graphics for relatedness research and log-ratio PCA biplot. Plots show the CEU sample of the 1000G project. IBS/IBD statistics were calculated over a set of 26,081 complete, LD-pruned autosomal SNPs with MAF above 0.4, and HWE exact test *p*-value above 0.05. **(A)** Scatterplot of the mean and standard deviation of the number of IBS alleles. **(B)** Scatterplot of the fraction of variants sharing two (p_2) against the fraction sharing zero (p_0) IBS alleles. **(C)** Scatterplot of the estimated probability of sharing one ($\hat{k}_1$) against the estimated probability of sharing zero ($\hat{k}_0$) IBD alleles. **(D)** log-ratio PCA biplot.

(Abecasis et al., 2001), or the proportions of variants sharing 0, 1, and 2 IBS alleles (denoted p_0, p_1, and p_2 respectively, Rosenberg, 2006) can be plotted (see **Figures 1A,B**). These plots reveal characteristic clusters corresponding to MZ, PO, FS, UN, and other pairs. Alternatively, in an IBD based approach, the probability of sharing 0, 1, or 2 IBD alleles for a pair of individuals (usually denoted by k_0, k_1, and k_2 and referred to as Cotterman's coefficients) can be represented in a scatterplot (see **Figure 1C**, Nembot-Simo et al., 2013). The Cotterman coefficients can be estimated by the method of moments (Purcell et al., 2007), maximum likelihood (Thompson, 1991; Milligan, 2003; Weir et al., 2006), or robust estimation methods (the KING program, Manichaikul et al., 2010). In IBD studies, reference values for the standard relationships are available (see **Table 2**). Related pairs can also be distinguished, albeit at lower resolution, by using the co-ancestry coefficient defined as $\theta = k_1/2 + k_2$ or the kinship coefficient defined as $\phi = \theta/2$. Galván-Femenía et al. (2017) give an overview of graphics used in relatedness research. **Figure 1** 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 1.000 Genomes project. These plots distinguish UN, PO, FS, and second degree pairs. 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 McPeek and Sun (2000; the Prest-plus program). Throughout this paper we employ the classical notion of degree of relationship, shown in the second column of **Table 2**, with PO and FS being considered first degree, HS, GG and AV, second degree, FC third degree, first cousins once removed fourth degree, second cousins fifth degree and second cousins once removed sixth degree, and so on.

TABLE 1 | Number of IBS alleles for possible combinations of genotypes.

	AA	AB	BB
AA	2	1	0
AB	1	2	1
BB	0	1	2

2.2. Log-Ratio Principal Component Analysis

Aitchison (1983) proposed log-ratio principal component analysis (PCA) for the exploration of compositional data. Many successful applications of log-ratio PCA have been described in

TABLE 2 | IBD probabilities for standard relationships.

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

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).

the literature, notably in geology. We briefly summarize log-ratio PCA and biplot construction (see Pawlowsky-Glahn et al., 2015 for a comprehensive account). Log-ratio PCA is usually performed by applying the centered log-ratio transformation to the compositional data, and we will follow that approach here. Let $\mathbf{X}$ be a matrix with n compositions in its rows, and having D parts (columns). Compositional data can be defined as strictly positive vectors for which the information of interest is in the ratios between the components (Aitchison, 1986). We consider the centered log-ratio transformation (clr) of a composition $\mathbf{x}$ (a row of $\mathbf{X}$) given by

$$\mathrm{clr}(\mathbf{x}) = \left[\ln\left(\frac{x_1}{\mathrm{g_m}(\mathbf{x})}\right), \ln\left(\frac{x_2}{\mathrm{g_m}(\mathbf{x})}\right), \cdots, \ln\left(\frac{x_D}{\mathrm{g_m}(\mathbf{x})}\right)\right], \quad (1)$$

where $\mathrm{g_m}(\mathbf{x})$ is the geometric mean of the components of the composition $\mathbf{x}$. Let $\mathbf{X}_\ell$ be the log transformed compositions, that is $\mathbf{X}_\ell = \ln(\mathbf{X})$ with the natural logarithmic transformation applied element-wise. The clr transformed data can be obtained by just centering the rows of this matrix, using the centering matrix $\mathbf{H}_r = \mathbf{I} - \frac{1}{D}\mathbf{1}\mathbf{1}'$. Then

$$\mathbf{X}_{\mathrm{clr}} = \mathbf{X}_\ell \mathbf{H}_r, \quad (2)$$

The rows of $\mathbf{X}_{\mathrm{clr}}$ are subject to a zero sum constraint because $\mathbf{H}_r\mathbf{1} = \mathbf{0}$. If there are no additional linear constraints, then $\mathbf{X}_{\mathrm{clr}}$ will have rank $D-1$. We now column-center the clr transformed data, producing a double-centered data matrix that has zero column and row means:

$$\mathbf{X}_{c\mathrm{clr}} = \mathbf{H}_c\mathbf{X}_{\mathrm{clr}} = \mathbf{H}_c\mathbf{X}_\ell\mathbf{H}_r, \quad (3)$$

where $\mathbf{H}_c$ is the centering matrix $\mathbf{H}_c = \mathbf{I} - (1/n)\mathbf{1}\mathbf{1}'$. Matrix $\mathbf{X}_{\mathrm{cclr}}$ is used as the input for a classical principal component analysis. We perform PCA by the singular value decomposition:

$$\mathbf{X}_{\mathrm{cclr}} = \mathbf{U}\mathbf{D}\mathbf{V}' = \mathbf{F}_p\mathbf{G}_s', \quad (4)$$

with $\mathbf{F}_p = \mathbf{U}\mathbf{D}$ and $\mathbf{G}_s = \mathbf{V}$. Matrix $\mathbf{F}_p$ contains the principal components, and its first two columns contain the biplot coordinates of the compositions. The columns of $\mathbf{G}_s$ are the eigenvectors of the covariance matrix of $\mathbf{X}_{\mathrm{cclr}}$, its first two columns contain the biplot coordinates of the parts of the compositions. We use sub-indexes p and s to distinguish principal and standard biplot coordinates. We will need to project supplementary compositions onto a given biplot (see section 3). This can be accomplished by regression (Graffelman and Aluja-Banet, 2003). The biplot coordinates, $\tilde{\mathrm{F}}_p$, of a matrix of supplementary compositions, $\mathbf{Y}$, can be found as

$$\tilde{\mathrm{F}}_p = \left(\mathbf{G}_s'\mathbf{G}_s\right)^{-1}\mathbf{G}_s'\mathbf{Y}_{c\mathrm{clr}}, \quad (5)$$

where $\mathbf{Y}_{c\mathrm{clr}}$ contains the clr-transformed supplementary compositions, but centered with respect to the compositions in $\mathbf{X}$, that is

$$\mathbf{Y}_{c\mathrm{clr}} = \mathbf{Y}_{\mathrm{clr}} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}_{\mathrm{clr}}. \quad (6)$$

We will construct a biplot of genotypic reference compositions by using Equation (4), and project empirical genotype compositions onto the biplot by using Equations (5) and (6).

2.3. Log-Ratio PCA of Genotype Sharing Data

For bi-allelic variants with alleles A and B, there exist six possible pairs of genotypes whose counts over k variants can be laid out in a triangular array shown in **Table 3**, 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. Consequently, each pair can be represented by a vector of six counts which can be expressed as a composition by division by its total (closure):

$$\mathbf{x} = (k_{00}, k_{10}, k_{20}, k_{11}, k_{21}, k_{22})/k. \quad (7)$$

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, almost never observed due to the existence of some mutations and genotyping error. Given n individuals, we construct matrix $\mathbf{X}$ with $q = \frac{1}{2}n(n-1)$ pairs in its rows, and propose to study relatedness by a log-ratio PCA of this $q \times 6$ matrix of compositions. This will allow the construction of a biplot, where each pair of individuals is represented by a point, and each part of the clr transformed composition by a vector. A drawback of the representation of pairs of individuals in a log-ratio PCA biplot is that the type of relationship cannot be inferred if it is undocumented. Without additional analysis one does not know for sure whether observed clusters correspond to FS, HS, or other pairs. 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 by constructing pedigrees from this subset, applying the Mendelian inheritance rules. For example, 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

TABLE 3 | Lower triangular matrix layout with counts for all possible genotype pairs.

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

All possible genotype pairs for a bi-allelic genetic variant. k_{ij} represents the number of genetic variants with i and j B alleles for a pair of individuals.

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 a re-sampling the alleles of the observed individuals. 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 generated conditionally on the allele frequencies of the observed sample. We now first apply log-ratio PCA to the pairs of the reference set (**X**), and construct a biplot of the reference set. The empirically observed pairs (**Y**) are projected onto this PCA biplot 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 (LDA) (Johnson and Wichern, 2002), using the simulated pairs as a training set.

3. RESULTS

In this section we first validate the proposed methodology with some simulations, comparing the log-ratio PCA approach with the well-known aforementioned (m, s), (p_0, p_2), and $(\hat{k}_0, \hat{k}_1)$ plots, and then show two examples with empirical genetic data.

3.1. Simulations

We simulated 35,000 independent genetic bi-allelic variants by sampling from a multinomial distribution under the Hardy-Weinberg assumption, using a minor allele frequency (MAF) of 0.5 for all variants. Using Mendelian inheritance rules, 100 independent pairs of each type of relationship were simulated. We assume a homogeneous population without mutation and genotyping error, generating simulated data sets that are free of Mendelian inconsistencies. The classical plots and the log-ratio PCA biplot of a simulation are shown in **Figure 2**. This figure shows that first and second degree pairs are easily identified by all methods. We will therefore focus on third and higher degree relationships which are harder to distinguish as they tend to blur in the plots. We investigated the effect of MAF and number of SNPs on the classification rate of our procedure, using different numbers of principal components for classification of third through sixth degree pairs (100 of each). **Figure 3** shows the classification rates obtained as a function of the minor allele frequency (MAF), the number of SNPs and the number of principal components used. These figures show, as expected, that the classification rate increases with the MAF and the number of SNPs. The simulations show that all five components are needed at low MAF, where more components increase the classification rate. At high MAF (0.40–0.50) there is little or no benefit in using more than two components. With 35,000 SNPs at 0.50 MAF the classification rate is around 95% irrespective of the number of components. With 35,000 SNPs at 0.10 MAF the classification rate varies from below 50% with one component through 93% using all five components. We report the false positive rates in **Table S1**; No UN or 6th degree individuals were misclassified as 4th degree or lower, and only 1.8% of the 5th degree pairs are misclassified as 4th degree. The simulations show that IBS based log-ratio PCA can discriminate higher degree relationships if a sufficient number of independent highly polymorphic variants is available. In the light of the simulations, we decided to use three principal components for classification with high MAF variants for the empirical data sets described in section 3.3.

3.2. Method Comparison

We compare our method with aforementioned classical procedures for identification of related pairs. **Figure 4** shows the classification rate as a function of the number variants with MAF 0.50 for four methods: the two IBS-based methods, the (m, s) plot and the (p_0, p_2) plot; one IBD-based method, the $(\hat{k}_0, \hat{k}_1)$ plot, using the KING estimator (Manichaikul et al., 2010); and the log-ratio PCA approach proposed in this paper. These classification rates were obtained by averaging over 25 replicates of the simulations, for each value of the MAF and the number of variants. It is clear that the log-ratio PCA approach (using three principal components) gives the best classification rates for all relationships. There is little difference in classification rate for third degree relationships, which are relatively more easy to classify. Interestingly, in terms of classification rate the (m, s) and (p_0, p_2) plots are seen to be fully equivalent, as they have exactly the same classification rate profile. Posteriorly, we found these statistics to be related by the equations $m = 1 - p_0 + p_2$ and $s = \sqrt{p_0(1-p_0) + p_2(1-p_2) + 2p_0p_2}$. As expected, classification rate increases with the number of variants. The results suggest that for all four methods 25,000 variants with MAF 0.50 are sufficient to almost perfectly classify PO, FS, second, third, and fourth degree relationships. The difference in classification rate between the log-ratio PCA approach and the conventional methods is larger for the more remote relationships. 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 therefore also investigated the effect of the MAF on the discriminatory power of the different methods, by simulating variants with different MAFs. **Figure 5** shows how the classification rate varies as a function of the MAF, using a fixed number of 5.000 bi-allelic polymorphisms. The log-ratio PCA approach, using five principal components, is seen to outperform the classical plots over the full MAF range.

3.3. Empirical Data Sets

In this section we use log-ratio PCA for a relatedness study of two genomic data sets. We use the CEU population of

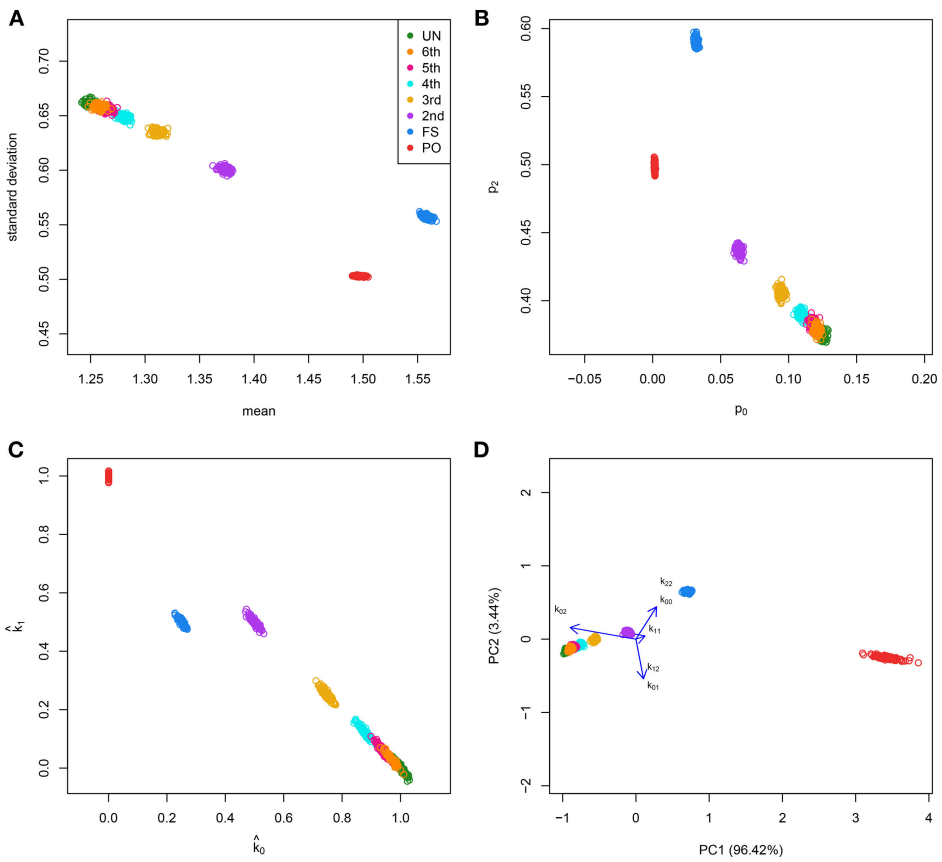

FIGURE 2 | Classical graphics and log-ratio PCA biplot for simulated samples. 100 pairs of each type of relationship [UN, sixth, fifth, fourth, third (FC), second (HS), FS, and PO] were generated using 35,000 independent bi-allelic variants with minor allele frequencies of 0.5, assuming Hardy-Weinberg equilibrium. **(A)** Scatterplot of the mean and standard deviation of the number of IBS alleles. **(B)** Scatterplot of the fraction of variants sharing two (p_2) against the fraction sharing zero (p_0) IBS alleles. **(C)** Scatterplot of the estimated probability of sharing one ($\hat{k}_1$) against the estimated probability of sharing zero ($\hat{k}_0$) IBD alleles. **(D)** log-ratio PCA biplot.

the 1,000 genomes project (www.internationalgenome.org, The 1000 Genomes Project Consortium, 2015), whose family relationships have been analyzed in detail by Pemberton et al. (2010), Kyriazopoulou-Panagiotopoulou 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, filtering and IBD estimation, and R (R Core Team, 2014) for log-ratio PCA and discriminant analysis.

3.3.1. The CEU Sample

First and second degree relationships for the CEU population were documented by Pemberton et al. (2010) using IBS methods, and confirmed by Kyriazopoulou-Panagiotopoulou 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). We detail the analysis of the CEU panel using log-ratio PCA. Variants were filtered according to missingness (only variants genotyped for all individuals were used), MAF ($>$ 0.40) and Hardy-Weinberg equilibrium test result (exact test mid p-value $>$ 0.05, Graffelman and Moreno, 2013). Variants were LD-pruned with Plink using a sliding window of 50 SNPs with an overlap of 5 SNPs between successive windows, and SNPs are removed from the window until no variants remain that have a squared correlation above 0.20 (Plink 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 classical plots of the allele sharing statistics were shown previously in **Figure 1**, including a log-ratio PCA biplot of all pairs (**Figure 1D**). We now illustrate the log-ratio PCA approach, using an iterative peel and zoom procedure. **Figure 1D** showed PO pairs to be outlying in the first dimension, for having a low k_{02}/k_{00} ratio. Theoretically, this ratio is zero for PO pairs, though with large numbers of variants it is non-zero due to mutations and genotyping errors. In fact, the 96 reported PO pairs are easily identified and excluded from the data by filtering with $k_{02} < 0.005$. Log-ratio PCA biplots, obtained by simulation with unrelated individuals of the CEU sample, are shown in **Figure 6**. The simulated pairs of given relationships are represented by convex hulls, and the projected empirical pairs by open dots that are colored according to their predicted

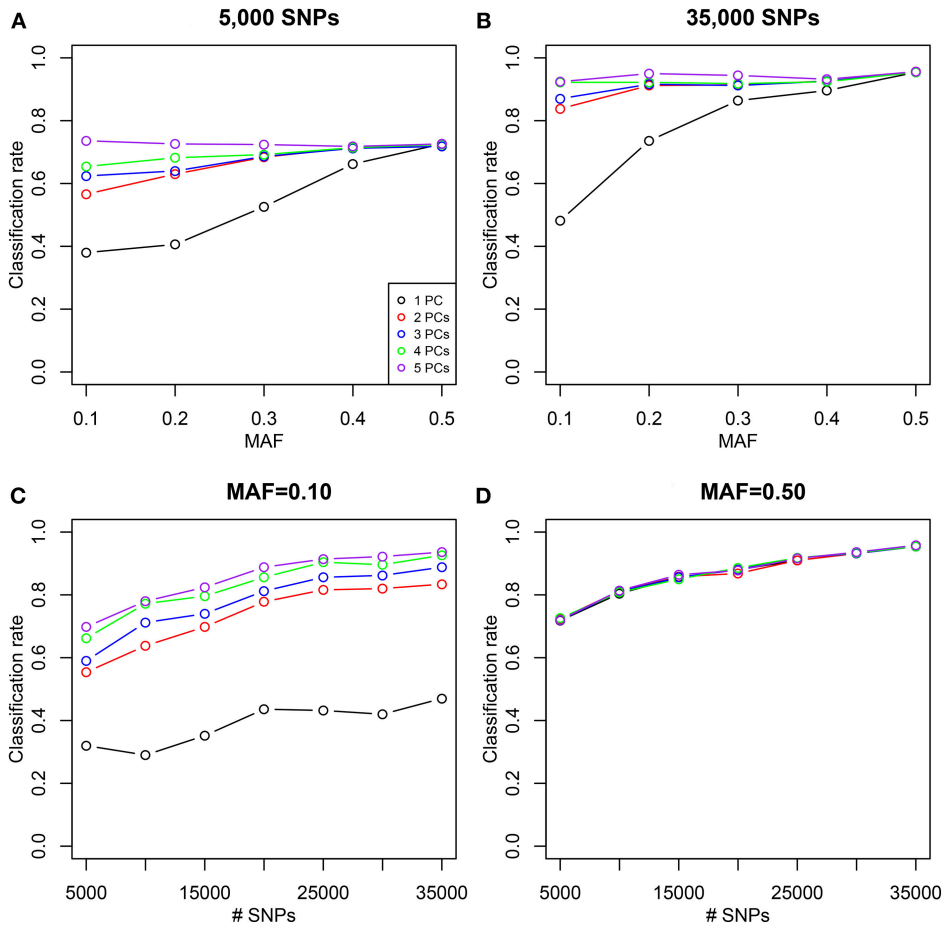

FIGURE 3 | Classification rate of log-ratio PCA combined with LDA for simulated samples. Classification rate for a varying number of principal components (PCs). Classification rates were obtained using 100 pairs of each type of relationships (UN, sixth, fifth, fourth, and third) using independent variants simulated assuming Hardy-Weinberg equilibrium. **(A,B)** Classification rates are shown as a function of the MAF for 5,000 and 35,000 SNPs. **(C,D)** Classification rates are shown as a function of the number of SNPs of a given MAF (0.10 and 0.50).

relationship, where the latter are inferred from the posterior probabilities obtained in LDA. The convex hulls delimit the cloud of the positions of the simulated UN, sixth, fifth, fourth, and third degree pairs (using 100 pairs of each). The overall classification rate of the simulated data was 91.4%, using three principal components. Classification rates for third, fourth, fifth, sixth, and UN were, respectively 100, 100, 90, 77, and 90%. Results in **Figures 1**, **6** suggest 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 and sixth degree pairs that merge with UN pairs. The analysis without PO pairs in **Figure 6A** shows the documented FS and AV pairs as outliers in the first dimension, for having high k_{00}/k_{02} and k_{22}/k_{02} ratios. Re-analysis after removal of the FS pair gives **Figure 6B**, showing the two AV pairs now as strong outliers in the first dimension. Peeling these two pairs, we obtain **Figure 6C**, with the single documented third degree pair being now the most prominent outlier. Five additional pairs are seen to separate from the UN cloud, and are classified as fourth degree pairs. Re-analysis after peeling off the third degree pair gives a plot with a more clear separation of the fourth degree pairs (**Figure 6D**). Another set of pairs, presumably of fifth degree, is seen to bud off from the UN cloud more clearly, once the fourth degree pairs are removed from the analysis (**Figure 6E**), and additional pairs, classified as sixth degree, separate out partly in the third dimension of this analysis. An exploration of the data up to the fifth dimension of the analysis, after peeling the most obvious PO, FS, AV, third, and fourth degree outliers, is shown in **Figure S1**. These graphs suggest there is information on relatedness up to and including at least the third dimension of the analysis.

The classification of the empirical pairs by k_{02} filtering followed 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-Panagiotopoulou et al. (2011). First and second degree relationships in the CEU sample are easily and almost certainly identified. Much more uncertainty resides in relationships of the third and higher degrees, and for these relationships conflicting inferences are reported in the literature. We therefore carried out a linear discriminant analysis with a simulated training sample containing pairs with a third through sixth degree relationship, as well as UN pairs,

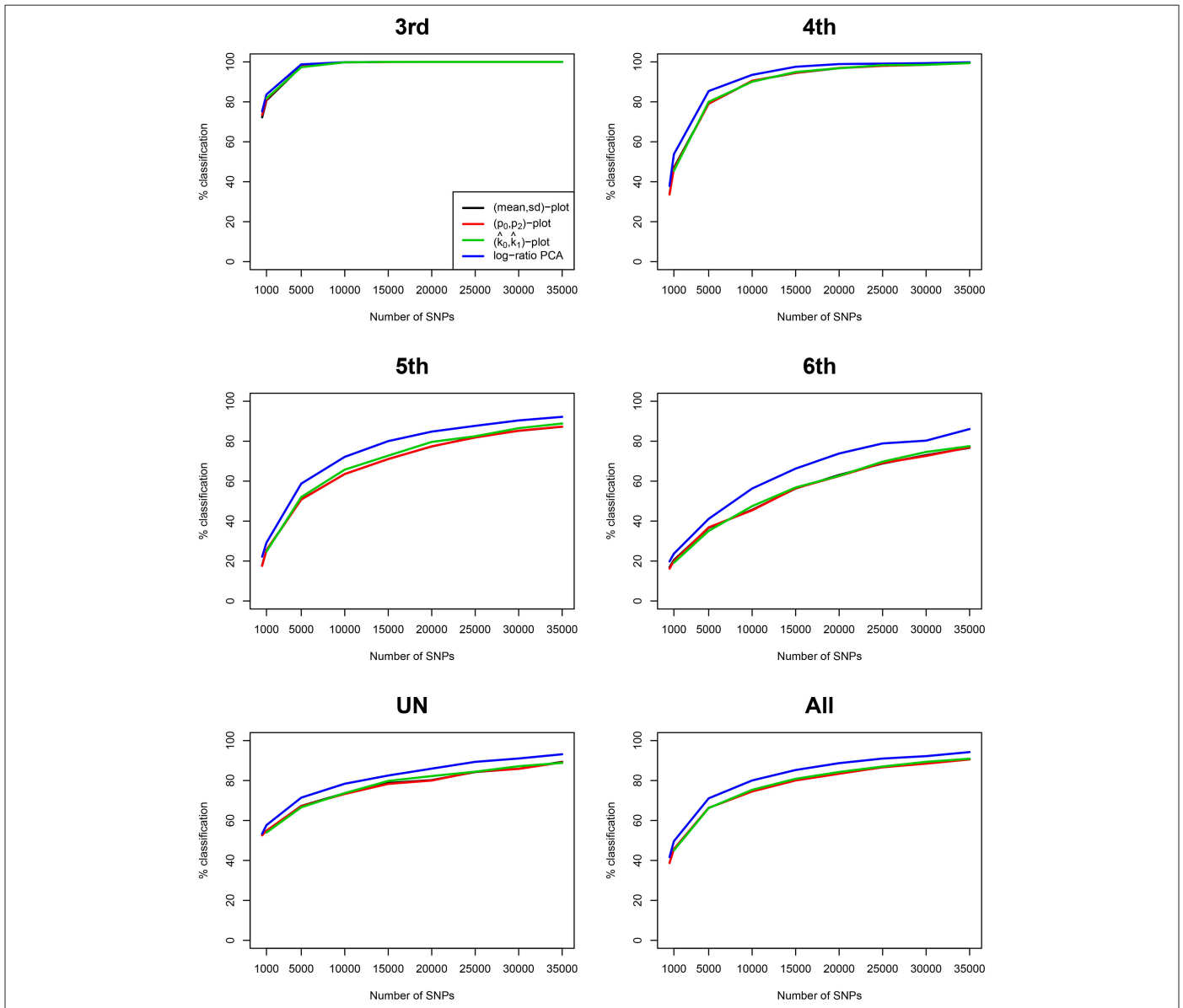

FIGURE 4 | Classification rates for different methods vs. number of SNPs. Classification rates for the different degrees of relationship (third, fourth, fifth, sixth, UN, and All) are shown for four methods, using five principal components. Classification rate profiles for the (m, s) plot and the (p_0, p_2) plot virtually coincide. The last panel All refers to the classification rate for third through UN relationships jointly. Rates are shown as a function of the number of SNPs with MAF 0.50, and were obtained by linear discriminant analysis. 100 pairs of each type of relationship (UN, fifth, fourth, third, second, FS, and PO) were generated assuming Hardy-Weinberg equilibrium.

and classified all empirical pairs which clearly had no first or second degree relationship. Third and fourth degree relationships uncovered by Kyriazopoulou-Panagiotopoulou et al. (2011) are reported in **Table 4**, together with the posterior probabilities obtained in our log-ratio PCA approach. We extended **Table 4** with additional fifth degree pairs uncovered by log-ratio PCA, for which LDA gave the highest posterior probability. In total, 18 pairs were classified as fifth degree relationship pairs, of which 10 had a posterior probability above 0.95 (marked in bold in **Table 4**). We tentatively suggest the CEU panel to contain at least ten fifth degree pairs. We found 1,285 sixth degree pairs, but do not report all these pairs in the light of the overlap with the UN cluster and the somewhat poorer classification rate of the sixth degree observed in the simulations.

Our results confirm a third degree pair (pair 1 in **Table 4**) reported by Kyriazopoulou-Panagiotopoulou 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 or sixth degree. 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

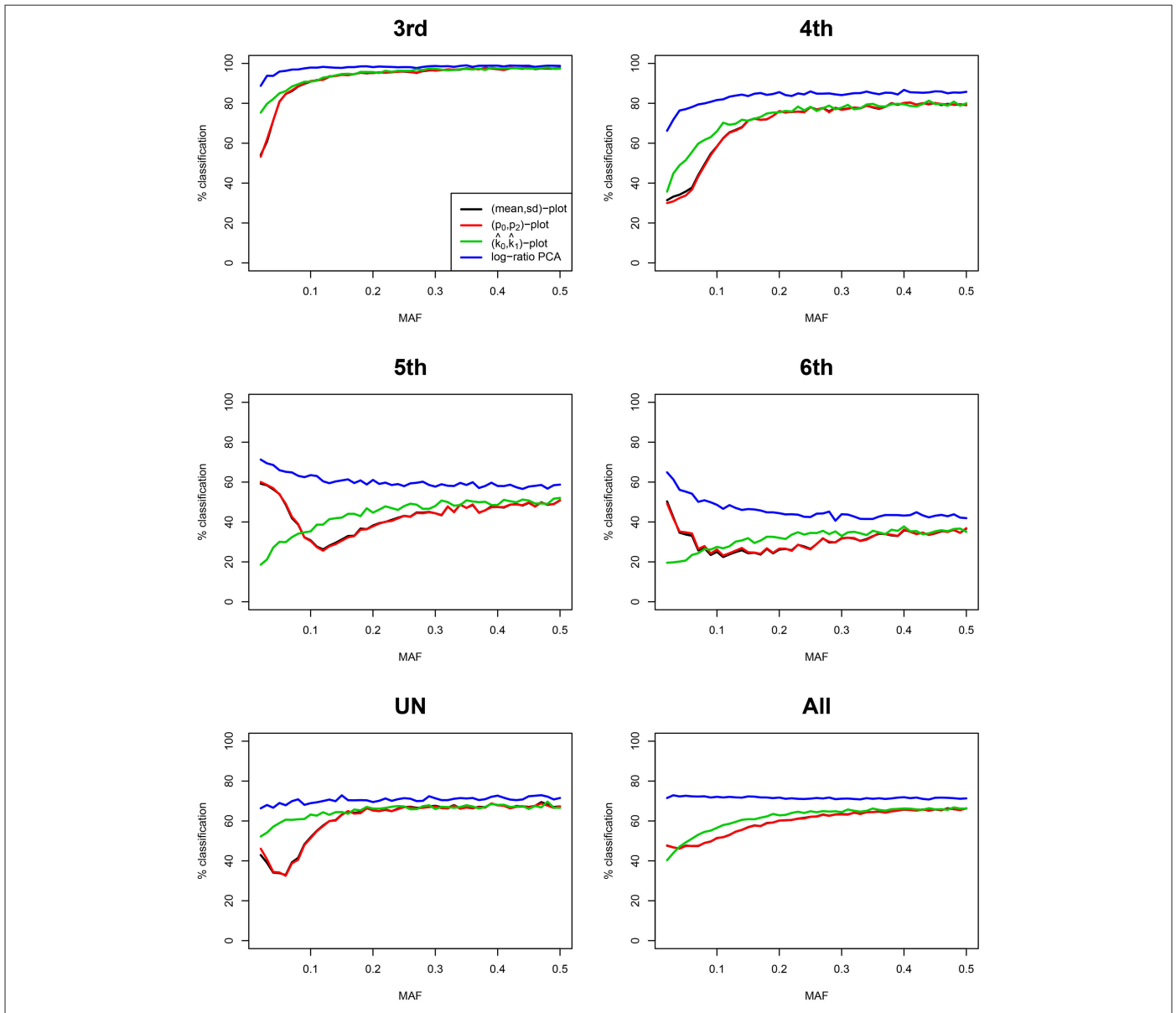

FIGURE 5 | Classification rates for different methods vs. MAF. Classification rates for the different degrees of relationship (third, fourth, fifth, sixth, UN, and All) are shown for four methods, using three principal components. Classification rate profiles for the (m, s) plot and the (p_0, p_2) plot virtually coincide. The last panel All refers to the classification rate for third through UN relationships jointly. Rates are shown, using 5,000 SNPs, as a function of the MAF, and were obtained by linear discriminant analysis. 100 pairs of each type of relationship (UN, sixth, fifth, fourth, and third) were generated assuming Hardy-Weinberg equilibrium.

degree pairs and one fifth degree pair reported by the latter authors, though we found two additional fourth degree pairs, and several fifth degree pairs, which are not confirmed by Huff et al. (2011).

3.3.2. The GCAT Sample

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 of 5,075 GCAT Spanish participants from Caucasian origin using 736,223 SNPs that passed quality control (Galván-Femenía et al., 2018). Inferred relatives of first and second degree were confirmed by the BST public agency, for pairs sharing one surname (PO, second 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

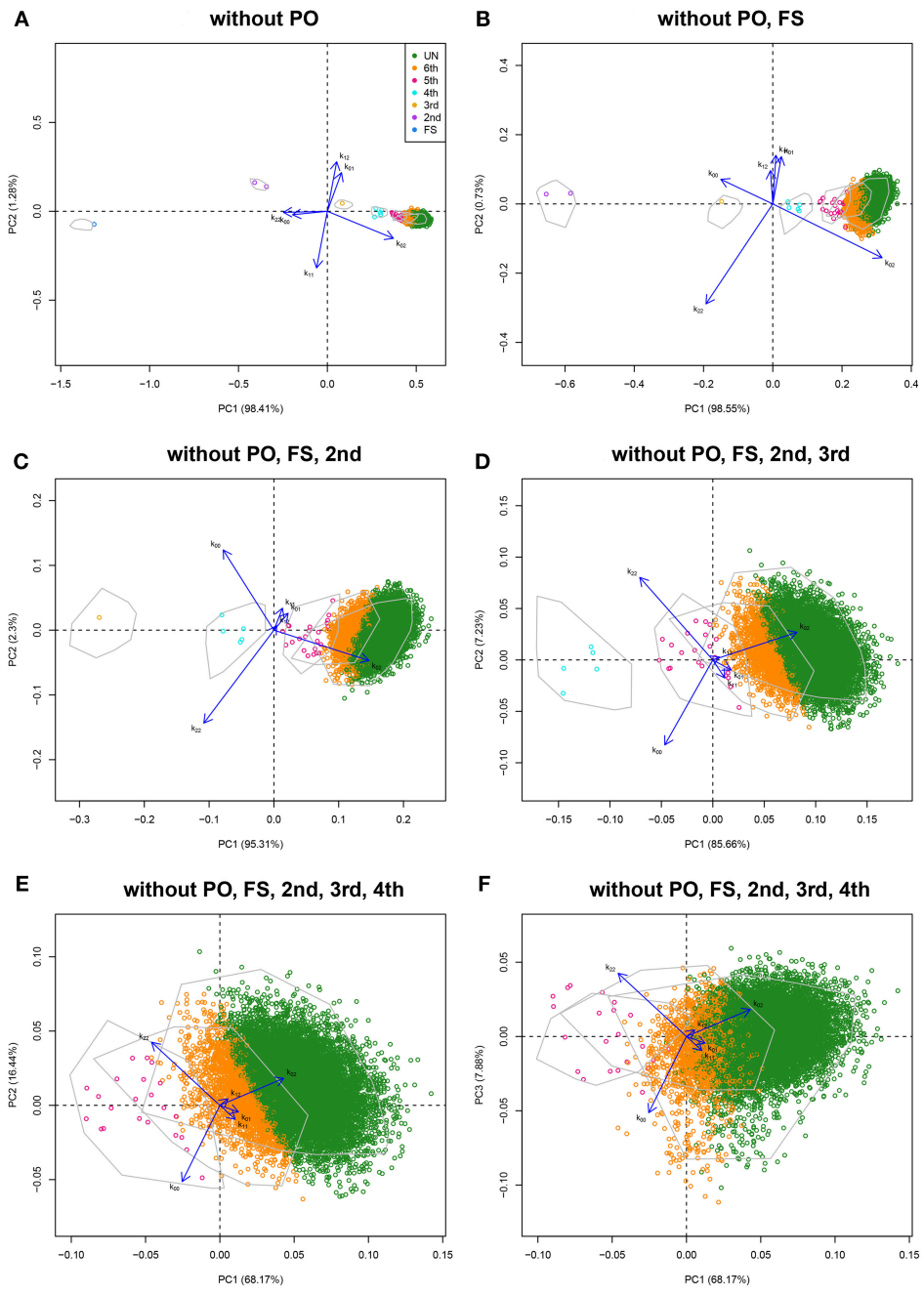

FIGURE 6 | Log-ratio PCA biplots for the CEU sample obtained by peeling and zooming. **(A)** log-ratio PCA biplot, PO pairs excluded. **(B)** PO and FS pairs excluded; **(C)** PO, FS, and AV pairs excluded; **(D)** PO, FS, AV, and third degree pairs excluded; **(E)** PO, FS, AV, third and fourth degree pairs excluded (PC1 vs. PC2); **(F)** PO, FS, AV, third and fourth degree pairs excluded (PC1 vs. PC3). Convex hulls delimit the region of the pairs obtained by simulation.

for relatedness analysis. PO and MZ pairs potentially having structural zeros were filtered with $k_{02} < 0.005$. Log-ratio PCA biplots representing over twelve million pairs, combined with the classification of the individuals by LDA, and using the peel and zoom procedure, are shown in **Figure 7**. This analysis shows the different relationships have in general, a larger variability than expected according to the simulated pairs. The FS cluster has a particular high variability, with pairs apparently less related than FS, and pairs stronger related than FS, in comparison with the FS hull. One apparent FS pairs is actually classified as second degree (**Figure 7A**). This fusion of FS and second degree pairs suggested us that three-quarter siblings might exist in the database and we therefore re-analyzed the data using a training set that included three-quarter siblings. Three-quarter siblings (3/4S) share more IBD alleles than second degree pairs but fewer than FS. 3/4S have one common parent, while their unshared parents can be FS or PO (see **Figure S2**). Three-quarter siblings have IBD probabilities $k_0 = 3/8, k_1 = 1/2$, and $k_2 = 1/8$, such that their

TABLE 4 | Predicted relationships of third (3rd), fourth (4th), and fifth (5th) degree pairs of the CEU sample.

Pair	ID1	Sex	ID2	Sex	Pem.	Kyr.	Ste.	Huf.	Predicted	Posterior probabilities					$\hat{k}_0$	$\hat{k}_1$	$\hat{k}_2$	$\hat{\phi}$
										3rd	4th	5th	6th	UN				
1	NA06997	F	NA12801	M	–	FC	FC	–	3rd	1.000	0.000	0.000	0.000	0.000	0.724	0.276	0.000	0.069
2	NA06993	M	NA07022	M	–	4th	–	4th	4th	0.000	1.000	0.000	0.000	0.000	0.870	0.127	0.003	0.033
3	NA06993	M	NA07056	F	–	4th	–	4th	4th	0.000	1.000	0.000	0.000	0.000	0.870	0.130	0.000	0.033
4	NA07031	F	NA12043	M	–	4th	–	–	4th	0.000	1.000	0.000	0.000	0.000	0.845	0.155	0.000	0.039
5	NA12155	M	NA12264	M	–	4th	–	4th	4th	0.000	1.000	0.000	0.000	0.000	0.867	0.133	0.000	0.033
6	NA12760	M	NA12830	F	–	FC	–	–	4th	0.000	1.000	0.000	0.000	0.000	0.855	0.133	0.012	0.039
7	NA06989	F	NA10831	F	–	–	–	–	5th	0.000	0.000	**0.965**	0.035	0.000	0.966	0.026	0.008	0.011
8	NA06989	F	NA12155	M	–	4th	–	–	5th	0.000	0.028	**0.972**	0.000	0.000	0.912	0.088	0.000	0.022
9	NA06991	F	NA07022	M	–	4th	–	–	5th	0.000	0.016	**0.983**	0.000	0.000	0.898	0.102	0.000	0.025
10	NA06994	M	NA12878	F	–	–	–	–	5th	0.000	0.000	0.814	0.185	0.000	0.951	0.041	0.008	0.014
11	NA06994	M	NA12892	F	–	4th	–	5th	5th	0.000	0.000	**0.997**	0.002	0.000	0.925	0.075	0.000	0.019
12	NA07014	F	NA12043	M	–	4th	–	–	5th	0.000	0.000	**0.966**	0.034	0.000	0.950	0.043	0.008	0.015
13	NA07029	M	NA12892	F	–	–	–	–	5th	0.000	0.000	0.563	0.437	0.000	0.942	0.056	0.002	0.015
14	NA07031	F	NA12752	M	–	–	–	–	5th	0.000	0.000	**0.980**	0.020	0.000	0.942	0.053	0.005	0.016
15	NA07031	F	NA12761	F	–	4th	–	–	5th	0.000	0.000	**0.991**	0.009	0.000	0.890	0.110	0.000	0.028
16	NA07055	F	NA10852	F	–	–	–	–	5th	0.000	0.000	0.853	0.147	0.000	0.959	0.040	0.001	0.011
17	NA10830	M	NA12842	M	–	–	–	–	5th	0.000	0.000	0.826	0.174	0.000	0.940	0.060	0.000	0.015
18	NA10852	F	NA10853	M	–	–	–	–	5th	0.000	0.000	0.731	0.269	0.000	0.964	0.033	0.003	0.010
19	NA10852	F	NA11843	M	–	–	–	–	5th	0.000	0.000	0.575	0.425	0.000	0.978	0.019	0.003	0.006
20	NA10863	F	NA12155	M	–	4th	–	–	5th	0.000	0.000	**0.959**	0.041	0.000	0.941	0.054	0.005	0.016
21	NA11843	M	NA11994	M	–	–	–	–	5th	0.000	0.000	0.781	0.219	0.000	0.945	0.055	0.000	0.014
22	NA11992	M	NA12778	F	–	–	–	–	5th	0.000	0.000	0.682	0.318	0.000	0.951	0.050	0.000	0.012
23	NA12752	M	NA12830	F	–	4th	–	–	5th	0.000	0.000	**0.997**	0.003	0.000	0.894	0.106	0.000	0.026
24	NA12760	M	NA12818	F	–	4th	–	–	5th	0.000	0.000	**0.998**	0.002	0.000	0.926	0.074	0.000	0.019
25	NA10831	F	NA12264	M	–	4th	–	–	6th	0.000	0.000	0.094	0.896	0.010	0.963	0.036	0.001	0.010
26	NA11931	F	NA12748	M	–	4th	–	–	6th	0.000	0.000	0.467	0.532	0.001	0.927	0.067	0.006	0.020
27	NA12752	M	NA12818	F	–	4th	–	–	6th	0.000	0.000	0.026	0.946	0.029	0.977	0.022	0.001	0.006

Third (3rd) and fourth (4th) degree pairs of the CEU sample of the 1000G project as reported by Kyriazopoulou-Panagiotopoulou et al. (2011) and additional detected fifth (5th) degree pairs. Posterior probabilities according to log-ratio PCA combined with LDA. Coding and abbreviations used: sex M = male, F = female; a hyphen (–) indicates the corresponding pair is not annotated or regarded unknown by the corresponding authors; FC, first cousin; Pem., Pemberton et al. (2010); Kyr., Kyriazopoulou-Panagiotopoulou et al. (2011); Ste., Stevens et al. (2012); Huf., Huff et al. (2011).

kinship coefficient is $\phi = 3/16$, below the value $\phi = 1/4$ of full siblings. In the re-analysis in **Figure 7B**, we found 63 FS pairs, 12 2nd pairs, and eight pairs were indeed classified as three-quarter siblings with large posterior probability (see **Table 5**). Two of these pairs (67, 71) had their kinship coefficient very close to the expected value of $\phi = 3/16$. Because Spanish people have both paternal and maternal surnames, three-quarter siblings share both surnames just as siblings do. The pairs classified as 3/4 siblings shared indeed both surnames, confirming these pairs are actually not second degree. Peeling siblings and three-quarter siblings reveals apparent second degree pairs more clearly (**Figure 7C**). Tentatively peeling second degree pairs brings the third degree pairs in focus (**Figure 7D**), and in this analysis we find 174 third, 66 fourth, 31 fifth, and 3,517 sixth degree pairs. Further peeling is difficult as the different clusters increasingly merge. In log-ratio PCA the clusters representing the different relationships have more elliptical shapes that separate better. Note that the number of pairs classified as sixth degree decreases as the lower degree relationships are peeled in the analysis.

For all simulated and empirical data sets studied above, the first principal component in the log-ratio PCA's is seen to strongly correlate with the kinship coefficient. The corresponding scatterplots and correlation coefficients are shown in **Figure S3**. The first principal component is clearly interpretable as a relatedness index. In **Figures 6A**, **7A** (without PO), the biplot vectors show that the first component separates homogeneous homozygote pairs (AA & AA; BB & BB) from heterogeneous homozygote pairs (AA & BB). The second principal component separates double heterzygote pairs from single heterozygote pairs. When FS pairs are removed, the second principal component changes, and reflects a contrast between pairs with heterozygotes and without heterozygotes.

4. DISCUSSION

We have developed a log-ratio PCA based procedure that can be used for uncovering cryptic relatedness in homogeneous populations. Simulations show the procedure has a better classification rate than the classical IBS and IBD based

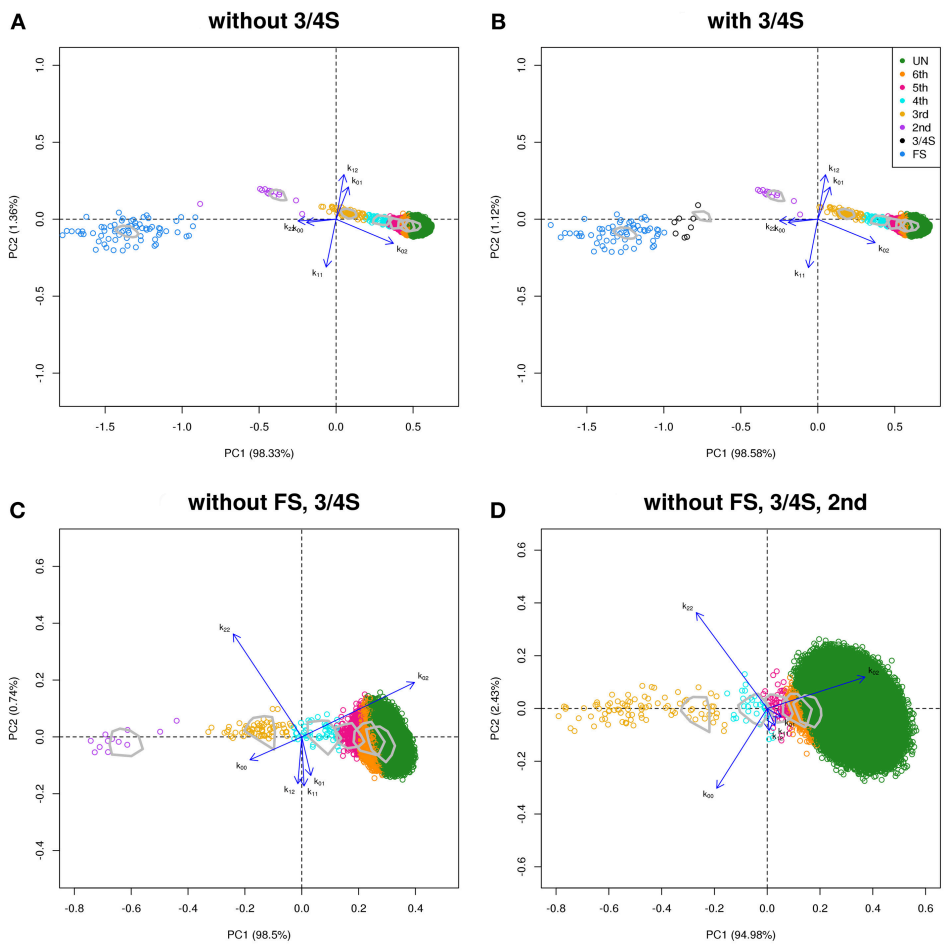

FIGURE 7 | Log-ratio PCA biplot of GCAT sample obtained by peeling and zooming. **(A)** log-ratio PCA biplot, PO and 3/4S pairs excluded. **(B)** 3/4S pairs included; **(C)** FS and 3/4S pairs excluded; **(D)** FS, 3/4S, and second degree pairs excluded. Convex hulls delimit the region of the pairs obtained by simulation.

approaches. The log-ratio PCA approach exploits the compositional nature of genotype sharing counts over variants, and can potentially use five dimensions for analysis, whereas the classical approaches collapse the data in two dimensions. The analysis of the CEU sample has led to the identification of a set of hitherto unreported pairs for which a fifth degree relationship is highly plausible (**Table 4**). Our conclusion is that log-ratio PCA, combined with LDA, increases the resolution of relationship discrimination. The classification rate for 6th degree pairs can still be improved if more than 35,000 independent MAF 0.50 variants would be used (see **Figure 4**). The $(p_0, p_2), (m, s)$, and $(\hat{k}_0, \hat{k}_1)$ scatterplots display estimates in a constrained space (Galván-Femenía et al., 2017), where Euclidean distances between points cannot be safely interpreted. This is particularly true for the higher degree relationships that merge toward the vertex of the triangular region inside the scatterplot. Log-ratio PCA, besides using more dimensions, frees the data of the unit sum constraint, and clearly enhances the discrimination of the higher degree relationships. We have compared our log-ratio based procedure with some basic procedures used in relatedness research. Its performance could be further explored in a more extensive comparison that includes IBD-segment based methods, such as the comprehensive study reported by Ramstetter et al. (2017).

The analysis of the GCAT samples shows, for almost all relationship categories, larger variability in the relationship clusters 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.

The simulated reference data sets were obtained by resampling genetic variants independently, and this does not take linkage disequilibrium (LD) and recombination into account (Hill and Weir, 2011). If the genotype data is phased, a biologically more realistic simulated data set can be obtained by sampling

TABLE 5 | Predicted FS and 3/4S relationships of the GCAT sample.

						Posterior probabilities											
Pair	**ID1**	**Sex**	**ID2**	**Sex**	**Predicted**	**FS**	**3/4S**	**2nd**	**3rd**	**4th**	**5th**	**6th**	**UN**	$\hat{k}_0$	$\hat{k}_1$	$\hat{k}_2$	$\hat{\phi}$
1	REL_00339	F	REL_02473	F	FS	1	0	0	0	0	0	0	0	0.254	0.479	0.266	0.253
2	REL_04741	F	REL_02513	F	FS	1	0	0	0	0	0	0	0	0.187	0.518	0.295	0.277
3	REL_00601	M	REL_02989	F	FS	1	0	0	0	0	0	0	0	0.190	0.508	0.303	0.278
4	REL_02339	M	REL_02391	M	FS	1	0	0	0	0	0	0	0	0.267	0.442	0.290	0.256
5	REL_03977	M	REL_01080	M	FS	1	0	0	0	0	0	0	0	0.222	0.538	0.240	0.255
6	REL_03220	F	REL_04615	F	FS	1	0	0	0	0	0	0	0	0.311	0.460	0.229	0.230
7	REL_04475	F	REL_04218	M	FS	1	0	0	0	0	0	0	0	0.248	0.514	0.237	0.247
8	REL_01150	F	REL_04384	F	FS	1	0	0	0	0	0	0	0	0.258	0.490	0.253	0.249
9	REL_01285	M	REL_03761	F	FS	1	0	0	0	0	0	0	0	0.237	0.496	0.267	0.257
10	REL_04693	F	REL_00797	F	FS	1	0	0	0	0	0	0	0	0.310	0.471	0.220	0.228
11	REL_00383	F	REL_03293	M	FS	1	0	0	0	0	0	0	0	0.254	0.530	0.216	0.241
12	REL_03212	M	REL_02516	F	FS	1	0	0	0	0	0	0	0	0.275	0.526	0.199	0.231
13	REL_00282	F	REL_04918	F	FS	1	0	0	0	0	0	0	0	0.247	0.440	0.313	0.267
14	REL_04616	F	REL_02777	F	FS	1	0	0	0	0	0	0	0	0.279	0.471	0.250	0.243
15	REL_00792	F	REL_00954	M	FS	1	0	0	0	0	0	0	0	0.262	0.509	0.229	0.242
16	REL_03627	F	REL_03315	F	FS	1	0	0	0	0	0	0	0	0.148	0.549	0.302	0.288
17	REL_00872	F	REL_01784	F	FS	1	0	0	0	0	0	0	0	0.252	0.528	0.221	0.242
18	REL_03442	F	REL_04510	F	FS	1	0	0	0	0	0	0	0	0.216	0.512	0.272	0.264
19	REL_01924	F	REL_00727	M	FS	1	0	0	0	0	0	0	0	0.236	0.449	0.315	0.270
20	REL_04704	F	REL_00804	M	FS	1	0	0	0	0	0	0	0	0.168	0.523	0.308	0.285
21	REL_04494	M	REL_00931	M	FS	1	0	0	0	0	0	0	0	0.280	0.492	0.228	0.237
22	REL_04439	F	REL_01640	F	FS	1	0	0	0	0	0	0	0	0.264	0.430	0.306	0.260
23	REL_00504	M	REL_04718	F	FS	1	0	0	0	0	0	0	0	0.243	0.505	0.252	0.252
24	REL_01624	F	REL_00750	F	FS	1	0	0	0	0	0	0	0	0.191	0.508	0.301	0.278
25	REL_01524	F	REL_03272	F	FS	1	0	0	0	0	0	0	0	0.232	0.511	0.257	0.256
26	REL_00769	M	REL_04746	F	FS	1	0	0	0	0	0	0	0	0.225	0.566	0.208	0.246
27	REL_01654	M	REL_03485	M	FS	1	0	0	0	0	0	0	0	0.282	0.432	0.285	0.251
28	REL_01564	F	REL_03827	F	FS	1	0	0	0	0	0	0	0	0.316	0.427	0.258	0.236
29	REL_03944	M	REL_03475	F	FS	1	0	0	0	0	0	0	0	0.231	0.542	0.227	0.249
30	REL_01888	M	REL_04360	M	FS	1	0	0	0	0	0	0	0	0.247	0.543	0.210	0.241
31	REL_00824	F	REL_00213	F	FS	1	0	0	0	0	0	0	0	0.221	0.446	0.332	0.278
32	REL_03838	F	REL_02496	F	FS	1	0	0	0	0	0	0	0	0.310	0.446	0.245	0.234
33	REL_00122	M	REL_01902	F	FS	1	0	0	0	0	0	0	0	0.286	0.494	0.220	0.233
34	REL_04592	F	REL_04600	F	FS	1	0	0	0	0	0	0	0	0.305	0.485	0.211	0.227
35	REL_00284	M	REL_02444	F	FS	1	0	0	0	0	0	0	0	0.278	0.511	0.211	0.233
36	REL_03395	F	REL_02694	F	FS	1	0	0	0	0	0	0	0	0.224	0.522	0.254	0.257
37	REL_02718	M	REL_02913	M	FS	1	0	0	0	0	0	0	0	0.218	0.479	0.303	0.271
38	REL_00968	M	REL_01577	F	FS	1	0	0	0	0	0	0	0	0.257	0.451	0.292	0.259
39	REL_01502	M	REL_03665	M	FS	1	0	0	0	0	0	0	0	0.312	0.477	0.211	0.225
40	REL_03904	F	REL_04994	F	FS	1	0	0	0	0	0	0	0	0.250	0.502	0.248	0.249
41	REL_02208	F	REL_03486	F	FS	1	0	0	0	0	0	0	0	0.231	0.460	0.310	0.270
42	REL_02208	F	REL_01630	F	FS	1	0	0	0	0	0	0	0	0.177	0.516	0.307	0.283
43	REL_03486	F	REL_01630	F	FS	1	0	0	0	0	0	0	0	0.170	0.502	0.327	0.289
44	REL_00340	F	REL_04294	F	FS	1	0	0	0	0	0	0	0	0.210	0.525	0.265	0.264
45	REL_02899	M	REL_01707	F	FS	1	0	0	0	0	0	0	0	0.285	0.454	0.261	0.244
46	REL_03001	F	REL_04111	F	FS	1	0	0	0	0	0	0	0	0.230	0.481	0.289	0.265
47	REL_00634	M	REL_03507	M	FS	1	0	0	0	0	0	0	0	0.203	0.508	0.289	0.272
48	REL_02905	F	REL_02575	F	FS	1	0	0	0	0	0	0	0	0.252	0.517	0.231	0.245
49	REL_01016	M	REL_00887	M	FS	1	0	0	0	0	0	0	0	0.243	0.496	0.260	0.254
50	REL_03151	M	REL_02204	F	FS	1	0	0	0	0	0	0	0	0.235	0.503	0.263	0.257

(Continued)

TABLE 5 | Continued

Pair	ID1	Sex	ID2	Sex	Predicted	Posterior probabilities								$\hat{k}_0$	$\hat{k}_1$	$\hat{k}_2$	$\hat{\phi}$
						FS	3/4S	2nd	3rd	4th	5th	6th	UN				
51	REL_04466	F	REL_02680	F	FS	1	0	0	0	0	0	0	0	0.313	0.427	0.260	0.237
52	REL_03607	M	REL_00319	F	FS	1	0	0	0	0	0	0	0	0.299	0.491	0.210	0.228
53	REL_01083	F	REL_01704	F	FS	1	0	0	0	0	0	0	0	0.182	0.567	0.251	0.267
54	REL_04427	F	REL_02635	F	FS	1	0	0	0	0	0	0	0	0.264	0.545	0.191	0.232
55	REL_01546	M	REL_03566	F	FS	1	0	0	0	0	0	0	0	0.212	0.525	0.263	0.263
56	REL_01450	M	REL_01960	M	FS	1	0	0	0	0	0	0	0	0.259	0.514	0.227	0.242
57	REL_03310	M	REL_03659	F	FS	1	0	0	0	0	0	0	0	0.259	0.559	0.182	0.231
58	REL_03880	M	REL_04789	F	FS	1	0	0	0	0	0	0	0	0.271	0.503	0.226	0.239
59	REL_01264	M	REL_04751	F	FS	1	0	0	0	0	0	0	0	0.183	0.518	0.299	0.279
60	REL_04529	F	REL_04492	F	FS	1	0	0	0	0	0	0	0	0.279	0.498	0.223	0.236
61	REL_03388	F	REL_02608	F	FS	1	0	0	0	0	0	0	0	0.216	0.497	0.287	0.268
62	REL_00009	F	REL_02335	F	FS	1	0	0	0	0	0	0	0	0.233	0.548	0.218	0.246
63	REL_04405	M	REL_03949	M	FS	1	0	0	0	0	0	0	0	0.262	0.523	0.215	0.238
64	REL_02752	F	REL_04859	F	3/4S	0	1	0	0	0	0	0	0	0.342	0.457	0.201	0.215
65	REL_01344	M	REL_02408	F	3/4S	0	1	0	0	0	0	0	0	0.361	0.439	0.200	0.210
66	REL_00083	M	REL_02333	M	3/4S	0	1	0	0	0	0	0	0	0.326	0.520	0.154	0.207
67	REL_03803	F	REL_02343	M	3/4S	0	1	0	0	0	0	0	0	0.349	0.510	0.140	0.198
68	REL_03924	M	REL_03023	F	3/4S	0	1	0	0	0	0	0	0	0.366	0.464	0.170	0.201
69	REL_04189	M	REL_00775	M	3/4S	0	1	0	0	0	0	0	0	0.367	0.427	0.206	0.210
70	REL_03150	F	REL_01804	F	3/4S	0	1	0	0	0	0	0	0	0.323	0.505	0.172	0.212
71	REL_03969	M	REL_00271	M	3/4S	0	1	0	0	0	0	0	0	0.342	0.560	0.098	0.189

FS and 3/4S pairs of the GCAT sample. Predicted relationships and posterior probabilities according to a log-ratio PCA combined with LDA. Coding and abbreviations used: sex M, male; F, female; $\hat{\phi}$, estimated kinship coefficient.

haplotypes. We have avoided this issue by LD pruning the data base prior to resampling, so removing tightly correlated markers. The reference data set is therefore constructed on the basis of a subset of variants that can expected to be approximately independent. This subset is then used as the basis for relationship estimation. This procedure has the advantage that it avoids potential additional uncertainty generated by using a phasing algorithm. However, the proposed procedure may be improved in the future by accounting for haplotype structure and recombination. The pruning threshold used in our method (0.20) is a compromise between precision and satisfying the independence assumption. A larger value will admit more variants and can increase the resolution, but due to correlation between variants it will invalidate the independence assumption used to generate the reference set of related pairs.

The proposed method for classifying pairs combining log-ratio PCA 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 in the database. We therefore suggest the method to be used for large samples with thousands of individuals, 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. Note that with only one hundred simulated pairs of each relationship, we build a classifier that can be used to classify millions of pairs. Our method is computationally feasible for over 5,000 individuals and 26,000 variants like in the GCAT sample. Most of the computation time is spent on the projection of the empirical pairs onto the reference structure, and these computations could easily be parallelized. Many public repositories of genomic data are currently available, but without recruitment and relatedness information, and for which the relatedness techniques discussed in this paper could be usefully applied.

The log-ratio transformation in Equation (1) does not admit zeros for the genotype sharing counts. In theory MZ pairs have $k_{10} = k_{20} = k_{21} = 0$, and PO pairs have $k_{20} = 0$. In practice, due to the summing over large numbers of variants, zeros are almost never observed as a consequence of some genotyping error and incidental mutations. If a few zero counts are observed, a replacement by 1 or 0.5 can eventually be used in order to proceed with the analysis. If there is a substantial amount of zeros, a ratio-preserving multiplicative replacement (Fry et al., 2000; Martín-Fernández et al., 2003) or a Bayesian procedure (Martin-Fernandez et al., 2015) are recommended. The zero problem is well-known in compositional data analysis, and a distinction is usually drawn between structural and rounding zeros (Martín-Fernández et al., 2003, 2011). In principle, MZ and PO pairs have structural zeros. However, MZ and PO pairs are the most easily detected relationships, and are easily dealt with separately, prior to applying the log-ratio transformation to the data. Exclusion

of the relationships up to the second or third degree is in fact desirable if possible, as it will allow the study of the more remote relationships at higher resolution.

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 decision is mostly made graphically by inspecting a (p_0, p_2) plot in IBS studies, or a $(\hat{k}_0, \hat{k}_1)$ 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 a resampling of the observed data.

Applications of IBD based methods typically employ three Cotterman coefficients that are constrained to sum one, and therefore represent relatedness in only two dimensions. However, IBD based methods can estimate additional Jacquard coefficients (Milligan, 2003) and thus potentially exploit more dimensions than is usually done in practice.

The current paper is focused on homogeneous populations. If population substructure exists, then log-ratio PCA can be expected to separate the different populations in its biplot. Methods that address substructure (distant relatedness) and family relationships (recent relatedness) jointly have been developed (Manichaikul et al., 2010; Conomos et al., 2015). Population substructure can be accounted for by using only variants with low weights on the first components for a relatedness analysis, as is done in the UK Biobank project (Bycroft et al., 2018), as the first components mostly capture substructure. In future work, the usefulness of the log-ratio PCA approach for the joint study of remote and recent relatedness could be further explored.

SOFTWARE AVAILABILITY

R code (R Core Team, 2014) implementing the logratio kinship biplot proposed in this paper is available online at github.com/ivangalvan/LR-kinbiplot.

ETHICS STATEMENT

Our study does use data from human subjects, but concerns data that is available in public repositories.

AUTHOR CONTRIBUTIONS

JG and IG contributed equally to this paper, where JG conceived the methodology and wrote the paper. IG developed computer programs, ran simulations, and performed data analysis. RdC supervised GCAT data analysis. RdC and CBV proof-read the manuscript. All authors contributed to the improvement of the paper.

FUNDING

This work was partially supported by grants MTM2015-65016-C2-2-R (JG), MTM2015-65016-C2-1-R (IG and CBV) and ADE 10/00026 (RdC) (MINECO/FEDER) of the Spanish Ministry of Economy and Competitiveness and European Regional Development Fund, by grants SGR1269 and 2017 SGR529 (RdC) of the Generalitat de Catalunya, by grant R01 GM075091 (JG) from the United States National Institutes of Health, and by the Ramon y Cajal action RYC-2011-07822 (RdC).

ACKNOWLEDGMENTS

We are grateful for the publicly available data sets of the 1,000 Genomes project, available at www.internationalgenome.org. This study makes use of data generated by the GCAT Genomes for Life Cohort study of the Genomes of Catalonia, IGTP. A full list of the investigators who contributed to the generation of the data is available from www.genomesforlife.com. IGTP is part of the CERCA Program of the Generalitat de Catalunya.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00341/full#supplementary-material

REFERENCES

Abecasis, G., Cherny, S., Cookson, W., and Cardon, L. (2001). GRR: graphical representation of relationship errors. *Bioinformatics* 17, 742–743. doi: 10.1093/bioinformatics/17.8.742

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika* 70, 57–65. doi: 10.1093/biomet/70.1.57

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Caldwell, NJ: The Blackburn Press.

Anandan, A., Anumalla, M., Pradhan, S., and Ali, J. (2016). Population structure, diversity and trait association analysis in rice (*Oryza sativa* L.) germplasm for early seedling vigor (esv) using trait linked ssr markers. *PLoS ONE* 11:e0152406. doi: 10.1371/journal.pone.0152406

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L., Sharp, K., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209. doi: 10.1038/s41586-018-0579-z

Conomos, M., Miller, M., and Thornton, T. (2015). Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* 39, 276–293. doi: 10.1002/gepi.21896

Cotterman, C. (1941). Relative and human genetic analysis. *Sci. Monthly* 53, 227–234.

Epstein, M., Duren, W., and Boehnke, M. (2000). Improved inference of relationship for pairs of individuals. *Am. J. Hum. Genet.* 67, 1219–1231. doi: 10.1016/S0002-9297(07)62952-8

Fry, J., Fry, T., and Mclaren, K. (2000). Compositional data analysis and zeros in micro data. *Appl. Econ.* 32, 953–959. doi: 10.1080/000368400322002

Galván-Femenía, I., Graffelman, J., and Barceló Vidal, C. (2017). Graphics for relatedness research. *Mol. Ecol. Resour.* 17, 1271–1282. doi: 10.1111/1755-0998.12674

Galván-Femenía, I., Obón-Santacana, M., Piñeyro, D., Guindo-Martinez, M., Duran, X., Carreras, A., et al. (2018). Multitrait genome association analysis identifies new susceptibility genes for human

anthropometric variation in the GCAT cohort. *J. Med. Genet.* 55, 765–778. doi: 10.1136/jmedgenet-2018-105437

Gower, J., Gardner Lubbe, E., and Le Roux, N. (2011). *Understanding Biplots*. Chichester: John Wiley.

Graffelman, J., and Aluja-Banet, T. (2003). Optimal representation of supplementary variables in biplots from principal component analysis and correspondence analysis. *Biometr. J.* 45, 491–509. doi: 10.1002/bimj.200390027

Graffelman, J., and Moreno, V. (2013). The mid *p*-value in exact tests for Hardy-Weinberg equilibrium. *Stat. Appl. Genet. Mol. Biol.* 12, 433–448. doi: 10.1515/sagmb-2012-0039

Hill, W., and Weir, B. (2011). Variation in actual relationship as a consequence of mendelian sampling and linkage. *Genet. Res.* 93, 47–64. doi: 10.1017/S0016672310000480

Huff, C., Witherspoon, D., Simonson, T., Xing, J., Watkins, W., Zhang, Y., et al. (2011). Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Res.* 21, 768–774. doi: 10.1101/gr.115972.110

Jakobsson, M., Scholz, S., Scheet, P., Gibbs, J., VanLiere, J., Fung, H., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451, 998–1003. doi: 10.1038/nature06742

Johnson, R. A., and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis, 5th Edn*. Upper Saddle River, NJ: Prentice Hall.

Kyriazopoulou-Panagiotopoulou, S., Kashef-Haghighi, D., Aerni, S., Sundquist, A., Bercovici, S., and Batzoglou, S. (2011). Reconstruction of genealogical relationships with applications to Phase III of HapMap. *Bioinformatics* 27, i333–i341. doi: 10.1093/bioinformatics/btr243

Manichaikul, A., Mychaleckyj, J., Rich, S., Daly, K., Sale, M., and Chen, W. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873. doi: 10.1093/bioinformatics/btq559

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. London: Academic Press.

Martín-Fernández, J., Barceló-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math. Geol.* 35, 253–278. doi: 10.1023/A:1023866030544

Martin-Fernandez, J., Hron, K., Templ, M., Filzmoser, P., and Palarea-Albaladejo, J. (2015). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Stat. Model.* 15, 134–158. doi: 10.1177/1471082X14535524

Martín-Fernández, J., Palarea-Albaladejo, J., and Olea, R. (2011). "Dealing with zeros," in *Compositional Data Analysis: Theory and Applications*, eds V. Pawlowsky-Glahn and A. Buccianti (Chichester: John Wiley & Sons), 43–58.

McPeek, M., and Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.* 66, 1076–1094. doi: 10.1086/302800

Milligan, B. (2003). Maximum-likelihood estimation of relatedness. *Genetics* 163, 1153–1167.

Nembot-Simo, A., Graham, J., and McNeney, B. (2013). CrypticIBD check: an R package for checking cryptic relatedness in nominally unrelated individuals. *Source Code Biol. Med.* 8:5. doi: 10.1186/1751-0473-8-5

Obón-Santacana, M., Vilardell, M., Carreras, A., Duran, X., Velasco, J., Galván-Femenía, I., et al. (2018). GCAT|Genomes for Life: a prospective cohort study of the genomes of catalonia. *BMJ Open* 8:e018324. doi: 10.1136/bmjopen-2017-018324

Oliehoek, P., Windig, J., van Arendonk, J., and Bijma, P. (2006). Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics* 173, 483–496. doi: 10.1534/genetics.105.049940

Pandit, E., Tasleem, S., Barik, S., Mohanty, D., Nayak, D., Mohanty, S., et al. (2017). Genome-wide association mapping reveals multiple qtls governing tolerance response for seedling stage chilling stress in indica rice. *Front. Plant Sci.* 8:552. doi: 10.3389/fpls.2017.00552

Pawlowsky-Glahn, V., Egozcue, J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Chichester: John Wiley & Sons.

Pemberton, T., DeGiorgio, M., and Rosenberg, N. (2013). Population structure in a comprehensive genomic data set on human microsatellite variation. *Genes Genomes Genet.* 3, 891–907. doi: 10.1534/g3.113.005728

Pemberton, T. J., Wang, C., Li, J. Z., and Rosenberg, N. A. (2010). Inference of unexpected genetic relatedness among individuals in hapmap phase iii. *Am. J. Hum. Genet.* 87, 457–464. doi: 10.1016/j.ajhg.2010.08.014

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., et al. (2007). Plink: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ramstetter, M., Dyer, T., Lehman, D., Curran, J., Duggirala, R., Blangero, J., et al. (2017). Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics* 207, 75–82. doi: 10.1534/genetics.117.1122

Rosenberg, N. A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity cell line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Ann. Hum. Genet.* 70, 841–847. doi: 10.1111/j.1469-1809.2006.00285.x

Sabatti, C., Service, S., Hartikainen, A., Pouta, A., Ripatti, S., Brodsky, J., et al. (2009). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* 41, 35–46. doi: 10.1038/ng.271

Sharma, S., MacKenzie, K., McLean, K., Dale, F., Daniels, S., and Bryan, G. (2018). Linkage disequilibrium and evaluation of genome-wide association mapping models in tetraploid potato. *G3 (Bethesda)* 8, 3185–3202. doi: 10.1534/g3.118.200377

Stevens, E., Baugher, J., Shirley, M., Frelin, L., and Pevsner, J. (2012). Unexpected relationships and inbreeding in HapMap Phase III populations. *PLoS ONE* 7:e49575. doi: 10.1371/journal.pone.0049575

Stevens, E., Heckenberg, G., Roberson, E., Baugher, J., Downey, T., and Pevsner, J. (2011). Inference of relationships in population data using indentity-by-descent and identity-by-state. *PLoS Genet.* 7:e1002287. doi: 10.1371/journal.pgen.1002287

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Thompson, E. (1975). The estimation of pairwise relationships. *Ann. Hum. Genet.* 39, 173–188. doi: 10.1111/j.1469-1809.1975.tb00120.x

Thompson, E. (1991). "Estimation of relationships from genetic data," in *Handbook of Statistics*, Vol. 8, eds C. Rao and R. Chakraborty (Amsterdam: Elsevier Science), 255–269.

Voight, B., and Pritchard, J. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genet.* 1:e32. doi: 10.1371/journal.pgen.0010032

Wang, C., Szpiech, Z., Degnan, J., Jakobsson, M., Pemberton, T., Hardy, J., et al. (2010). Comparing spatial maps of human population-genetic variation using procrustes analysis. *Stat. Appl. Genet. Mol. Biol.* 9:13. doi: 10.2202/1544-6115.1493

Wang, J. (2018). Effects of sampling close relatives on some elementary population genetics analyses. *Mol. Ecol. Resour.* 18, 41–54. doi: 10.1111/1755-0998.12708

Weir, B. S., Anderson, A. D., and Hepler, A. B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* 7, 771–780. doi: 10.1038/nrg1960

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

4.3 Heredity

The third article (submitted) accomplish with the objectives Obj. 1 and Obj. 3 described in 2.1. In summary, we propose a likelihood ratio approach to distinguish three-quarter siblings (3/4S) from full-siblings and second degree relatives. We show that this approach is useful to infer 3/4S instead of plotting the IBD probabilities.

This article has been submitted to *Heredity* journal.
Submitted: May 2020.
Impact factor: 3.179 (Q2). Journal Citation Reports Ranking: 46/171 (Genetics & Heredity); 15/49 (Evolutionary Biology); 31/158 (Ecology).

A likelihood ratio approach for identifying three quarter siblings in genetic databases

Iván Galván-Femenía[1,2], Carles Barceló-i-Vidal[1], Lauro Sumoy[3], Victor Moreno[4,5,6,7], Rafael de Cid[2,#] and Jan Graffelman[8,9,#]

[1]Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona; Girona, Spain

[2]Genomes For Life - GCAT lab, Institute for Health Science Research Germans Trias i Pujol (IGTP), Can Ruti Campus, Badalona, Barcelona, Spain

[3]High Content Genomics and Bioinformatics Unit, Institute for Health Science Research Germans Trias i Pujol (IGTP), Can Ruti Campus, Badalona, Barcelona, Spain

[4]Oncology Data Analytics Program, Catalan Institute of Oncology (ICO)

[5]ONCOBELL Program, Bellvitge Biomedical Research Institute (IDIBELL)

[6]Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP)

[7]Department of Clinical Sciences, University of Barcelona, Barcelona, Spain

[8]Department of Statistics and Operations Research, Universitat Politècnica de Catalunya; Barcelona, Spain

[9]Department of Biostatistics, University of Washington; Seattle, WA, USA

[#] Corresponding authors: jan.graffelman@upc.edu; rdecid@igtp.cat

Abstract

The detection of family relationships in genetic databases is of interest in various scientific disciplines such as genetic epidemiology, population and conservation genetics and forensic science. Nowadays, screening genetic databases for related individuals forms an important aspect of standard quality control procedures. Relatedness research is usually based on an allele sharing analysis of identity by state (IBS) or identity by descent (IBD) alleles. Existing IBS/IBD methods mainly aim to identify first degree relationships (parent-offspring or full-siblings) and second degree (half-siblings, avuncular or grandparent-grandchild). Little attention has been paid to the detection of in-between first and second degree relationships such as three-quarter siblings (3/4S) who share fewer alleles than first degree relationships but more alleles than second degree relationships. With the progressively increasing sample sizes used in genetic research, it becomes more likely that such relationships are present in the database under study. In this paper we extend existing likelihood ratio methodology to accurately infer the existence of 3/4S, distinguishing them from full-siblings and second degree relatives. Our proposal accounts for linkage disequilibrium (LD) by using marker pruning, and we validate our methodology with a pedigree-based simulation study accounting for both LD and recombination. A empirical genome-wide array dataset from the GCAT Genomes for Life cohort project is used to illustrate the method.

Keywords: SNP, identity by state/descent, family relationships, likelihood ratio, three quarter siblings.

1 Introduction

The detection of related individuals in genetic databases is of great interest in various areas of genetic research. Most obviously, it is of interest in forensic studies aiming at identifying relationships between individuals such as paternity tests (Evett and Weir, 1998). In conservation genetics, careful selection of unrelated individuals for breeding programs is needed (Oliehoek et al., 2006), requiring the estimation of pairwise genetic relationships. In genome wide association studies (GWAS) that have become popular during the past two decades (Visscher et al., 2017), standard quality control filters are applied prior to genetic association analysis. The presence of cryptic relatedness violates the assumption of independent individuals in association modeling. For this reason, removing related individuals in the genetic database prior to the GWAS analysis is a common quality control step (Anderson et al., 2010).

Many methods for relatedness research are described in the literature. Most of them are based on the principle of the allele sharing. Two individuals can share 0, 1 or 2 alleles for a diploid genetic marker. These alleles can be identical by state (IBS) or identical by descent (IBD). A scatterplot of

the mean ($\bar{x}_{IBS}$) and standard deviation (s_{IBS}) of the number of IBS alleles over variants can be used to identify related pairs (Abecasis et al., 2001). Alternatively, a scatterplot of the proportions of sharing 0, 1 or 2 IBS alleles (p_0, p_1, p_2) is also often used to detect related pairs (Rosenberg, 2006). In genetic studies, the probabilities of sharing 0, 1 and 2 IBD alleles (k_0, k_1, k_2) can be estimated and used for relationship inference, since their theoretically expected values are known for the standard relationships (see Table 1). For example, parent-offspring pairs have (k_0, k_1, k_2) = (0, 1, 0) and full-siblings have (k_0, k_1, k_2) = (0.25, 0.50, 0.25). For a given pair of individuals, these probabilities can be estimated by maximum likelihood (Milligan, 2003; Thompson, 1975, 1991), by the method of moments (Purcell et al., 2007) or with robust estimators (Manichaikul et al., 2010). From these probabilities, the kinship coefficient, defined as $\phi = k_1/4 + k_2/2$, can be obtained. The kinship coefficient can be used to remove individuals with first degree (parent-offspring (PO) or full-siblings (FS)) and second degree relationships (half-siblings, avuncular or grandparent-grandchild) by retaining only pairs with $\phi < 1/16$. Additionally, third degree relationships (first cousins (FC)) can be eliminated by retaining only pairs with $\phi < 1/32$ (Anderson et al., 2010). All these methods have in common that the inference of the family relationships is based on the judgement of the analyst of the point estimates ($\hat{k}_0, \hat{k}_1, \hat{k}_2, \hat{\phi}$) or of a graphical representation (($\bar{x}_{IBS}$,s_{IBS}),(p_0, p_1, p_2) or ($\hat{k}_0, \hat{k}_1, \hat{k}_2$)) (Galván-Femenía et al., 2017).

The sample size used in genetic studies, GWAS in particular, is progressively increasing due to large human sequencing projects that involve genetic data from hundreds of thousands of individuals such as UK Biobank (Bycroft et al., 2018), gnomAD (Karczewski et al., 2019), TOPMed (Taliun et al., 2019) and DiscovEHR (Staples et al., 2018) among others. With such large databases, it becomes increasingly likely that in-between 1st and 2nd degree, and in-between 2nd and 3rd degree relationships are found. Such in-between relationships are mostly ignored in a relatedness analysis, which typically mostly focus on 1st, 2nd and 3rd degree relationships. In this paper we therefore develop a likelihood ratio approach that will allow us to identify three-quarter siblings (3/4S), a family relationship whose individuals share fewer alleles than 1st degree relationships but more alleles than 2nd degree relatives (Table 1). A 3/4S pair has one common parent, while their unshared parents have a first degree relationship (FS or PO; see Graffelman *et al.*, 2019 Fig. S2). The IBD probabilities for 3/4S are $(k_0, k_1, k_2) = (3/8, 1/2, 1/8)$ and their kinship coefficient is $\phi = 3/16$. A detailed derivation of these probabilities is shown in Appendix A. A 3/4S relationship is not very common, but is more likely to be present in GWAS studies with ever increasing sample sizes. The 3/4S relationship has received very little attention in the literature, and the aim of this paper is to develop tools that distinguish 3/4S from full-siblings and second degree relatives.

The remainder of this paper is structured as follows. Section 2 develops

a likelihood ratio approach for identifying three quarter siblings. Section 3 evaluates the likelihood ratio approach in a simulation study. Section 4 illustrates our approach with genome-wide SNP array data from the GCAT Genome for Life project cohort. Finally, we end the article with a discussion of the proposed methodology.

2 Methods and materials

2.1 Overview of the likelihood of a relationship

A detailed derivation of the likelihood of having a given relationship is given by Wagner et al. (2006). Briefly, let n be the number of individuals in a non-inbred homogeneous population and assuming absence of population structure. We consider bi-allelic genetic variants with alleles A and B having allele frequencies p and q respectively. Let G_1/G_2 be the genotypes for a pair of individuals, let k_m with $m = 0, 1, 2$ be their IBD probabilities (shown in Table 1) and let R be their family relationship. Then, the probability of observing G_1/G_2, given R is:

$$\begin{aligned} P(G_1/G_2|R) = P(G_1/G_2|m=0)k_0 \\ +P(G_1/G_2|m=1)k_1 \\ +P(G_1/G_2|m=2)k_2. \end{aligned} \tag{1}$$

The terms $P(G_1/G_2|m=0), P(G_1/G_2|m=1)$ and $P(G_1/G_2|m=2)$ are the

probabilities of observing each pair of genotypes given the number of IBD alleles (Table 2).

2.2 The likelihood ratio approach for identifying three quarter siblings

The likelihood ratio (LR) approach has been widely used for relatedness research during the last decades (Boehnke and Cox, 1997; Heinrich et al., 2016; Katki et al., 2010; Kling and Tillmar, 2019; Thompson, 1986; Weir et al., 2006). Briefly, the LR approach is based on the contrast of two hypotheses, one in the numerator, H_i; and the other one in the denominator, H_j. The larger the LR, the more plausible is H_i; whereas the smaller the LR, the more plausible is H_j. For relatedness research, we consider the ratio of the probabilities from Equation 1 according to the hypothesis of the R_i and R_j relationships. That is:

$$LR(R_i, R_j|G_1/G_2) = \frac{P(G_1/G_2|R_i)}{P(G_1/G_2|R_j)} \tag{2}$$

Here we consider the FS, 3/4S, 2nd and UN relationships and calculate three LR having FS, 3/4S or 2nd in the numerator and having the UN relationship in the denominator. The common denominator makes the LR values comparable in order to distinguish 3/4S from FS and 2nd degree. Inference of relatedness for each pair of individuals is based on the largest LR value in the FS∼UN, 3/4S∼UN and 2nd∼UN ratios. The LR are shown in Table 3,

depending on the observed genotypes of a pair of individuals. Most of these ratios are derived in Heinrich et al. (2016), and the new results for 3/4S are shown in Appendix B. The e parameter from the PO∼UN ratio in Table 3 is a small number (i.e. 0.001) used to account for genotype errors and *de novo* mutations if the genotype combination does not occur. In this way, the LR cannot be zero. For S biallelic SNPs, the LR can be obtained by multiplying the LR_s across markers and by dividing by the number of SNPs. It is convenient to work in a logarithmic scale such that:

$$\log_{10}(LR) = \frac{1}{S}\log_{10}\left(\prod_{s=1}^{S} LR_s(R_i, R_j|G_1/G_2)\right) = \frac{1}{S}\sum_{s=1}^{S}\log_{10}\left(LR_s(R_i, R_j|G_1/G_2)\right), \tag{3}$$

which corresponds to the logarithm of the geometric mean of the likelihood ratios.

2.3 Materials

We test our method for detecting 3/4S with data from the GCAT Genome for Life cohort project (Obón-Santacana et al., 2018). Briefly, the GCAT project is a prospective study that includes ∼20K participants recruited from the general population of Catalonia, a Western Mediterranean region in the Northeast of Spain. A subset of 5459 participants were genotyped using the Infinium Expanded Multi-Ethnic Genotyping Array (MEGAEx) (ILLU-

MINA, San Diego, California, USA). In the present work, we consider 5,075 GCAT participants from Caucasian ancestry and 756,003 SNPs that passed strict quality control (Galván-Femenía et al., 2018). A previous relatedness research analysis of this dataset reported 63 FS, 8 3/4S and 12 2nd degree candidate pairs (Graffelman et al., 2019).

3 Simulations

In this section we evaluate the likelihood ratio approach to distinguish 3/4S from FS and 2nd relationships by using simulated data. Pedigrees were simulated from the genetic data of the individuals of the GCAT project, using the ped-sim method of Caballero et al. (2019). We apply this method in order to account for recombination by using sex-specific genetic maps (Bhérer et al., 2017) and also a crossover interference model (Campbell et al., 2015). The simulations were carried out as follows. First, we identified 4,147 potentially unrelated individuals with kinship coefficient < 0.025. From these individuals, we retained 537,488 autosomal SNPs with minor allele frequency (MAF) > 0.01, Hardy-Weinberg exact mid p-value > 0.05 (Graffelman and Moreno, 2013) and missing call rate zero. Genotypes of the unrelated individuals were phased with SHAPEIT4 (Delaneau et al., 2018) and were used as input for the ped-sim method. Then, we simulated 500 pedigrees containing one FS pair and 500 pedigrees containing one 3/4S pair (Supplementary Figures S1 and S2). In total, we used 3000 random GCAT individuals as

founders to generate 3000 artificial individuals. The number of simulated related pairs were 4,000 PO, 500 FS, 500 3/4S and 3,500 2nd degree from a total of 17,997,000 of pairs. To estimate the IBD probabilities and the kinship coefficient for these simulated pairs we used 27,087 SNPs obtained by retaining variants with MAF > 0.40 and by linkage disequilibrium (LD) pruning, requiring markers to have low pair-wise correlation ($r^2 < 0.20$).

Figure 1 shows the $(\hat{k}_0, \hat{k}_1)$-plot for these simulated pairs of individuals. The IBD probabilities were estimated with the PLINK software (Purcell et al., 2007). As expected, the estimated IBD probabilities are close to the expected theoretical values from Table 1 for most pairs of individuals. In Figure 1, the 3/4S relationships show good separation from 2nd degree relationships but mix to some extent with FS pairs. Estimated IBD probabilities appear to be centered on their expected values for FS, 3/4S and 2nd degree pairs, and have larger variance then PO and UN pairs. The discriminative power of our method crucially depends on the variance of these estimated probabilities (Hill and Weir, 2011).

Boxplots of the kinship estimator recently proposed by Goudet & Weir (Figure 2; Goudet et al. (2018); Weir and Goudet (2017)) show clearly a difference in median for 3/4S and 1st and 2nd degree relationships, though the distribution of the kinship coefficient of the 3/4S overlaps with those of 1st and 2nd degree pairs. Also, kinship coefficients can be identical for different rela-

tionships, as is the case for PO and FS. Therefore, according to the Equation (3), we calculate the FS∼UN, 3/4S∼UN and 2nd∼UN likelihood ratios for 500 2nd, 500 3/4S and 500 FS simulated pairs. Figure 3 shows that FS pairs mostly have the largest LR values in the FS∼UN ratio, 3/4S pairs mostly have the largest LR values in the 3/4S∼UN ratio and 2nd degree pairs mostly have largest LR in the 2nd∼UN. Note the plotted data profile shaped in a 'greater-than' sign (">") pattern suggesting the inference of 3/4S for most 3/4S pairs. In fact, the correct classification rate of the LR approach for the 2nd, 3/4S and FS simulated pairs is $\frac{500}{500} = 1$, $\frac{479}{500} = 0.958$ and $\frac{475}{500} = 0.95$ respectively. These simulations show the proposed LR approach to be useful for distinguishing 3/4S relationships from FS and 2nd degree relationships.

4 Case study

In this section we apply the likelihood ratio approach in a genome-wide SNP array data from the aforementioned GCAT project. Graffelman *et al.* (2019, Table 5 and Figure 7) suggested this database to contain eight 3/4S pairs using a log-ratio biplot approach combined with discriminant analysis (LR-kinbiplot). Figures 4 and 5 show the $(\hat{k}_0, \hat{k}_1)$-plot and boxplots of kinship estimates of the GCAT data. The IBD probabilities were estimated with the PLINK software, whereas the kinship coefficient was estimated by the estimator proposed by Weir and Goudet (2017). The colors for the FS, 3/4S and

2nd degree pairs in both Figures were assigned according to the relationship inferred by the LR approach. Figure 4 shows, like the simulations, a larger variance for FS pairs, and smaller variances for PO and UN pairs.

Figure 6 shows the LR ratio values (FS∼UN, 3/4S∼UN and 2nd∼UN ratios) for the presumably FS, 3/4S and 2nd pairs from the GCAT project. The LR analysis reveals eight 3/4S pairs (black color) that have the 'greater-than' sign (">") shaped pattern. All inferred FS pairs (blue color) have the monotonously increasing "/" shaped pattern and all 2nd degree pairs have the monotonously decreasing "\" pattern. Table 4 shows the LR values for each pair which confirm that there are eight 3/4S pairs in concordance with the LR-kinbiplot approach.

5 Discussion

In this paper we show that the likelihood ratio approach is useful for distinguishing three quarter siblings from FS and 2nd degree relationships. Figure 4 shows that in a standard $(\hat{k}_0, \hat{k}_1)$-plot, 3/4S can easily go unnoticed as FS pairs. The LR approach can be of great help to detect such cases. The LR approach developed in this paper confirmed eight 3/4S pairs previously uncovered by a log-ratio biplot (LR-kinbiplot) approach (Graffelman et al., 2019) for genome wide SNP array data from the GCAT cohort.

The estimated relationships for the GCAT cohort were to some extent confirmed by an analysis of the surnames of the participants, respecting their privacy. In Spain, people have a double surname, usually the first from the father and the second from the mother. This implies that FS and 3/4S pairs share two surnames, whereas 2nd degree relationships share only one. All identified 3/4S pairs were confirmed to share two surnames, supporting that these pairs are not 2nd degree.

The proposed LR approach multiplies the likelihoods over loci, under the assumption of independence. The existence of linkage disequilibrium (LD) between variants violates this assumption. In order to approximately meet the requirement of independence, LD pruning of neighbouring variants in a window is therefore recommended (Kling and Tillmar, 2019). This pruning can be done in PLINK (Purcell et al., 2007) or with other software (Calus and Vandenplas, 2018). A future improvement of the LR approach could use Markov chain algorithms (Abecasis and Wigginton, 2005; Kling et al., 2015) that allow efficient likelihood computations on blocks of tightly linked markers.

The LR approach developed in this paper assumes known allele frequencies and non-inbred individuals. The first assumption seems reasonable given the large sample size used in this study. Inbreeding could be accounted for by the

use of nine condensed Jacquard coefficients (Hanghøj et al., 2019; Jacquard, 1974). The $(\hat{k}_0, \hat{k}_1)$-plot of the GCAT data in Figure 4 also shows some pairs in-between a 2nd a 3rd degree relationship. In future work, the likelihood ratio approach presented in this paper could be further refined to identify the relationship of these pairs more precisely. In-between relationships, like the 3/4S relationship studied in this paper, essentially stress that relatedness is a continuous rather than a discrete concept.

Acknowledgements

This study makes use of data generated by the GCAT Genomes for Life Cohort study of the Genomes of Catalonia, IGTP. A full list of the investigators who contributed to the generation of the data is available from www.genomesforlife.com. We thank the CERCA Program of the Generalitat de Catalunya for institutional support. We are also very grateful to Bruce S. Weir for his comments on the manuscript as well as the computer resources and technical expertise provided by Daniel Matías-Sánchez, Jordi Valls-Margarit and David Torrents-Arenales from the Life Sciences - Computational Genomics group of the Barcelona Supercomputing Center - Centro Nacional de Supercomputación (BSC-CNS).

Funding

This work was partially supported by grants RTI2018-095518-B-C22 (JG), RTI2018-095518-B-C21 (IGF and CBV) and ADE 10/00026 (RdC) (MCIU/AEI/FEDER) of the Spanish Ministry of Science, Innovation and Universities and the European Regional Development Fund, by grants SGR1269 and 2017 SGR529 (RdC) of the Generalitat de Catalunya, by grant R01 GM075091 (JG) from the United States National Institutes of Health, by the Ramon y Cajal action RYC-2011-07822 (RdC), by Agency for Management of University and Research Grants (AGAUR) of the Catalan Government grant 2017SGR723 (VM), and by the Spanish Association Against Cancer (AECC) Scientific Foundation, grant GCTRA18022MORE (VM).

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abecasis, G. R. and Wigginton, J. E. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *The American Journal of Human Genetics*, 77(5):754–767, 2005.

Abecasis, G. R., Cherny, S. S., Cookson, W., and Cardon, L. R. GRR: graphical representation of relationship errors. *Bioinformatics*, 17(8):742–743, 2001.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. Data quality control in genetic case-control association studies. *Nature protocols*, 5(9):1564, 2010.

Bhérer, C., Campbell, C. L., and Auton, A. Refined genetic maps reveal sexual dimorphism in human meiotic recombination at multiple scales. *Nature communications*, 8(1):1–9, 2017.

Boehnke, M. and Cox, N. J. Accurate inference of relationships in sib-pair linkage studies. *The American Journal of Human Genetics*, 61(2):423–429, 1997.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., OConnell, J., et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562 (7726):203, 2018.

Caballero, M., Seidman, D., Qiao, Y., Sannerud, J., Dyer, T., Lehman, D., Curran, J., Duggirala, R., Blangero, J., Carmi, S., and Williams, A. Crossover interference and sex-specific genetic maps shape identical by descent sharing in close relatives. *PLoS Genet*, 15(12):e1007979, 2019. doi: doi.org/10.1371/journal.pgen.1007979.

Calus, M. P. and Vandenplas, J. SNPrune: an efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. *Genetics Selection Evolution*, 50(1):34, 2018.

Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D., and Auton, A. Escape from crossover interference increases with maternal age. *Nature communications*, 6:6260, 2015.

Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J., and Dermitzakis, E. Integrative haplotype estimation with sub-linear complexity. *bioRxiv*, page 493403, 2018.

Evett, I. W. and Weir, B. S. *Interpreting DNA evidence*. Sinauer Associates, Inc., 1998.

Galván-Femenía, I., Graffelman, J., and Barceló-i Vidal, C. Graphics for relatedness research. *Molecular Ecology Resources*, 17(6):1271–1282, 2017. doi: 10.1111/1755-0998.12674.

Galván-Femenía, I., Obón-Santacana, M., Piñeyro, D., Guindo-Martinez, M., Duran, X., Carreras, A., Pluvinet, R., Velasco, J., Ramos, L., Aussó,

S., Mercader, J. M., Puig, L., Perucho, M., Torrents, D., Moreno, V., Sumoy, L., and de Cid, R. Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort. *Journal of medical genetics*, 55(11):765–778, 2018.

Goudet, J., Kay, T., and Weir, B. S. How to estimate kinship. *Molecular ecology*, 27(20):4121–4135, 2018.

Graffelman, J. and Moreno, V. The mid p-value in exact tests for Hardy-Weinberg equilibrium. *Statistical Applications in Genetics and Molecular Biology*, 12(4):433–448, 2013.

Graffelman, J., Galván-Femenía, I., de Cid, R., and Barceló-i Vidal, C. A log-ratio biplot approach for exploring genetic relatedness based on identity by state. *Frontiers in genetics*, 10:341, 2019.

Hanghøj, K., Moltke, I., Andersen, P. A., Manica, A., and Korneliussen, T. S. Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *GigaScience*, 8(5), 2019.

Heinrich, V., Kamphans, T., Mundlos, S., Robinson, P. N., and Krawitz, P. M. A likelihood ratio-based method to predict exact pedigrees for complex families from next-generation sequencing data. *Bioinformatics*, 33(1): 72–78, 2016.

Hill, W. and Weir, B. Variation in actual relationship as a consequence of

mendelian sampling and linkage. *Genetics research (Camb)*, 93(1):47–64, 2011. doi: 10.1017/S0016672310000480.

Jacquard, A. *The genetic structure of populations*. Springer-Verlag, 1974.

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *BioRxiv*, page 531210, 2019.

Katki, H. A., Sanders, C. L., Graubard, B. I., and Bergen, A. W. Using DNA fingerprints to infer familial relationships within NHANES III households. *Journal of the American Statistical Association*, 105(490):552–563, 2010.

Kling, D. and Tillmar, A. Forensic genealogy-a comparison of methods to infer distant relationships based on dense SNP data. *Forensic science international. Genetics*, 42:113–124, 2019. doi: 10.1016/j.fsigen.2019.06.019.

Kling, D., Tillmar, A., Egeland, T., and Mostad, P. A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations. *International journal of legal medicine*, 129 (5):943–954, 2015.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

366 Milligan, B. G. Maximum-likelihood estimation of relatedness. *Genetics*, 163
367 (3):1153–1167, 2003.

368 Obón-Santacana, M., Vilardell, M., Carreras, A., Duran, X., Velasco, J.,
369 Galván-Femenía, I., Alonso, T., Puig, L., Sumoy, L., Duell, E., Perucho,
370 M., Moreno, V., and de Cid, R. GCAT-Genomes for life: a prospective
371 cohort study of the genomes of Catalonia. *BMJ open*, 8(3):e018324, 2018.

372 Oliehoek, P., Windig, J., van Arendonk, J., and Bijma, P. Estimating relat-
373 edness between individuals in general populations with a focus on their use in
374 conservation programs. *Genetics*, 173(1):483–496, 2006. doi: 10.1534/ge-
375 netics.105.049940.

376 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender,
377 D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., et al. PLINK: a tool
378 set for whole-genome association and population-based linkage analyses.
379 *The American journal of human genetics*, 81(3):559–575, 2007.

380 Rosenberg, N. A. Standardized subsets of the hgdp-ceph human genome
381 diversity cell line panel, accounting for atypical and duplicated samples
382 and pairs of close relatives. *Annals of human genetics*, 70(6):841–847,
383 2006.

384 Staples, J., Maxwell, E. K., Gosalia, N., Gonzaga-Jauregui, C., Snyder, C.,
385 Hawes, A., Penn, J., Ulloa, R., Bai, X., Lopez, A. E., et al. Profiling and

386 leveraging relatedness in a precision medicine cohort of 92,455 exomes. *The*
387 *American Journal of Human Genetics*, 102(5):874–889, 2018.

388 Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres,
389 R., Taliun, S. A. G., Corvelo, A., Gogarten, S. M., Kang, H. M., et al.
390 Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.
391 *BioRxiv*, page 563866, 2019.

392 Thompson, E. The estimation of pairwise relationships. *Annals of human*
393 *genetics*, 39(2):173–188, 1975.

394 Thompson, E. Likelihood inference of paternity. *American journal of human*
395 *genetics*, 39(2):285, 1986.

396 Thompson, E. Estimation of relationships from genetic data. *Handbook of*
397 *statistics*, 8:255–269, 1991.

398 Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., Mccarthy, M. I., Brown,
399 M. A., and Yang, J. 10 Years of GWAS Discovery: Biology, Function,
400 and Translation. *The American Journal of Human Genetics*, 101(1):5–22,
401 2017. ISSN 0002-9297. doi: 10.1016/j.ajhg.2017.06.005.

402 Wagner, A., Creel, S., and Kalinowski, S. Estimating relatedness and re-
403 lationships using microsatellite loci with null alleles. *Heredity*, 97(5):336,
404 2006.

405 Weir, B. S. and Goudet, J. A unified characterization of population structure
406 and relatedness. *Genetics*, 206(4):2085–2103, 2017.

407 Weir, B. S., Anderson, A. D., and Hepler, A. B. Genetic relatedness analysis:
408 modern data and new challenges. *Nature Reviews Genetics*, 7(10):771,
409 2006.

Tables and Figures

			Probability of IBD Sharing		
Type of Relative	R	ϕ	k_0	k_1	k_2
Monozygothic twins (MZ)	0	1/2	0	0	1
Parent-offspring (PO)	1	1/4	0	1	0
Full-siblings (FS)	1	1/4	1/4	1/2	1/4
Three-quarter siblings (3/4S)	-	3/16	3/8	1/2	1/8
Half-siblings/ grandchild-grandparent/ niece/nephew-uncle/aunt (2nd)	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).

G_1/G_2	$m=0$	$m=1$	$m=2$
AA/AA	p^4	p^3	p^2
AA/AB	$2p^3q$	p^2q	0
AA/BB	p^2q^2	0	0
AB/AB	$4p^2q^2$	pq	$2pq$

Table 2: Possible pairs of biallelic genotypes and the probability of each pair given the number of alleles identical by descent (m). We assume that the order of the genotypes is irrelevant, i.e. the probabilities for G_1/G_2 and G_2/G_1 are the same.

LR	AA/AA	AA/AB	AB/AB	AA/BB
PO∼UN	$\frac{1}{p}$	$\frac{1}{2p}$	$\frac{1}{4pq}$	$\frac{e}{p^2q^2}$
FS∼UN	$\frac{1}{4}+\frac{1}{2p}+\frac{1}{(2p)^2}$	$\frac{1}{4}+\frac{1}{4p}$	$\frac{1}{4}+\frac{1}{4pq}$	$\frac{1}{4}$
3/4S∼UN	$\frac{3}{8}+\frac{1}{2p}+\frac{1}{8p^2}$	$\frac{3}{8}+\frac{1}{4p}$	$\frac{3}{8}+\frac{3}{16pq}$	$\frac{3}{8}$
2nd∼UN	$\frac{1}{2}+\frac{1}{2p}$	$\frac{1}{2}+\frac{1}{4p}$	$\frac{1}{2}+\frac{1}{8pq}$	$\frac{1}{2}$
FC∼UN	$\frac{3}{4}+\frac{1}{2p}$	$\frac{3}{4}+\frac{1}{4p}$	$\frac{3}{4}+\frac{1}{16pq}$	$\frac{3}{4}$

Table 3: Likelihood ratio (LR) for relatedness research for biallelic SNPs. The considered LR are PO, FS, 3/4S, 2nd or FC relationships in the numerator and the UN relationship in the denominator. The LR values depend on the observed genotypes of a pair of individuals and the allele frequencies p and q of the population under study. The e parameter is used to account for genotype errors and *de novo* mutations if the genotype combination does not occur (Heinrich et al., 2016). We assume that the order of the genotypes is irrelevant, i.e. the LR for G_1/G_2 and G_2/G_1 is the same.

Table 4: Likelihood ratio inference (LR approach) for the presumably 2nd, 3/4S and FS pairs from the GCAT cohort. FS $\sim$ UN, 3/4S $\sim$ UN and 2nd $\sim$ UN are the LR values for each pair. LR-kinbiplot is the inferred relationship from Graffelman et al. (2019). $\hat{\phi}$: estimated kinship coefficient. $\hat{k}_0$, $\hat{k}_1$ and $\hat{k}_2$: estimated IBD probabilities.

pair	IID	sex	IID	sex	$\hat{k}_0$	$\hat{k}_1$	$\hat{k}_2$	$\hat{\phi}$	LR-kinbiplot	FS $\sim$ UN	3/4S $\sim$ UN	2nd $\sim$ UN	LR approach
1	REL_00178	F	REL_01132	F	0.61	0.36	0.04	0.107	2nd	-0.0165	0.0027	**0.0092**	2nd
2	REL_02227	F	REL_00865	M	0.57	0.43	0.00	0.109	2nd	-0.0164	0.0035	**0.0109**	2nd
3	REL_04137	F	REL_03163	M	0.51	0.49	0.00	0.122	2nd	-0.0103	0.0082	**0.0142**	2nd
4	REL_04126	F	REL_02089	F	0.50	0.50	0.00	0.126	2nd	-0.0106	0.0080	**0.0143**	2nd
5	REL_04141	F	REL_02030	M	0.49	0.50	0.01	0.129	2nd	-0.0072	0.0101	**0.0152**	2nd
6	REL_02092	M	REL_00587	F	0.48	0.52	0.00	0.129	2nd	-0.0073	0.0104	**0.0158**	2nd
7	REL_02212	M	REL_04828	F	0.47	0.53	0.00	0.132	2nd	-0.0061	0.0111	**0.0161**	2nd
8	REL_00603	F	REL_00189	F	0.47	0.53	0.00	0.134	2nd	-0.0076	0.0101	**0.0156**	2nd
9	REL_03666	M	REL_02902	M	0.47	0.53	0.00	0.134	2nd	-0.0057	0.0112	**0.0160**	2nd
10	REL_00132	F	REL_00707	M	0.45	0.55	0.00	0.137	2nd	-0.0059	0.0113	**0.0164**	2nd
11	REL_02058	F	REL_03610	F	0.45	0.55	0.00	0.139	2nd	-0.0041	0.0125	**0.0170**	2nd
12	REL_01692	F	REL_00010	F	0.44	0.56	0.00	0.139	2nd	-0.0041	0.0127	**0.0173**	2nd
13	REL_03969	M	REL_00271	M	0.34	0.56	0.10	0.189	3/4S	0.0260	**0.0328**	0.0279	3/4S
14	REL_03803	F	REL_02343	M	0.35	0.51	0.14	0.198	3/4S	0.0317	**0.0361**	0.0287	3/4S
15	REL_03924	M	REL_03023	F	0.37	0.46	0.17	0.201	3/4S	0.0365	**0.0393**	0.0301	3/4S
16	REL_00083	M	REL_02333	M	0.33	0.52	0.15	0.207	3/4S	0.0377	**0.0403**	0.0313	3/4S
17	REL_01344	M	REL_02408	F	0.36	0.44	0.20	0.210	3/4S	0.0402	**0.0412**	0.0304	3/4S
18	REL_04189	M	REL_00775	M	0.37	0.43	0.21	0.210	3/4S	0.0422	**0.0428**	0.0314	3/4S
19	REL_03150	F	REL_01804	F	0.32	0.51	0.17	0.212	3/4S	0.0411	**0.0426**	0.0322	3/4S
20	REL_02752	F	REL_04859	F	0.34	0.46	0.20	0.215	3/4S	0.0441	**0.0443**	0.0325	3/4S
21	REL_01502	M	REL_03665	M	0.31	0.48	0.21	0.225	FS	**0.0482**	0.0469	0.0339	FS
22	REL_04592	F	REL_04600	F	0.30	0.48	0.21	0.226	FS	**0.0511**	0.0493	0.0358	FS
23	REL_04693	F	REL_00797	F	0.31	0.47	0.22	0.228	FS	**0.0520**	0.0498	0.0357	FS
24	REL_03607	M	REL_00319	F	0.30	0.49	0.21	0.228	FS	**0.0501**	0.0484	0.0350	FS
25	REL_03220	F	REL_04615	F	0.31	0.46	0.23	0.230	FS	**0.0532**	0.0505	0.0360	FS
26	REL_03212	M	REL_02516	F	0.28	0.53	0.20	0.231	FS	**0.0548**	0.0526	0.0386	FS
27	REL_03310	M	REL_03659	F	0.26	0.56	0.18	0.231	FS	**0.0496**	0.0484	0.0358	FS
28	REL_04427	F	REL_02635	F	0.26	0.54	0.19	0.232	FS	**0.0502**	0.0487	0.0358	FS
29	REL_00122	M	REL_01902	F	0.29	0.49	0.22	0.233	FS	**0.0542**	0.0513	0.0368	FS
30	REL_00284	M	REL_02444	F	0.28	0.51	0.21	0.233	FS	**0.0517**	0.0494	0.0356	FS
31	REL_03838	F	REL_02496	F	0.31	0.45	0.24	0.234	FS	**0.0561**	0.0523	0.0367	FS
32	REL_01564	F	REL_03827	F	0.32	0.43	0.26	0.236	FS	**0.0571**	0.0528	0.0365	FS
33	REL_04529	F	REL_04492	F	0.28	0.50	0.22	0.236	FS	**0.0555**	0.0522	0.0373	FS
34	REL_04494	M	REL_00931	M	0.28	0.49	0.23	0.237	FS	**0.0560**	0.0525	0.0373	FS
35	REL_04466	F	REL_02680	F	0.31	0.43	0.26	0.237	FS	**0.0576**	0.0531	0.0367	FS
36	REL_04405	M	REL_03949	M	0.26	0.52	0.22	0.238	FS	**0.0557**	0.0525	0.0376	FS
37	REL_03880	M	REL_04789	F	0.27	0.50	0.23	0.239	FS	**0.0566**	0.0529	0.0376	FS
38	REL_00383	F	REL_03293	M	0.25	0.53	0.22	0.241	FS	**0.0574**	0.0538	0.0385	FS
39	REL_01888	M	REL_04360	M	0.25	0.54	0.21	0.241	FS	**0.0566**	0.0532	0.0383	FS
40	REL_00792	F	REL_00954	M	0.26	0.51	0.23	0.242	FS	**0.0585**	0.0543	0.0385	FS
41	REL_00872	F	REL_01784	F	0.25	0.53	0.22	0.242	FS	**0.0598**	0.0556	0.0398	FS
42	REL_01450	M	REL_01960	M	0.26	0.51	0.23	0.242	FS	**0.0586**	0.0544	0.0386	FS
43	REL_04616	F	REL_02777	F	0.28	0.47	0.25	0.243	FS	**0.0604**	0.0553	0.0386	FS
44	REL_02899	M	REL_01707	F	0.28	0.45	0.26	0.244	FS	**0.0618**	0.0562	0.0389	FS
45	REL_02905	F	REL_02575	F	0.25	0.52	0.23	0.245	FS	**0.0604**	0.0557	0.0394	FS
46	REL_00769	M	REL_04746	F	0.23	0.57	0.21	0.246	FS	**0.0606**	0.0564	0.0406	FS
47	REL_00009	F	REL_02335	F	0.23	0.55	0.22	0.246	FS	**0.0603**	0.0558	0.0399	FS
48	REL_04475	F	REL_04218	M	0.25	0.51	0.24	0.247	FS	**0.0615**	0.0564	0.0397	FS
49	REL_01150	F	REL_04384	F	0.26	0.49	0.25	0.249	FS	**0.0639**	0.0580	0.0403	FS
50	REL_03944	M	REL_03475	F	0.23	0.54	0.23	0.249	FS	**0.0618**	0.0568	0.0403	FS
51	REL_03904	F	REL_04994	F	0.25	0.50	0.25	0.249	FS	**0.0631**	0.0573	0.0400	FS
52	REL_01654	M	REL_03485	M	0.28	0.43	0.29	0.251	FS	**0.0660**	0.0588	0.0398	FS
53	REL_00504	M	REL_04718	F	0.24	0.50	0.25	0.252	FS	**0.0645**	0.0582	0.0404	FS
54	REL_00339	F	REL_02473	F	0.25	0.48	0.27	0.253	FS	**0.0651**	0.0584	0.0400	FS
55	REL_01016	M	REL_00887	M	0.24	0.50	0.26	0.254	FS	**0.0661**	0.0594	0.0411	FS
56	REL_03977	M	REL_01080	M	0.22	0.54	0.24	0.255	FS	**0.0644**	0.0583	0.0408	FS
57	REL_02339	M	REL_02391	M	0.27	0.44	0.29	0.256	FS	**0.0688**	0.0608	0.0411	FS
58	REL_01524	F	REL_03272	F	0.23	0.51	0.26	0.256	FS	**0.0674**	0.0604	0.0419	FS
59	REL_01285	M	REL_03761	F	0.24	0.50	0.27	0.257	FS	**0.0670**	0.0597	0.0410	FS

60	REL_03395	F	REL_02694	F	0.22	0.52	0.25	0.257	FS	**0.0680**	0.0609	0.0423	FS
61	REL_03151	M	REL_02204	F	0.23	0.50	0.26	0.257	FS	**0.0683**	0.0610	0.0421	FS
62	REL_00968	M	REL_01577	F	0.26	0.45	0.29	0.259	FS	**0.0744**	0.0654	0.0445	FS
63	REL_04439	F	REL_01640	F	0.26	0.43	0.31	0.260	FS	**0.0721**	0.0630	0.0421	FS
64	REL_01546	M	REL_03566	F	0.21	0.53	0.26	0.263	FS	**0.0701**	0.0621	0.0428	FS
65	REL_03442	F	REL_04510	F	0.22	0.51	0.27	0.264	FS	**0.0714**	0.0630	0.0431	FS
66	REL_00340	F	REL_04294	F	0.21	0.53	0.26	0.264	FS	**0.0710**	0.0628	0.0432	FS
67	REL_03001	F	REL_04111	F	0.23	0.48	0.29	0.265	FS	**0.0727**	0.0636	0.0430	FS
68	REL_00282	F	REL_04918	F	0.25	0.44	0.31	0.267	FS	**0.0748**	0.0648	0.0430	FS
69	REL_01083	F	REL_01704	F	0.18	0.57	0.25	0.267	FS	**0.0715**	0.0634	0.0439	FS
70	REL_03388	F	REL_02608	F	0.22	0.50	0.29	0.268	FS	**0.0739**	0.0645	0.0436	FS
71	REL_01924	F	REL_00727	M	0.24	0.45	0.32	0.270	FS	**0.0769**	0.0663	0.0440	FS
72	REL_02208	F	REL_03486	F	0.23	0.46	0.31	0.270	FS	**0.0769**	0.0665	0.0444	FS
73	REL_02718	M	REL_02913	M	0.22	0.48	0.30	0.271	FS	**0.0765**	0.0662	0.0443	FS
74	REL_00634	M	REL_03507	M	0.20	0.51	0.29	0.272	FS	**0.0754**	0.0656	0.0443	FS
75	REL_04741	F	REL_02513	F	0.19	0.52	0.30	0.277	FS	**0.0783**	0.0676	0.0455	FS
76	REL_00601	M	REL_02989	F	0.19	0.51	0.30	0.278	FS	**0.0802**	0.0689	0.0462	FS
77	REL_01624	F	REL_00750	F	0.19	0.51	0.30	0.278	FS	**0.0790**	0.0680	0.0456	FS
78	REL_00824	F	REL_00213	F	0.22	0.45	0.33	0.278	FS	**0.0815**	0.0693	0.0456	FS
79	REL_01264	M	REL_04751	F	0.18	0.52	0.30	0.279	FS	**0.0795**	0.0684	0.0459	FS
80	REL_02208	F	REL_01630	F	0.18	0.52	0.31	0.283	FS	**0.0826**	0.0706	0.0473	FS
81	REL_04704	F	REL_00804	M	0.17	0.52	0.31	0.285	FS	**0.0829**	0.0707	0.0472	FS
82	REL_03627	F	REL_03315	F	0.15	0.55	0.30	0.288	FS	**0.0838**	0.0714	0.0478	FS
83	REL_03486	F	REL_01630	F	0.17	0.50	0.33	0.289	FS	**0.0873**	0.0738	0.0488	FS

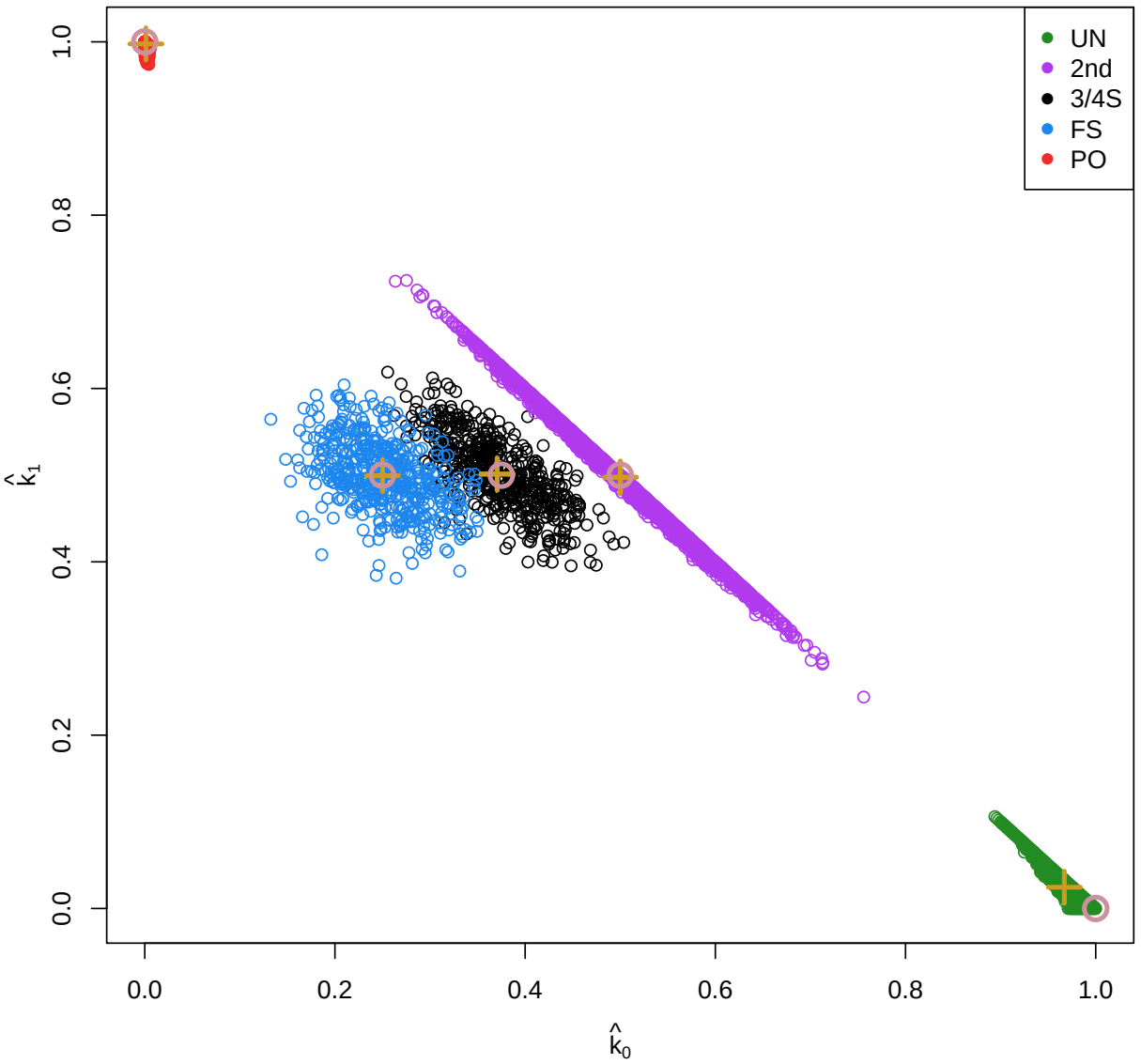

Figure 1: $(\hat{k}_0, \hat{k}_1)$-plot of ∼18 million pairs of simulated individuals using 27,087 SNPs. UN: unrelated; 2nd: second degree relationships; 3/4S three-quarter siblings. FS: full-siblings; PO: parent-offspring. Brown open dots represent theoretical IBD probabilities; brown + signs the average of the corresponding group.

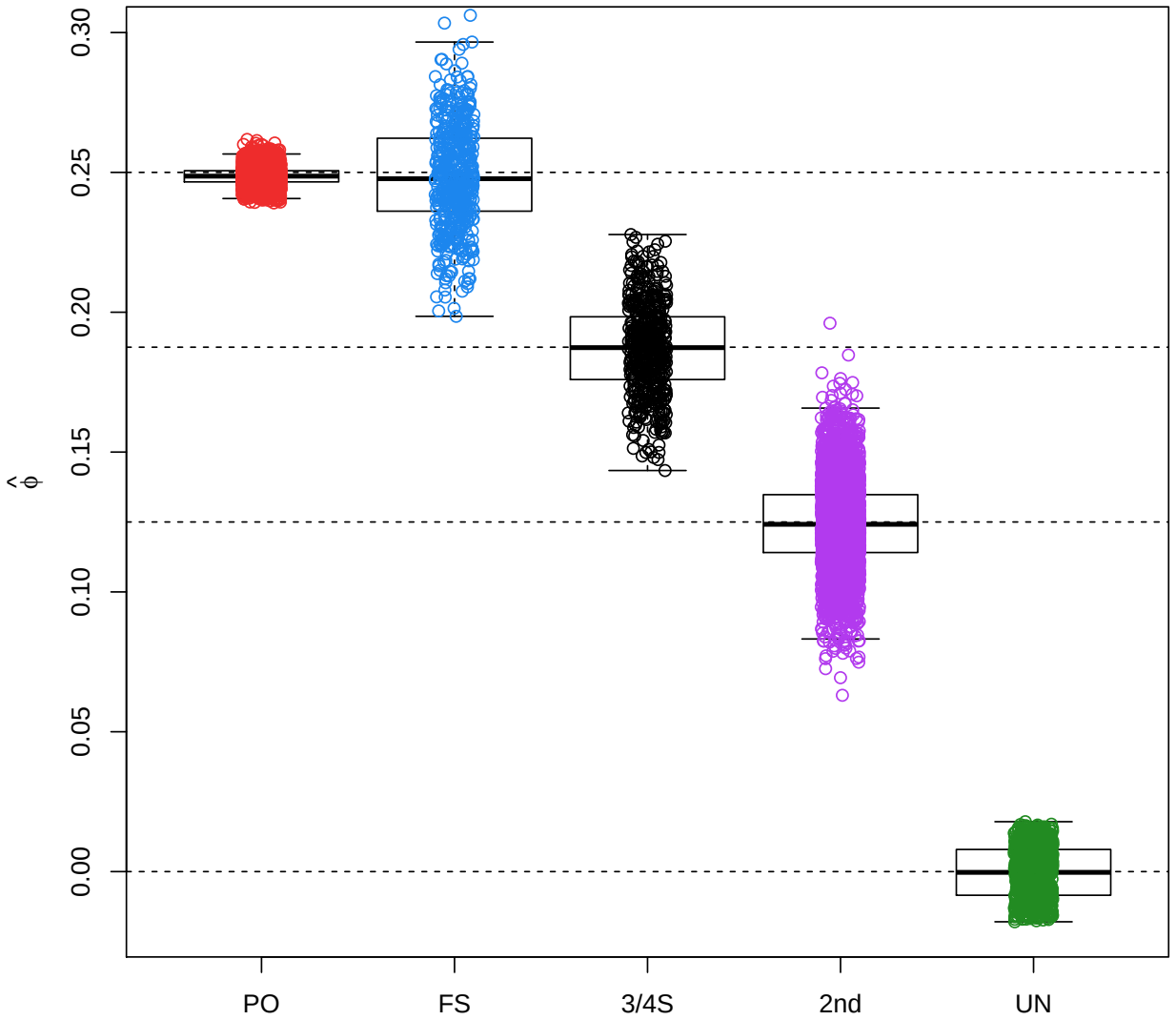

Figure 2: Boxplot of kinship estimates of ∼18 million pairs of simulated individuals using 27,087 SNPs.

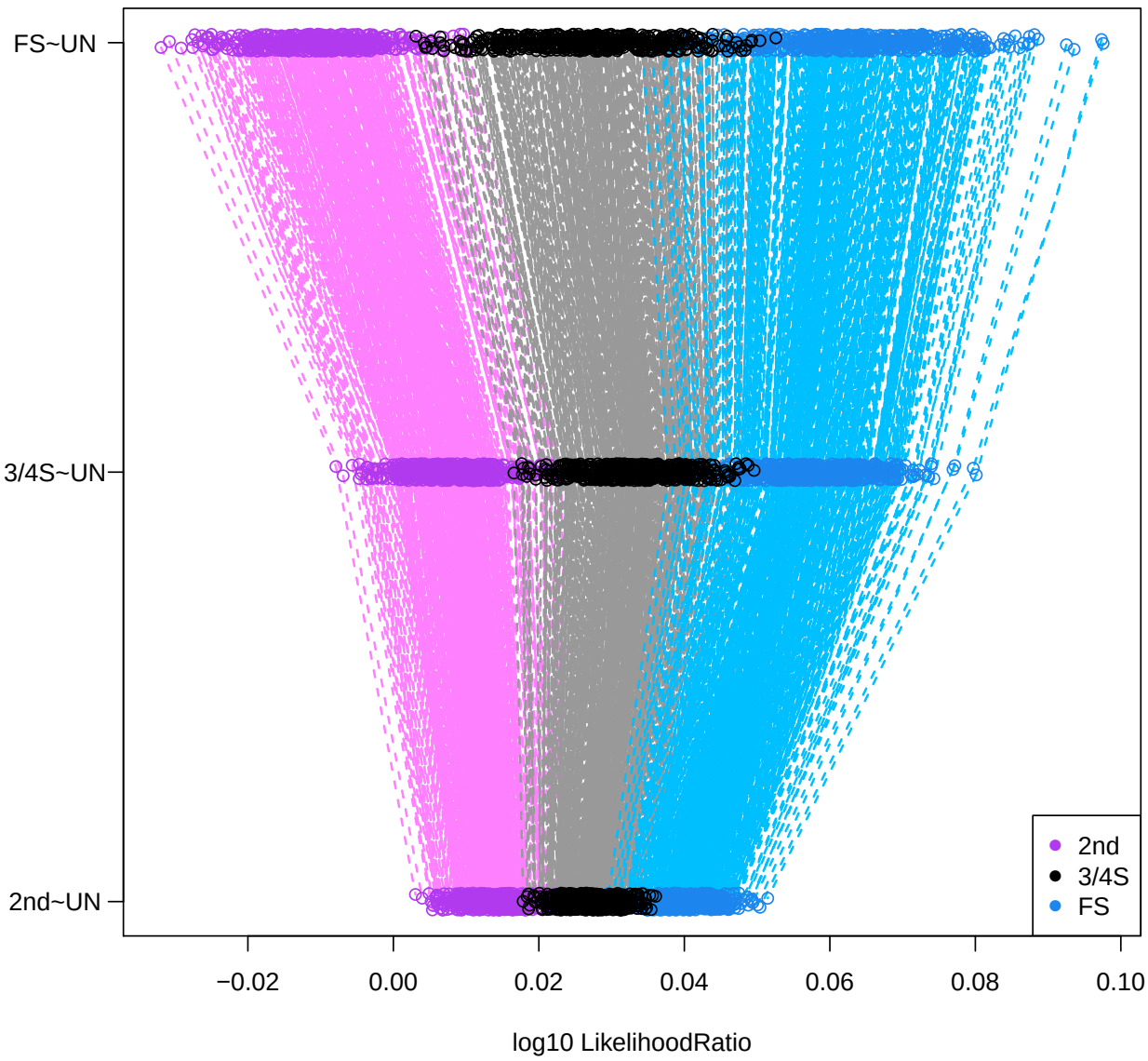

Figure 3: Log10 likelihood ratio approach of the simulated 2nd, 3/4S and FS pairs (500 for each relationship) using 27,087 SNPs. Note the larger than sign shaped (“>”) pattern (gray dashed lines) for most 3/4S pairs.

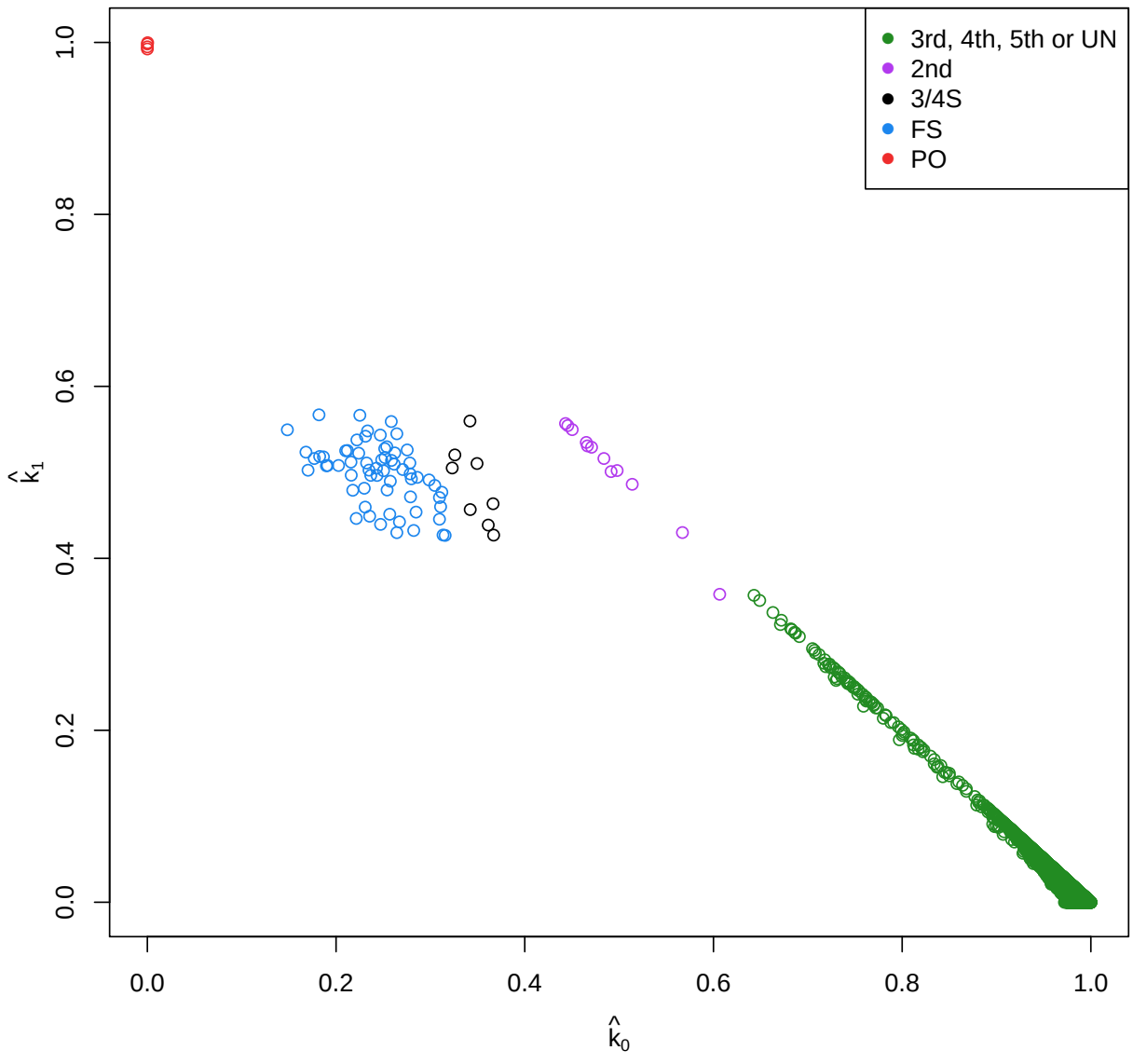

Figure 4: $(\hat{k}_0, \hat{k}_1)$-plot of the GCAT cohort for 5,075 individuals and 26,006 SNPs (MAF > 0.40, LD-pruned, HWE exact mid p-value > 0.05, and missing call rate 0). UN: unrelated; 5th, 4th, 3rd, 2nd: fifth, fourth, third, and second degree relationships; 3/4S: three quarter siblings; FS: full siblings; PO: parent-offspring.

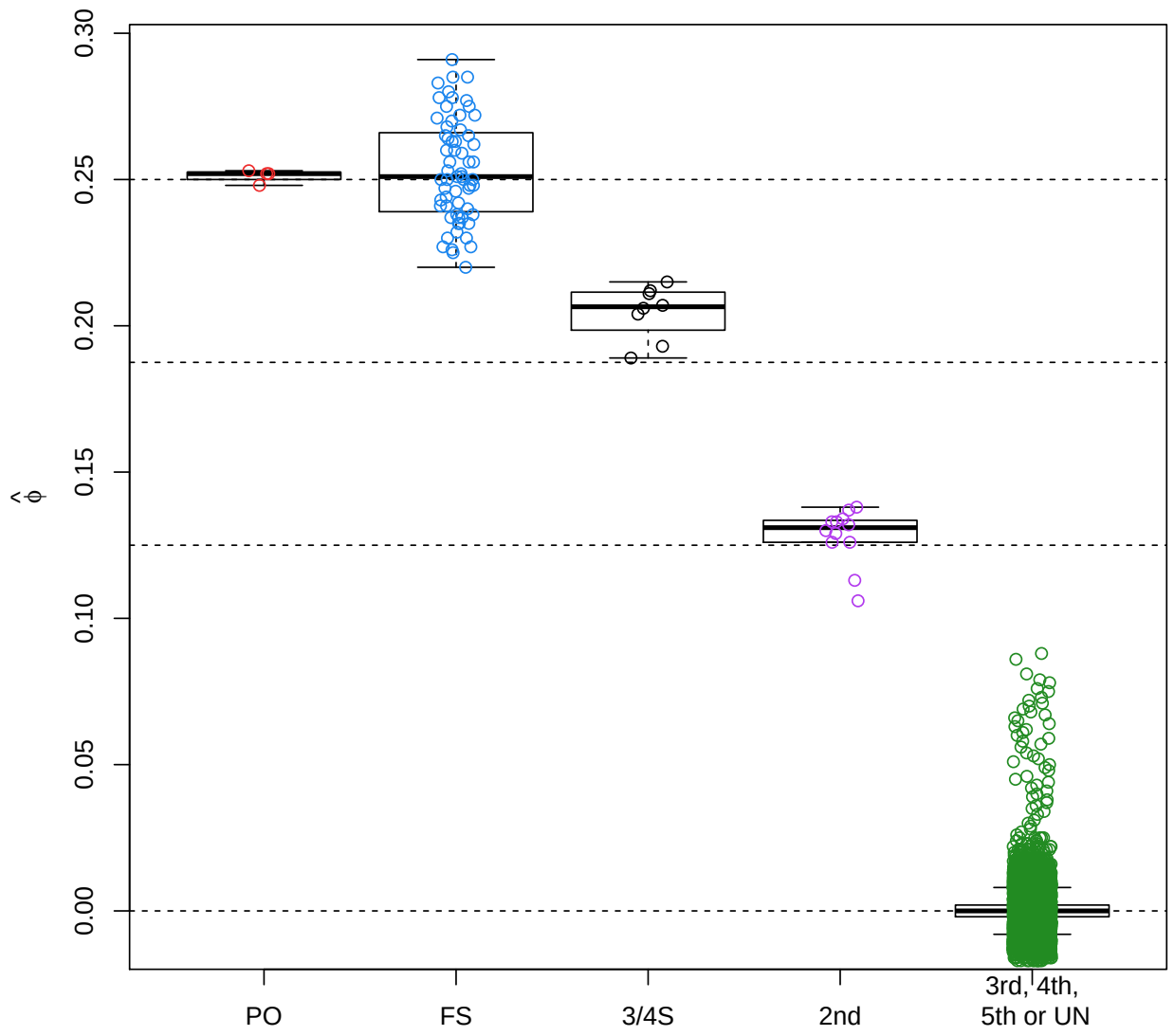

Figure 5: Boxplot of kinship estimates of the GCAT cohort for 5,075 individuals and 26,006 SNPs (MAF > 0.40, LD-pruned, HWE exact mid p-value > 0.05, and missing call rate 0).

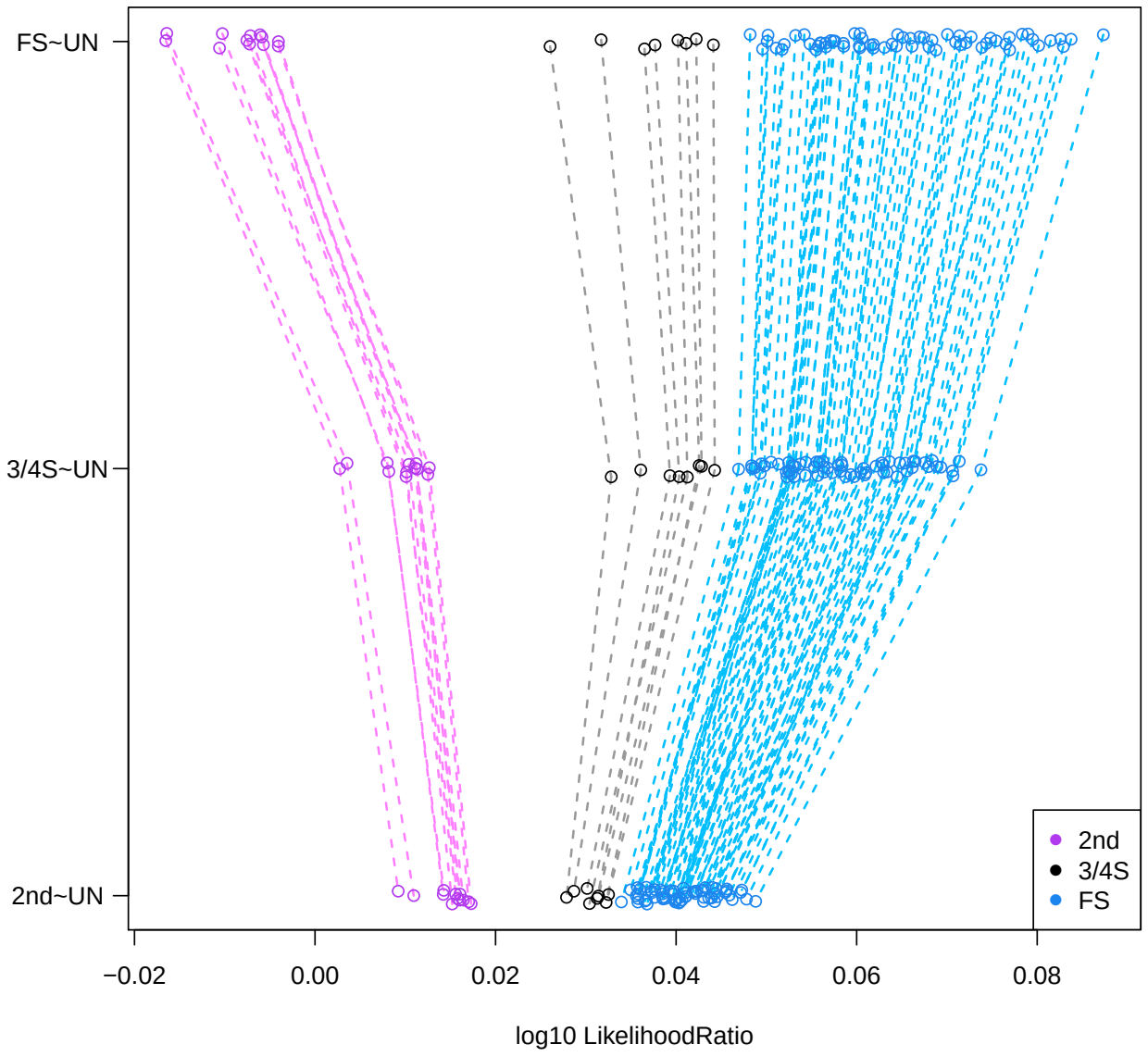

Figure 6: Log10 likelihood ratio approach of the presumably 2nd, 3/4S and FS pairs from the GCAT cohort using 26,006 SNPs (MAF > 0.40, LD-pruned, HWE exact mid p-value > 0.05, and missing call rate 0).

412 # Supplementary Figures

413

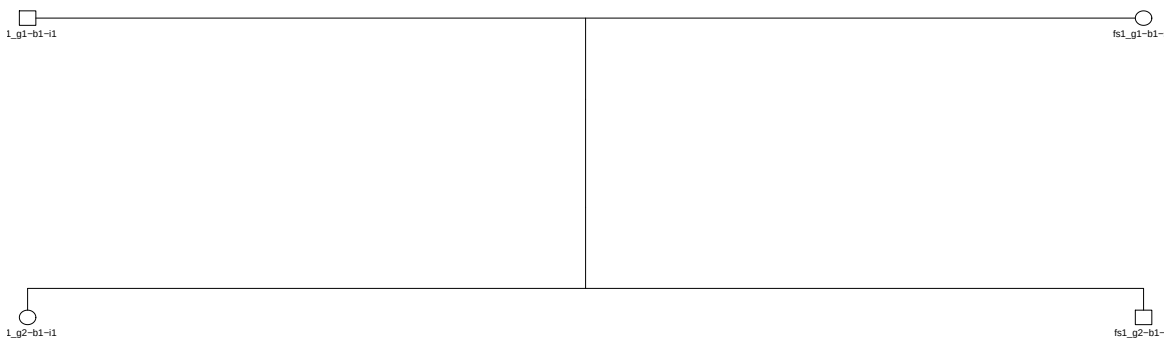

Figure S1: Pedigree simulated with ped-sim including one FS pair.

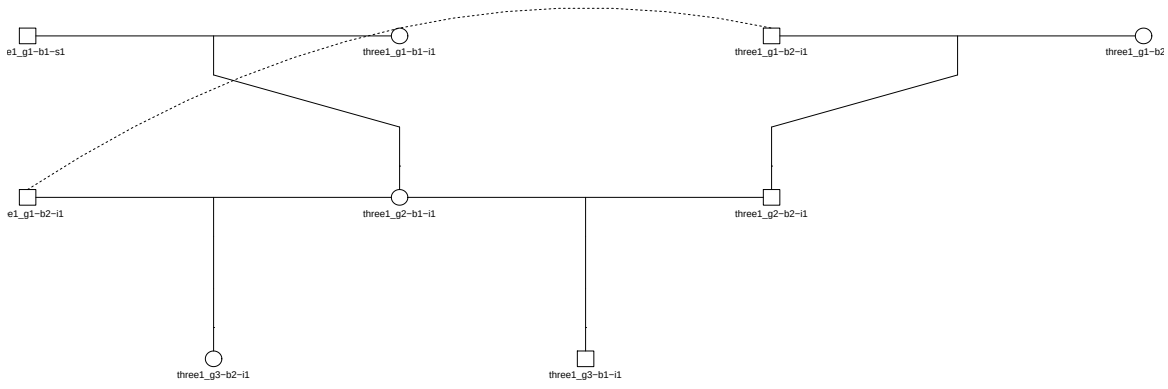

Figure S2: Pedigree simulated with ped-sim including one 3/4S pair.

Appendix A

We derive the IBD probabilities for three-quarter siblings (3/4S) in the case that a pair of individuals have one parent in common while their unshared parents are full-siblings (FS) (Figure S3). In the case that the unshared parents have a parent-offspring relationship, the IBD probabilities can be derived analogously.

Let $\delta\gamma$ be the genotype of the common parent of a 3/4S pair, and $\alpha\beta$, αB, $A\beta$ and AB the possible genotypes of a FS pair. Then, all the possible genotypes and the IBD alleles shared for a 3/4S pair are shown in Table S1.

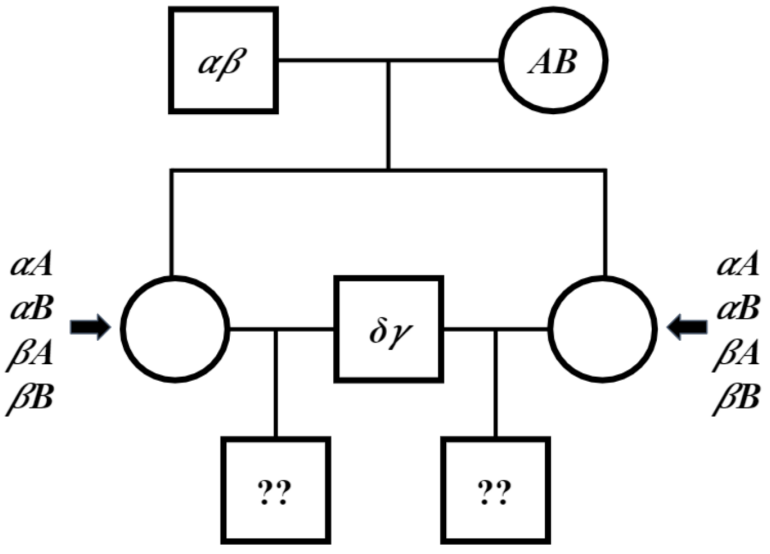

Figure S3: Pedigree of a 3/4S pair where their unshared parents are FS.

	$\alpha\delta$	$\alpha\gamma$	$A\delta$	$A\gamma$	$\beta\delta$	$\beta\gamma$	$B\delta$	$B\gamma$
$\alpha\delta$	2	1	1	0	1	0	1	0
$\alpha\gamma$	1	2	0	1	0	1	0	1
$A\delta$	1	0	2	1	1	0	1	0
$A\gamma$	0	1	1	2	0	1	0	1
$\beta\delta$	1	0	1	0	2	1	1	0
$\beta\gamma$	0	1	0	1	1	2	0	1
$B\delta$	1	0	1	0	1	0	2	1
$B\gamma$	0	1	0	1	0	1	1	2

Table S1: Number of IBD alleles for all possible pairs of 3/4S where their unshared parents are FS.

From Table S1, the IBD probabilities for 3/4S are:

$k_0 = P(IBD = 0) = \frac{24}{64} = 3/8$

$k_1 = P(IBD = 1) = \frac{32}{64} = 1/2$

$k_2 = P(IBD = 2) = \frac{8}{64} = 1/8$

And their kinship coefficient is:

$\phi = k_1/4 + k_2/2 = \frac{1}{2}\frac{1}{4} + \frac{1}{8}\frac{1}{2} = 3/16$

Appendix B

Here we show the LR of 3/4S∼UN for a biallelic SNP whose alleles are A and B. Let p and q be the allele frequencies for A and B of the population under study. For a pair of individuals, we show the LR computation for four genotype pairs: AA/AA, AA/AB, AA/BB and AB/AB. The LR for the remaining genotype pairs (AB/AA, AB/BB, BB/AA, BB/AB and BB/BB) are equivalent or can be obtained analogously.

The IBD probabilities for 3/4S are $(k_0, k_1, k_2) = (3/8,\ 1/2,\ 1/8)$ and for UN pairs are $(k_0, k_1, k_2) = (1,\ 0,\ 0)$. Then, according to the Tables 1 and 2 and Equations 1 and 2, the LR for 3/4S∼UN is derived as follows:

AA/AA case:

$$LR = \frac{\frac{3}{8}p^4 + \frac{1}{2}p^3 + \frac{1}{8}p^2}{p^4} = \frac{3}{8} + \frac{1}{2p} + \frac{1}{8p^2}$$

AA/AB case:

$$LR = \frac{\frac{3}{8}2p^3q + \frac{1}{2}p^2q}{2p^3q} = \frac{3}{8} + \frac{1}{4p}$$

451

452

453 AA/BB **case:**

$$LR = \frac{\frac{3}{8}p^2q^2}{p^2q^2} = \frac{3}{8}$$

454

455

456 AB/AB **case:**

$$LR = \frac{\frac{3}{8}4p^2q^2 + \frac{1}{2}pq + \frac{1}{8}2pq}{4p^2q^2} = \frac{3}{8} + \frac{3}{16pq}$$

457

458

Chapter 5

Results and discussion

In this chapter, we review and discuss the main results derived from this doctoral thesis. Mainly, we show that the statistical methods from Compositional Data analysis (ternary diagram visualization, isometric log-ratio transformation and log-ratio biplot) make a valuable contribution to the field of relatedness research. Furthermore, we develop a likelihood ratio approach to detect three quarter siblings in genetic databases.

5.1 Compositional graphics for relatedness research

To illustrate the compositional approach for relatedness research, we consider 377 microsatellite markers genotyped for 25 individuals from the Maya population of the HGDP-CEPH diversity panel described in Section 3.2.3. Using this genetic database, we study and review the usual graphics in allele sharing studies based on identity by state/descent alleles.

Regarding IBS studies (section 3.2.1), Figures 5.1 (a) and (b) show the scatterplot of means ($\bar{x}$) and standard deviations (s) of the IBS alleles and the scatterplot of the proportion of sharing 0 and 2 IBS alleles (p_0, p_2, respectively). Regarding IBD studies (section 3.2.2), Figure 5.2 (a) shows the scatterplot of the estimated probabilities of sharing 0 and 1 IBD alleles ($\hat{k}_0, \hat{k}_1$, respectively). Hereafter, we refer to these graphics as $(\bar{x}, s)$, (p_0, p_2) and $(\hat{k}_0, \hat{k}_1)$-plots respectively. The pairs of individuals are colored according to the family relationships that were reported by Rosenberg (2006) and inferred using the RELPAIR program (Boehnke and Cox, 1997; Epstein *et al.*, 2000). To confirm these reported relationships, we plot convex hulls of simulated pairs of individuals with known artificial pedigrees and therefore we evaluate if each relationship falls into his expected simulated convex hull.

Instead of using the (p_0, p_2) and $(\hat{k}_0, \hat{k}_1)$-plots, Sun (2012) uses the (p_0, p_1)-plot and Moltke and Albrechtsen (2014) use the $(\hat{k}_1, \hat{k}_2)$-plot. In fact, any combination of the three IBS/IBD probabilities could be plotted for relatedness research. We refer to these combinations of IBS/IBD graphics as (p_i, p_j) and $(\hat{k}_i, \hat{k}_j)$-plots (for $i, j = 0, 1, 2$ and $i < j$) where p_i and $\hat{k}_i$ correspond to the X-axis of each plot and p_j and $\hat{k}_j$ to the Y-axis.

We propose the use of the ternary diagram (section 3.1.1) as an alternative of the (p_i, p_j) and $(\hat{k}_i, \hat{k}_j)$-plots in order to represent the full set of IBS or IBD probabilities. In fact, the ternary diagram takes into account the compositional nature of the allele sharing data and is able to represent the three IBS/IBD probabilities simultaneously as is shown in Figures 5.1 (c) and 5.2 (b).

However, the main limitation of the $(\bar{x}, s)$, (p_i, p_j) and $(\hat{k}_i, \hat{k}_j)$-plots and the ternary diagram is that they occupy a constrained space and the Euclidean distance interpretation in these graphics

is not adequate as described in section 1.2. Specifically, the $(\bar{x}, s)$-plot is constrained by the "umbrella" shaped space as is shown in Figure 5.1 (a). This "umbrella" space is obtained by simulating pairs of individuals and genetic markers containing all the possible combinations of 0, 1 and 2 IBS alleles. Otherwise, the (p_0, p_2) and $(\hat{k}_0, \hat{k}_1)$-plots are constrained by the unit sum ($p_0 + p_1 + p_2 = 1$, $k_0 + k_1 + k_2 = 1$) which is represented by the line $y = 1 - x$ as is shown in Figures 5.1 (b) and 5.2 (a). Despite the $(\bar{x}, s)$-plot does not have the unit sum constraint characteristic of Compositional Data, we found that both $(\bar{x}, s)$ and (p_0, p_2) statistics are related by the equations: $\bar{x} = 1 - p_0 + p_2$ and $s = \sqrt{p_0(1-p_0) + p_2(1-p_2) + 2p_0p_2}$.

To overcome the limitation of the Euclidean distance interpretation in the constrained classical graphics and the ternary diagram, we propose to use the isometric log-ratio transformation of the IBS and IBD probabilities. In this way, Euclidean distances in the log-ratio transformed space are the same as the Aitchison distances defined in the simplex (principle of working in coordinates, (Mateu-Figueras *et al.* (2011), sections 1.2, 3.1.3 and 3.1.5). Figures 5.1 (d) and 5.2 (c) show the isometric log-ratio transformation of the IBS and IBD probabilities defined as: $\left(z_{11} = \frac{1}{\sqrt{2}} \ln\left(\frac{p_2}{p_0}\right), z_{12} = \frac{1}{\sqrt{6}} \ln\left(\frac{p_0 p_2}{p_1^2}\right)\right)$ and $\left(z_{11} = \frac{1}{\sqrt{2}} \ln\left(\frac{\hat{k}_2}{\hat{k}_0}\right), z_{12} = \frac{1}{\sqrt{6}} \ln\left(\frac{\hat{k}_0 \hat{k}_2}{\hat{k}_1^2}\right)\right)$ respectively.

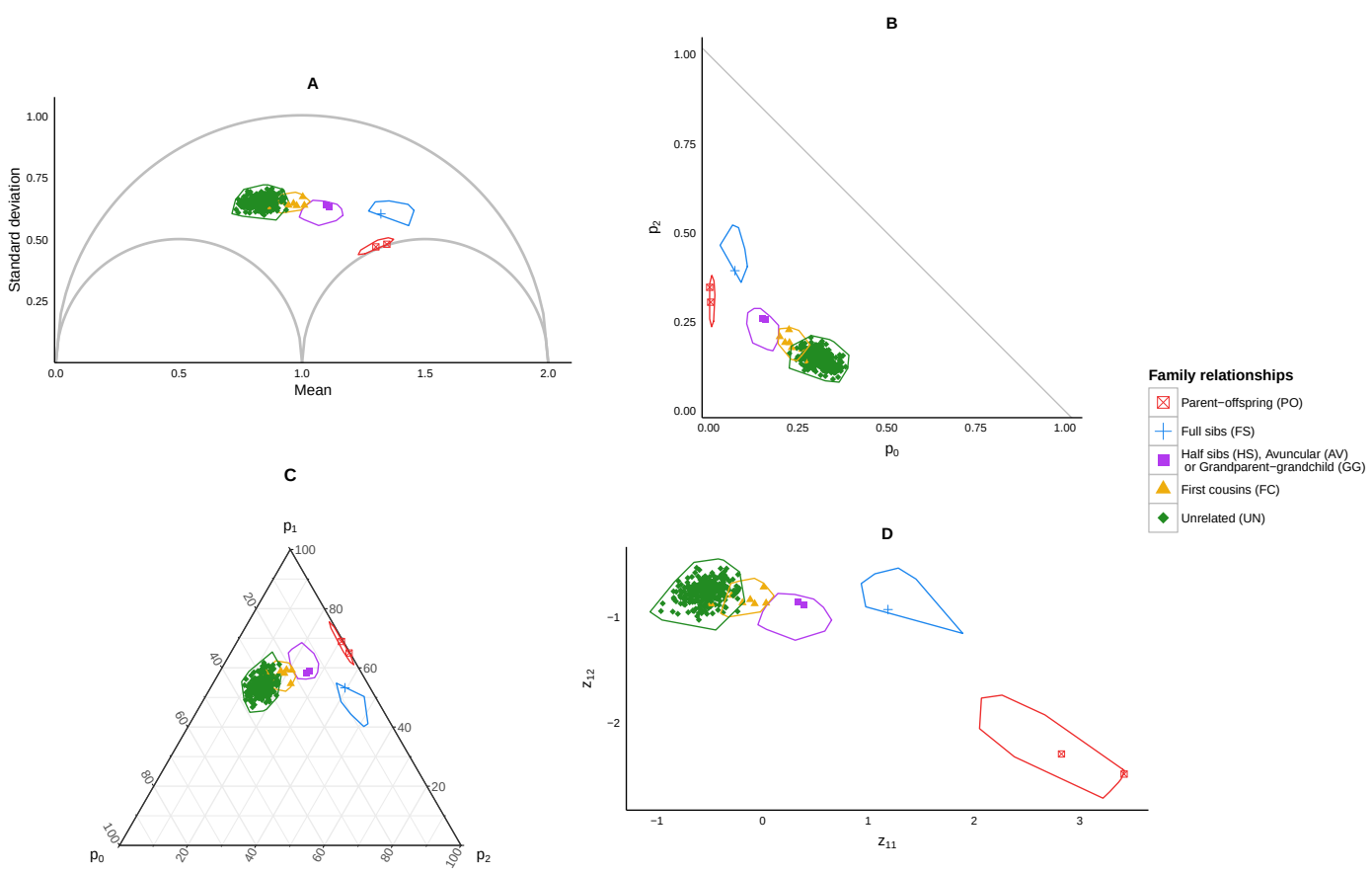

Figure 5.1: Identical by state (IBS) alleles for all the pairs of individuals from the Maya population. (a) Plot of means versus standard deviations. (b) (p_0, p_2)-plot. (c) Ternary diagram. (d) Ilr-coordinates: (z_{11}, z_{12}). The convex hulls are obtained by simulating artificial children from a subset of unrelated individuals from the Maya population.

Regarding IBS graphics, the $(\bar{x}, s)$-plot (Figure 5.1, a) shows two PO pairs located close to the gray curve with the smallest standard deviations. First and second degree relationships have

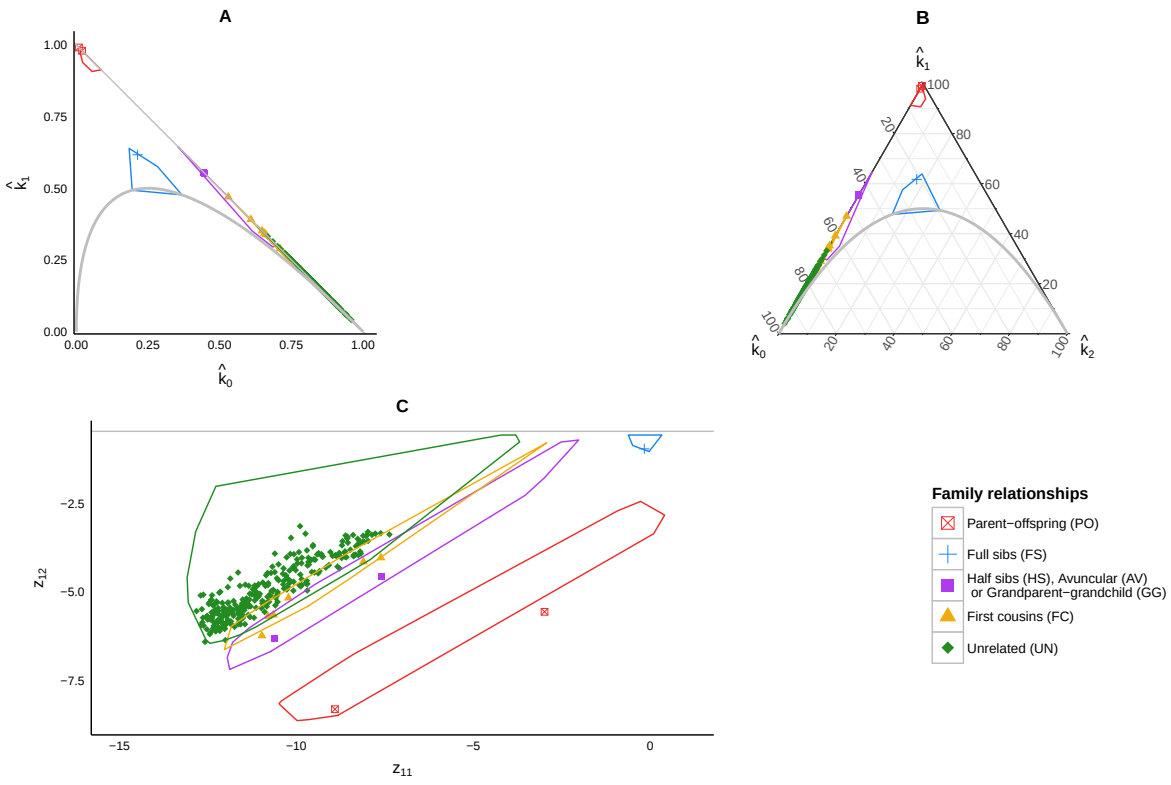

Figure 5.2: Identical by descent (IBD) alleles for all the pairs of individuals from the Maya population. (a) $(\hat{k}_0, \hat{k}_1)$-plot. (b) Ternary diagram. (c) Ilr-coordinates: (z_{11}, z_{12}).

a mean above 1. The (p_0, p_2)-plot (Figure 5.1, b) shows the two PO pairs close to the p_2-axis and the FS pair have the larger p_2 values. In the ternary diagram (Figure 5.1, c), PO pairs are points on the opposite side of the vertex p_0, meaning that the p_0 is close to 0. The FS pair is the point closest to the vertex p_2, which has the largest p_2. In Figure 5.1 (d), the first ilr-coordinate (z_{11}) clearly discriminates first-degree relatives from UN pairs. Pairs with larger values for z_{11} are more likely to correspond to related individuals. PO pairs are extreme outliers because they have p_0 values close to 0 which increase the first coordinate of the corresponding log-ratio. The scatterplot of the log-ratios is seen to produce a larger degree of separation between FS and PO pairs, and between first-degree pairs and all other pairs. The convex hulls for the simulated related pairs in Figure 5.1 are seen to enclose the sample estimates of the PO, FS, HS and FC pairs and so confirm the assigned relationships.

Regarding IBD graphics, the $(\hat{k}_0, \hat{k}_1)$-plot (Figure 5.2, a) separates the first, second and some pairs of third degree of relationship. In the ternary diagram (Figure 5.2, b), it is easy to identify PO pairs at the vertex of $\hat{k}_1$, a FS pair close to the barycenter of the triangle and other family relationships of second degree on the opposite side of the $\hat{k}_2$ vertex. UN pairs are on the $\hat{k}_0 - \hat{k}_1$ edge and tend towards the $\hat{k}_0$ vertex. Third-degree pairs are mixed with unrelated individuals. In Figure 5.2 (c), the pairs with a close family relationship tend to have larger values of z_{11}. The ilr-coordinates clearly separates out FS and PO relationships from all other pairs. Notice that Figures 5.2 (a) and (b) show only one pair with a second degree relationship (the violet point), whereas in Figure 5.2 (c), there are two visible violet pairs.

The IBS and IBD graphics (Figures 5.1 and 5.2) were amplified with convex hulls of artificially generated pedigrees to show the approximate expected positions for the different relationships. These hulls mainly confirm the assigned relationships. However, in ilr-coordinates, PO hulls do

not capture all observed PO pairs. The accuracy of the convex hulls depends on the sample size, the number of genetic markers, and in particular on the number of UN individuals in the sample from which it is generated. More accurate convex hulls may be obtained if linkage disequilibrium is taken into account and artificial pairs are generated by sampling from haplotypes instead of by sampling individual markers independently. Furthermore, the position of a PO pair in the ilr-coordinates of the IBD probabilities has a high variability and depends on the tolerance and initial point used in the maximization of the likelihood (Graffelman and Galván-Femenía, 2016). If the sample size is small, or the number of simulated pairs is small, the PO hull many not cover the full area compatible with PO pairs. It is worth remarking that PO and FS convex hulls do not intersect each other and do not overlap with the rest of the hulls, having a valuable discriminative power. We think the current simulated convex hulls are helpful to assess uncertainty but of limited value and see a clear need for methods of formal statistical inference on relationships by means of hypothesis testing and confidence regions (García-Magariños *et al.*, 2015).

Everything discussed during this section represents the content of the article: Galván-Femenía, I., Graffelman, J. & Barceló-i-Vidal, C. (2017). **Graphics for relatedness research**. *Molecular Ecology Resources*, 17(6), 1271-1282. doi:10.1111/1755-0998.12674.

Once we introduced the compositional approach of the classical IBS/IBD graphics for relatedness research, we deepened our understanding of the nature of the allele sharing data. We are aware that taking into account genotype sharing information instead of allele sharing information, we can interpret data with higher dimensionality. Specifically, the genotype sharing data has six counts instead of the allele sharing data that has only three counts. We explore the genotype sharing data by using log-ratio principal component analysis (PCA) (section 3.1.7). Furthermore, we introduce statistical inference of family relationships using linear discriminant analysis (LDA) (Johnson *et al.*, 2002) in the log-ratio biplot space. The training set used for linear discriminant analysis is composed of artificial pedigrees generated from the population under study (section 3.2.4). This approach is presented in the following section and is part of the contents of the article: Graffelman, J., Galván-Femenía, I., de Cid, R. & Barceló-i-Vidal, C. (2019). **A log-ratio biplot approach for exploring genetic relatedness based on identity by state**. *Frontiers in Genetics*, 10, 341. doi:10.3389/fgene.2019.00341.

5.2 Log-ratio biplot for relatedness research

To illustrate the log-ratio biplot approach for relatedness research, we consider simulated data and individuals from the GCAT Genomes for Life Cohort study of the Genomes of Catalonia (section 3.2.3).

In this approach, we consider genotype sharing data instead of classical allele sharing data (section 3.2). Briefly, given bi-allelic variants with alleles A and B, we consider six possible pairs of genotypes whose counts over k variants can be laid out in a triangular array shown in Table 5.1. The counts k_{ij} refers to the number of variants that have i B alleles for one individual, and j B alleles for the other individual.

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

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

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

Table 5.1: Lower triangular matrix layout with counts for all possible genotype pairs.

The total number of variants is given by $k = \sum_{i \geq j} k_{ij}$. Given n individuals, we construct matrix X with $q = \frac{1}{2}n(n-1)$ pairs in its rows, and propose to study relatedness by a log-ratio PCA of this q x 6 matrix of compositions (section 3.1.7). This will allow the construction of a biplot, where each pair of individuals is represented by a point, and each part of the clr transformed composition by a vector. A limitation of the representation of pairs of individuals in a log-ratio PCA biplot is that the type of relationship cannot be inferred if it is undocumented. Without additional analysis one does not know for sure whether observed clusters correspond to FS, PO, or other pairs. We resolve this by first identifying a subset of approximately unrelated individuals in the database, having a kinship coefficient with other individuals that is below 0.05. We next simulate pairs of related individuals of known relationships by constructing artificial pedigrees from this subset, applying Mendelian inheritance rules (section 3.2.4). The artificially generated data set forms a reference set against which the empirically observed data can be compared. This reference set is generated conditionally on the allele frequencies of the observed sample. We now first apply log-ratio PCA to the pairs of the reference set (X), and construct a biplot of the reference set. The empirically observed pairs (Y) are projected onto this PCA biplot (section 3.1.7) 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 (LDA) (Johnson *et al.*, 2002), using the simulated pairs as a training set.

Following the former approach, we simulated 35,000 independent genetic bi-allelic variants by sampling from a multinomial distribution under the Hardy-Weinberg assumption, using a minor allele frequency (MAF) of 0.5 for all variants. Considering Mendelian inheritance rules, 100 independent pairs of each type of relationship were simulated (section 3.2.4). Using this simulated data, we apply the log-ratio PCA approach and depict the log-ratio biplot jointly with the classical $(\bar{x}, s)$, (p_0, p_2) and $(\hat{k}_0, \hat{k}_1)$ plots in Figure 5.3. This Figure shows that first and second degree pairs are easily identified by all methods and consequently, using linear discriminant analysis, we obtain a classification rate of these groups equals to 1. For this reason, we focus on third and higher degree relationships which are harder to distinguish as they tend to blur in the plots. Therefore, we investigated the effect of the number of SNPs used for the classification rate of the artificial simulated relationships of third through sixth degree pairs (100 of each). Figure 5.4 shows the classification rate as a function of the number variants with MAF 0.50 for the four aforementioned methods. These classification rates were obtained by averaging over 25 replicates of the simulations, for each number of variants. It is clear that the log-ratio PCA approach gives the best classification rates for all relationships. There is little difference in classification rate for third degree relationships, which are relatively more easy to classify. As expected, classification rate increases with the number of variants. The results suggest that for all four methods 25,000 variants with MAF 0.50 are sufficient to almost perfectly classify PO, FS, second, third, and fourth degree relationships. The difference in classification rate between the log-ratio PCA approach and the conventional methods is larger for the more remote relationships. 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.

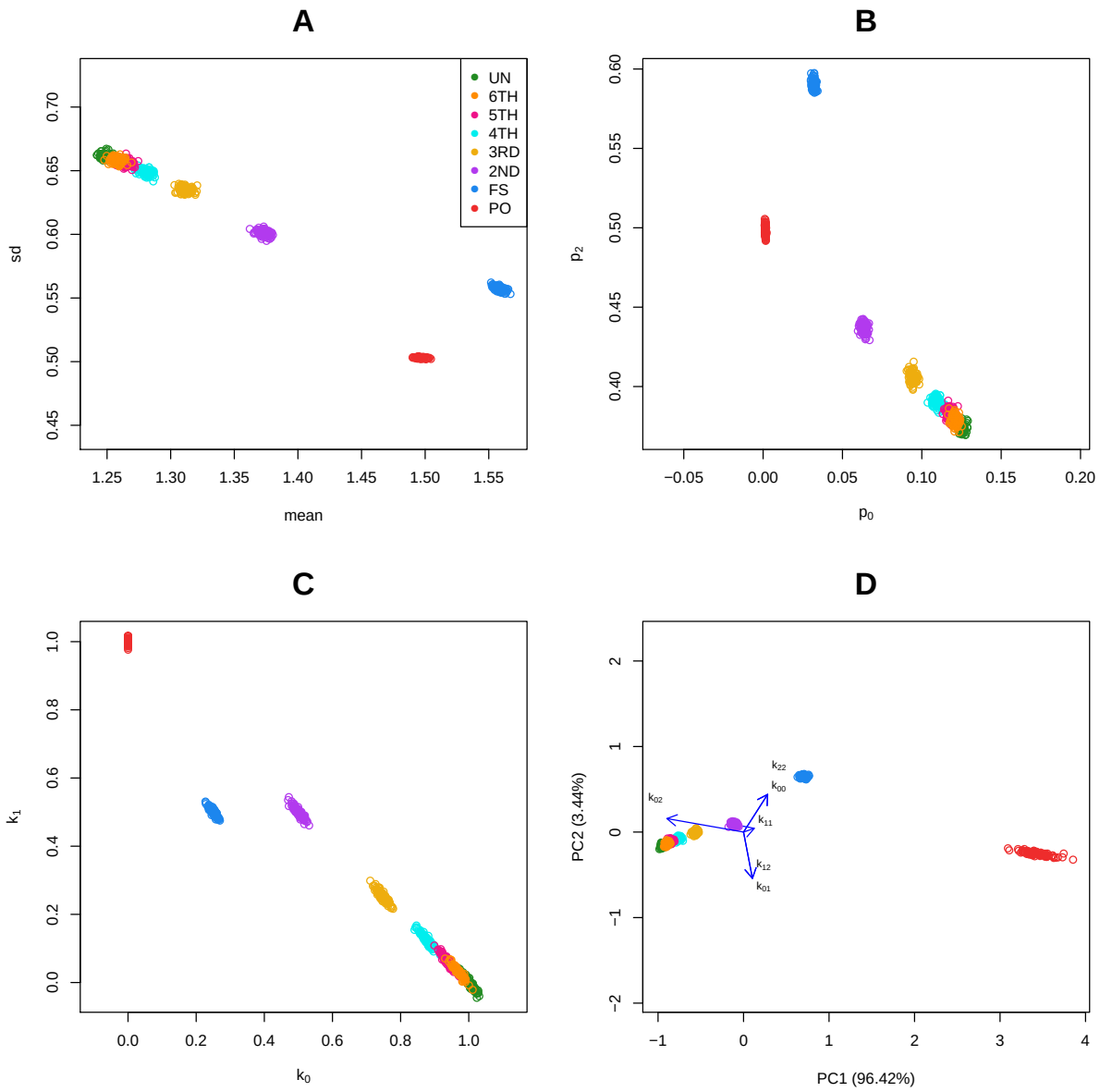

Figure 5.3: Classical graphics and log-ratio PCA biplot for simulated samples. 100 pairs of each type of relationship (UN, sixth, fifth, fourth, third, second, FS and PO) were generated using 35,000 biallelic variants with minor allele frequencies of 0.5, assuming Hardy Weinberg equilibrium. (a) $(\bar{x}, s)$-plot. (b) (p_0, p_2)-plot. (c) (k_0, k_1)-plot. (d) Log-ratio PCA biplot.

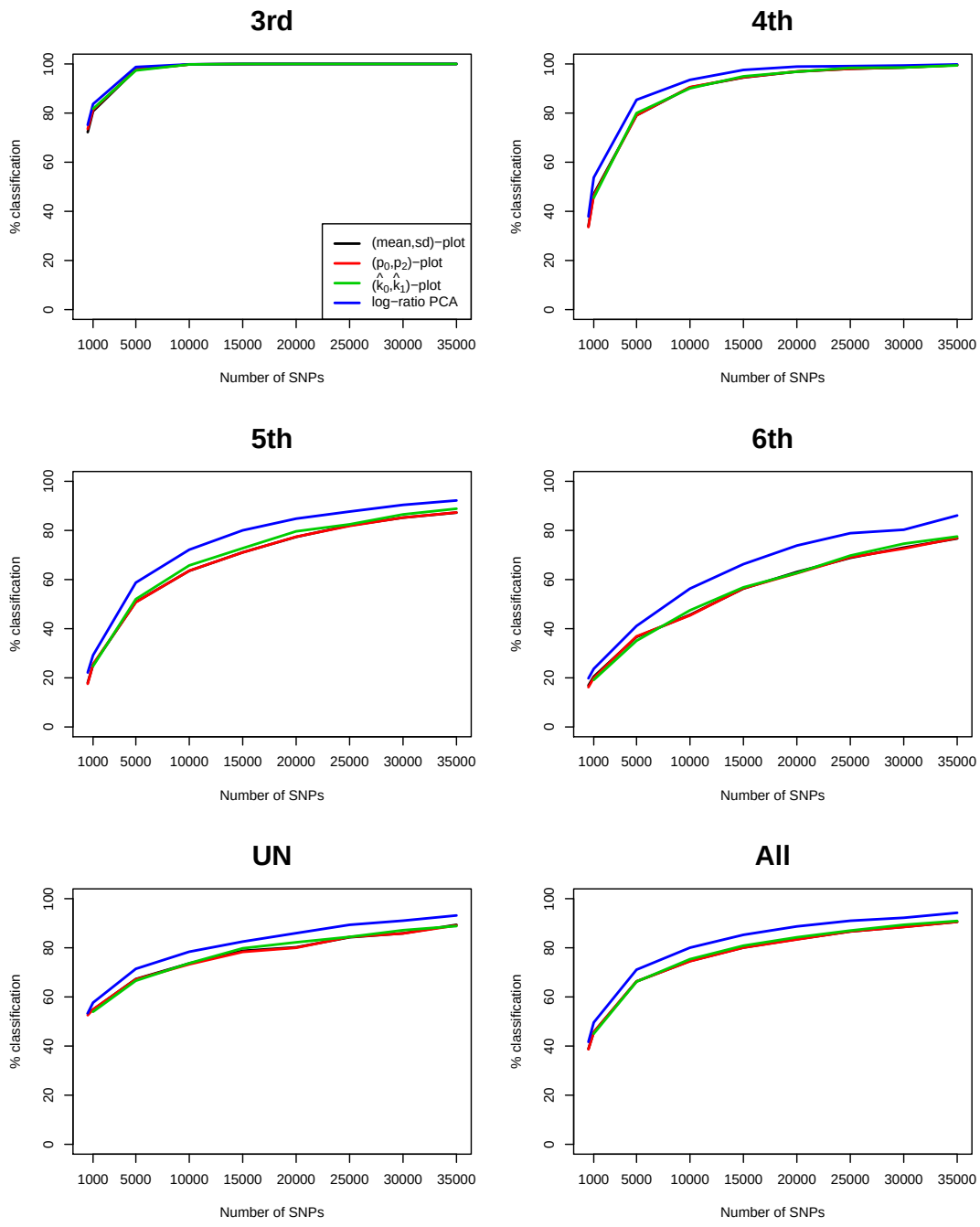

Figure 5.4: Classification rates for different methods vs. number of SNPs. Classification rates for the different degrees of relationship (third, fourth, fifth, sixth, UN, and All) are shown for four methods. Classification rate profiles for the $(\bar{x}, s)$-plot and the (p_0, p_2)-plot virtually coincide. The last panel All refers to the classification rate for third through UN relationships jointly. Rates are shown as a function of the number of SNPs with MAF 0.50, and were obtained by linear discriminant analysis. 100 pairs of each type of relationship were generated assuming Hardy-Weinberg equilibrium.

We therefore show an application of the log-ratio biplot approach in the GCAT Genomes for life project, a prospective cohort study of the genomes of Catalonia (`www.genomesforlife.com/en_index`). GCAT 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 of 5,075 GCAT self-reported Spanish participants from Caucasian origin using 736,223 SNPs that passed quality control (Galván-Femenía *et al.*, 2018). Inferred relatives of first and second degree were confirmed by the BST public agency, for pairs sharing one surname (PO, second degree pairs) or two surnames (FS pairs), respecting the privacy of the participants. Variants were filtered according to missingness (only variants genotyped for all individuals were used), MAF (> 0.40) and Hardy-Weinberg equilibrium test result (exact test mid p-value$>$0.05, Graffelman and Moreno (2013)). Variants were LD-pruned with PLINK software using a sliding window of 50 SNPs with an overlap of 5 SNPs between successive windows, and SNPs are removed from the window until no variants remain that have a squared correlation above 0.20 (PLINK option `indep-pairwise 50 5 0.2`). After applying all these filtering criteria, a total of 26,006 SNPs remained for relatedness analysis. Log-ratio PCA biplots representing over twelve million pairs, combined with the classification of the individuals by LDA are shown in Figure 5.6. This analysis shows the different relationships have in general, a larger variability than expected according to the simulated pairs. The FS cluster has a particular high variability, with pairs apparently less related than FS, and pairs stronger related than FS, in comparison with the FS hull. One apparent FS pairs is actually classified as second degree (Figure 5.6 (a)). This fusion of FS and second degree pairs suggested us that three-quarter siblings might exist in the database and we therefore re-analyzed the data using a training set that included three-quarter siblings. Three-quarter siblings (3/4S) share more IBD alleles than second degree pairs but fewer than FS. 3/4S have one common parent, while their unshared parents can be FS or PO (see Figure 5.5). Three-quarter siblings have IBD probabilities $k_0 = 3/8$, $k_1 = 1/2$, and $k_2 = 1/8$, such that their kinship coefficient is $\phi = 3/16$, below the value $\phi = 1/4$ of full siblings (Table 3.2). In the re-analysis in Figure 5.6 (b), we found 63 FS pairs, 12 2nd pairs, and eight pairs were indeed classified as three-quarter siblings with large posterior probability. Two of these pairs had their kinship coefficient very close to the expected value of $\phi = 3/16$. Because Spanish people have both paternal and maternal surnames, three-quarter siblings share both surnames just as siblings do. The pairs classified as 3/4 siblings shared indeed both surnames, confirming these pairs are actually not second degree. Peeling siblings and three-quarter siblings reveals apparent second degree pairs more clearly (Figure 5.6 (c)). Tentatively peeling second degree pairs brings the third degree pairs in focus (Figure 5.6 (d)), and in this analysis we find 174 third, 66 fourth, 31 fifth, and 3,517 sixth degree pairs. Further peeling is difficult as the different clusters increasingly merge.

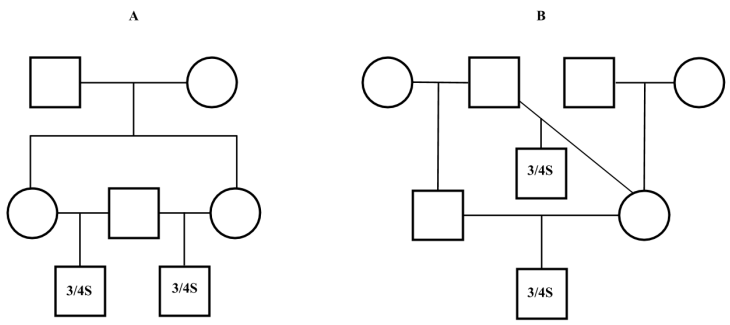

Figure 5.5: (a) Pedigree of a 3/4S where their unshared parents are FS. (b) Pedigree of a 3/4S where their unshared parents are PO.

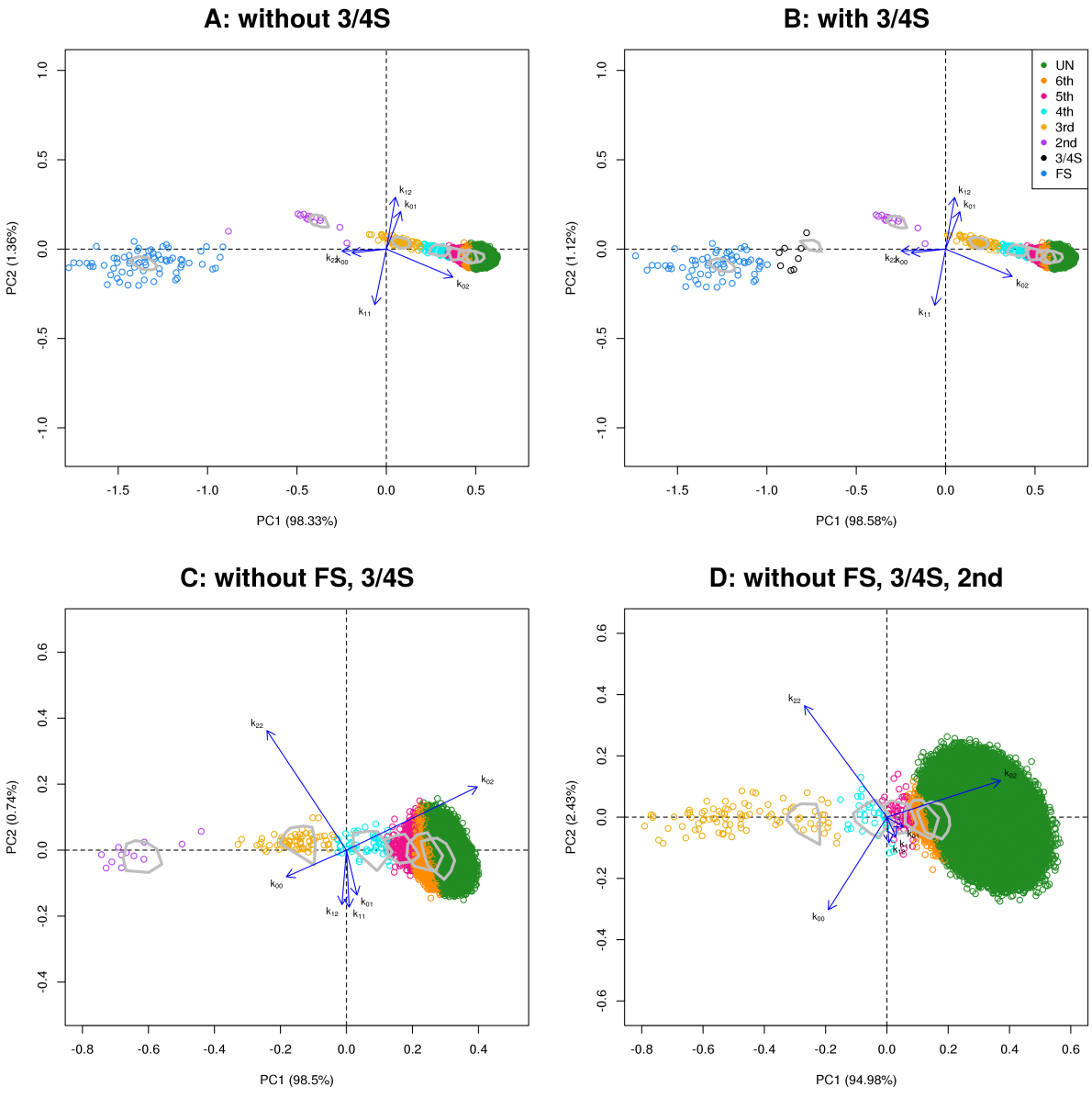

Figure 5.6: Log-ratio PCA biplot of GCAT sample obtained by peeling and zooming. (a) log-ratio PCA biplot, PO and 3/4S pairs excluded. (b) 3/4S pairs included; (c) FS and 3/4S pairs excluded; (d) FS, 3/4S, and second degree pairs excluded. Convex hulls delimit the region of the pairs obtained by simulation.

The simulated reference data set used in the log-ratio PCA biplot (Figure 5.6) was obtained by resampling genetic variants independently, and this does not take linkage disequilibrium (LD) and recombination into account. If haplotype data is estimated, a biologically more realistic simulated data set can be obtained by sampling haplotypes instead of genotypes. We have avoided this issue by LD pruning the data base prior to resampling, so removing tightly correlated markers. The reference data set is therefore constructed on the basis of a subset of variants that can expected to be approximately independent. This subset is then used as the basis for relationship estimation. This procedure has the advantage that it avoids potential additional uncertainty generated by the haplotype estimation algorithm. However, the proposed procedure may be improved in the future by accounting for haplotype structure and recombination. The pruning threshold used in our method (0.20) is a compromise between precision and satisfying the independence assumption. A larger value will admit more variants and can increase the resolution, but due to correlation between variants it will invalidate the independence assumption used to generate the reference set of related pairs.

The log-ratio PCA biplot approach is focused on homogeneous populations. If population substructure exists, then log-ratio PCA can be expected to separate the different populations in its biplot. Methods that address substructure and family relationships jointly have been developed (Manichaikul *et al.*, 2010; Conomos *et al.*, 2016). Population substructure can be accounted for by using only variants with low weights on the first components for a relatedness analysis, as is done in the UK Biobank project (Bycroft *et al.*, 2018), as the first components mostly capture substructure. In future work, the usefulness of the log-ratio PCA approach for the joint study of remote and recent relatedness could be further explored.

The analysis of the GCAT samples suggested that eight three-quarter siblings pairs exist in this database. For this reason, we develop an additional statistical methodology such as the likelihood ratio approach in order to confirm these eight three-quarter siblings pairs. This approach is presented in the following section and is part of the content of the article: Galván-Femenía, I., Barceló-i-Vidal, C., Sumoy, L., Moreno, V., de Cid, R. & Graffelman, J. (2020). **A likelihood ratio approach for identifying three quarter siblings in genetic databases**. *Heredity*. Submitted.

5.3 Likelihood ratio approach for identifying three quarter siblings

The likelihood ratio (LR) approach has been widely used for relatedness research during the last decades (Thompson, 1986; Boehnke and Cox, 1997; Weir *et al.*, 2006; Katki *et al.*, 2010; Heinrich *et al.*, 2016). Briefly, the LR approach is based on the contrast of two hypotheses, one in the numerator, H_i; and the other one in the denominator, H_j. The larger the LR, the more plausible is H_i; whereas the smaller the LR, the more plausible is H_j. For relatedness research, we consider the ratio of the probabilities from Equation 3.1 according to the hypothesis of the R_i and R_j relationships. That is:

$$LR(R_i, R_j|G_1/G_2) = \frac{P(G_1/G_2|R_i)}{P(G_1/G_2|R_j)}. \tag{5.1}$$

Here we focus on the FS, 3/4S, 2nd degree and UN relationship categories. We calculate three LRs having FS, 3/4S or 2nd in the numerator and having the UN relationship in the denominator. The common denominator makes the LR values comparable in order to distinguish 3/4S from FS and 2nd degree. Inference of relatedness for each pair of individuals is based on the largest LR value in the FS∼UN, 3/4S∼UN and 2nd∼UN ratios. The LR are shown in Table 5.2, depending

on the observed genotypes of a pair of individuals. Most of these ratios are derived in Heinrich *et al.* (2016). For S biallelic SNPs, the LR can be obtained by multiplying the LR_s across markers and by dividing by the number of SNPs. It is convenient to work in a logarithmic scale such that:

$$\log_{10}(LR) = \frac{1}{S}\log_{10}\left(\prod_{s=1}^{S} LR_s(R_i, R_j|G_1/G_2)\right) = \frac{1}{S}\sum_{s=1}^{S}\log_{10}\left(LR_s(R_i, R_j|G_1/G_2)\right). \tag{5.2}$$

LR	AA/AA	AA/AB	AB/AB	AA/BB
FS∼UN	$\frac{1}{4}+\frac{1}{2p}+\frac{1}{(2p)^2}$	$\frac{1}{4}+\frac{1}{4p}$	$\frac{1}{4}+\frac{1}{4pq}$	$\frac{1}{4}$
3/4S∼UN	$\frac{3}{8}+\frac{1}{2p}+\frac{1}{8p^2}$	$\frac{3}{8}+\frac{1}{4p}$	$\frac{3}{8}+\frac{3}{16pq}$	$\frac{3}{8}$
2nd∼UN	$\frac{1}{2}+\frac{1}{2p}$	$\frac{1}{2}+\frac{1}{4p}$	$\frac{1}{2}+\frac{1}{8pq}$	$\frac{1}{2}$

Table 5.2: Likelihood ratio (LR) for relatedness research for biallelic SNPs. The considered LR are FS, 3/4S, 2nd relationships in the numerator and the UN relationship in the denominator. The LR values depend on the observed genotypes of a pair of individuals and the allele frequencies p and q of the population under study. We assume that the order of the genotypes is irrelevant, i.e. the LR for G_1/G_2 and G_2/G_1 is the same.

Therefore, according to the Equation 5.2, we calculate the FS∼UN, 3/4S∼UN and 2nd∼UN ratios for the FS, 3/4S and 2nd degree pairs detected in the GCAT cohort using the log-ratio biplot PCA approach (Figure 5.6). Figure 5.7 shows the LR ratio values and reveals eight 3/4S pairs (black color) that have the larger than sign shaped (“>”) pattern. All inferred FS pairs (blue color) have the monotonously increasing “/” shaped pattern and all 2nd degree pairs have the monotonously decreasing “\” pattern. All the relationships are colored and inferred according to the relationship category of the numerator of the largest LR value. This inference coincide with the log-ratio PCA biplot based on LDA and we confirm the existence of eight 3/4S pairs in the GCAT database.

The main motivation of using the proposed LR approach instead of the classical $(\hat{k}_0, \hat{k}_1)$-plot is that 3/4S can easily go unnoticed as FS pairs. Figure 5.8 shows the $(\hat{k}_0, \hat{k}_1)$-plot of the GCAT cohort, the pairs of individuals are colored according to the LR inference approach. It can be shown that 3/4S pairs are close to the FS pairs and can be misclassified as FS. Thus, the LR approach can be of great help to identify such cases.

The proposed LR approach multiplies the likelihoods over loci (Equation 5.2), under the assumption of independence. The existence of linkage disequilibrium (LD) between variants violates this assumption. In order to approximately meet the requirement of independence, we LD pruned neighbouring variants in a window using the PLINK software. However, independence is not always satisfied for related individuals since variants are highly correlated due to the biological inheritance patterns. Further improvement of the LR calculation can be achieved if LD and genetic recombination maps are taken into account.

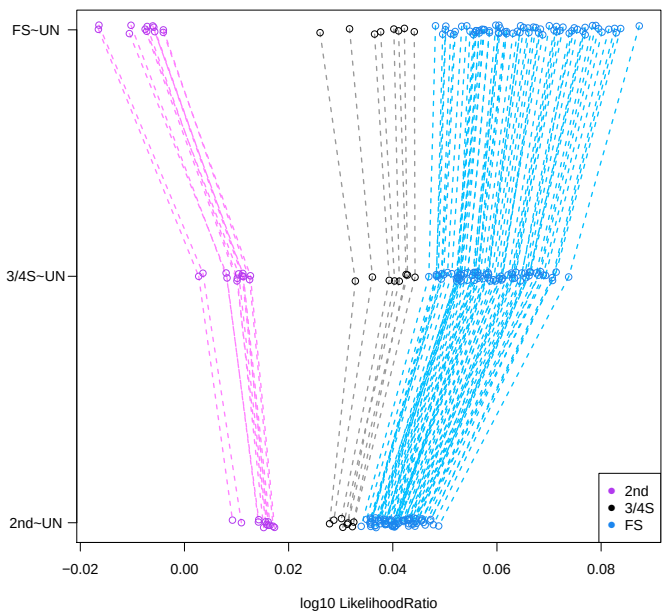

Figure 5.7: Log10 likelihood ratio approach of the presumably 2nd, 3/4S and FS pairs from the GCAT cohort using 26,006 SNPs (MAF $>$ 0.40, LD-pruned, HWE exact mid p-value $>$ 0.05, and missing call rate 0).

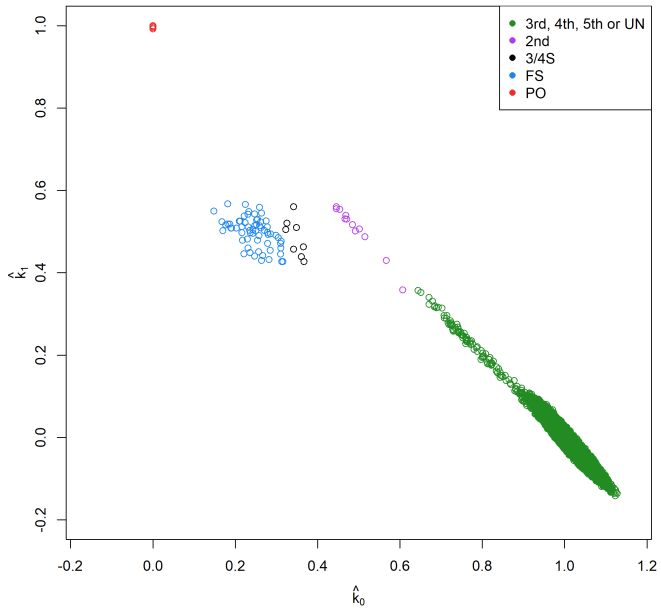

Figure 5.8: $(\hat{k}_0, \hat{k}_1)$-plot of the GCAT cohort for 5,075 individuals and 26,006 SNPs (MAF $>$ 0.40, LD-pruned, HWE exact mid p-value $>$ 0.05, and missing call rate 0). UN: unrelated; 5th, 4th, 3rd, 2nd: fifth, fourth, third, and second degree relationships; 3/4S: three quarter siblings; FS: full siblings; PO: parent-offspring.

5.4 Conclusions

The statistical methods from the field of Compositional Data used in the first and second work of this doctoral thesis have enriched the classical methods used for relatedness research.

The first article shows that the identity by state/descent probabilities can be formally defined in the simplex. These probabilities can be considered as 3-part compositions and can be represented in ilr-coordinates. This approach allows to interpret the distances between the different family relationships properly.

The second article of this thesis shows that identity by state has higher dimensionality if the analyst take into account genotype sharing data instead of allele sharing data. Therefore, genotype sharing data can be considered as 6-part compositions. For this reason, we explore these data using log-ratio PCA biplots. This approach jointly with linear discriminant analysis shows accurate inferences up to fourth degree relationships.

The log-ratio PCA biplot uncovered the detection of three-quarter siblings. Thus, this thesis finishes with a third work where an additional statistical method based on likelihood ratios is developed in order to identify three-quarter siblings relationships.

5.5 Further research

During the development of this thesis, new lines of research arose to improve the methodology presented in this compendium of three research articles.

- All the three articles are supported with simulations of artificial pedigrees. However these simulations do not take into account linkage disequilibrium (LD) and genetic recombination. This can explain the large variability of the empirical FS pairs in the simulated FS convex hull of the log-ratio PCA biplots. Therefore, the simulations can be made more realistic by accounting for both biological phenomena. Simulations taking into account haplotype estimation algorithms such as SHAPEIT (Delaneau *et al.*, 2012) can overcome this limitation.

- Additional applications to other genetic databases, such as the UK Biobank (Bycroft *et al.*, 2018), the Qatari Genome (Fakhro *et al.*, 2016) and the San Antonio Mexican American Family Studies (SAMAFS) (Ramstetter *et al.*, 2017), will enrich the statistical methodology described in this doctoral thesis. The first database is considered for having almost half million of individuals, the second for being an endogamous population and the third for having confirmed family relationships and a benchmarking of 12 state-of-the-art pairwise relatedness methods. We think that it is worth to apply the log-ratio biplot methodology and to analyze the existence of three-quarter siblings in all three databases. Applications of the log-ratio biplot methodology to these new datasets will likely give further insight in its possibilities and limitations for relatedness research.

- All the three approaches presented in this thesis are illustrated in homogeneous populations. In the presence of population substructure, statistical approaches such as removing non-ancestry genetic markers prior to relatedness research can be considered to deal with this phenomenon (Bycroft *et al.*, 2018). This thesis has shown that the log-ratio approach is able to detect related individuals in genetic databases. Most likely, it will also be useful to detect population substructure (non-homogenous samples).

- In the second article of this thesis, we represent the 6-part composition of the genotype sharing data in the log-ratio biplot space. Consequently, the statistical inference of family relationships is based on linear discriminant analysis in these log-ratio biplot space. Our hypothesis is that the inference can be done directly in the linear discriminant space. That is, it is not necessary to represent previously the data in the log-ratio biplot PCA. It could lead a better separation of the simulated convex hulls of artificial pedigrees.

- In the third article, LD and genetic recombination can be also taken into account for the calculation of likelihood ratios, since there is a strong assumption of independent genetic markers. We LD-pruned genetic variants to approximately accomplish with this assumption. However, the independence of genetic markers is not always satisfied by LD-pruning in the presence of related individuals.

- The identity by descent probabilities considered in the likelihood ratio approach are based on the assumption of non-inbreeding populations. The three IBD probabilities can be extended to nine probabilities (Jacquard coefficients, Jacquard (1974)) and therefore accounting also for inbreeding populations (Milligan, 2003).

- The analysis of the GCAT database reveals the existence of three-quarter siblings, a family relationship of degree between first (full siblings) and second degree (half siblings). The log-ratio biplot approach also detects some pairs in-between a second and third degree relationship. The presence of a non discrete degree of relationships in human populations tentatively suggests that the kinship coefficient is a continuous rather than a discrete concept.

Bibliography

1000 Genomes Project Consortium *et al.* (2015). A global reference for human genetic variation. *Nature*, **526**(7571), 68.

Abecasis, G. R., Cherny, S. S., Cookson, W., and Cardon, L. R. (2001). GRR: graphical representation of relationship errors. *Bioinformatics*, **17**(8), 742–743.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **44**(2), 139–160.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, **70**(1), 57–65.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall, London, UK.

Aitchison, J. (1992). On criteria for measures of compositional difference. *Mathematical Geology*, **24**(4), 365–379.

Aitchison, J. and Greenacre, M. (2002). Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **51**(4), 375–392.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., and Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Mathematical Geology*, **32**(3), 271–275.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, **5**(9), 1564.

Barceló-Vidal, C. and Martín-Fernández, J.-A. (2016). The mathematics of compositional analysis. *Austrian Journal of Statistics*, **45**(4), 57–71.

Boehnke, M. and Cox, N. J. (1997). Accurate inference of relationships in sib-pair linkage studies. *The American Journal of Human Genetics*, **61**(2), 423–429.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., *et al.* (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, **562**(7726), 203.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, **4**(1), 7.

Comas Cufí, M. (2019). *Aportacions de l'anàlisi composicional a les mixtures de distribucions*. Ph.D. thesis, Universitat de Girona.

Conomos, M. P., Reiner, A. P., Weir, B. S., and Thornton, T. A. (2016). Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics*, **98**(1), 127–148.

Cotterman, C. W. (1941). Relatives and Human Genetic Analysis. *The Scientific Monthly*, **53**(3), 227–234.

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nature Methods*, **9**(2), 179.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2006). Simplicial geometry for compositional data. *Geological Society, London, Special Publications*, **264**(1), 145–159.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, **35**(3), 279–300.

Epstein, M. P., Duren, W. L., and Boehnke, M. (2000). Improved inference of relationship for pairs of individuals. *The American Journal of Human Genetics*, **67**(5), 1219–1231.

Fakhro, K. A., Staudt, M. R., Ramstetter, M. D., Robay, A., Malek, J. A., Badii, R., Al-Marri, A. A.-N., Khalil, C. A., Al-Shakaki, A., Chidiac, O., *et al.* (2016). The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Human Genome Variation*, **3**, 16016.

Foulkes, A. S. (2009). *Applied statistical genetics with R: for population-based association studies*. Springer Science & Business Media.

Galván-Femenía, I., Obón-Santacana, M., Piñeyro, D., Guindo-Martinez, M., Duran, X., Carreras, A., Pluvinet, R., Velasco, J., Ramos, L., Aussó, S., Mercader, J. M., Puig, L., Perucho, M., Torrents, D., Moreno, V., Sumoy, L., and de Cid, R. (2018). Multitrait genome association analysis identifies new susceptibility genes for human anthropometric variation in the GCAT cohort. *Journal of Medical Genetics*, **55**(11), 765–778.

García-Magariños, M., Egeland, T., López-de Ullibarri, I., Hjort, N. L., and Salas, A. (2015). A parametric approach to kinship hypothesis testing using identity-by-descent parameters. *Statistical Applications in Genetics and Molecular Biology*, **14**(5), 465–479.

Ghalanos, A. and Theussl, S. (2015). *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*. R package version 1.16.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, **8**, 2224.

Graffelman, J. and Aluja-Banet, T. (2003). Optimal representation of supplementary variables in biplots from principal component analysis and correspondence analysis. *Biometrical Journal*, **45**(4), 491–509.

Graffelman, J. and Galván-Femenía, I. (2016). An application of the isometric log-ratio transformation in relatedness research. In *Martín-Fernández J., Thió-Henestrosa S. (eds) Compositional Data Analysis. CoDaWork 2015. Springer Proceedings in Mathematics & Statistics*, volume 187, pages 75–84. Springer, Cham.

Graffelman, J. and Moreno, V. (2013). The mid p-value in exact tests for Hardy-Weinberg equilibrium. *Statistical Applications in Genetics and Molecular Biology*, **12**(4), 433–448.

Heinrich, V., Kamphans, T., Mundlos, S., Robinson, P. N., and Krawitz, P. M. (2016). A likelihood ratio-based method to predict exact pedigrees for complex families from next-generation sequencing data. *Bioinformatics*, **33**(1), 72–78.

Jacquard, A. (1974). *The genetic structure of populations*. Prentice hall Upper Saddle River, NJ.

Johnson, R. A., Wichern, D. W., *et al.* (2002). *Applied multivariate statistical analysis*, volume 5. Prentice hall Upper Saddle River, NJ.

Katki, H. A., Sanders, C. L., Graubard, B. I., and Bergen, A. W. (2010). Using DNA fingerprints to infer familial relationships within NHANES III households. *Journal of the American Statistical Association*, **105**(490), 552–563.

Laird, N. M. and Lange, C. (2010). *The fundamentals of modern statistical genetics*. Springer Science & Business Media.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**(22), 2867–2873.

Martín-Fernández, J. A. (2001). *Medidas de Diferencia y Clasificación Automática no Paramétrica de Datos Composicionales*. Ph.D. thesis, Universitat Politècnica de Catalunya.

Mateu-Figueras, G. (2003). *Models de distribució sobre el Símplex*. Ph.D. thesis, Universitat Politècnica de Catalunya.

Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J. J. (2011). The principle of working on coordinates. *Compositional Data Analysis*, pages 29–42.

Milligan, B. G. (2003). Maximum-likelihood estimation of relatedness. *Genetics*, **163**(3), 1153–1167.

Moltke, I. and Albrechtsen, A. (2014). RelateAdmix: a software tool for estimating relatedness between admixed individuals. *Bioinformatics*, **30**(7), 1027–1028.

Obón-Santacana, M., Vilardell, M., Carreras, A., Duran, X., Velasco, J., Galván-Femenía, I., Alonso, T., Puig, L., Sumoy, L., Duell, E., Perucho, M., Moreno, V., and de Cid, R. (2018). GCAT-Genomes for life: a prospective cohort study of the genomes of Catalonia. *BMJ Open*, **8**(3), e018324.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2011). Lecture Notes on Compositional Data Analysis.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons.

Pearson, K. (1897). Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, **60**(359-367), 489–498.

Pemberton, T. J., Wang, C., Li, J. Z., and Rosenberg, N. A. (2010). Inference of unexpected genetic relatedness among individuals in HapMap Phase III. *The American Journal of Human Genetics*, **87**(4), 457–464.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**(3), 559–575.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramstetter, M. D., Dyer, T. D., Lehman, D. M., Curran, J. E., Duggirala, R., Blangero, J., Mezey, J. G., and Williams, A. L. (2017). Benchmarking relatedness inference methods with genome-wide data from thousands of relatives. *Genetics*, **207**(1), 75–82.

Rosenberg, N. A. (2006). Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of Human Genetics*, **70**(6), 841–847.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, **298**(5602), 2381–2385.

Sun, L. (2012). Statistical Human Genetics: Methods and Protocols. Chapter 2. pages 25–46.

Thompson, E. (1975). The estimation of pairwise relationships. *Annals of Human Genetics*, **39**(2), 173–188.

Thompson, E. (1986). Likelihood inference of paternity. *American Journal of Human Genetics*, **39**(2), 285.

Thompson, E. (1991). Estimation of relationships from genetic data. *Handbook of Statistics*, **8**, 255–269.

Vieira, M. L. C., Santini, L., Diniz, A. L., and Munhoz, C. d. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and Molecular Biology*, **39**(3), 312–328.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, **101**(1), 5–22.

Vives-Mestres, M. (2014). *Gràfic de control T^2 de Hotelling per a Dades Composicionals*. Ph.D. thesis, Universitat de Girona.

Voight, B. F. and Pritchard, J. K. (2005). Confounding from cryptic relatedness in case-control association studies. *PLoS Genetics*, **1**(3), e32.

Wagner, A., Creel, S., and Kalinowski, S. (2006). Estimating relatedness and relationships using microsatellite loci with null alleles. *Heredity*, **97**(5), 336.

Weir, B. S., Anderson, A. D., and Hepler, A. B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics*, **7**(10), 771.