Aus dem Deutsches Krebsforschungszentrum
(Kommissarischer Wissenschaftlicher Stiftungsvorstand: Prof. Dr. rer. nat. Michael Boutros)

Abteilung Epigenetik
(Abteilungsleiter: Prof. Dr. Frank Lyko)

In Zusammenarbeit mit dem Institut für Angewandte Mathematik der Universität Heidelberg
(Geschäftsführender Direktor: Prof. Dr. Rainer Dalhaus)

Abteilung Numerische Optimierung
(Abteilungsleiterin: Prof. Dr. Ekaterina Kostina)

Exploring and controlling for underlying structure in genome and microbiome case-control association studies

Inauguraldissertation
Zur Erlangung des Doctor scientiarum humanarum (Dr.sc.hum.)
An der
Medizinischen Fakultät Heidelberg
Der
Ruprecht-Karls-Universität

Vorgelegt von
Carine Legrand

Aus
La Garenne-Colombes (Frankreich)
2016

Dekan: Prof. Dr. Wolfgang Herzog

Doktorvater: Prof. Dr. Frank Lyko

To all who gave me a chance and helped me going forward,

To difficulties and setbacks, from which I learned,

To my family and friends,

To sweet Charente, and to the lively, challenging Paris-Banlieue,

To cultural and civilization stronghold Germany,

To everything that made me as I am,

To everything well and good that happened, which I try and repay.

TABLE OF CONTENTS

ABBREVIATIONS AND SYMBOLS

AIC	Akaike Information's Criterion
BLP	Bad Leverage Point
CO	Contextual Outlier (or Conditional Outlier)
DO	Distance Outlier
DOd	Distance Outlier with one Distant point
e.g.	for example (*exampli gratia*)
et al.	and a group of collaborators (*et alii*)
EPIC	European Prospective Investigation Into Cancer and nutrition
HMP	Human Microbiome Project
i.e.	that is to say (*id est*)
IBS	Identity By State
LAR	Least Absolute Residuals
MAD	Median Absolute Deviation
MAF	Minor Allele Frequency
MCD	Minimum Covariance Determinant
MDS	Multi-Dimensional Scaling
cMDS	classical Multi-Dimensional Scaling
nmMDS	non-metric Multi-Dimensional Scaling
MDSm	Multi-Dimensional Scaling on Manhattan distances
MDSe	Multi-Dimensional Scaling on Euclidean distances
qMDS	Quantile Multi-Dimensional Scaling
PCA	Principal Components Analysis
RMDS	Robust Multi-Dimensional Scaling
RSMDS	Robust Multi-Dimensional Scaling for Structure outliers
SNP	Single Nucleotide Polymorphism
SPH	SPHerical Principal Components Analysis
WCS	Weighted Corrected Similarity

1 INTRODUCTION

1.1 Case-control association studies: medical research and the role of confounding, outliers and stratification

Case-control association studies rely on the observation of a disease status, or a symptom and on collected characteristics such as medical data, smoking status, or else, of samples. Association between samples characteristics and affection status or symptoms are sought, in order to find a significant link. Early studies of this type (Lane-Claypon 1926, Doll 1950) brought about decisive advances in their time, and case-control studies are still indispensable in fields where randomized controlled trials are hardly possible, such as in human genetics or in the human microbiome fields.

Association between genetic predictors (generally, single nucleotide polymorphisms, SNPs, are typed for each control and each case) or microbiological predictors (for example, presence and abundance of a specific microbe in each sample individual) and a trait yields the base for valuable information to the clinician, such as risk prediction and treatment outcome prognosis. Well-known examples are SNPs in the BRCA gene, which correspond to an increased risk of breast, ovarian or prostate cancer, allowing to plan tailored screening or prevention actions. Treatment outcome has also become an important aim of association studies, for example to evaluate toxicity or not after chemotherapy (recent examples are Montassier et al. 2016 - microbiome ; Ozkavruk et al. 2016 - genome). In both cases the goal is the emergence of personalized medicine. To this aim, predictive or prognostic biomarkers are looked for, such as specific SNPs or sets of SNPs (to name only a few recent examples of conditions where biomarker are studied: diabetes, cancer, asthma, attention disorder, rheumatoid arthritis and more generally complex diseases - Buus et al. 2016, Komi et al. 2016, Bonvicini et al. 2016, Hollman et al. 2016, Saad et al. 2016, Müller et al. 2016). Recently, many associations between the microbiome and human health were highlighted, such as in gastrointestinal diseases and obesity (Malinen et al. 2010, Turnbaugh et al. 2006),

but also in complex diseases such as cancer and Crohn's disease (Kostic et al. 2012, Zou et al. 2016, Ferrarelli 2016, Dey et al. 2013), or psychological traits (a short review on links between microbiota and diseases is given in Gilbert et al. 2016).

However, association studies are plagued by confounding, and -as a consequence- by failed replications, which affect the genetic associations field since its beginnings (Ioannidis et al. 2001, Cardon and Bell 2001, Hirschhorn et al. 2002, Marchini et al. 2004) and is only an emerging topic in microbiome associations (Blaser et al. 2013, Sinha et al. 2015). Confounding happens when unknown (or unaccounted covariates) factors impair validity of the association tested, either if the association is spurious (false positive, inflated type I error), or if a true association fails to be detected (false negative, masking). For both genetic or microbiome associations, this has been imputed for a part to population stratification (Pritchard and Rosenberg 1999, Cardon and Bell 2001, Ioannidis et al. 2001, Hirschhorn et al. 2002 ; Blaser et al. 2013, Vogtmann and Goedert 2016). On the other hand, underlying structure or population stratification can be useful, for example to define groups-specific associations (Larson et al. 2000, Li et al. 2016 ; Nasidze et al. 2009, Holmes et al. 2011, Blaser et al. 2013, Falony et al. 2016), and therefore be a part of personalized medicine.

Lastly, it is important to stress that confounding can take different forms. Stratification is a situation where distinct groups or clusters are present. Additionally or alternatively, a continuous structure, for example a varying ancestry component or cryptic relatedness, can be present. Finally, outliers are confounders since they influence the estimates although they do not correspond to the pattern taken by the majority of the data.

1.2 Genetic case-control association studies and population structure

Genetic case-control association studies aim at finding a significant association between allele frequencies in a collection of genetic variants (SNPs) sequenced in a large number of case and control individuals. The dataset usually consists of a matrix containing the sum of minor alleles, for each SNP and for each individual. Confounding through population structure has been identified in such studies early on, and has long motivated the use of family-based studies instead of population-based studies (Cardon and Bell 2001). Beginning in 1999, methods to detect and correct for population structure emerged, in particular the STRUCTURE software based on a set of unlinked SNPs (Pritchard and Rosenberg 1999, Pritchard and Donnelly 2000) and the genomic control method (Devlin and Roeder 1999), which allows to adjust p-values for residual population stratification by looking at allele

frequencies differences. Intensive discussion was led and a wealth of new methods appeared at the time (Reich and Goldstein 2001, Satten et al. 2001, Zhang et al. 2001, Thomas and de Witte 2002, Wacholder et al. 2002, Hoggart et al. 2003). In parallel, concerns about replication were aired (Cardon and Bell 2001, Ioannidis et al. 2001). In particular, Hirschhorn et al. (2002) reviewed 166 genetic association studies and found that only 6 were "consistently replicable", even if half of the remaining associations could be replicated sometimes, but not always.

As a consequence, this pioneering era was followed by increased awareness that a cautious study design was needed in genetic case-control studies (Freedman et al. 2004, Marchini et al. 2004), that an increased number of SNPs from a few dozens to thousands were needed (Freedman et al. 2004), even if ancestry marker SNPs were used, and finally that increased sample sizes were necessary (Hirschhorn and Daly 2004), which lead to the development of genome-wide association studies (GWAS - 1st GWAS in Ozaki et al. 2002, early and later reviews in Hirschhorn and Daly 2004 and Kingsmore et al. 2008, respectively). Further care was still taken to detect and correct for population structure, or to use alternative methods (genomic control, ancestry matching of cases and control groups). In particular, Marchini et al. (2004) revealed that even subtle population structure inside a population leads to confounding, when sample size increases. Accordingly to the increase in significance and in size of the GWAS, the practicable and efficient PCA method (Eigenstrat) was developed by Price et al. 2006, which has been applied broadly (Price et al. 2010, Clarke et al. 2011, Bouaziz et al. 2011, Wu et al. 2011) ever since, while genomic control was deemed less appropriate (Freedman et al. 2004, Price et al. 2010, Bouaziz et al. 2011, Wu 2011).

Nevertheless, concern still arose about confounding through finer structure or cryptic relatedness (Voight and Pritchard 2005, Astle and Balding 2009, Sillanpää 2011, Liu et al. 2011, Shibata et al. 2013), which can be viewed as a form of continuous structure and occurs when unknown moderate relatedness is present between samples which are taken to be unrelated, and which is not explicitly taken into account by STRUCTURE or PCA (Eigenstrat). Enduring concerns about insufficient correction of population stratification were also expressed (Salanti et al. 2005, Saito et al. 2006, Choudhry et al. 2006). Several methods were proposed to tackle both issues of population stratification and cryptic relatedness, such as mixed models (Price et al. 2010, Yang et al. 2014, Chen et al. 2016) which make use of a kinship matrix to account for relatedness. Further tailored PCA-derived methods like ROADTRIPS, EMMAX or PCAiR (Thornton et al. 2010, Kang et al. 2010, Conomos et al. 2015, respectively) were also developed. However, PCA was found to perform at least as well

as mixed models in the presence of admixture (Liu et al. 2011), and mixed models are mostly relevant when only relatedness and no stratification occurs (Yang et al. 2014), which is however seldom the case. Moreover, methods ROADTRIPS and EMMAX did underperform as compared to PCA (Eigenstrat) in Wu et al. (2011) and there is little information yet on method PCAiR, which is for now rarely applied (Dunn et al. 2016). Furthermore, multidimensional scaling methods were applied, either on a matrix of covariance or of identity-by-descent between samples, but the results obtained were found similar to PCA (Wang et al. 2009, Zhang et al. 2014).

To this day, PCA (Eigenstrat) remains the most broadly applied method to detect and correct for population structure (Clarke et al. 2011, Bouaziz et al. 2011, Wu et al. 2011), in agreement with the fact that authors insist PCA is necessary (Widmer et al. 2014) or at least recommended (Liu et al. 2011), although concerns about fine structure or cryptic relatedness still exist ("There is no certainty that population stratification is completely controlled for in large-scale meta-analyses", Robinson et al. 2016). Some amount of confusion even exists, as some authors sometimes revert to using genomic control (Fuchsberger et al. 2016) though it may impair finding of true associations, or come back to using family-based designs complementary to population structure correction (Robinson et al. 2016, Okbay et al. 2016), though family-based studies are also prone to false positives and are underpowered in comparison to case-control studies (Hirschhorn and Daly 2004).

1.3 Microbiome case-control association studies and underlying structure

The microbiome is constituted of microbes which coat the surfaces of the human body, such as the gastrointestinal tract or the skin. A few early studies in 1985 (Lane et al. 1985) pioneered the analysis of this microbiome by genomic sequencing of the set of microbes, and in particular by typing the highly variable 16S ribosomal RNA part. As a result, microbes can be classified in phylogenies, and the abundance of each of these phylogenies is recorded. A microbiome dataset generally consists of a matrix of abundances for each phylogeny and for each sample. Microbiome analysis have developed fast after 2002 due to improved sequencing possibilities, which has led to the first microbiome association studies (Breitbart et al. 2002, Tyson et al. 2004) and to numerous studies up to now. Many potentially useful associations in a large array of conditions ranging from gastro-intestinal diseases to cancer, and as far as autism or depression (a review can be found in Gilbert et al. 2016).

Microbiome associations, similarly to genetic associations, are interesting in medical research because they provide risk factors, predictive or prognostic biomarkers useful in precision medicine. They are similar to genetic association studies in the sense that a careful analysis is needed and that confounding factors can be present and impair replication (Blaser et al. 2013, Gilbert et al. 2016, Vogtmann and Goedert 2016). First ideas point to population structure confounding, which might either correlate to host genotype favoring specific microbiomes, or to diet habits (same authors). However, underlying structure in microbiome is not only a confounding factor but also an expected advantage, because it can help in prediction or tailoring treatment (Nasidze et al. 2009, Holmes et al. 2011). Many authors express the need to identify patient stratification, for example classify patients in subgroups of individuals corresponding to different levels of risk (Nasidze et al. 2009, Holmes et al. 2011, Blaser et al. 2013, Falony et al. 2016, Wang and Jia 2016, Zmora et al. 2016), which allows ultimately the clinician to choose the option best adapted to each subgroup. Furthermore, the recognition of a definite structure in the microbiome, for example a deviation of the microbiome from a healthy or normal microbiome can inform on the course or the grading of a disease (Cho and Blaser 2012, Ash 2016, Lloyd-Price et al. 2016). For example, it is hoped that cancer diagnosis can be accelerated and treatment toxicity reduced by taking into account information derived from microbiome analyses (Vogtmann and Goedert, citing Wallace et al. 2010 and Iida et al. 2013).

The limited knowledge on confounding factors on one hand, and the potential expected from a better defined microbiome structure motivates the exploration of microbiome datasets with dimension reduction techniques like PCA or MDS (Gilbert et al. 2016). Many such methods were proposed, for example MDS using different distance metrics (Gilbert et al. 2016) or robust PCA based on the least absolute distances ($\ell 1$-norm, Brooks et al. 2013) instead of least squared distances ($\ell 2$ norm). However, no consensus has been reached so far although it is necessary in order to obtain results comparable between studies (Blaser et al. 2013). As microbiome association studies have moved towards large-scale consortia (MWAS, HMP project Turnbaugh et al. 2007), and as more integration between different omics is called for (Blaser et al. 2013, GWAS with microbiome as an outcome as in Folseraas et al. 2012, Goodrich et al. 2016), a sound choice of these exploration methods is essential.

1.4 Stable estimates in the presence of confounding structure

Confounding and subsequent replication exist in both genome and microbiome association studies (genome: Ioannidis et al. 2001, Hirschhorn et al. 2002, Marchini et al. 2004, Salanti et al. 2005, Saito et al. 2006, Choudhry et al. 2006, Nilsson et al. 2013, Li and Meyre 2013, Liu et al. 2013 ; microbiome: Harley and Karp 2012, Blaser et al. 2013, Sinha et al. 2015, Gilbert et al. 2016). For both, population stratification is still thought to be responsible for these issues, at least partly (Nilsson et al. 2013, Gilbert et al. 2016).

However, robust PCA or MDS, which aim at producing stable estimates of underlying structure and hence enhance reproducibility, are seldom applied. This is probably because most robust methods were developed in other contexts or are not always directly applicable to genomic or microbiome datasets (for example minimum covariance determinant MCD, Rousseeuw and van Driessen 1999, was developed for the chemometrics field). The few methods claiming to be robust in the genetics field actually do remove outliers (Liu et al. 2013), or use the term 'robust' for the resilience to a specific aspect only (for instance to relatedness, which is a continuous structure, but not to outliers, in Conomos et al. 2015).

Currently, attempts to make analyses more reliable consist in applying a sound and consensus quality control (for example: protocols Clarke et al. 2011, Turner et al. 2011). Additionally, it is almost always recommended to remove outliers from the dataset (Bellenguez et al. 2010, Shen et al. 2010, Clarke et al. 2011, Turner et al. 2011, Liu et al. 2013, Callahan et al. 2016). However, this practice is questionable, since these outliers are not necessarily false points but may contain relevant or sometimes decisive information (Yu et al. 2002, Scott et al. 2005, McMurdie and Holmes 2014, Ray 2014, Friend and Schadt 2014). Furthermore, they are detected based on a perceived, but not necessarily relevant difference with respect to the majority of the points (van den Broeck 2005, Bakker et al. 2014), so that it might not be straightforward to decide which points are or not outliers and if they should be excluded. Robust statistics aim precisely at resolving this issue by accommodating outliers, that is to say it takes advantage of the information contained in these, while maintaining stable estimates.

1.5 Objectives and organisation of this work

Considering the largely unmet need in detecting underlying structure reliably and robustly, the aim of this thesis is the robust exploration and control of underlying structure in genetic and microbiome case-control association studies. This is carried out by comparing a large array of standard and robust methods in terms of representativity of the underlying structure, and in terms of robustness to outliers, among which several new improved methods are proposed.

PCA- and MDS- based methods commonly applied in the genome and microbiome fields are included, as well as a variety of robust extensions, including MCD, spherical PCA (Locantore et al. 1999), MDS based on Manhattan metric (Barile and Weisstein 2016) or non-metric MDS (Kruskal 1964a, 1964b). My contributions include robustifications of existing methods, adaptation to the genomic or microbiome field, and a new method for dimension exploration and reduction, nSimplices, applicable in both contexts.

The comparisons were carried out on two synthetic datasets, on genetic SNP data from the EPIC initiative (European Prospective Investigation Into Cancer and nutrition, Riboli et al. 2002, Campa et al. 2011), and finally on microbiome abundance data from the Human Microbiome Project (HMP, Turnbaugh et al. 2007). Outliers were introduced in varying proportions in synthetic example 2 and in the two real datasets. As a result, 10 principal components or axes were kept for each method and for each percentage of outliers included. Then, these estimated axes were assessed in terms of representativity. This was measured by the quality of the fit of these axes to the known underlying structure. Additionally, the robustness of the 10 axes disturbed by the inclusion of outliers was evaluated. This was assessed by how well the disturbed 10 axes could fit each one of the original axes.

2 MATERIAL AND METHODS

2.1 Datasets

The different methods were compared on two synthetic examples and two real world datasets. Simple synthetic examples served to assess methods at a well-controlled and tractable scale. Genetic variants and microbiome abundances datasets were used to assess all methods in a realistic context in higher dimension. Different classes of outliers were introduced, based on the principle of their relevance with regard to different types of data. Indeed, most genetic or microbiome datasets contain outliers but their published versions are extensively cleaned from outliers. This makes necessary to reintroduce some, in order to reproduce a realistic situation where some outliers are present.

2.1.1 Synthetic dataset 1: Underlying structure with small variance

This first dataset is designed to mimic a situation where a relevant or confounding structure is present in the data, but where this confounding structure is hard to detect because of its small variance. The structure is represented by one underlying factor with two distinct but close groups. One group of 1000 samples is drawn from a normal distribution with mean 1. and standard deviation 0.5 ($\mathcal{N}(1,0.5)$), and a second group of 1000 samples from $\mathcal{N}(-2.,1.)$.

Further 15 independent underlying factors devoid of any structure are generated for the same samples by drawing 2000 values from $\mathcal{N}(\mu, \sigma)$. The mean μ is itself drawn from $\mathcal{N}(0,10)$, and σ, the standard deviation, in the set $\{0.2,0.9,1.71,5,10\}$ (recycled three times). This set was chosen to represent standard deviations ranging from values smaller, equal or greater to 1.71, which corresponds to the standard deviation of the first, confounding factor. This leads to a set of 16 underlying factors, which constitute a 2000-by-16 matrix X_{factor}. These

underlying factors were linearly combined with a mixing matrix A, so as to obtain a more realistic dataset, where a relevant factor can be spread over several measured variables. The mixing matrix A is a square matrix of order 16. Elements of A are picked from a uniform distribution between -0.5 and 3 (U[-0.5,3]). The synthetic dataset 1 with underlying structure is the 2000-by-16 matrix X calculated as follows: $X = X_{factor} \cdot A$.

2.1.2 Synthetic dataset 2: Underlying structure with noise and outliers

This second dataset is an extension of an example dataset described previously (Spence and Lewandowski 1989, Forero and Giannakis 2012). Briefly, 25 points are spread evenly along two orthogonal segments of length 12 (arbitrary units), which intersect at their middle point. These distances are contaminated with noise drawn from a normal distribution $\mathcal{N}$(0,0.3) (truncated so that resulting distances are positive). In addition, a varying percentage in {0.25,0.5,1,2,5,7,10}% of the distances are picked at random and turned into outliers by adding a value drawn from U[0,20] (this is similar but not fully identical to Forero and Giannakis 2012). These contaminated configurations are denoted DO (distance outliers).

I added two configurations to strain further the subsequent multidimensional scaling (MDS) calculations. The idea here is that, though MDS is non-robust with respect to large disparities (see for example Spence and Lewandowsky 1989, Forero and Giannakis 2012), a distant, but in-plane point, should degrade accuracy but should not modify the general shape of the solution. As a consequence, a single distant point is a simple test for overall consistency of a MDS-based method. In the first additional configuration I introduced one such distant point (coordinates (100,100) in the axis system defined by the two segments with origin at their intersection). Noise and distance outliers are added as described above. This additional configuration is denoted DOd (DO with distant point).

Finally, contextual outliers are introduced. Contextual (or conditional) outliers correspond to observations which could be considered correct or aberrant depending on the context (see a description for example in Song et al. 2007). They can be visualized geometrically, as points which have an additional coordinate along an axis orthogonal to the subspace where most points lie. These outliers are generated by adding a fixed contribution δ_{suppl} drawn from

$U[0,20]$ along an axis orthogonal to the main subspace. This translates simply in the following formula : $\delta^2_{context} = \delta^2_{hyperspace} + \delta^2_{suppl}$. The intuition here is that these outliers might represent a more drastic perturbation for a MDS-based method than a distant point or noisy observations. This is because MDS would try and mirror the additional, outlying distance, thus affecting all coordinates, without taking into account the fact that the outlying distance regards only one or a few points. Contextual outliers are added to the noisy distances, in proportions of {4,8,12}% of the points (which correspond to 1, 2 and 3 outliers). No additional distance outliers are added. This last setting is denoted CO (contextual outlier).

2.1.3 EPIC

The EPIC (European Prospective Investigation Into Cancer and nutrition) dataset used here comprises genome-wide single nucleotide polymorphisms (SNPs) from 432 prostate cancer cases and 428 controls from the EPIC cohort (see published description of the EPIC cohort in Riboli et al. 2002). Before quality control, 591 841 SNPs are present. Among these, some are filtered out using PLINK v1.07 (Purcell et al. 2007) if minor allele frequency (MAF) is less than 5%, or if the p-value for Hardy-Weinberg equilibrium is less than 0.001, or again if missingness in SNPs exceeds 10%. Remaining SNPs are further pruned out if linkage disequilibrium is present (option "--indep 50 5 2" in PLINK). After quality control, 123 174 SNPs remain, among which 100 000 are randomly selected for further analyses. The resulting (n,p)=(100 000,860) matrix X_{EPIC} is used for further analyses. In this matrix, each genotype is coded as the number of minor alleles: 0 (2 major alleles), 1 (1 minor allele) or 2 (2 minor alleles).

A varying percentage ({0.25,0.5,1,2,5,7,10}% of the original individuals) of individual outliers are generated according to the bad leverage points (BLP) definition (outliers with detrimental influence because of large components along main axe - Rousseeuw and van Zomeren 1990), by introducing large values along principal components 1 and 2, drawn from $\mathcal{N}(10,1)$, and a random small component on remaining principal components, drawn from $\mathcal{N}(0,0.0001)$. The new individuals thus generated in principal components base are

subsequently transformed in a SNPs-by-individual matrix, using the following equation, after which they are rounded to the nearest integer values 0, 1 or 2:

$$I_{SNP} = \sqrt{f \cdot (1-f) \cdot p} \cdot (I_{PC} \cdot P \cdot X_{EPIC}) + E[X_{EPIC}],$$

where f is the unbiased minor allele frequency $f = (1 + n \cdot E[X_{EPIC}]) \,/\, (2 + 2 \cdot n)$, $E[X_{EPIC}]$ is the mean of X_{EPIC} along rows (i.e., SNPs), and P is the (p,p) square matrix containing eigenvectors (in columns). This formula corresponds to de-centering ($+E[X_{EPIC}]$) and de-standardisation ($\sqrt{f \cdot (1-f) \cdot p}$), which is applied to the matrix of outliers in principal components base (I_{PC}) transformed back into a SNPs base ($I_{PC} \cdot P \cdot X_{EPIC}$) (this is simply derived from singular value decomposition equations, described for example in Lieser and Mehrmann 2015). The eigenvectors used here in matrix P are calculated using Eigenstrat (Price et al. 2006, see also brief description in §2.2). The resulting matrix of outlier individuals is joined to X_{EPIC} to constitute an extended, synthetic outlier-containing dataset.

Alternatively, real individual outliers are drawn randomly from non-European populations from the 1000 Genomes public database (1000 Genomes Consortium et al. 2012), and added in varying percentages {0.25,0.5,1,2,5,7,10}% of the original EPIC individuals. The subset of SNPs corresponding unambiguously to the EPIC SNPs was used (same name, same position, same reference and alternative alleles). The resulting additional individuals were joined to X_{EPIC}, which constitutes a real outlier-containing dataset.

2.1.4 HMP

The HMP dataset is constituted of human microbial abundances at several body sites, downloaded from the Human Microbiome Project (HMP, Turnbaugh et al. 2007). The source data are the high quality phylotype counts, classified using the microbial 16S ribosomal variable region V13, as made available by the HMP consortium (URL: http://hmpdacc.org/HMMCP/). Then, quality control filters are applied to exclude samples or phylogenies with less than 10 strictly positive counts. The dataset is used subsequently either unchanged, as a (2 255,425) samples-by-phylogenies counts matrix, or transformed in proportions so that phylogenies proportions add up to 1 for each sample.

Outlier samples are generated according to the bad leverage points (BLP) definition, as described above, by creating large coordinates along the main principal components and transforming the data back to a samples-by-phylogenies matrix. The eigenvectors used in this transformation derive from multidimensional scaling on a distance matrix (Torgerson 1957, see brief description in §2.2).

2.2 Available methods to summarize underlying structure

A subset of the following methods was applied to the synthetic examples and to the EPIC and HMP datasets. Of the principal components, or top axes, produced, the 10 top ones were kept for the evaluation of representativity, or of robustness. These axes were also used in the subsequent association test in the EPIC dataset. Unless otherwise stated, each method was implemented in the Python language (van Rossum 1995). As the second example contains only 2 dimensions, only 2 components or axes are examined in this case.

A brief summary of each method is presented here. A list of methods applied to each dataset is provided in §2.4.

2.2.1 Methods based on Principal Components Analysis

- **ICA.** Independent Component Analysis (ICA) might be useful to detect an underlying structure, or outlyingness, because it separates non-gaussian components. Indeed, it was designed to separate independent but mixed signals (Hérault and Ans 1984, Comon 1994). This situation can be compared with a structure concealed in a high number of dimensions. As independence implies non-gaussianity (see for example Hyvärinen and Oja 2000), this could be useful in the uncovering of outliers. Here, the algorithm fastICA of Hyvärinen and Oja 2000, as implemented in R package *fastICA* (Marchini et al. 2013), was used. ICA was applied as an illustration and proof of principle to the synthetic example 1 only. Indeed, ICA does not provide ordered components, contrary to principal components analysis (PCA, Pearson 1901). Therefore, retrieving the first 10 components from ICA does not necessarily mean that these components are the most meaningful. As a consequence, the use of ICA to

explore a large-dimensional dataset such as EPIC or HMP is not straightforward and hasn't been done here.

- **PCA and Eigenstrat (denoted EIG in the following).** Principal Components Analysis (PCA, Pearson 1901) is applied either on a centered and standardized matrix (synthetic dataset 1, using R) or following the Eigenstrat method. Eigenstrat was suggested and described in Price et al. 2006. Eigenstrat is a popular approach to correct for population stratification and is therefore used here as a reference. Briefly, a (n,p) SNP-by-individual data matrix A, is first centered and normalized as follows:

$$b_{ij} = (a_{ij} - 1/p \cdot \sum_{j=1}^{p} a_{ij}) / \sqrt{f_i \cdot (1 - f_i)}\,,$$

where b_{ij} are the elements of matrix B, (i,j) are the indices of rows and columns of A or B, and $f_i = (1 + \sum_{j=1}^{p} a_{ij})/(2 \cdot (p+1))$. This step is denoted "Eigenstrat normalisation" in the following. B is then used to build a covariance matrix $C = 1/n\, B^T.B$. Principal components are computed by eigen-decomposition of C.

- **SPH.** Spherical PCA was proposed by Locantore et al. 1999. This procedure is similar to Eigenstrat, but instead of performing an Eigenstrat normalization, the following normalization towards a unity sphere is done:

$$b_{ij} = (a_{ij} - 1/n \cdot \sum_{i=1}^{n} a_{ij}) \Big/ \sqrt{\sum_{j=1}^{p} (a_{ij} - 1/n \cdot \sum_{i=1}^{n} a_{ij})^2}\,,$$

where matrices A and B follow the same definition as in EIG.

- **MCD.** The Minimum Covariance Determinant (MCD) procedure was introduced by Rousseeuw (1984) and improved in Rousseeuw and van Driessen (1999). The implementation used here comes from Python package *scikit-learn* (Pedregosa et al. 2011), which was adapted for the needs of this thesis (tolerance to missingness, acceleration). The principle at the heart of this method is to select the core of the data by iterating on the covariance matrix's determinant. The size of the core subset can vary between 50% and 100% of the original dataset. Here this size is set to 60%. The data matrix X_{EPIC} had to be restricted to 20 000 SNPs instead of 100 000 to guarantee that all calculations were tractable. Furthermore, a

small amount of jitter was applied to X_{EPIC} in order to ensure that subset-based covariance matrices have almost always full rank. This is better than removing fully identical SNPs, because there are only very few of them, so that their removal would have only marginally helped to guarantee full rank matrices.

- **IBS.** This approach consists in PCA applied to the Identity-By-State (IBS) matrix between individuals, instead of a covariance matrix, as done in PLINK v1.07 (Purcell et al. 2007). IBS is defined as follows: $IBS_{jk,i} = 1 - \left|\left(g_{j1,i} + g_{j2,i}\right) - \left(g_{k1,i} + g_{k2,i}\right)\right|/2$, where $g_{j1,i}$ is the genotype of individual j (1st allele copy) at SNP i.

- **L1PCA.** (ℓ1-norm)-PCA was proposed by Brooks et al. (2013) and is based on the least absolute distances, which correspond to the ℓ1-norm. Here, a Python implementation, based on the algorithm of Brooks et al. (2013) and on the R function *l1pcastar* provided by the authors is used. The calculation relies on linear program optimization, here performed by the Fortran routine rqfnc.f (Koenker and Portnoy 1997), which was integrated into Python using F2PY (Peterson 2009).

2.2.2 Methods relying on Multidimensional Scaling

Multidimensional scaling (MDS) aims at finding the coordinates of a set of points along 1, 2 or more main axes, based on the pairwise distances between them. MDS takes a distance matrix, which contains the pairwise distances, as input, and generates the coordinates along the main axes for all the points. This is done by minimizing the residual errors between the input distances and the distances recalculated based on the coordinates (Hastie et al. 2001).

- **cMDS.** Classical multidimensional scaling (cMDS) was presented by Torgerson 1957. This method is closely related to PCA since it relies on an eigen-decomposition, which is applied to a double-centered distance matrix instead of being applied to a covariance or correlation matrix (Hastie et al. 2001).

- **MDSm and MDSe.** Both methods apply cMDS to a distance matrix, where the distances are either calculated using the Manhattan distance (Barile and Weisstein 2016), denoted MDSm,

or Euclidean distance, denoted MDSe, on pyhlogenies proportions. In synthetic dataset 1, MDSe is carried out in R.

- **LAR.** The Least Absolute Residuals (LAR) method is a MDS method which relies on the $\ell 1$-norm instead of the $\ell 2$-norm. This means in practice that squared distance residuals are replaced by absolute distance residuals. The LAR algorithm proposed by Heiser (1988) was implemented and used here.

- **RMDS and RSMDS.** The methods Robust MDS (RMDS) and Robust MDS for Structured Outliers (RSMDS) were introduced by Forero and Giannakis (2012). Both are iterative procedures which aim to regularize distance outliers and therefore accommodate for them. RSMDS additionally aims to take advantage of outliers sparsity. Both methods need a tuning parameter. Wherever necessary, an extensive domain was explored in order to find out the best tuning parameter (lowest final MDS stress - only the best tuning parameter was used to produce the results presented here).

- **nmMDS.** Non-metric MDS was proposed in Kruskal (1964a and 1964b) and considers ranked distances only, instead of least squares residuals (Hastie et al. 2001). The input used for nmMDS is the matrix of Euclidean distances between phylogenies (normalized as proportions). The implementation used here is the *manifold* function (package *scikit-learn*, Pedregosa et al. 2011), except regarding synthetic dataset 1, for which nmMDS is carried out in R.

2.3 Proposed improved methods to explore underlying structure

In this section, I present six new own methods to improve structure detection and analysis of genome-wide SNP data, or of microbiome abundance data.

2.3.1 gMCD

In genetic MCD (gMCD), the MCD procedure is applied to unbiased estimates of inter-individual relatedness (Oliehoek et al. 2006, Weighted Corrected Similarity - WCS), instead

of being applied to a covariance matrix. Genetic relatedness between individuals j and k can be written as follows (simplified from Oliehoek et al. 2006, with only two alleles possible):

$$r_{jk} = \frac{4}{W} \cdot \sum_{i=1}^{n} w_i \cdot S_{jk,i} - 2,$$

where $w_i^{-1} = 3/4 \cdot (maf^2 + (1 - maf)^2) \cdot (1 - (maf^2 + (1 - maf)^2)$, W is the sum of weights $W = \sum_{i=1}^{n} w_i$, maf is the minor allele frequency: $maf = \left(1 + \sum_{j=1}^{p} a_{ij}\right) / (2 \cdot (1 - p))$, and the elementary genetic similarity between individuals j and k at SNP i is: $S_{jk,i} = 1/4 \cdot [I_{j1,k1} + I_{j1,k2} + I_{j2,k1} + I_{j2,k2}]$, with $I_{jx,ky}$ equal to 1 if allele copy x in individual j and allele copy y in individual k are identical, equal to 0 otherwise (same definition as in Oliehoek et al. 2006).

2.3.2 gMDS

Genetic Multidimensional scaling (gMDS) corresponds to cMDS applied on a genetic distances matrix. The genetic distance between individuals j and k, g_{jk}, is defined as follows: $g_{jk} = 1 - r_{jk}/2$, where r_{jk} is the genetic relatedness proposed by Oliehoek et al. (2006) described in previous paragraph.

2.3.3 rgMDS

Robust genetic multidimensional scaling is a method based on gMDS, where classical MDS, which relies on the minimization of the squared residuals, is replaced by the minimization of robust residuals:

$$\min_X \sum_{j,k\ k \neq j} \rho(x_{jk}),$$

where $\rho(\cdot)$ is the Huber function (Huber 1964):

$$\rho: u \mapsto \begin{cases} \frac{1}{2} \cdot u^2 & if\ |u| \geq \gamma \\ \gamma \cdot \left(|u| - \frac{1}{2}\gamma\right) & if\ |u| < \gamma \end{cases},$$

where x_{jk} is the residual $x_{jk} = \left(\delta_{jk} - d_{jk}(M)\right)$, δ_{jk} is the fixed distance calculated on the SNP-by-individual data matrix, and $d_{jk}(M)$ is the distance calculated on a (MDS-axes)-by-individual matrix M. This procedure was initialized using the cMDS solution, further optimization towards the minimum was conducted using IPOPT (Wächter and Biegler 2006)

and linear solvers from HSL (HSL. A collection of Fortran codes for large scale scientific computation. http://www.hsl.rl.ac.uk/).

2.3.4 wMDS

Weighted corrected MDS (wMDS) corresponds to a transposition of the Weighted Corrected Similarity method from Oliehoek et al. 2006 to count data, here microbiome abundances. The similarity $S_{jk,i}$ between samples j and k at SNP i is replaced by the distance $z_{jk,i}$ between samples j and k, at phylogeny k. The key principle in weighted corrected similarity is to weigh each SNP by the inverse of genetic variance to ensure that each SNP contributes equally to the relatedness estimate. Here, the weight applied to phylogeny i is taken as the inverse of the variance of distances at phylogeny i. The resulting weighted distance $z_{jk,i}$ can be written:

$$Z_{jk,i} = 1/W \cdot \sum_{i=1}^{p} w_i \cdot z_{jk,i}$$

where $W = \sum_{i=1}^{p} w_i$, with $w_i = \frac{1}{variance(z_{jk,i})}$ for j,k in {1,n}.

2.3.5 quantile-MDS

The idea behind this method is that a full equivalent to a genetic distance must take into account the statistical distribution of the abundance data under study. Whereas genetic data can be considered binomial with the minor allele frequency as unique parameter, microbiome data are highly skewed towards 0 and are more likely to be well modelled by a Negative Binomial distribution, Poisson, or Exponential distribution (Holmes et al. 2012 ; phylogenies may be modelled by an exponential, see for example Jeraldo et al. 2012).

The Poisson, Negative Binomial and Exponential distributions were examined by fitting each of them to a subset of samples (these regressions were done using function *fitdistr*, in R). This results in parameters r_P, μ_{NB}, s_{NB} and λ_E. Akaike Information's Criterion (AIC) derived from each fit was used to compare and choose appropriate model distributions. Then, fitted parameters were used to transform abundances into quantiles of the corresponding distribution. These quantile values (instead or counts or proportions) are used to build a distance matrix, using Euclidean distance, after which cMDS is applied to this quantile

distance matrix. The method is denoted q^P-MDS, q^{NB}-MDS or q^E-MDS, for the Poisson, Negative Binomial or Exponential distribution.

Remark: The Cumulative Sum Scaling (CSS) proposed and described in Paulson et al. (2013) presents a similar idea than quantile-MDS, although these methods have been initiated and developed independently. However it appeared important to apply the CSS method in the frame of this work (using Bioconductor package *metagenomeseq*, Paulson et al. 2015), so as to provide a comparison between CSS and quantile-MDS.

2.3.6 nSimplices

Here, I present a method which, to my knowledge, is entirely original. Allegedly, simplices, which are a generalization of triangles and tetrahedrons in any dimension n, have already been used extensively in other contexts (for example to investigate data depth in Liu 1990, tumorigenesis in Roman et al. 2015, protein geometry statistical analysis in Tropsha et al. 1996, localization in Thomas and Ros 2005). However, it seems that their use to perform dimension reduction or to accommodate outliers is new. Briefly, I developed the four following tasks with the aim of improving robustness in high-dimensional genetic or microbiome abundance distance matrix analysis:

- detect dimension,
- detect and accommodate outliers,
- perform dimension reduction,
- apply classical MDS at the reduced dimension.

The main concept, and how each of these tasks is carried out, are presented in the following.

- **Influence of contextual outliers.** The intuition behind this method is that multivariate outliers do not necessarily have large components along the main axes, but could have influential contributions along less important axes. Here, this category is denoted "contextual outliers" (CO). The term "contextual" (or sometimes "conditional") refers to the fact that these points are aberrant or not depending on the context, or depending on a condition (Song et al. 2007). Fig.1 presents a schematic view of this type of outliers, where most points lie in a plane of dimension 2. However, one point has a coordinate along a third dimension. Thus, this

point could be considered aberrant if the supplementary dimension is considered, or as a normal measurement if one looks at its projection on the plane. The nSimplices method aims at identifying and accommodating such outliers, where, in the simple situation described in Fig.1, accommodation would correspond to projecting the outlier on the plane.

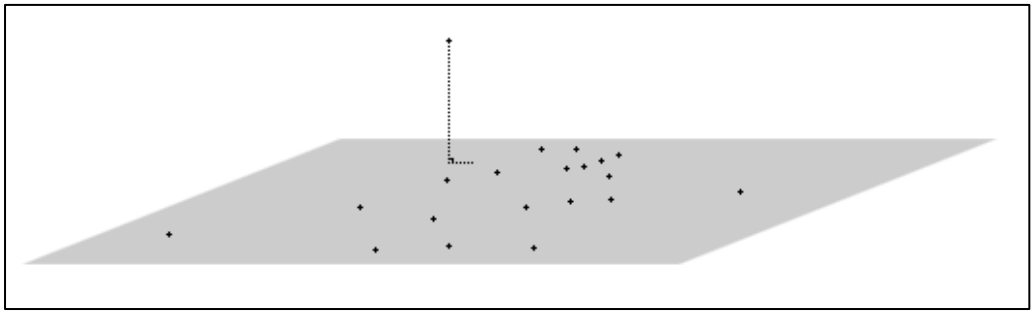

Figure 1. Schematic view of a contextual outlier

- **Non-zero volumes inform on the relevant dimension in a dataset.** Considering the schematic situation in Fig.1, three points chosen at random would form a triangle with a non-zero area, which will be denoted (n=2)-volume. The (n=2)-volumes formed by different sets of 3 points take various values which reflect the specific positions of the points. The fact that these volumes are positive intuitively relates to the fact that the dimension of this dataset is n=2 at least. Then, if four points are chosen at random, a tetrahedron is formed, which would have a (n=3)-volume. Most of these (n=3)-volumes would be equal to zero, since they would be included in the plane. From this derives, that the relevant dimension of the dataset in Fig.1 is n=2, and not n=3, since the vast majority of (n=3)-volumes are zero.

This rationale can be extended to any dimension n=4, 5, etc., using the notion of n-simplex. A n-simplex is a generalization of a triangle or tetrahedron in dimension n, constituted by (n+1) distinct points (Sommerville 1929). As such, a triangle can be seen as a 2-simplex, in dimension 2, formed by 3 points ; a tetrahedron is a 3-simplex, in dimension 3, formed by 4 points ; if n=10, a 10-simplex would be a volume in dimension 10, and is formed by 11 points. The volume of a n-simplex can be calculated on the matrix of pairwise distances between the points, using the Cayley-Menger formula (Sommerville 1929):

$$V_n{}^2 = 2^n \cdot (n!)^2 \cdot det\left(\begin{bmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 0 & \delta_{1,2}^2 & \dots & \delta_{1,n+1}^2 \\ 1 & \delta_{1,2}^2 & 0 & \dots & \delta_{2,n+1}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \delta_{1,n+1}^2 & \delta_{2,n+1}^2 & \dots & 0 \end{bmatrix}\right),$$

where V_n is the n-volume of the n-simplex constituted by n+1 points, det() is the matrix determinant, and δ_{jk} is the distance between points j and k.

- **Contextual outliers form a non-zero outlying volume.** Still considering Fig.1, the (n=3)-volume of a tetrahedron formed by the unique point outside the plane and any three other points would be non-zero, contrary to most (n=3)-volumes which would be constituted by four points in the plane, and therefore are equal to 0. This observation can be used to detect the point outside the plane as a contextual outlier. This can also be generalized to any dimension n using the Cayley-Menger formula mentioned previously.

- **In practice, volume in dimension n can be replaced by the height in dimension n.** In dimension n, the height of a specific point belonging to a n-simplex is the distance between this point and the base of the simplex, exactly like the height in a tetrahedron is the distance of one point relative to the base formed by the 3 other points (Sommerville 1929). Then, as the base is itself a (n-1)-simplex, in dimension (n-1), formed by n points, the height can be expressed simply as $h = n \cdot V_n / V_{n-1}$ (Sommerville 1929). A zero or non-zero volume in dimension n corresponds to a zero or non-zero height, respectively, so that the relevant dimension in a dataset can be determined using heights instead of volumes. Similarly, a contextual outlier can be revealed by its strictly positive height in dimension n, whereas other points would have zero height in dimension n.

- **In real datasets, the notion of strictly positive height can be replaced by the notion of non-normal height distribution.** In real datasets, neither a flat volume as contained in the plane in figure 1 would be strictly equal to zero, nor the corresponding height, because of the imperfect measurements in the data, which will make the volume or height calculations not fully exact. These deviations can be considered as random variables, and it is assumed here that they follow a normal (or Gaussian) distribution. This assumption is reasonable because

the normal distribution is encountered in many situations in nature or human sciences, though not always (Bortz 2005). Based on this hypothesis, zero height in a perfect case would correspond in a real framework to normal-distributed heights. Conversely, strictly positive heights in a perfect case in dimension n, indicating that this dimension n is still relevant, correspond to a real situation where heights do not follow a normal distribution. This can be visualized by considering that in a still relevant dimension, heights mirror - at least partly - the configuration of points (dimension 1 or dimension 2 in Fig.1), or else said the structure of the points, whereas in the first non-relevant dimension, heights reflect some noise, which is assumed normal-distributed here (dimension 3 in Fig.1, with the exception of the contextual outlier).

- **Decide if dimension n is relevant as compared to dimension n-1**. Dimension n is here taken as relevant if heights do not follow a normal distribution. This is done in practice by studying the distribution of height over a large number B (for example B=1000) of randomly sampled n-simplices replicates. This set of height is denoted $\{h_b\}_{b=1,..,B}$, or abbreviated as h. Normality or non-normality in the distribution of h can be measured by using a gaussianity or non-gaussianity index (see for example Hyvärinen and Oja 2000). A convenient choice is the kurtosis K, which is equal to μ^4/σ^4, where μ is the mean and σ is the standard deviation. Following Moors' interpretation (Moors 1986), kurtosis takes large positive values if the distribution is concentrated around its mean value (for example a normal distribution has a kurtosis equal to 3). I hypothesize here that positive values of K roughly correspond to noise or perturbations, thereby extending to distributions close to the normal distribution. Whereas, if K takes negative values, then the distribution is less concentrated around the mean (Moors 1986). In the method nSimplices, negative values of K are interpreted as actual structure in the data, as opposed to noise. The steep increase of K, and ultimately when K takes a positive value, is interpreted as the passage of a relevant dimension (n) to a non-relevant one (n+1).

Since kurtosis is calculated by taking the mean and standard deviation to the power of four, it is highly sensitive to outliers. For this reason the robust kurtosis formula proposed by Moors (1988) is used: $K = ((E7 - E5) + (E3 - E1))/(E6 - E2) - 1.23$, where Ei is the ith octile in the set $\{h_b\}_{b=1,..,B}$. As h is unsigned (so that only half of the distribution is observable), this

has to be simplified into the following equation: $K = (Q3 - Q1)/Q2 - 1.23$, where Qi is the i^{th} quartile in the set $\{h_b\}_{b=1,..,B}$.

Here is the algorithm I suggest and that is implemented in method nSimplices, to detect the relevant dimension of a dataset:

> for n in {n_{min},...,n_{max}}:
> for point i in {1,...,N}:
> for b in {1,...,B}:
> Pick a set of n points distinct from point i, without replacement
> Calculate V_n, volume of the n-simplex S_n formed by the set of n points and point i
> Calculate V_{n-1}, volume of the (n-1)-simplex S_{n-1} formed by the set of n points alone
> Calculate $h_i(b) = n \cdot V_n / V_{n-1}$
> Calculate $K_i^{(n)}$ on $\{h_i(b)\}_{b=1,..,B}$
> (optional: Take $h_i = median(\{h_b\}_{b=1,..,B})$)
> (optional: Take $hs_i = standard\ deviation(trim(\{h_b\}_{b=1,..,B}))$)
> Stop if $median(\{K_i^{(n)}\}_{i=1,..,N})/median(\{K_i^{(n-1)}\}_{i=1,..,N}) < c$

The cutoff c on the kurtosis index can for example be set to 0.5, to detect a large increase in kurtosis (this corresponds to a doubling of the kurtosis, when the initial kurtosis is negative). The ratio $median(\{K_i^{(n)}\}_{i=1,..,N})/median(\{K_i^{(n-1)}\}_{i=1,..,N})$ is denoted "kurtosis index".

- **Detect and accommodate outliers.** A point can be identified as outlier for example if the ratio of its height to the sub-space relative divided by the median distance between this point and other points, $hm_i = h_i/median(\{\delta_{ij}\}_{j=1,..,N})$, exceeds a threshold. This threshold can for example be set at $median(hm_i) + 3 \cdot median.absolute.deviation(hm_i)$. Remark 1: there is a bias when including δ_{ii} in the calculation, but it is here neglected since the median absolute deviation is highly robust. Remark: Other choices for outlier detection would be possible. For example, a point i_0 could be identified as outlier if $K_{i_0}^{(n)}/median(\{K_i^{(n-1)}\}_{i=1,..,N})$ is still below the cutoff c, while the kurtosis index is above it.

The accommodation of outlier i_0 is done by removing the component which lies outside of the relevant n dimensions. The height h calculated in dimension (n+1) corresponds to this outside component. Therefore the accommodation can be described by the following equation: ${}^{corr}\delta_{i0j}^2 = \delta_{i0j}^2 - h_{i0}^2$. Remark: instead of removing h_{i0}^2, it could make sense to remove the excess outlyingness $(h_{i_0} - median(h_{i_0}))^2$.

- **Perform dimension reduction.** In case the relevant dimension n is much higher than the number of dimensions desired, then a procedure similar to outlier accommodation can be performed, by considering that additional components in larger dimensions are a form of outlyingness of each point. Provided that h_i has been calculated for all points i=1,…,N, then distances are corrected by removing the highest outlying component. This can be written as follows: ${}^{corr}\delta_{ij}^2 = \delta_{ij}^2 - \max(h_i - h_j)^2$. Remark: possibly, outlying components of two points i and j lie along orthogonal components. Then the correction should be written $\delta_{ij}^2 - h_i^2 - h_j^2$. It is possible to check for this, for example by calculating the height of j in dimension n+2 relative to the simplex formed in dimension n+1 by i and n further points. This relative height is simply derived from the formula of height in dimension n (Sommerville 1929): $h_j(i) = (n+2) \cdot V_{n+2}/V_{n+1}$, with V_{n+1} including point i and a set of n+1 distinct other points, and with V_{n+2} including point j and all points of V_{n+1}. However, this calculation is intensive and does not necessarily improve the dimension reduction, which is why it was not carried out here.

- **Apply classical MDS at the reduced dimension.** Outlier accommodation or dimension reduction procedures lead to a distance matrix of the same size as before correction. Classical MDS (or other forms of MDS) can be applied to this matrix. As for other methods, the 10 top axes resulting from MDS are selected, in order to assess their representativity and their robustness.

2.4 List of methods applied to synthetic, EPIC and HMP datasets

The rationale and list of methods applied are detailed in §2.4.1 to §2.4.4 and summarized in table 1.

2.4.1 Exploration of synthetic dataset 1

The synthetic example 1 is constituted of random components interwoven with a structure component characterized by a small variance. This dataset is thought of as a proof of principle for the detection of small variance axes. Therefore, ICA and the main categories of methods were applied, as given in the following list: PCA, MCD, MDSe and nmMDS.

2.4.2 Exploration of synthetic dataset 2

This example was designed to illustrate RMDS performances and served as a motivation to develop nSimplices. Therefore, RMDS, nSimplices, and complementary MDS-based methods MDSe and nmMDS, are applied.

2.4.3 Analysis of the EPIC dataset

The following standard and robust methods were applied to the EPIC dataset:

- PCA-based methods: Eigenstrat, SPH-PCA, MCD, IBS
- MDS-based methods: LAR, RMDS, RSMDS, nmMDS
- own methods: gMCD, gMDS, rgMDS, nSimplices.

It should be noted that classical MDS (cMDS) has not been applied to the EPIC dataset, since cMDS has been compared to Eigenstrat for the discovery of genetic population structure in Wang et al. (2009), who found that both methods produced similar structure estimates (while Eigenstrat-derived principal components performed marginally better in association tests). More robust MDS-based procedures are nevertheless included, since they were not considered in Wang et al. (2009).

2.4.4 Analysis of the HMP dataset

It is useful to note that microbiome counts are sparse and therefore most raw-counts-based methods, used for SNP or other types of medical data are not appropriate. However, $\ell 1$-based PCA proved useful in microbiome analysis (Brooks et al. 2013) and was therefore included. Otherwise, microbiome count data can be transformed in a distance matrix, and analyzed by MDS-based methods. The methods used are:

- PCA-based methods: L1PCA,
- MDS-based methods: MDSm, MDSe, LAR, RMDS, RSMDS,
- own methods: wMDS, quantile-MDS, nSimplices,

Additionally, the method CSS (Paulson et al. 2013) was included.

Table 1. Methods contributed and applied to synthetic examples, EPIC dataset, and HMP dataset.

	Description	Cont.	SE1	SE2	EPIC	HMP
CSS	Multidimensional scaling based on standardized proportions					●
gMCD	Minimum Covariance Determinant procedure on unbiased genetic relatedness estimates	X			●	
gMDS	Multidimensional scaling on unbiased genetic relatedness estimates	X			●	
IBS	Multidimensional scaling on genetic Identity-by-State				●	
ICA	Independent Component Analysis		●			
L1PCA	Pure ℓ1-norm principal component analysis					●
LAR	Multidimensional scaling with least absolute residuals					●
PCA	Principal components analysis (Eigenstrat for EPIC)		●		●	●
MCD	Minimum covariance determinant, iterative procedure to find the main core of a dataset				●	●
MDSe	Multidimensional scaling with Euclidean distance		●	●		●
MDSm	Multidimensional scaling with Manhattan (taxicab) distance					●
nmMDS	Non-metric multidimensional scaling			●		●
nSimplices	Outlier-robust dimension reduction method to apply before e.g. MDS	X		●	●	●
qMDS	Multidimensional scaling on phylogenies counts normalized in empirical distribution quantiles	X				●
rgMDS	Robust multidimensional scaling using Huber loss function in an optimization procedure	X			●	
RMDS	Robust multidimensional scaling based on graphic networks and outlier regularization			●		●
RSMDS	Robust multidimensional scaling based on graphic networks and sparsely structured outlier regularization					●
SPH	PCA on a data matrix normalized to the unity sphere				●	
wMDS	Multidimensional scaling on standardized phylogenies counts	X				●

Cont.: contributed ; SE1, SE2: Synthetic example 1, 2.

2.5 Structure fit quality and Association testing

For all datasets, assessment of how good a method does represent the underlying structure is carried out by applying a regression (in R software, R core team 2015), where the structure or phenotype is considered as the outcome Y, and the main 10 axes are considered as the predictors. The regression is either linear or logistic (Hastie et al. 2001) depending if the outcome Y is continuous or categorical, respectively. I used AIC-based model selection (Hastie et al. 2001, and function *stepAIC* in R) to find out a subset of most representative predictors, and assess the quality of the model. The starting model for AIC-based model selection can be written as follows:

$$Y \sim PC_1 + PC_2 + \ldots + PC_{10},$$

where PC_i is the principal component with the ith highest variance, or the top ith axis yielded by MDS or other methods. The quantity used to compare quality of the model provided by each method is the AIC. Additionally, association testing is conducted on the EPIC dataset. In this case, principal components are still present in the model as covariates, and additionally each SNP is included as predictor.

For synthetic example 1, the outcome Y is categorical and corresponds to the underlying structure, a vector of length 2000 containing the group identifier, coded 0 or 1:

$$group \sim PC_1 + PC_2 + \ldots + PC_{10}\,.$$

In the synthetic example 2, Y is continuous and corresponds to the coordinates x and y:

$$x \sim PC_1 + PC_2 \quad \text{and} \quad y \sim PC_1 + PC_2$$

In the EPIC dataset, the structure is either the origin of the sample (country, genotyping center), in which case a logistic regression is performed, or its ancestry (CEU, Utah residents with Northern and Western European ancestry from the CEPH collection ; YRI, Yoruba in Ibadan, Nigeria ; ASA, Asian ancestry), in which case a linear regression is performed. Association testing consists in the logistic regression of the disease status on the candidate

SNPs rs520820, rs3783501, rs4955720, rs546950, rs388372, etc. from Campa et al. (2011), including principal components as covariates. This can be written:

$$disease.status \sim X + PC_1 + PC_2 + \ldots + PC_{10},$$

where the disease status is a vector of size N coded 0 (control) or 1 (case) for all N individuals and the variable X is a vector of size N, containing the sum of minor alleles for each individual at one of the candidate SNPs.

In the HMP dataset, the representativity of underlying structure is investigated by a logistic regression between body site and the samples. This can be written as follows:

$$site \sim PC_1 + PC_2 + \ldots + PC_{10},$$

where site is a vector of length N containing the site coded as a factor containing the multiple sites, or as a 0/1 binary variable for each site.

2.6 Evaluation of robustness

Robustness is assessed in EPIC and HMP datasets by applying a regression where each axis calculated on the original dataset is taken as the outcome, and where the 10 axes perturbed by the inclusion of outliers are included as predictors. This can be written as follows:

$$PC_i \sim PC_1^o + PC_2^o + \ldots + PC_{10}^o ,$$

where PC_i is the ith original axis, and PC_l^o is the lth perturbed axis, i and l are integers between 1 and 10. Methods are compared using adjusted R^2 (denoted simply R^2 in the following). The use of AIC is not necessary here, since the number of predictors is fixed. Further robustness measures, the influence function and the breakdown point are considered (see for example their description in Héritier et al. 2009). The empirical influence of outliers on each axis is taken as the value of R^2 of the top axis. The empirical breakdown point ε^* is taken as the proportion of outliers for which R^2 drops below the threshold value 0.70.

3 RESULTS

3.1 Detection of confounding structure with small variance

The synthetic dataset was investigated using ICA, PCA, metric and non-metric MDS, and the resulting axes were used as predictors of the underlying structure. The goodness-of-fit measure after variable selection, Akaike Information's criterion, is indicated for each method in table 2.

Table 2. Structure inference in synthetic example 1

	ICA	**PCA**	**MDSe**	**nmMDS**
AIC	434.7	1743.7	656.5	656.6

ICA performs best, with the lowest AIC value of 434.72 corresponding to 5 predictor axes selected, among which only one is highly significant (axis 7, p-value<2.2e-16). PCA selects 6 axes and performs worst with an AIC of 1743.7. MDSe and nmMDS achieve similar, intermediate values of AIC of 656.51 and 656.58, respectively and both select 8 axes.

Figure 2 presents the axes kept after variable selection. The number of axes selected is 5, 6 and 8 in the ICA, PCA and MDS methods, respectively, which confirms the advantage represented by ICA, since a lower number of predictors describes best the underlying confounding structure. Furthermore, ICA axis number 7 clearly displays the 2 groups, while no such structure is visible on scatterplots nor on histograms (visible on the diagonal of Fig.2B) of PCA axes. The axes derived from MDSe and nmMDS exhibit little visible difference. Axis 8 in MDSe and nmMDS seems the closest to the underlying structure, in

agreement with the lower AIC shown in table 2, relatively to PCA. This is further confirmed by calculating the Pearson correlation coefficient ρ_P between the group status corresponding to the underlying structure and the selected axis with the lowest p-value. Indeed, ρ_P is highest in ICA with a value of 0.88 (axis 7), lowest in PCA (ρ_P=0.55, PC 9) and intermediate in both MDSe and nmMDS (ρ_P=0.71, axis 8).

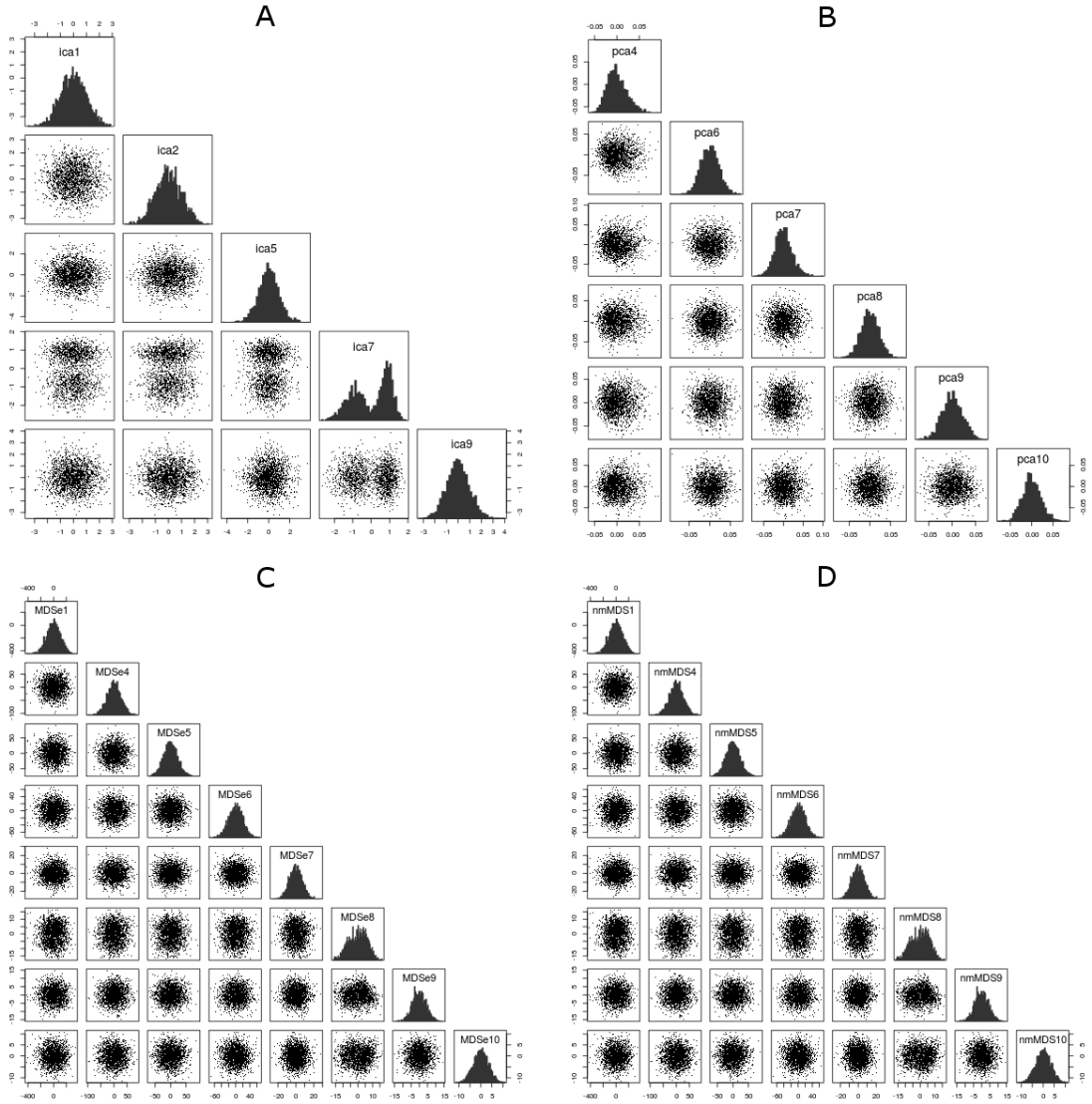

Figure 2. Pairwise scatterplots and histograms of the predictor axes chosen by a variable selection procedure, for synthetic example 1 (underlying 2-groups structure with low variance). Predictor axes are derived from the following methods: (A): ICA (Independent Component Analysis) ; (B): PCA (Principal Components Analysis) ; (C): MDSe (Multi-Dimensional Scaling on Euclidean distances) ; (D): nmMDS (non-metric Multi-Dimensional Scaling)

3.2 Detection of structure with noise and outliers

The synthetic dataset 2 studied here is a simple linear structure, shown in figure 4, contaminated with noise and with outliers. These outliers are either distance outliers, distance outliers and a distant point, or contextual outliers. RMDS and nSimplices are compared to classical and non-metric MDS (MDSe and nmMDS). The AIC of each method and each configuration is given in table 3.

Table 3. Structure detection accuracy (AIC) in synthetic example 2

Distance outliers								
%	MDSe		nmMDS		RMDS		nSimplices	
	x	y	x	y	x	y	x	y
0	-10.3	-11.7	-30.3	-29.1	72.4	74.3	-10.3	-11.7
0.25	80.6	75.7	23.1	5.7	66.3	104.5	80.6	75.7
0.5	111.0	51.8	11.3	9.9	65.0	85.2	111.0	51.8
1	101.7	125.0	6.1	0.3	77.5	95.6	101.7	125.0
2.5	104.9	102.9	38.4	29.0	103.5	88.0	104.9	102.9
5	105.0	103.3	68.0	45.2	86.4	101.4	105.0	103.3
7.5	120.7	118.8	69.9	52.0	107.9	107.5	120.7	118.8
10	123.3	119.6	80.4	74.1	94.7	107.6	123.3	119.6
Distance outliers and distant point								
0	12.0	11.4	-38.0	-56.5	153.0	158.9	12.0	11.4
0.25	103.6	106.0	5.1	-18.6	150.4	157.7	103.6	106.0
0.5	97.8	98.1	-21.0	-3.5	150.7	154.9	97.8	98.1
1	76.5	76.9	-15.3	2.2	151.5	158.8	76.5	76.9
2.5	79.4	72.7	31.7	10.9	148.9	158.0	79.4	72.7
5	99.0	100.0	108.4	121.8	159.0	148.3	99.0	100.0
7.5	102.2	99.7	105.6	125.1	162.7	148.6	102.2	99.7
10	104.8	103.6	94.0	132.1	155.4	152.8	104.8	103.6
Contextual outliers								
0	-10.3	-11.7	-30.3	-29.1	72.4	74.3	-10.3	-11.7
4	33.8	-11.7	59.3	6.0	65.2	110.7	7.3	-11.5
8	77.0	26.1	82.8	24.8	86.2	71.3	-9.0	25.1
12	100.0	112.9	87.1	84.5	98.2	103.7	-8.7	69.0

Remark: AIC can be negative because the likelihood density function (used in R for the calculation of AIC in linear regressions) can take values larger than 1. However this does not affect the comparison between models.

In the first configuration, distance outliers are present in proportions from 0% to 10%. For all methods, AIC increases when more outliers are introduced, in agreement with the fact that the structure is increasingly difficult to detect correctly when more and more outliers are

introduced. nmMDS performs best with an AIC of 80.4 and 74.1 for axes x and y, respectively, when 10% of outliers are present. RMDS performs less good than MDSe when no outliers are present (AIC of 72.4 and 74.3 in RMDS,), but it presents an advantage when 1% or more outliers are present, with an AIC of 94.7 (x) and 107.6 (y) in RMDS, as compared to 123.3 and 119.6 in MDSe. However, nmMDS performs in all cases better than both RMDS and MDSe. The method nSimplices correctly detects that the actual dimension is 2, in spite of noise and up to 10% outliers (table 4a). Indeed, kurtosis takes values from -0.34 to -0.17 for n=2, and values from -0.09 to 1.46 when n=3. The kurtosis index (ratio of kurtosis in dimension 3 to dimension 2) lies between -9.1 and -0.2, which is less than the cutoff c=0.5, leading to the correct detection of n=2 as the relevant dimension. In agreement with the tested configuration, no contextual outlier is detected. As no correction for outlyingness takes place, the inferred axes are the same as the subsequent method used, which is here MDSe. Accordingly, AIC takes the same values for nSimplices than for MDSe, which are second best when no outlier is present (AIC of -10.3 and -11.7 for axes x and y , respectively), and which performs worst when 10% of outliers are present (AIC of 123.3 and 119.6).

Table 4a. Dimension and outlier detection in nSimplices, configuration with distance outliers

	Distance outliers			**Distance outliers and distant point**		
%	**$K^{(2)}$**	**$K^{(3)}$**	**detected n**	**$K^{(2)}$**	**$K^{(3)}$**	**detected n**
0	-0.34	-0.09	2	-0.33	-0.11	2
0.25	-0.33	-0.06	2	-0.33	-0.10	2
0.5	-0.34	-0.08	2	-0.33	-0.06	2
1	-0.34	-0.06	2	-0.32	-0.06	2
2.5	-0.35	-0.06	2	-0.28	0.18	2
5	-0.23	0.77	2	-0.18	0.92	2
7.5	-0.21	0.92	2	-0.14	1.06	2
10	-0.17	1.46	2	-0.16	1.00	2

Table 4b. Dimension and outlier detection in nSimplices, with contextual outliers

Contextual outliers					
%	**$K^{(2)}$**	**$K^{(3)}$**	**detected n**	**detected CO**	**actual CO (h.)**
0	-0.34	-0.09	2	None	None
4	-0.35	-0.11	2	Point 5	5 (5.32)
8	-0.39	-0.21	3	Point 3	3 (9.49), 21 (6.07)
12	-0.35	-0.17	2	Points 3, 15	3 (13.62), 15 (17.83), 21 (3.72)

h.: outlying component

In the second configuration, a single and unvarying distant point is added, along with a growing proportion of distance outliers, similarly to the first configuration. The added point affects MDSe and nmMDS in a limited manner. AIC is indeed lower, in spite of the presence of the distant point, when no or few distance outliers are present (nmMDS: AIC -21.0 and -3.5 on axes x and y, respectively, when 0.5% distance outliers are present in the dataset), as compared to the case with no distant point. The largest AIC for MDSe is 104.8 (axis x) and 103.6 (y), which is better than without the distant point (123.3 and 119.6), and the largest AIC in nmMDS is 132.1, which mirrors less well the confounding structure than in the configuration without any distant point (where AIC takes the maximum value of 80.4). RMDS is the most affected method, since AIC takes the largest values, between 148.3 and 162.7. AIC does not seem to increase with the growing proportion of outliers, indicating that RMDS is affected by the single distant point rather than by the outliers. This is to compare with values between 65.0 and 107.6 obtained by the RMDS method when no distant point is present. The nSimplices method evaluates correctly that the relevant dimension is n=2, since the measured kurtosis takes values -0.33 (no outlier) to -0.14 (10% of outliers) when n=2, and -0.1 to 1.1 when n=3. As a consequence, the kurtosis index takes values between -7.6 and -0.2 , which is less than the cutoff c=0.5, leading correctly to the detection of 2 as the relevant dimension, and 3 as an irrelevant dimension, in spite of the noise and outliers introduced. No contextual outlier is detected, again in agreement with the configuration. In particular, the distant point is not detected as a contextual outlier, which is correct. As above, since no correction occurs, AIC values are the same for nSimplices as for MDSe.

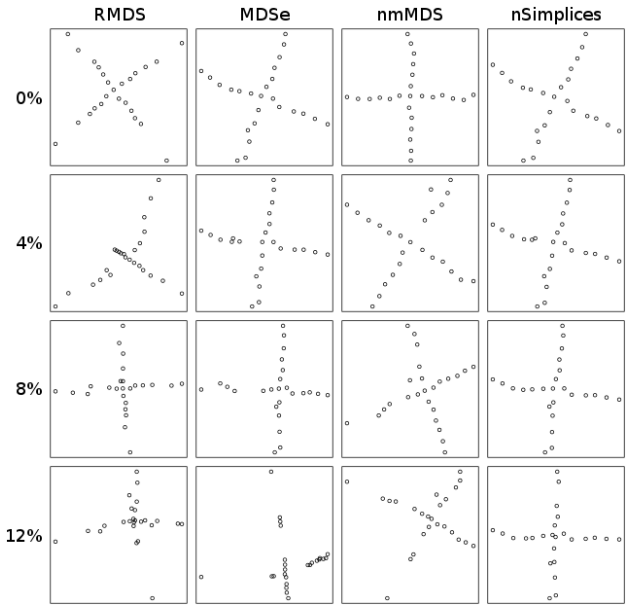

Figure 3. Robustness of axes inferred when contextual outliers are present.

The last configuration considers noisy distances affected by 1, 2 or 3 contextual outliers (4 to 12%). All methods are affected, as AIC reaches values of 112.9 for MDSe, 87.1 for nmMDS, 110.7 for RMDS, and 69.0 for nSimplices, for 12% of outliers. This is also visible in figure 3, which shows the detected structure for 0 to 12% of outliers. One of the branch of the structure seems shifted in method RMDS, when one contextual outlier is introduced. When more points are introduced, this shift disappears, but perturbations occur at each point of the structure. MDSe and nmMDS seem to resist to one or two contextual outliers, though some clear perturbations are visible, and to break down when three outliers are present (shifted branches in MDSe, uneven rendered distances in nmMDS). nSimplices is less affected, in agreement with the lowest AIC values as compared to the other methods, when 1 or more contextual outliers are introduced, indicating that the outlier accommodation part of the method does brings an advantage in this situation. Additionally, dimension assessment in nSimplices is still correct in spite of 4% or 12% of outliers. Contextual outliers are also identified by nSimplices and corrected for correctly, when one outlier is introduced. When 2 (3) outliers are introduced, 1 (2) outlier(s) are correctly detected and accommodated, whereas outlier 21 is neither identified nor corrected for. Figure 4 presents the standardized residual distances, for methods nSimplices and RMDS affected by contextual outliers. This shows that residuals in nSimplices lie between -1 and 1, with most residuals close to 0 and about 10 larger residual distances. These larger residuals correspond to the most affected pairs of points, for example

the second and third points from the top of the vertical branch, in figure 3. Regarding method RMDS, in agreement with AIC values, residuals take larger values, and for a larger number of points. At least one group of points seems shifted (indices 240 to 300), which is likely to correspond to a point with a large deviation, which in turn affects pairwise distances between this point and all the others (for example the point at the bottom of the vertical branch in figure 3).

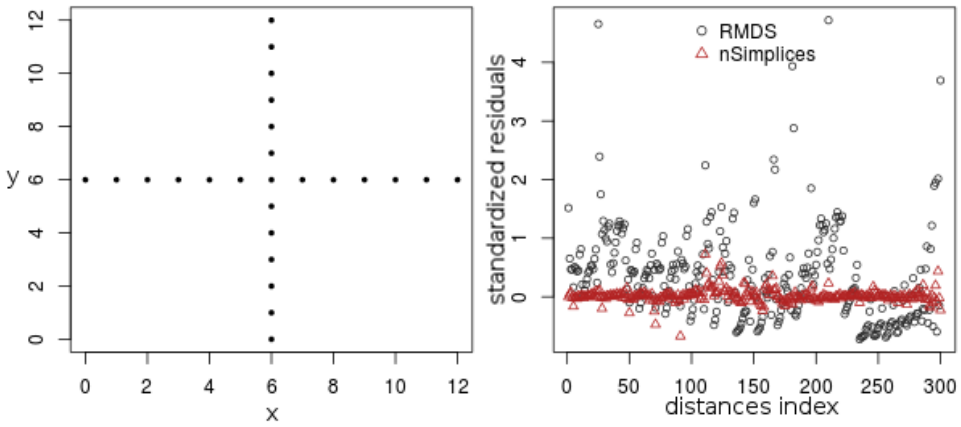

Figure 4. Underlying structure and standardized residuals with 3 contextual outliers.

3.3 Structure representativity and robustness, EPIC dataset

3.3.1 Structure representativity in EPIC dataset

The EPIC dataset comprises 860 individuals affected or not by prostate cancer. These individuals are stratified by country and center, and structured by ancestry components (European, African, Asian). How well each studied method reflects these underlying elements is reflected by how low AIC is, as given in table 5a (country, overall structure), table 5b (center, fine structure) and table 6 (ancestry components).

Most methods reflect stratification by country comparably, with for example an AIC between 316.3 and 372.5, when the structure examined is Denmark/other country. This is valid for most methods, i.e. SPH, EIG, nSimplices, rgMDS, gMDS, IBS, gMCD and MCD. The group of 4 non-metric and $\ell 1$-based methods comprising nmMDS, LAR, RMDS and RSMDS is clearly less representative of the underlying Denmark/other country structure, with AIC ranging from 980.0 to 1026.0. This pattern is similar for other countries, for example almost all methods have a very low AIC of 6.0 (or 8.0 for rgMDS) and thus are well representative of the dichotomy Spain / other country, whereas LAR, RMDS, RSMDS and nmMDS have a larger AIC lying between 244.4 and 349.3. There is one exception for Netherlands, where AIC takes comparable values of 321.4 to 348.2 for all methods. This seems to correspond to individuals which are in small numbers and which do not correspond to any cluster (according to the methods used), as can be seen in figure 5. Besides, the best method for each country is not often the best method overall. Indeed, the best AIC for Denmark, Germany, etc. corresponds to methods SPH, gMDS, IBS (which has most often the lowest AIC), nSimplices and EIG, whereas the best method overall is EIG, with an AIC of 782.8 (followed by SPH, gMDS, nSimplices, IBS, etc. with AIC between 789.6 and 938.4). Again, the 4 methods LAR, RMDS, RSMDS and nmMDS are less representative overall with AIC values of 2634.5 to 3145.7, which is consistent with the fact that they are less representative for each individual feature.

Table 5a. Detection of country stratification, EPIC dataset (AIC)

	EIG	SPH	MCD	IBS	gMCD	LAR	RMDS	RSMDS	nmMDS	gMDS	rgMDS	nSimplices
Denmark	321.9	316.3	372.5	332.4	362.9	1025.4	1022.5	1026.0	980.0	330.9	329.7	324.6
Germany	550.6	544.4	653.6	562.2	642.6	887.1	887.1	887.0	879.5	531.7	534.1	556.3
Greece	48.1	44.0	56.5	42.1	49.8	153.2	145.9	148.2	112.2	50.0	50.7	46.6
Italy	117.5	126.2	141.3	108.1	140.1	432.0	420.0	423.0	311.1	114.6	121.2	110.0
Netherlands	325.9	328.2	328.6	329.7	330.7	348.2	347.5	338.9	347.0	325.6	321.4	325.0
Spain	6.0	6.0	6.0	6.0	6.0	345.9	349.3	349.3	244.4	6.0	8.0	6.0
Sweden	6.0	6.0	8.0	6.0	19.3	643.0	638.8	622.2	424.3	6.0	8.0	4.0
UK	687.8	693.5	722.0	725.0	688.4	842.6	850.4	850.2	847.5	730.3	732.9	709.1
COUNTRY	782.8	789.6	938.4	829.7	864.9	3141.8	3145.7	3140.9	2634.5	797.8	854.8	802.9

Table 5b. Detection of center stratification, EPIC dataset (AIC)

	EIG	SPH	MCD	IBS	gMCD	LAR	RMDS	RSMDS	nmMDS	gMDS	rgMDS	nSimplices
Aarhus	378.5	379.8	384.6	387.6	383.6	560.9	557.2	561.1	542.7	381.1	385.0	379.9
Asturias	36.4	44.3	48.3	54.2	45.9	63.4	62.8	63.4	50.7	53.2	54.9	50.6
Bithoven	325.9	328.2	328.6	329.7	330.7	348.2	347.5	338.9	347.0	325.6	321.4	325.0
Cambridge	641.3	648.1	664.6	667.4	646.8	763.8	769.2	768.3	761.3	676.0	678.9	656.2
Copenhagen	493.7	488.3	520.2	493.4	511.8	819.4	815.3	819.4	804.1	491.6	494.4	489.1
Florence	114.6	112.5	132.4	111.2	117.8	274.6	271.3	272.8	216.9	120.2	121.1	110.7
Granada	12.0	27.2	12.0	12.0	16.0	30.3	30.3	22.7	10.0	22.3	23.8	8.0
Greece	48.1	44.0	56.5	42.1	49.8	153.2	145.9	148.2	112.2	50.0	50.7	46.6
Heidelberg	569.6	567.0	587.1	570.3	589.2	644.1	644.2	643.9	637.4	568.0	565.8	572.2
Navarra	63.5	63.5	62.8	57.7	60.3	91.5	92.3	87.0	81.7	62.1	63.7	61.6
Oxford	224.0	225.3	227.5	230.5	218.4	235.1	241.6	242.0	242.0	226.6	227.6	227.6
Potsdam	271.8	270.0	340.4	271.7	316.1	510.3	505.4	509.7	506.5	266.7	273.1	270.3
Ragusa	14.0	36.0	34.9	38.1	45.2	84.5	80.2	85.8	58.7	40.0	45.3	37.2
San Sebastian	60.7	64.3	65.7	68.7	62.1	251.8	255.6	253.4	182.0	70.8	70.5	62.5
Turin	90.5	93.5	103.8	98.6	99.3	134.2	127.3	123.9	131.3	97.9	103.0	93.2
Umea	6.0	6.0	8.0	6.0	19.3	643.0	638.8	622.2	424.3	6.0	8.0	4.0
Varese	53.4	54.1	52.8	49.7	52.9	64.5	73.5	73.5	60.0	46.1	47.2	51.4
CENTER	1729.9	1756.9	1907.5	1769.9	1813.6	4095.8	4096.7	4096.7	3636.0	1745.5	1897.0	1738.8

Stratification by center is also characterized by two groups of methods, where most have comparably low AICs, for example 4.0 to 19.3 for methods nSimplices to gMCD, regarding the genotyping center Umea, whereas methods nmMDS, LAR, RMDS and RSMDS are notably less representative, with AIC ranging from 424.3 to 643.0. This is true for many centers but not all, in particular the gap is less large for centers Granada and Oxford (AICs between 8.0 and 30.3, and between 218.4 and 242.0, respectively). The best representative

method is either EIG, SPH, IBS, gMDS, rgMDS or nSimplices when centers are taken individually, as already observed above, and the best method overall is likewise EIG (AIC=1729.9). The next best methods are nSimplices, gMDS, SPH and IBS (AIC from 1738.8 to 1769.9).

Table 6. Detection of ancestry structure, EPIC dataset (AIC)

	Ancestry		
	CEU	**ASA**	**YRI**
EIG	-4915.8	-4999.3	-5476.8
SPH	-4853.9	-4953.3	-5369.3
MCD	-4773.4	-4897.8	-5327.7
IBS	-4836.2	-4802.7	-5074.1
gMCD	-4946.1	-5018.4	-5550.2
LAR	-3460.1	-4171.4	-4034.7
RMDS	-3463.0	-4182.0	-4031.0
RSMDS	-3470.7	-4186.4	-4033.9
nmMDS	-3508.2	-4177.8	-4112.5
gMDS	-4846.4	-4999.2	-5458.8
rgMDS	-4859.4	-5010.6	-5519.1
nSimplices	-4627.2	-4749.9	-4918.2

Structure by ancestry as detected by the different methods is given in table 6. Comparably, most methods obtain low AICs for CEU (European) ancestry, ranging from -4946.1 (gMCD) to -4627.2 (nSimplices), whereas larger values are obtained in nmMDS, LAR, RMDS and RSMDS (-3508.2 to -3460.1). The best method for all three ancestry components is gMCD with AIC of -4946.1, -5018.4 and -5550.2 for CEU, ASA and YRI ancestry, respectively. Notably, the gMCD method did underperform most methods in terms of country or center stratification though it is best in terms of ancestry. Inversely, nSimplices underperforms many methods in terms of ancestry, whereas it performed good or sometimes best in terms of country and center stratification.

Regarding dimension detection and reduction by method nSimplices, the kurtosis calculated in each dimension n=2 to n=11 remains stable (table 6). As a consequence, the kurtosis index does not fall below the cutoff c=0.5, which indicates that the relevant dimension is likely more than n=10. This suggests that more than 10 axes would maybe be necessary to fully

describe the underlying structure in the EPIC dataset. This is consistent with the fact that some elements are well mirrored by the different axes (for example stratification for countries Spain and Sweden with AICs as low as 4.0 or 6.0, respectively), whereas others are substantially less well represented (e.g. Germany or UK with AICs no better than 534.1 and 687.8, respectively).

Table 7. Kurtosis index for dimension detection in method nSimplices, EPIC dataset.

n	2	4	6	8	10
kurtosis	-1.226	-1.227	-1.228	-1.228	-1.228

The main axes determined by methods EIG, SPH, gMCD and rgMDS are shown in pairwise scatterplots in figure 5. In agreement with the low AIC for extreme geographical locations shown in table 5 (for example Spain, Sweden, Umea), the structure of origin within Europe seems well represented by these axes. Though all depicted methods are well representative, they seem sensitive to a limited number of outliers, for example axis 4 of EIG and axis 1 of SPH but also gMCD on axis 1 or rgMDS on axis 4.

Regarding globally the presented results (tables 5 and 6), methods non-metric MDS and ℓ1-based methods did not seem to converge fully. In particular, the objective function which is used in RMDS and RSMDS iterative algorithm did not constantly decrease, whereas it should. These procedures had to be stopped after the maximum number of iterations was attained, but the non always decreasing objective function implies that the optimal solution may not have been found. Non-metric MDS also relies on an optimization and can have found a local minimum, which would explain the results obtained, which are less good than those produced by MDSe. New methods gMCD, gMDS and rgMDS bring some improvements in specific situations. For example, gMCD performs best in terms of ancestry. The optimization-based rgMDS improves the model quality as compared to gMDS, as indicated by AIC values decreased from -4846.4, -4999.2 and -5458.8 to -4859.4, -5010.6 and -5519.1, in CEU, ASA and YRI ancestry, respectively. The nSimplices method, which is applied to a genetic distance like gMCD, gMDS and rgMDS, brings an improvement in terms of fine-scale structure (for example, AIC by single country or center is sometimes better than the one produced by EIG), and brings a comparable or moderately improved result respectively to a method relying on

the same distance (gMDS, gMCD). For example AIC is decreased from 1745.5 (gMDS) to 1738.8 (nSimplices) for the overall structure by center. However the ancestry structure is better represented by gMDS or gMCD than by nSimplices.

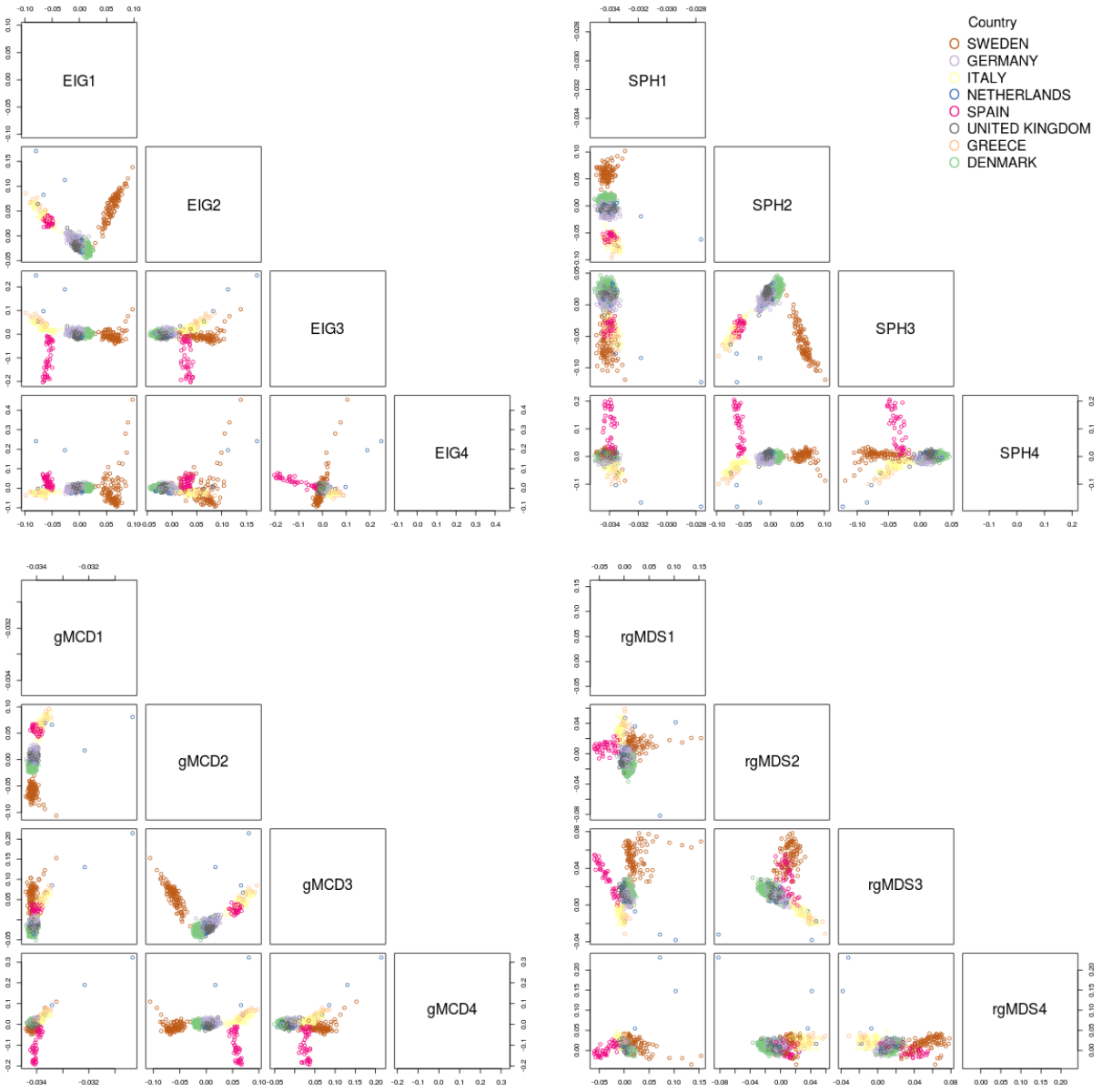

Figure 5. Pairwise scatterplots of axes 1,..,4 produced by a subset of methods.(A): EIG (Eigenstrat) ; (B): SPH (spherical PCA) ; (C): gMCD (genetic Multi-Dimensional Scaling ; (D): rgMDS (robust genetic Multi-Dimensional Scaling)

3.3.2 Robustness in EPIC dataset

Subsequently, synthetic outliers (bad leverage points) were introduced to strain the different methods in terms of robustness. The methods performing best, and MCD, are shown in Fig.6, panels A and B (whereas panels C and D correspond to real outliers and are described later). The full table which was used to produce this figure, including all methods, is given in table S1 in appendix. The influence of outliers on the top axis is represented in panel A, with SPH and IBS methods staying the most robust, since R^2 is 1.00 from 0 to 10% of outliers. Then, gMCD and nSimplices perform better or slightly better than the reference method EIG (R^2 of 0.96, 0.88 0.85, respectively, when 10% outliers are present). In contrast, MCD proves least robust in this subset of methods, with R^2=0.69. The method gMDS (table S1, not shown in figure 6A), is more robust than MCD but less than EIG, with a minimum R^2 of 0.81 for axis 1. Approaches LAR, RMDS and RSMDS maintain their first axis (minimum R^2 is 0.98, 0.75 and 0.97, respectively, table S1) in spite of the presence of 10% outliers. The method nmMDS (table S1) performs worst, since R^2 is 0.00 when 0.25% or more outliers are included.

Then, axes 1 to 10 are examined in terms of breakdown point ε^* (Fig.4B), corresponding to the situation when R^2 falls under a cutoff of 0.70. The breakdown of axis 1 occurs at more than 10% outliers for most methods represented, except MCD (where ε^* is not larger, but equal to 10%). A large discrepancy exists for axis 2, where SPH, IBS and gMCD maintain a breakdown point higher than 10%, but the other methods have $\varepsilon^* = 2\%$ (EIG, MCD, nSimplices). The information contained in axes 3 to 10 is not well maintained when outliers are introduced, as breakdown occurs at 1% (axis 3 and 4, except MCD), or less (axes 5 to 10). The method gMDS (table S1) has a breakdown at 2% or less for axes 2 to 10. Methods nmMDS, LAR, RMDS and RSMDS are least robust with a breakdown at 0.25% for axes 2 to 10.

Alternatively, outliers sampled from a real population were included in the dataset. The methods most robust to this configuration are depicted in figure 4, panels C and D, and the full data is provided in table S2 in appendix. The influence on axis 1 is shown on panel C. All

methods shown keep an almost identical first axis, even with 10% of outliers. SPH is slightly less good and MCD performs least well in this subset of methods, with R^2 of 0.98 and 0.95, respectively. Remaining methods also maintain their first axis well, except nmMDS, which presents with a R^2 of 0.00 when 0.25% or more outliers are present.

The breakdown point ε^* for the configuration with real outliers is presented in figure 4, panel D. Breakdown ε^* is higher than 10% for axes 1 to 5 in all methods in figure 4 except MCD (0.25% for axes 4 and 5). On axes 6 to 10, nSimplices and SPH keep the largest breakdown point with ε^* values of 10% (axes 6, 7) and equal or more than 2% (axes 8 and 9). Then, methods EIG, IBS and gMDS drop to 5% (axis 6) and 2% or less (axes 7 to 10). IBS (table S1) has a breakdown point at 5% for axes 6 and 7, and 0.25% for axes 8 to 10. Methods nmMDS, LAR, RMDS and RSMDS prove least robust with real outliers, similarly to the situation with synthetic outliers, with a breakdown point ε^* of 0.25% for axes 2 to 10.

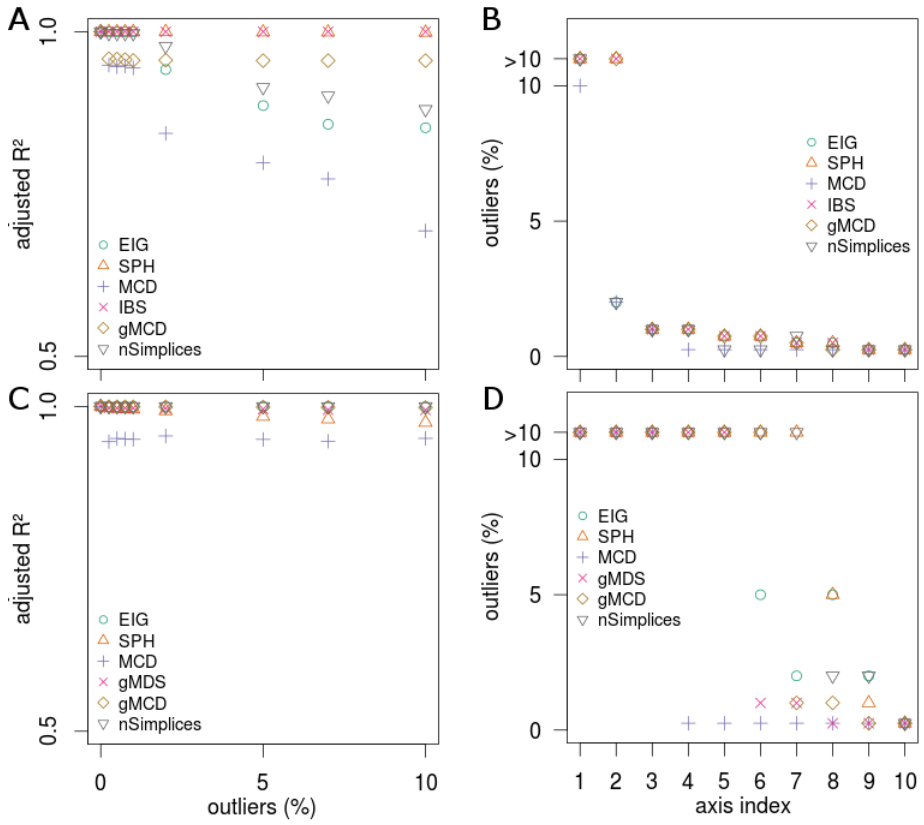

Figure 6. Adjusted R^2 for axes disturbed by synthetic outliers, for methods EIG, SPH, MCD, IBS, gMCD and nSimplices in panels (A): influence function ; (B): breakdown point ε^* for axes 1 to 10. Adjusted R^2 for real outliers from 1000genomes, methods EIG, SPH, MCD, gMCD, gMDS and nSimplices in panels (C): influence function ; (D): breakdown point ε^*.

3.3.3 Association test, EPIC dataset

Finally, candidate SNPs in EPIC prostate (Campa et al. 2011, table 2) were submitted to an association test with the prostate cancer disease status. The p-values corresponding to these tests are shown in table 8. Consistently with Campa et al. (2011), no significant association is found (with a predefined significance level at 0.05).

Table 8. Candidate SNPs and their p-values as given by different methods, EPIC dataset

	EIG	SPH	MCD	IBS	gMCD	gMDS	rgMDS	nSimplices
rs520820	0.60	0.61	0.64	0.74	0.57	0.63	0.64	0.61
rs3783501	0.11	0.11	0.11	0.12	0.10	0.13	0.13	0.10
rs4955720	0.30	0.32	0.30	0.34	0.31	0.35	0.36	0.30
rs546950	0.84	0.85	0.90	0.91	0.88	0.88	0.88	0.84
rs388372	0.99	0.99	0.97	0.99	0.89	0.98	0.99	0.96

Next, an extended set of candidate SNPs corresponding to supplementary table 1 of Campa et al. (2011) was also tested for association with prostate cancer status. Table 9 shows the 20 SNPs which obtain the lowest p-values in all methods. Though, when adjusting for multiple testing by the Bonferroni method (cutoff of 4.6e-5, corresponding to 0.05 divided by the number of tests, here 1 084), no p-value remains significant.

The top SNP, rs8071475, has been cited in the context of gene expression in colorectal cancer (Slattery et al. 2014) and in kinase activity-affecting genes (Wang et al. 2015). However, genome-wide studies did not report any significant association between prostate cancer and rs8071475 (www.gwascentral.org, Beck et al. 2013), although there were significant associations with other diseases. Similarly, the second and third top SNPs were reported for other disorders than prostate cancer, but not in prostate cancer.

Table 9. Top 20 SNPs with lowest p-values as given by several methods

	EIG	SPH	MCD	IBS	gMCD	gMDS	rgMDS	nSimplices
rs8071475	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
rs3799631	0.004	0.004	0.004	0.003	0.003	0.003	0.003	0.003
rs2589118	0.005	0.005	0.004	0.007	0.004	0.007	0.007	0.005
rs17239241	0.007	0.007	0.008	0.004	0.008	0.006	0.006	0.008
rs7778077	0.009	0.009	0.014	0.014	0.010	0.011	0.012	0.011
rs4425665	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
rs10235949	0.01	0.01	0.02	0.01	0.01	0.02	0.02	0.01
rs6443624	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
rs4648553	0.02	0.02	0.02	0.02	0.03	0.02	0.02	0.02
rs7789699	0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.02
rs12703162	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
rs3799619	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
rs11757572	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
rs7621329	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
rs7754623	0.03	0.03	0.03	0.04	0.02	0.03	0.03	0.03
rs4600802	0.03	0.03	0.03	0.05	0.03	0.03	0.03	0.03
rs3934597	0.03	0.03	0.05	0.04	0.03	0.04	0.04	0.03
rs6965834	0.03	0.03	0.04	0.04	0.03	0.03	0.03	0.04
rs1569462	0.05	0.05	0.04	0.06	0.05	0.06	0.06	0.04
rs3747636	0.05	0.06	0.05	0.06	0.05	0.06	0.06	0.05

3.4 Structure representativity and robustness, HMP dataset

3.4.1 Structure representativity in HMP dataset

Samples and phylogenies in the HMP dataset are structured by body site where the microbiome was collected from. Representativity of each method with respect to this structure is measured by the AIC shown in table 11 for all methods studied except L1PCA, since this last method demanded long calculations (more than 48 hours for one point, in spite of the 20-fold acceleration obtained by recoding the procedure in Python linked to Fortran). Before examining table 11, the model distribution on which the method qMDS relies has to be selected. This choice is based on the AIC of the fit of each considered distribution with the phylogenies proportions, for a subset of samples. These values are given in table 10 as the median AIC and its variability, measured by the median absolute deviation. The low AIC (488.7) obtained by the negative binomial distribution motivates its use, since it presents the

lowest AIC as well as the lowest variability. The exponential distribution presents a larger AIC of 602.6. However this distribution uses only one parameter where the negative binomial uses 2, so that it is always possible that the advantage obtained in the negative binomial is due to over-fitting. Therefore, the exponential distribution is also considered in the subsequent analyses. The Poisson distribution is not further used because it is least appropriate, since its AIC and corresponding variability are substantially larger. The same analyses conducted on a second subset of samples led to the same conclusions. qMDS with negative binomial (exponential) model distribution is denoted q^{NB}-MDS (q^{E}-MDS).

Table 10. Model distributions for qMDS method. AIC median and AIC MAD (median absolute deviation) are calculated from two independent subsets of 50 samples.

Distribution	**Median AIC (MAD)**			
	Set 1		Set 2	
Exponential	602.6	(235.6)	620.1	(198.6)
Negative Binomial	488.7	(170.7)	532.0	(176.9)
Poisson	26 549.6	(19 585.6)	24 882.6	(15 601.2)

Among methods presented in table 11, the most representative of individual body sites is nSimplices, combined either with q^{NB}-MDS, q^{E}-MDS or CSS (for example the sites throat, ears and mouth with AIC 628.0, 606.9 and 30.2, respectively). Then, a second group of methods performs less well with larger (though not exceedingly) values of AIC between 14.0 and 1153.7. This group comprises MDSm, MDSe, q^{NB}-MDS, q^{E}-MDS and CSS (these three latter methods without prior dimension reduction with nSimplices). Notably, MDS based either on Manhattan or Euclidean distance performed identically, with the same values of AIC. Besides, the CSS method performs most often better than MDSm and q-MDS to represent the fine structure (sites throat, nose, elbows, mouth). Methods LAR, RMDS, RSMDS, as well as the new method wMDS (based on proportions with standardized variance) do no perform well, since AIC is largely increased, for example AIC takes values of 1184.3 to 1276.1 at body site vagina, whereas most methods obtain an AIC of 10.1 to 95.1. This is true for all body sites.

The best method to represent the overall structure is nSimplices combined with q^E-MDS, with an AIC of 1587.4. The next best methods are nSimplices combined with CSS, CSS alone and nSimplices combined with q^{NB}-MDS, with AIC values of 1643.2, 1714.7 and 1742.7, respectively. In agreement with the performance achieved on individual features, overall AIC is larger in MDSm, MDSe, q^E-MDS alone and q^{NB}-MDS, and clearly inflated in LAR, RMDS, RSMDS and wMDS.

Table 11. AIC for HMP microbiome site on 10 top axes from standard and robust methods

	MDSm	MDSe	LAR	RMDS	RSMDS	wMDS	q_{NB}-MDS	q_E-MDS	CSS	nSimplices		
										q^{NB}-MDS	q^E-MDS	CSS
Throat	742.9	742.9	1016.7	1016.4	914.9	1017.9	661.7	700.8	654.0	628.0	696.9	661.1
Ears	633.6	633.6	1738.0	1753.7	1526.7	1752.8	1082.0	1153.7	937.5	927.5	606.9	699.4
Stool	469.7	469.7	1028.5	1029.5	735.9	1024.6	554.7	528.8	536.1	465.6	431.3	545.5
Nose	915.9	915.9	3127.0	3127.0	2830.7	2898.6	917.2	1080.9	830.7	785.1	780.7	722.7
Elbows	781.6	781.6	1596.1	1590.8	1415.4	1355.3	806.0	695.5	726.1	756.5	693.3	846.1
Mouth	50.4	50.4	1076.0	1079.1	966.7	585.9	53.1	102.6	49.5	32.7	51.5	30.2
Vagina	14.0	14.0	1275.9	1276.1	1184.3	1273.3	75.4	95.1	57.3	33.8	10.1	16.0
SITE	**1801.1**	**1801.1**	**7142.4**	**7146.7**	**5742.1**	**5729.3**	**1898.8**	**1863.2**	**1714.7**	**1742.7**	**1587.4**	**1643.2**

The kurtosis detected by the nSimplices method is shown on figure 7. Kurtosis does not increase and actually seems to decrease between n=2 and n=10. As a consequence, the relevant dimension is not detected for n≤10, and is likely to be higher than 10 in this dataset.

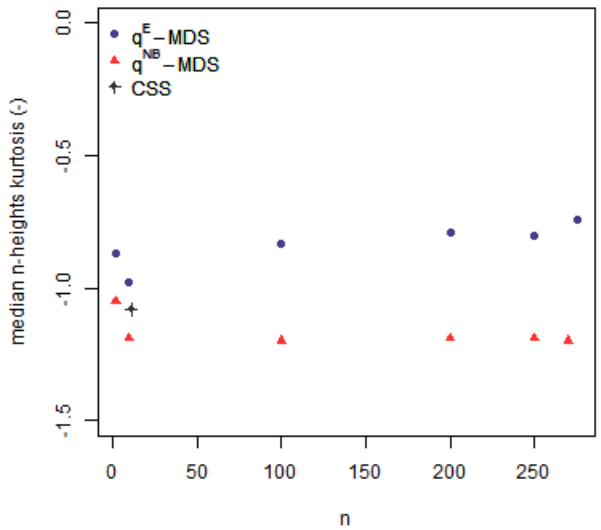

Figure 7. Kurtosis of heights in dimension n (median taken over a set of 1000 replicates).

For higher values of n, kurtosis stays stable in q^{NB}-MDS and increases slightly in q^{E}-MDS. The kurtosis index does not fall under the predefined cutoff c=0.5, so that a detection of relevant dimension also does not happen for n=10 to n=260.

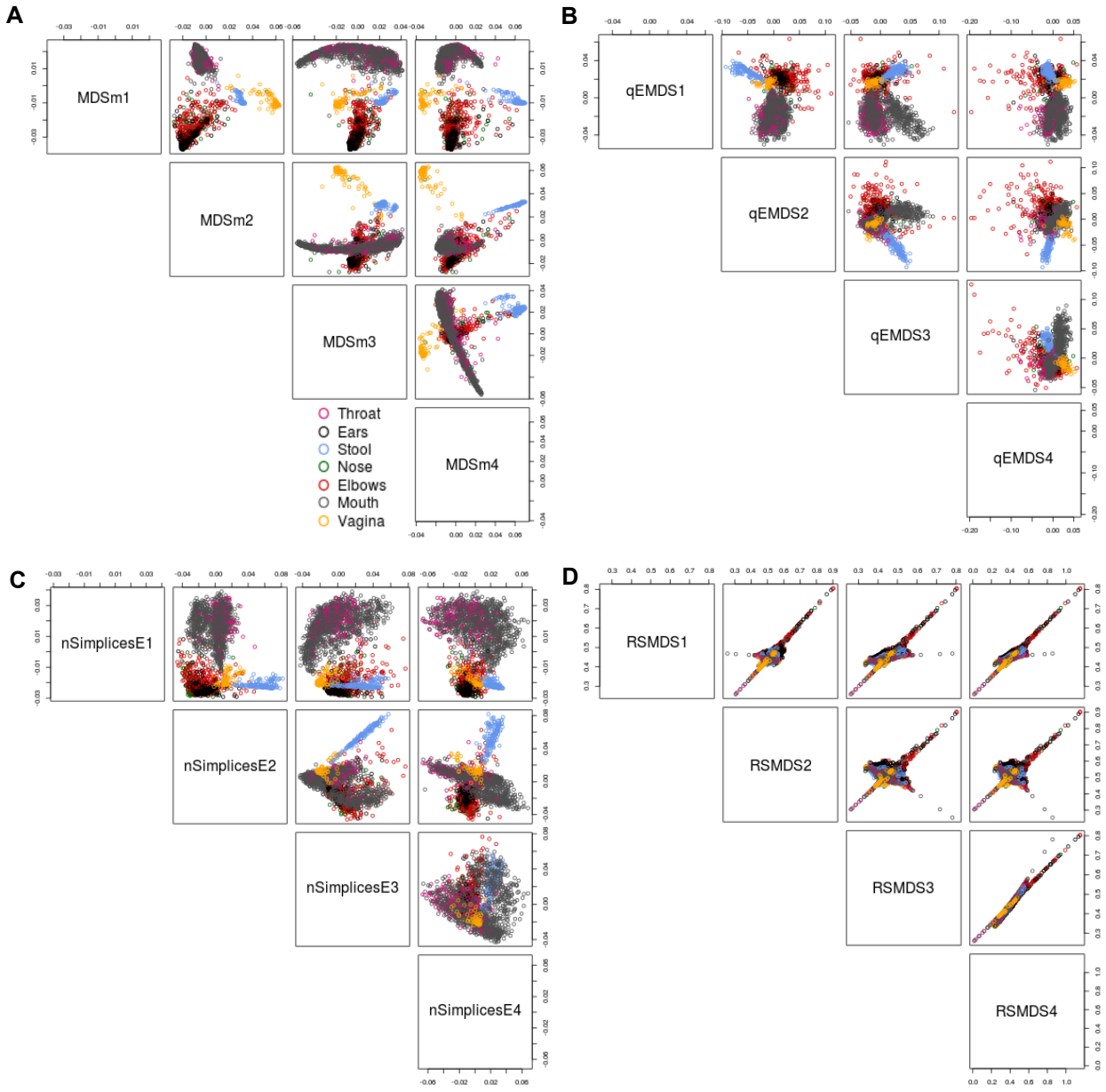

Figure 8. Structure restituted on 4 axes, in the HMP dataset, using methods (A) MDSm, (B) q^{E}-MDS, (C) nSimplices-q^{E}-MDS and (D) RSMDS.

The top axes produced by MDSm, q^E-MDS, nSimplices combined with q^E-MDS, and RSMDS are shown in figure 8. The way the underlying structure is reported is markedly different depending on the method. As an example, the mouth structure (in grey) forms a circular arc in MDSm, but is represented by two lobes in q^E-MDS (axes 1 and 3), while nSimplices exposes both aspects of this structure (2 lobes in axes1 and 2, circular arc in axes 1 and 3). MDSm seems to reveals other arc-shaped groups of points, for example on axes 2 and 3, whereas q^E-MDS exposes roughly ellipsoidal clusters of points. The combination of q^E-MDS and nSimplices leads to varied structures, which apparently better reflect the actual data structure, since AIC is lowest for this method (table 11). RSMDS axes show a cross-shaped structure and seem influenced by three points of the mouth body site (in grey), which take large coordinates along axes 2, 3 and 4.

3.4.1 Robustness in HMP dataset

In figure 9 is exposed the influence of outliers and the breakdown point of a subset of relevant methods applied to the HMP dataset. Exhaustive results are given in table S3, in appendix. All methods shown in figure 9 are robust, since R^2 stays close to 1 even with 10 % of outliers for axis 1, except for wMDS, for which R^2 drops to 0.47 (0.40) with 2.5% (10%) of outliers (table S3). Further methods MDSe, LAR, q^{NB}-MDS (alone or combined with nSimplices) and RMDS also exhibit little influence on the top axis, with R^2 equal to 1.00 (except LAR: R^2=0.98). Method RSMDS is strongly influenced since R^2 is 0.19 or less when 2.5% or more outliers are included.

Most methods (in figure 9) are equally robust in terms of breakdown point ε^* as in terms of influence curve, since this characteristic is larger than 10% for axes 1 to 9 for all methods except wMDS (2.5%). This is also true for MDSe and q^{NB}-MDS (alone or combined with nSimplices), but not for LAR, RMDS and RSMDS, for which ε^* is 2.5% for axes 2 to 9 (table S3). On axis 10, the methods q^E-MDS or q^{NB}-MDS, in combination with nSimplices or alone, are the only methods to stay robust, with a breakdown point ε^* still larger than 10%. CSS-nSimplices has a breakdown point of 5%, and remaining methods MDSm, CSS, wMDS, LAR, RMDS and RSMDS have a breakdown point no larger than 2.5%.

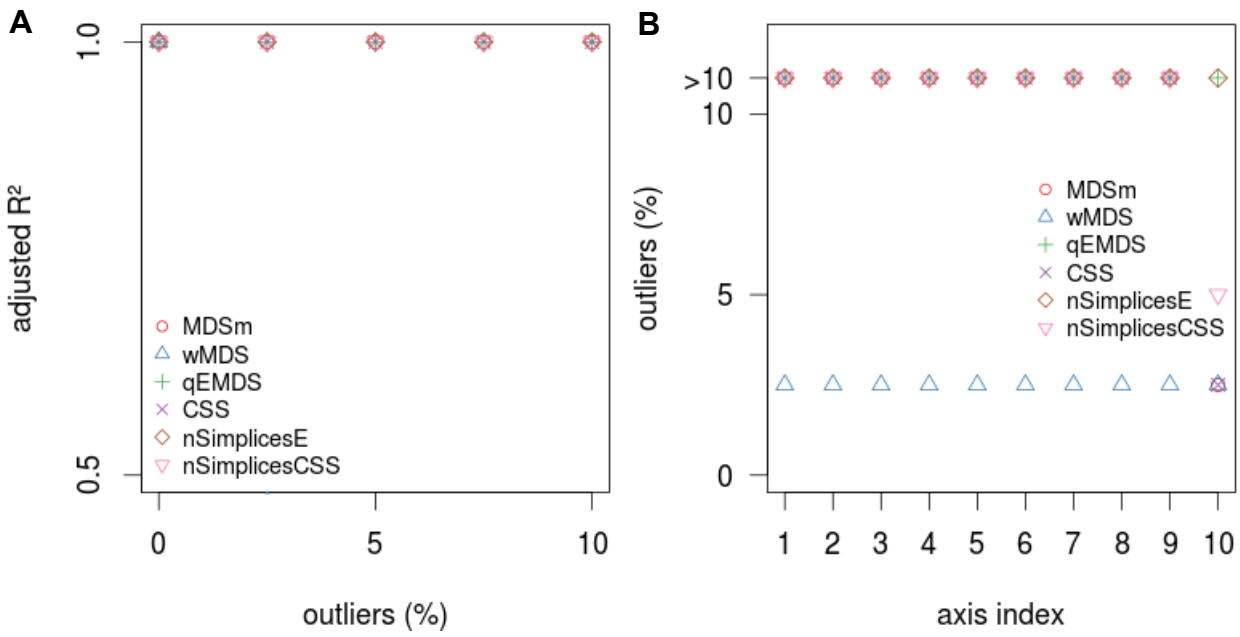

Figure 9. Robustness in HMP dataset (A) Influence function ; (B) Breakdown point ε^*.

4 DISCUSSION

4.1 Robust exploration and control of underlying structure

The objective of this work was the robust exploration and correction of underlying structure in case-control association studies. To this aim, several standard and robust PCA- and MDS-based methods, as well as some contributed improved methods, whose role is to mirror and to correct for the confounding structure and outliers, were compared. Each method produces a summary of the underlying structure on a limited number of axes, so that these axes can be included in the frame of genetic case-control association studies, of in the frame of microbiome association studies. The performance of each method was assessed in terms of representativity (meant here as the ability to summarize the confounding structure in a limited number of variables) and robustness (ability to produce stable estimates in the presence of perturbations, here outliers). The results obtained in the synthetic examples, in genetic SNP data from the EPIC study, and in microbiome data from the HMP project are discussed, in particular within the frame of currently published research, in §4.2.1 to §4.2.3. Then, a focus on advantages and limitations of method nSimplices is given in §4.3. Finally, the conclusions drawn from this thesis and an outlook are exposed in §4.4.

4.2 Confounding structure and outliers in synthetic, genomic and microbiome data

4.2.1 Structure identification in synthetic examples 1 and 2

Results obtained on synthetic example 1 highlight the fact that a structure with low variance could stay undetected in a situation where the number of samples is high when compared to

the number of observations. The only method able to detect the hidden signal of two underlying groups with close but distinct locations was ICA, which is in agreement with the fact that this method detects preferably non-gaussian signals. Methods based on least squares such as PCA or MDS could not detect this 2-groups structure. This is in agreement with a similar yet less extreme observation made by Günther et al. 2014, where 2 groups in an integrative biology analysis could be better distinguished using ICA than with PCA. This stresses the influence of the implied solution shape (low entropy or normal distribution in ICA and PCA, respectively) on the solution. This could be of incidence in the frame of a study design where groups, or more broadly a non-gaussian structure is of interest. This has been recognized for instance in the context of neuronal activity (Bell and Sejnowski 1997, Delorme et al. 2007) but is seldom considered yet outside of this field.

In the second example, method RMDS (and RSMDS in HMP data) showed a tendency to preferably account for a structure made of crossing lines (Figures 4 and 8), which represented the underlying structure less well than other methods (Tables 3 and 11). The fact that both RMDS and RSMDS rely on a graphical network might explain this characteristic, which proved detrimental here, but which could be useful in case such a representation is looked for.

Both presented examples about ICA/PCA and RMDS/RSMDS corroborate the principle that different methods are able to discover or account for different patterns, similarly to an experimentator who might select different probes to measure specific traits. This encourages to use and compare several complementary methods, and is further justified by the observations of e.g. Sampson et al. (2011), Günther et al. (2014) who showed the benefits of this strategy in proteomics and multi-omics analyses, respectively.

Regarding robustness to outliers in these examples, both distance outliers or contextual outliers degraded strongly the estimation of underlying structure, except when nmMDS or nSimplices methods were used, respectively (in example 2). The fact that MDS is sensitive to outliers and that non metric MDS is able to overcome this sensitivity is known and in agreement with e.g. Spence and Lewandowsky (1989) who developed and applied an instance of non-metric MDS in a setting with outliers, and with Liu et al. (2013), who observed

inflated false positive rates in MDS or PCA when outliers are introduced. These outliers could in principle be removed using MCD, as done in Liu et al. 2013 or in Song et al. 2007, but method nSimplices proved able to keep the contextual outliers in the dataset with no substantial loss in estimation quality.

4.2.2 Structure identification in genetic SNP data, EPIC study

Structure identification and association test in the EPIC study. PCA (Eigenstrat) and robust methods like SPH and IBS did allow to identify underlying structure well. Nevertheless, new contributions gMCD and nSimplices brought an advantage in terms of representativity (in some cases) and robustness, which comes at a small to moderate additional calculation cost (of the order of minutes for a PCA, tens of minutes for nSimplices, hours for gMCD). The MCD method applied to SNPs did not yield the anticipated advantages, probably because the matrix determinant is often very small, which impairs the convergence of MCD since it is precisely based on this determinant. However, the use of genetic relatedness estimates in gMCD did overcome this issue. Also unexpectedly, non-metric MDS (nmMDS) performed poorly in the EPIC dataset. As nmMDS relies on optimization, the solution found could still be a local minimum (Shinkareva et al. 2013, citing Groenen and Heiser 1996 and Hubert et al. 1992), which would explain this result. A possible way to address this issue could be to use nmMDS after dimension reduction with nSimplices. The gMDS method, though using genetic unbiased relatedness estimates as in gMCD, proved less robust than other methods, and slightly less representative than Eigenstrat or IBS. The unbiased relatedness estimates would however probably bring a benefit in a context when admixture or genetic relatedness is substantial, for example in Hispano-American populations (unpublished data on South-American populations from 1000genomes and the HGDP panel www.cephb.fr/en/hgdp_panel.php).

Association p-values between prostate cancer affection status and 1 084 SNPs examined in the EPIC publication by Campa et al. (2011) lead to comparable, non-significant p-values for a subset of 5 candidate SNPs. However, 20 other SNPs obtain significant p-values, which however do not stay significant after adjustment for multiple testing. As the three top SNPs

rs8071475 (colorectal cancer, Slattery et al. 2014, kinase activity-affecting genes Wang et al. 2015), rs3799631 and rs2589118 have been detected in other contexts than prostate cancer, but not in prostate cancer (www.gwascentral.org, Beck et al. 2013), these results possibly reflect a correlation of prostate cancer with an altered health condition, rather than a direct causal link between these SNPs and prostate cancer. The fact that these SNPs were not significant in Campa et al. (2011) may derive from the fact that the dataset is slightly different (here, a subset was studied), and because sample stratification was not accounted for in Campa et al., since the allele frequencies differences were small. Though, the absence of correction might have masked true associations (Marchini et al. 2004), which is corroborated by the fact that principal axes were able to capture a complex structure (tables 5 and 6), and by the indication that 10 might be a too small number of axes to describe the underlying structure (table 7, method nSimplices).

Link to current genetic case-control association studies research. Confounders in case-control association studies are most often detected and corrected for using Eigenstrat (Price et al. 2006) or mixed models (Price et al. 2010). This work supports the use of Eigenstrat as a reasonable method to correct for population structure in genome-wide association studies (when no more than 2% outliers are present, Fig. 6), in line with authors who recommend Eigenstrat over mixed models and other methods (Liu et al. 2011, Liu et al. 2013, Widmer et al. 2014, Yang et al. 2014). However, concerns about the detection of fine relatedness structure by PCA lead to the continued use of genomic control (Fuchsberger et al. 2016). This is not supported by the results obtained here, since fine structure was detected by PCA-based methods (Table 5b), similarly to what was observed by Wang et al. (2009) and Zhang et al. (2014), and in agreement with Mc Vean (2009), who showed theoretically that PCA had the ability to detect fine structure, provided that a large enough set of SNPs is used. Therefore, the use of genomic control to avoid cryptic relatedness seems questionable. Dropping genomic control should avoid masking of some true associations (Freedman et al. 2004, Price et al. 2010, Bouaziz et al. 2011, Wu et al. 2011).

The use of robust methods in this work was motivated by the need for stable and reproducible associations, as highlighted in Nilsson et al. (2013) and Li and Meyre (2013), which was

followed by only limited attempts to apply robust population structure correction (Liu et al. 2013 applied MCD an other methods to remove outliers, Conomos et al. 2015 developed a relatedness-robust method). This thesis is one of the very few, if not unique, systematic comparison of standard and robust methods to correct population structure. As a result, there is no direct comparison with published results possible, though the successful use of MCD in Liu et al. 2013 to remove outliers and therefore maintain stable estimates can be said to relate to the robustness of gMCD observed in this work (Fig. 6). This small number of studies on the influence of outliers and robust methods underscore the need for further investigations in this topic.

More broadly, though genome-wide case-control association studies have lead to tremendous new amounts of biological information in a very short period of time (Visscher et al. 2012), they can only partly explain disease heritability and functional mechanisms, as explained e.g. in Manolio et al. (2009). Missing heritability can indeed derive from many other factors, including epigenetic or environmental factors, gene-gene interactions, and gene-environment interactions (Manolio et al. 2009, Li and Meyre 2013). This motivated a move towards several new approaches, including multi-omics or integrative approaches, which aim at evaluating the contribution and interactions of SNPs and other factors (Kristensen et al. 2014, Günther et al. 2014, Iyengar et al. 2015, Miller et al. 2016, Zhang et al. 2016), or, meta-analyses of GWAS (advocated in Begum et al. 2012), or finally, extended GWAS (sometimes called 'post-GWAS' studies), where pathway analysis or other functional-oriented analyses are conducted (Hart and Kranzler 2015). Notably, all of these new strategies are still based on a case-control association framework, so that a reliable correction for underlying structure will continue to be relevant. Even more so, since the ability of PCA and related tools to summarize data and reduce dimension are a key feature of integrative analyses (Günther et al. 2014, Ritchie et al. 2015, Meng et al. 2016).

4.2.3 Structure identification in human microbiome (HMP) data

Structure identification in the HMP microbiome dataset. The methods qMDS, CSS and nSimplices performed best in the HMP dataset, while simpler methods MDSm and MDSe

performed reasonably well (table 11). All methods proved robust to 10% of outliers (Figure 9).

The better performance of distribution or quantile-based methods is in agreement with the fact that proportions might not reflect best the information contained in microbiome data, because they downweigh information, if an abundance outlier is large, or increase artificially all proportions, in case the data for one or more phylogenies is missing (Paulson et al. 2013). Moreover, the distribution of phylogenies is highly asymmetric, with density concentrated about zero and decreasing steeply for more elevated counts (Holmes et al. 2012). In order to take this into account, new methods wMDS and qMDS were developed and applied. The transposition of gMDS, applicable to genetic SNP data, to wMDS, for abundance data, did not bring the expected benefit. This is likely due to the fact that microbiome abundances do not follow a normal distribution, but -for example- an exponential or multinomial-derived distribution (Jeraldo et al. 2012, Holmes et al. 2012), whereas the underlying distribution of SNP data can be thought of as a normal distribution (because it corresponds to the sum of a large number of binomial variables, central limit theorem - Bortz 2005). The qMDS method selects an appropriate model distribution before performing normalization, and accordingly outperformed most other methods (table 11), though the CSS method of Paulson et al. (2013) was better. Finally, the procedure combining qMDS (using an exponential model) and nSimplices outperformed all the other methods, including CSS (table 11). The efficiency of this procedure is likely due to the ability of nSimplices to remove unuseful (because regarding only a single point, for example) information in each sample, therefore performing dimension reduction. This is in contrast with the way dimensionality reduction is done in PCA or MDS, since these methods select directions of lowest variance or distances, respectively. This makes them sensitive to single outlying values, and less sensitive if many points contain a small amount of private (e.g., regarding only one point) information, which might be the case in the HMP dataset.

A drawback has been the transposition of ℓ1-based methods L1PCA, LAR, RMDS and RSMDS to the HMP dataset. Indeed, these methods rely on an iterative procedure, which proved relatively slow, or did not converge to a relevant solution. In particular, L1PCA,

though it was shown to perform well for microbiome data in Brooks et al. 2013, proved very long. First attempts in R were accelerated 20-fold when the L1PCA was transposed in Python, but would have lasted more than 48 hours on a fast computer (3.4 GHz, 64bit, processors i7). However, implementations in R and Python, though advantageous because they come with rich packages of both classic and cutting-edge methods, cannot compare with more demanding but incomparably faster languages such as Fortran or C/C++ (http://benchmarksgame.alioth.debian.org/). The transposition of L1PCA in one of these languages would certainly make its use tractable and support its wide application. However, lack of convergence in RMDS and RSMDS seems to be due to an actual shortcoming of the algorithm.

In a nutshell, this work confirms the methods CSS and q^E-MDS - nSimplices as methods of choice to explore underlying structure in microbiome data. All top methods used were very robust, especially q^E-MDS and q^{NB}-MDS, combined or not with nSimplices . More research aiming at making multivariate $\ell 1$-methods more tractable would be necessary, In particular, faster implementations would strongly support their evaluation and use in clinical research.

Link to current microbiome research. The awareness about confounding structure in association or differential studies in the microbiome is only emerging (Blaser et al. 2013, Gilbert et al. 2016), while informative structure such as treatment tailoring based on microbiome profile has long ago been called for (Nasidze et al. 2009, Holmes et al. 2011). As Blaser et al. (2013) and Gilbert et al. 2016 pointed, no consensus on a preferable method exists yet, however some authors started evaluating or pointing flaws in current methods (Faust et al. 2012, Jiang et al. 2013, Schommer and Gallo 2013, McMurdie and Holmes 2014, Mandal et al. 2015).

The use of proportions lead to underperforming results (table 11, MDSe, MDSm) when compared with data-driven normalization methods (qMDS, CSS). This is in agreement with McMurdie and Holmes (2014), who claimed that simple proportions are an inefficient way to normalize microbiome data, because of unequal variances. McMurdie and Holmes additionally advocate the use of negative-binomial based methods, derived from the gene

expression field. Even if the negative binomial seemed the most appropriate model (table 10), the exponential distribution (q^E-MDS combined with nSimplices) led to better estimates, showing that the negative binomial might not be the best model distribution for phylogenies abundances. This observation is consistent with Faust et al. (2012), Schommer and Gallo (2013), Mandal et al. (2015), who argue that the microbiome composition derives from an ecosystem of competing phylogenies, which is not adequately reflected in binomial or multinomial-based distributions. Jiang et al. (2013) compared PCA with ISOMAP, a procedure comparable to MDSe (Hastie et al. 2001, p.573), on 45 marine samples and 33 human gut metagenomes. They found similar results for both methods, which complements the similar findings obtained here in MDSe and MDSm, so that PCA and MDS can be considered mostly equivalent in the analysis of microbiome data. This was also observed in genetic SNP data (Wang et al. 2009).

Finally, least absolute ($\ell 1$) methods have theoretical and practical benefits which have already led to the reinforced use of the median, univariate $\ell 1$ regression (a comprehensive review including historical aspects can be found in Tveite 1985), the LASSO (Tibshirani 1996), and many other $\ell 1$ methods (e.g. Huber 1964). However, high dimensional applications to microbiome data proved difficult here and remain rare in the literature (Brooks et al. 2013 is one example). Though, there is a high expected benefit from the application of $\ell 1$ in high-dimension datasets, as compared to $\ell 2$-based least squares (Aggarwal et al. 2001), which motivates making $\ell 1$ methods more practicable for high-dimensional datasets, possibly making them ultimately as widespread as the median or LASSO.

4.3 Advantages and limitations of method nSimplices

The method nSimplices allowed to explore the actual dimension of a distance matrix, to accommodate contextual outliers and to perform dimension reduction in an efficient manner (Tables 3, 5b and 11). Common methods for dimension reduction are based on an index, for instance this index is the variance in PCA, the distance-based cost function in MDS, the negentropy in ICA, or even any arbitrary index in projection pursuit (Friedman and Tukey

1974). An advantage of nSimplices is that it is free of such an index, since it relies on the volumes formed by the cloud of observations only. In other terms, this method directly identifies what lies outside of the relevant dimensional subspace (for example outside a 2-dimension space, see table 4), and removes this external component accordingly.

Method nSimplices did substantially outperform other methods in the HMP microbiome data and in synthetic example 2 with contextual outliers (tables 3 and 11). It further performed slightly better than other methods, in some cases, in the EPIC genetic SNPs dataset (table 5). In all these examples, nSimplices was among the most robust methods (figures 6 and 9). Dimension detection worked out well in synthetic example 2 (table 4), even in the most strained configurations, and hinted at a higher relevant dimension than 10 in EPIC and HMP datasets.

A limitation lies in the kurtosis index of additional heights, which is supposed to notably increase between dimension n, the last relevant dimension, and dimension n+1. This condition could be violated in datasets with large levels of noise (such as synthetic example 2 with added gaussian noise with a large standard deviation of the same order of magnitude as the points themselves). Another restriction is that distance outliers were not properly accommodated (table 3). However, a correction of triangle inequalities is in principle sufficient to address this issue, such as the triangle fixing approach proposed by Suvrit et al. (2005).

As a conclusion, this new approach proved efficient, or markedly advantageous, in a variety of settings. Method nSimplices should therefore be considered to explore any genomic, microbiome or other dataset which can be meaningfully presented as a distance matrix, as far as the level of noise in the data is not exceedingly large, and if one takes the precaution to correct triangle inequalities wherever necessary.

4.4 Conclusions

A systematic evaluation of several PCA and MDS-based methods applied to synthetic examples, to genetic SNP data from the EPIC study and to microbiome abundance data from HMP has been conducted. Each method yielded 10 axes, for which was assessed how well the underlying known structure was mirrored, and how robust these axes were after the inclusion of up to 10% outliers. The conclusions that can be drawn from this work are detailed in three sections: methods recommended in genetic data in §4.4.1, in microbiome data in §4.4.2, and recommendations valid in both categories, and which can be extended to further types of data, in §4.4.3. Briefly, it is possible to detect correctly and robustly an underlying confounder, for example using gMCD method (genetic SNPs), or using nSimplices (microbiome). Furthermore, different methods will report different types of structure, so that the use of two (or more) complementary methods is advisable. Finally, an original method for outlier detection and dimension reduction, nSimplices, has shown great promise and can be applied to genetic, microbiome or other categories of datasets.

4.4.1 Exploring and controlling for confounding structure in genetic SNP data

- To robustly control for population structure in genetic case-control association studies with up to 10% of outliers (real outliers, top axes), the methods SPH, IBS, gMCD and nSimplices should be best in terms of representativity and of robustness, as was the case here.

- Outliers removal based on a diverging ancestry might not be necessary, since 5 top axes are conserved for up to 10% of outliers for most methods used. This applies to the case of real individual outliers, sampled from unrelated 1000genomes individuals.

- To control for population structure in genetic case-control association studies with a low amount of outliers (up to 2%), Eigenstrat remains a reasonably efficient method.

- In the EPIC dataset, the SNPs rs520820, rs3783501, rs4955720, rs546950, rs388372 were characterized by non significant association p-values, in agreement with phase 2 of Campa et al. (2011). However, 20 candidate SNPs were significant (which however do not reach

significance after correcting for multiple testing): rs8071475, rs3799631, rs2589118, rs17239241, rs7778077, etc. (full list in table 9). These SNPs were detected in GWAS in other disorders and might correlate indirectly with the prostate cancer status.

4.4.2 Exploring and controlling for confounding structure in microbiome abundance data

- A recommendable method to account for overall and fine structure in microbiome data is nSimplices combined with qMDS approach (with an Exponential model of abundances). The qMDS method relies in the normalization of counts in quantiles of an exponential distribution. The nSimplices procedure comes as a pre-processing step to reduce dimensionality. Both approaches are tractable and fast (qMDS runs in minutes, nSimplices in 30 minutes, on the EPIC and HMP datasets used here).

- The CSS and qMDS methods performed second best here and should therefore still be meaningful, though suboptimal. These methods described both the overall and the detailed structure better than most methods.

- MDSe or MDSm were similarly representative and robust, but they underperformed markedly as compared to normalization-based methods. MDSe and MDSm would be advantageously replaced by nSimplices combined with q^E-MDS or CSS.

- Outliers in a microbiome dataset need not be removed. Indeed, nSimplices, qMDS and CSS proved robust to them. This is valid for up to 10% outliers, and for all 10 top axes (except CSS: axes 1 to 9).

4.4.3 Exploring and controlling for confounding structure in further types of datasets

The following considerations apply in principle to any dataset where a confounding structure could be present. This encompasses genetic SNP data and microbiome abundance data, but also further categories of datasets, for example multi-omics.

- The use of the nSimplices procedure developed and presented here should be taken in consideration for the exploration and dimension reduction of any high-dimensional dataset.

Indeed, nSimplices allowed to correctly detect and accommodate outliers (example 2), brought some advantage in genetic SNP data (EPIC cohort), and finally delivered a decisive advantage in the HMP dataset. This applies to any dataset which can be meaningfully represented as a distance matrix.

- The method shapes the solution: indeed, the approach chosen to detect confounding structure has itself and influence on the structure found, as observed in example 2 and in the EPIC and HMP datasets. The selected method could be compared to a probe with specificity for a particular target. It is recommended to apply several complementary methods to efficiently detect underlying structure and to avoid artefacts.

- More tractable procedures using least absolute residuals methods (ℓ1-norm approaches) would be needed, in particular high-performance implementations, since practical issues in this work undermined the expected potential of these methods.

5 SUMMARY

Case-control association studies in human genetics and microbiome pave the way to personalized medicine by enabling a personalized risk assessment, improved prognosis, or allowing an early diagnosis. However, confounding due to population structure, or other unobserved factors, can produce spurious findings or mask true associations, if not detected and corrected for. As a consequence, underlying structure improperly accounted for could explain lack of power or some unsuccessful replications observed in case-control association studies. Besides, points considered as outliers are commonly removed in such studies although they do not always correspond to technical errors. A wealth of methods exist to determine structure in genetic and microbiome association studies. However, there are few systematic comparisons between these methods in the frame of genetic or microbiome association studies, and even less attempts to apply robust methods, which produce stable estimates of confounding underlying structure, and which are able to incorporate information from outliers without degrading estimates quality.

Consequently, the aim of this thesis was to detect and control robustly for underlying confounding structure in genetic and microbiome data, by comparing systematically the most relevant standard and robust forms of principal components analysis (PCA) or multidimensional scaling (MDS) based methods, and by contributing new robust methods. Own contributions include robustification of existing methods, adaption to the genetic or to the microbiome framework, and a dimensionality exploration and reduction method, nSimplices. Analysed datasets include a first synthetic example with a low-variance 2-groups confounding structure, a second synthetic example with a simple linear underlying structure, genome-wide single nucleotide polymorphism (SNP) from 860 case and control individuals enrolled in the European Prospective Investigation into Cancer and nutrition (EPIC prostate), and finally, 2 255 microbiome samples from the human microbiome project (HMP). Synthetic or real outliers were added in the second example and in EPIC and HMP datasets. All meaningful existing and contributed methods were applied to the EPIC and HMP datasets, while a restricted set was applied to the synthetic, illustrative examples. The 10 principal components or top axes resulting from each method were kept for further analysis. Quality of a method was assessed by how well these axes summarized the underlying structure (using Akaike's information criterion -AIC- from the regression of the 10 axes on known underlying structure in the data), and by how robust the estimates stayed in the presence of outliers (adjusted R^2 from the regression of each outlier-disturbed axis on the original axis).

In synthetic example 1, only ICA was able to uncover the low-variance confounding structure, whereas PCA or MDS failed to do so, in agreement with the fact that these methods detect large rather than small variance or distance components. In synthetic example 2, non-metric MDS remained the most representative and robust method when distance outliers are included, while nSimplices combined with classical MDS was the only method to stay representative and robust if contextual outliers are present. In the EPIC dataset, Eigenstrat was the most representative method (AIC of 782.8) whereas sample ancestry was best captured by new method gMCD (unbiased genetic relatedness estimates used in a Minimum Covariance

Determinant procedure). Methods gMCD, spherical PCA, IBS (MDS on Identity-by-State estimates) and nSimplices were more robust than Eigenstrat, with a small to moderate loss in terms of representativity (AIC between 789.6 and 864.9). Association testing yielded p-values comparable with published values on candidate SNPs. Further SNPs rs8071475, rs3799631, rs2589118 with lowest p-value were identified, whose known role in other disorders could point to an indirect link with prostate cancer. In the HMP dataset, the new method nSimplices combined to data-driven normalization method qMDS mirrored best the underlying structure. The most robust method was qMDS (with nSimplices or alone), followed by CSS and MDS. Lastly, the original method nSimplices performed in all settings at least comparably (except ancestry in EPIC), and in some cases considerably better than other methods, while remaining tractable and fast in high-dimensional datasets.

The improved performance of gMCD and qMDS agrees with the fact that these methods use adapted measures (genetic relatedness, selected model distribution, respectively) and recognized robust approaches (minimum covariance determinant and quantiles). Conversely, wMDS is likely to have failed because variance is not an adequate parameter for microbiome data. More generally, different methods report the underlying structure differently and are advantageous in different settings, for example PCA or non-metric MDS were best in some settings but failed in other. Finally, the original method nSimplices proved useful or markedly better in a variety of settings, with the exception of highly noisy datasets, and provided that distance outliers are corrected.

Current genetic case-control association studies tend to integrate several types of data, for example clinical and SNP data, or several omics datasets. These approaches are promising but could be subject to increased inaccuracies or replication issues, by the mere combination of several sources of data. This motivates a reinforced use of robust methods, which are able to mirror accurately and steadily genetic information, such as gMCD, nSimplices or spherical PCA. Nevertheless, results on Eigenstrat show this stays a reasonable method. Results in microbiome confirmed that MDS based on proportions is a suboptimal method, and suggested the exponential distribution should be considered instead of multinomial-based distributions, certainly because the exponential better represents the inherent competitiveness between phylogenies in the microbiome. Moreover, illustrative and real world examples showed that methods could capture relevant, but different information, encouraging to apply several complementary methods when starting to explore a dataset. In particular, a low-variance confounder could stay undetected in some methods. Additionally, methods based on least absolute residuals revealed several shortcomings in spite of their utility in a univariate frame, but their expected benefit in a multivariate setting should motivate the development of more tractable implementations.

Finally, SPH, IBS, gMCD are recommended methods in a genetic SNP dataset, while Eigenstrat should perform best if no more than 2% outliers are present. To mirror structure in a microbiome dataset, nSimplices (combined with qMDS, or with CSS) can be expected to perform best, whereas MDS on proportions is likely to underperform. Method nSimplices proved beneficial or largely better in various situations and should therefore be considered to analyse datasets including, but not limited to, genetic SNP and microbiome abundances.

6 REFERENCES

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**:56-65
[Online:] URL: http://www.1000genomes.org/ [Stand: 1000 genomes phase 3, 27.05.2015, 00:00]

Aggarwal CC, Hinneburg A, Keim DA (2001) On the surprising behavior of distance metrics in high dimensional spaces, pp. 420-434
In: Proceedings of the 8th International Conference on Database Theory. Ed. Springer, Heidelberg

Amaratunga D, Cabrera J (2001) Analysis of data from viral DNA microchips. *J Am Stat Assoc* **96**:1161

Astle W, Balding DJ (2009) Population Structure and Cryptic Relatedness in Genetic Association Studies. *Stat Sci* **24**:451-471

Ash C (2016) "Normal" for the gut microbiota. *Science* **352**:546

Bakker M, Wichert JM (2014) Outlier Removal and the Relation with Reporting Errors and Quality of Psychological Research. *PLoS One* **9**:e103360

Barile M and Weisstein E (2016) Taxicab Metric. From MathWorld--A Wolfram Web Resource
[Online:] URL: http://mathworld.wolfram.com/TaxicabMetric.html [Stand: 14.07.2016, 08:18]

Beck T, Hastings RK, Gollapudi S, Free RC, Brookes AJ (2013) GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur J Hum Genet* **22**:949-952
[Online:] URL: http://www.gwascentral.org [Stand: 12.07.2016, 07:14]

Begum F, Ghosh D, Tseng GC, Feingold E (2012) Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Res 40:3777-3784

Bell AJ, Sejnowski TJ (1997) The "Independent Components" of Natural Scenes are Edge Filters. *Vision res* **37**:3327-3338.

Bellenguez C, Strange A, Freeman C, Wellcome Trust Case Control Consortium, Donnelly P, Spencer CCA (2010) A robust clustering algorithm for identifying problematic samples in genome-wide association studies. *Bioinformatics* **28**:134-135

Blaser M, Bork PB, Fraser C, Knight R, Wang J (2013) The microbiome explored: recent insights and future challenges. *Nat Rev Microbiol* **11**:213-217

Bonvicini C, Faraone SV, Scassellati C (2016) Attention-deficit hyperactivity disorder in adults: A systematic review and meta-analysis of genetic, pharmacogenetic and biochemical studies. *Mol Psychiatry* **21**:872-884

Bortz J (2005) Wahrscheinlichkeitstheorie und Wahrscheinlichkeitsverteilungen, p76 and p94.
In: Bortz J (2005) Statistik für Human- und Sozialwissenschaftler. 6th edition, Springer, Heidelberg

Bouaziz M, Ambroise C, Guedj M (2011) Accounting for population stratification in practice: a comparison of the main strategies dedicated to genome-wide association studies. *PLoS One* **6**:e28845

Bray JR, Curtis JT (1957) An ordination of upland forest communities of southern Wisconsin. *Ecol Monog* **27**:325-349

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F (2002) Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**:14250-14255

Brooks JP, Dulà JH, Boone EL (2013) A pure L1-norm principal component analysis. *Comput Stat Data Anal* **61**:83-98
[Online:] URL: https://cran.r-project.org/web/packages/pcaL1/ [Stand: 17.08.2015, 20:25]

Buus R, Sestak I, Kronenwett R, Denkert C, Dubsky P, Krappmann K, Scheer M, Petry C, Cuzick J, Dowsett M (2016) Comparison of EndoPredict and EPclin With Oncotype DX Recurrence Score for Prediction of Risk of Distant Recurrence After Endocrine Therapy. *J Natl Cancer Inst* **108**:djw149

Callahan BJ, Sankaran K, Fukuyama JA, McMurdie PJ, Holmes SP (2016) Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Research* **5**:1492

Campa D, Hüsing A, Stein A, Dostal L, Boeing H, Pischon T, Tjønneland A, Roswall N, Overvad K, Nautrup Østergaard J, Rodríguez L, Sala S, Sánchez MJ, Larrañaga N, Huerta JM, Barricarte A, Khaw KT, Wareham N, Travis RC, Allen NE, Lagiou P, Trichopoulou A, Trichopoulos D, Palli D, Sieri S, Tumino R, Sacerdote

C, van Kranen H, Bueno-de-Mesquita HB, Hallmans G, Johansson M, Romieu I, Jenab M, Cox DG, Siddiq A, Riboli E, Canzian F, Kaaks R (2011) Genetic variability of the mTOR pathway and prostate cancer risk in the European Prospective Investigation on Cancer (EPIC). *PLoS ONE* **6**:e16914

Cardon LR, Bell JI (2001) Association study designs for complex diseases. *Nat Rev Genet* **2**:91-99

Chaufan C, Joseph J (2013) The 'missing heritability' of common disorders: should health researchers care? *Int J Health Serv* **43**:281-303

Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, Szpiro AA, Chen W, Brehm JM, Celedón JC, Redline S, Papanicolaou GJ, Thornton TA, Laurie CC, Rice K, Lin X (2016) Control for population structure and relatedness for binary traits in genetic association studies via Logistic Mixed Models. *Am J Hum Gen* **98**:653-666

Cho I, Blaser MJ (2012) The human microbiome: at the interface of health and disease. *Nat Rev Genet* **13**:260-270

Choudhry S, Coyle NE, Tang H, Salari K, Lind D, Clark SL, Tsai HJ, Naqvi M, Phong A, Ung N, Matallana H, Avila PC, Casal J, Torres A, Nazario S, Castro R, Battle NC, Perez-Stable EJ, Kwok PY, Sheppard D, Shriver MD, Rodriguez-Cintron W, Risch N, Ziv E, Burchard EG; Genetics of Asthma in Latino Americans GALA Study (2006) Population stratification confounds genetic association studies among Latinos. *Hum Genet* **118**:652-664

Clarke G, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT (2011) Basic statistical analysis in genetic case-control studies. *Nat Protoc* **6**:121-133

Comon P (1994) Independent component analysis - a new concept? *Signal Processing* **36**:287-314

Conomos M, Miller MB, Thornton TA (2015) Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol* **39**:276-293

Cox, DR (1958) The regression analysis of binary sequences (with discussion). *J Roy Stat Soc B* **20**:215-242

Delorme A, Sejnowski T, Makeig S (2007) Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage* **34**:1443-1449

Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* **55**: 997-1004

Dey N, Soergel DA, Repo S, Brenner SE (2013) Association of gut microbiota with post-operative clinical course in Crohn's disease. *BMC Gastroenterol* **13**:131

Dhankani V, Gibbs DL, Knijnenburg T, Kramer R, Vockley J, Niederhuber J, Shmulevich I, Bernard B (2016) Using incomplete trios to boost confidence in family based association studies. *Front Genet* **7**:34

Doll R, Hill AB (1950) Smoking and carcinoma of the lung; preliminary report. *Br Med J* **2**:739-748

Dunn EC, Sofer T, Gallo LC, Gogarten SM, Kerr KF, Chen CY, Stein MB, Ursano RJ, Guo X, Jia Y, Qi Q, Rotter JI, Argos M, Cai J, Penedo FJ, Perreira K, Wassertheil-Smoller S, Smoller JW (2016) Genome-wide association study of generalized anxiety symptoms in the Hispanic Community Health Study/Study of Latinos. *Am J Med Genet B Neuropsychiatr Genet* **[Epub ahead of print]**

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**:446-450

Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, Kurilshikov A, Bonder MJ, Valles-Colomer M, Vandeputte D, Tito RY, Chaffron S, Rymenans L, Verspecht C, De Sutter L, Lima-Mendez G, D'hoe K, Jonckheere K, Homola D, Garcia R, Tigchelaar EF, Eeckhaudt L, Fu J, Henckaerts L, Zhernakova A, Wijmenga C, Raes J (2016) Population-level analysis of gut microbiome variation. *Science* **352**:560-564

Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C (2012) Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol* **8**:e1002606

Ferrarelli LK (2016) Homegrown factors in colon cancer. *Sci Signal* **9**:114

Folseraas T, Melum E, Rausch P, Juran BD, Ellinghaus E, Shiryaev A, Laerdahl JK, Ellinghaus D, Schramm C, Weismüller TJ, Gotthardt DN, Hov JR, Clausen OP, Weersma RK, Janse M, Boberg KM, Björnsson E, Marschall HU, Cleynen I, Rosenstiel P, Holm K, Teufel A, Rust C, Gieger C, Wichmann HE, Bergquist A, Ryu E, Ponsioen CY, Runz H, Sterneck M, Vermeire S, Beuers U, Wijmenga C, Schrumpf E, Manns MP, Lazaridis KN, Schreiber S, Baines JF, Franke A, Karlsen TH (2012) Extended analysis of a genome-wide association study in primary sclerosing cholangitis detects multiple novel risk loci. *J Hepatol* **57**:366-375

Forero PA, Giannakis GB (2012) Sparsity-exploiting robust multidimensional scaling. *IEEE Trans Signal Process* **60**:4118-4134

Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D, Hirschhorn JN, Altshuler D (2004) Assessing the impact of population stratification on genetic association studies. *Nat Genet* **4**:388-393

Friedman JH, Tukey JW (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Trans Comput C* **23**:881-890

Friend SH, Schadt EE (2014) Clues from the resilient. *Science* **344**:970-972

Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, Ma C, Fontanillas P, Moutsianas L, McCarthy DJ, Rivas MA, Perry JR, Sim X, Blackwell TW, Robertson NR, Rayner NW, Cingolani P, Locke AE, Tajes JF, Highland HM, Dupuis J, Chines PS, Lindgren CM, Hartl C, Jackson AU,

Chen H, Huyghe JR, van de Bunt M, Pearson RD, Kumar A, Müller-Nurasyid M, Grarup N, Stringham HM, Gamazon ER, Lee J, Chen Y, Scott RA, Below JE, Chen P, Huang J, Go MJ, Stitzel ML, Pasko D, Parker SC, Varga TV, Green T, Beer NL, Day-Williams AG, Ferreira T, Fingerlin T, Horikoshi M, Hu C, Huh I, Ikram MK, Kim BJ, Kim Y, Kim YJ, Kwon MS, Lee J, Lee S, Lin KH, Maxwell TJ, Nagai Y, Wang X, Welch RP, Yoon J, Zhang W, Barzilai N, Voight BF, Han BG, Jenkinson CP, Kuulasmaa T, Kuusisto J, Manning A, Ng MC, Palmer ND, Balkau B, Stančáková A, Abboud HE, Boeing H, Giedraitis V, Prabhakaran D, Gottesman O, Scott J, Carey J, Kwan P, Grant G, Smith JD, Neale BM, Purcell S, Butterworth AS, Howson JM, Lee HM, Lu Y, Kwak SH, Zhao W, Danesh J, Lam VK, Park KS, Saleheen D, So WY, Tam CH, Afzal U, Aguilar D, Arya R, Aung T, Chan E, Navarro C, Cheng CY, Palli D, Correa A, Curran JE, Rybin D, Farook VS, Fowler SP, Freedman BI, Griswold M, Hale DE, Hicks PJ, Khor CC, Kumar S, Lehne B, Thuillier D, Lim WY, Liu J, van der Schouw YT, Loh M, Musani SK, Puppala S, Scott WR, Yengo L, Tan ST, Taylor HA Jr, Thameem F, Wilson G, Wong TY, Njølstad PR, Levy JC, Mangino M, Bonnycastle LL, Schwarzmayr T, Fadista J, Surdulescu GL, Herder C, Groves CJ, Wieland T, Bork-Jensen J, Brandslund I, Christensen C, Koistinen HA, Doney AS, Kinnunen L, Esko T, Farmer AJ, Hakaste L, Hodgkiss D, Kravic J, Lyssenko V, Hollensted M, Jørgensen ME, Jørgensen T, Ladenvall C, Justesen JM, Käräjämäki A, Kriebel J, Rathmann W, Lannfelt L, Lauritzen T, Narisu N, Linneberg A, Melander O, Milani L, Neville M, Orho-Melander M, Qi L, Qi Q, Roden M, Rolandsson O, Swift A, Rosengren AH, Stirrups K, Wood AR, Mihailov E, Blancher C, Carneiro MO, Maguire J, Poplin R, Shakir K, Fennell T, DePristo M, Hrabé de Angelis M, Deloukas P, Gjesing AP, Jun G, Nilsson P, Murphy J, Onofrio R, Thorand B, Hansen T, Meisinger C, Hu FB, Isomaa B, Karpe F, Liang L, Peters A, Huth C, O'Rahilly SP, Palmer CN, Pedersen O, Rauramaa R, Tuomilehto J, Salomaa V, Watanabe RM, Syvänen AC, Bergman RN, Bharadwaj D, Bottinger EP, Cho YS, Chandak GR, Chan JC, Chia KS, Daly MJ, Ebrahim SB, Langenberg C, Elliott P, Jablonski KA, Lehman DM, Jia W, Ma RC, Pollin TI, Sandhu M, Tandon N, Froguel P, Barroso I, Teo YY, Zeggini E, Loos RJ, Small KS, Ried JS, DeFronzo RA, Grallert H, Glaser B, Metspalu A, Wareham NJ, Walker M, Banks E, Gieger C, Ingelsson E, Im HK, Illig T, Franks PW, Buck G, Trakalo J, Buck D, Prokopenko I, Mägi R, Lind L, Farjoun Y, Owen KR, Gloyn AL, Strauch K, Tuomi T, Kooner JS, Lee JY, Park T, Donnelly P, Morris AD, Hattersley AT, Bowden DW, Collins FS, Atzmon G, Chambers JC, Spector TD, Laakso M, Strom TM, Bell GI, Blangero J, Duggirala R, Tai ES, McVean G, Hanis CL, Wilson JG, Seielstad M, Frayling TM, Meigs JB, Cox NJ, Sladek R, Lander ES, Gabriel S, Burtt NP, Mohlke KL, Meitinger T, Groop L, Abecasis G, Florez JC, Scott LJ, Morris AP, Kang HM, Boehnke M, Altshuler D, McCarthy MI (2016) The genetic architecture of type 2 diabetes. *Nature* **536**:41-47

Gallitano AL, Tillman R, Dinu V, Geller B (2012) Family-based association study of early growth response gene 3 with child bipolar I disorder. *J Affect Disord* **138**:387-396

Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, Jansson JK, Dorrestein PC, Knight R (2016) Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature* **535**:94-103

Goodrich JK, Davenport ER, Waters JL, Clark AG, Ley RE (2016) Cross-species comparisons of host genetic associations with the microbiome. *Science* **352**:532-535

Groenen PJF, Heiser WJ (1996) The tunneling method for global optimization in multidimensional scaling. *Psychometrika* **61**:529-550

Günther OP, Shin H, Ng RT, McMaster WR, McManus BM, Keown PA, Tebbutt SJ, Lê Cao KA (2014) Novel Multivariate Methods for Integration of Genomics and Proteomics Data: Applications in a Kidney Transplant Rejection Study. *OMICS* **18**:682-695

Harley ITW, Karp CL (2012) Obesity and the gut microbiome: Striving for causality. *Mol metab* **1**:21-31

Hart AB, Kranzler HR (2015) Alcohol dependence genetics: Lessons learned from genome-wide association studies (GWAS) and post-GWAS analyses. Alcoholism, clinical and experimental research. 39:1312-1327

Hastie T, Tibshirani R, Friedman J (2001) Multidimensional scaling, pp. 570-572.
In: Hastie T, Tibshirani R, Friedman J: The Elements of statistical learning. 2nd ed. Springer, New York

Heiser WJ (1988) Multidimensional scaling with least absolute residuals, pp. 455-462.
In: Bock HH: Classification and related methods of data analysis. Amsterdam.

Hérault J, Ans B (1984) Réseau de neurones à synapses modifiables: décodage de messages sensoriels composites par apprentissage non supervisé et permanent. *C R Acad Sci III* **299**:525-528

Héritier S, Cantoni E, Copt S, Victoria-Feser MP (2009) Key measures and results, pp. 17-20.
In: Héritier S, Cantoni E, Copt S, Victoria-Feser MP: Robust methods in biostatistics, 1st ed. Wiley, New York

Hirschhorn JN, Daly MJ (2004) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6**:95-108

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002) A comprehensive review of genetic association studies. *Genet Med* **4**:45-61

Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM (2003) Control of confounding of genetic associations in stratified populations. *Am J Hum Genet* **72**:1492-1504

Hollman AL, Tchounwou PB, Huang HC (2016) The Association between Gene-Environment Interactions and Diseases Involving the Human GST Superfamily with SNP Variants. *Int J Environ Res Public Health* **13**:E379

Holmes E, Li JV, Athanasiou T, Ashrafian H, Nicholson JK (2011) Understanding the role of gut microbiome-host metabolic signal disruption in health and disease. *Trends Microbiol* **19**:349-359

Holmes I, Harris K, Quince C (2012) Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PLoS One* **7**:e30126

Huber PJ (1964) Robust Estimation of a Location Parameter. **Ann Stat** *53*:73-101

Hubert L, Arabie P, Hesson-McInnis M (1992) Multidimensional scaling in the city-block metric: a combinatorial approach. *J Classif* **9**:211-236

Hyvärinen A, Oja E (2000) Independent component analysis: Algorithms and applications. *Neural Networks* **13**:411-430

Iida N, Dzutsev A, Stewart CA, Smith L, Bouladoux N, Weingarten RA, Molina DA, Salcedo R, Back T, Cramer S, Dai RM, Kiu H, Cardone M, Naik S, Patri AK, Wang E, Marincola FM, Frank KM, Belkaid Y, Trinchieri G, Goldszmid RS (2013) Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. *Science* **342**:967-970

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. *Nat Genet* **29**:306-309

Iyengar R, Altman RB, Troyanskya O, FitzGerald GA (2015) MEDICINE. Personalization in practice. *Science* **350**:282-283

Jeraldo P, Sipos M, Chia N, Brulc JM, Dhillon AS, Konkel ME, Larson CL, Nelson KE, Qu A, Schook LB, Yang F, White BA, Goldenfeld N (2012) Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proc Natl Acad Sci U S A* **109**:9692-9698

Jiang X, Hu X, Xu W, He T, Park EK (2013) Comparison of dimensional reduction methods for detecting and visualizing novel patterns in human and marine microbiome. *IEEE Trans Nanobioscience* **12**:199-205

Jones E, Oliphant E, Peterson P, and SciPy developers (2001-) SciPy: Open Source Scientific Tools for Python. [Online:] https://www.scipy.org [Stand: version 0.13.3, 05.11.2013, 09:08]

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**:348-354

Kingsmore SF, Lindquist IE, Mudge J, Gessler DD, Beavis WD (2008) Genome-wide association studies: progress and potential for drug discovery and development. *Nat Rev Drug Discov* **7**:221-230

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **308**:385-389

Koenker R, Portnoy S (1997) The Gaussian Hare and the Laplacian Tortoise: Computability of squared-error vs. absolute-error estimators, with discussion. *Stat Sci* **12**:279-300
[Online:] URL: https://cran.r-project.org/src/contrib/quantreg_5.26.tar.gz [Stand: 03.09.2015, 15:04]

Komi DE, Kazemi T, Bussink AP (2016) New Insights Into the Relationship Between Chitinase-3-Like-1 and Asthma. *Curr Allergy Asthma Rep* **16**:57

Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, Ojesina AI, Jung J, Bass AJ, Tabernero J, Baselga J, Liu C, Shivdasani RA, Ogino S, Birren BW, Huttenhower C, Garrett WS, Meyerson M (2012) Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res* **22**:292-298

Kristensen VN, Lingjærde OC, Russnes HG, Vollan HK, Frigessi A, Børresen-Dale AL(2014) Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer 14:299-313

Kruskal J (1964a) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**:1-27

Kruskal J (1964b) Nonmetric multidimensional scaling: a numerical method. Psychometrika, 29:115-129

Laivuori H (2013) Pitfalls in setting up genetic studies on preeclampsia. *Pregnancy Hypertens* **3**:60

Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A* **82**: 6955-6959

Lane-Claypon JE (1926) A further report on cancer of the breast : with special reference to its associated antecedent conditions, pp. 19-53.
In: Reports on public health and medical subjects 32. Ed: His Majesty's Stationery Office, London

Larson N, Hutchinson R, Boerwinkle E (2000) Lack of association of 3 functional gene variants with hypertension in African Americans. *Hypertension* **35**:1297-1300

Li A, Meyre D (2013) Challenges in reproducibility of genetic association studies: lessons learned from the obesity field. *Int J Obes* **37**:559-567

Li C, Wang B, Lu D, Jin JY, Gao Y, Matsunaga K, Igawa Y, Nijem I, Lu M, Strasak A, Chernyukhin N, Girish S (2016) Ethnic sensitivity assessment of the antibody-drug conjugate trastuzumab emtansine (T-DM1) in patients with HER2-positive locally advanced or metastatic breast cancer. *Cancer Chemother Pharmacol* **[Epub ahead of print]**

Liesen J, Mehrmann V (2015) The Singular Value Decomposition, pp. 295-302.
In: Liesen J, Mehrmann V: Linear Algebra, 1st ed. Springer, Cham.
Liu L, Zhang D, Liu H, Arendt C (2013) Robust methods for population stratification in genome wide association studies. *BMC Bioinformatics* **14**(132):1-12
Liu N, Zhao H, Patki A, Limdi NA, Allison DB (2011) Controlling Population Structure in Human Genetic Association Studies with Samples of Unrelated Individuals. *Stat Interface* **4**:317-326
Liu R (1990) On a notion of data depth based on random simplices. *Ann Stat* **18**:405-414
Lloyd-Price J, Abu-Ali G, Huttenhower C (2016) The healthy human microbiome. *Genome Med* **8**:51
Locantore N, Marron J , Simpson D, Tripoli N, Zhang J, Cohen K (1999) Robust principal component analysis for functional data. *Test* **8**:1-73
Malinen E, Krogius-Kurikka L, Lyra A, Nikkilä J, Jääskeläinen A, Rinttilä T, Vilpponen-Salmela T, von Wright AJ, Palva A (2010) Association of symptoms with gastrointestinal microbiota in irritable bowel syndrome. *World J Gastroenterol* **16**:4532-4540
Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol in Health Dis* **26**:27663
Marchini J, Cardon LR, Phillips MS, Donnely P (2004) The effects of human population structure on large genetic association studies. *Nat Genet* **36**:512-517
Marchini JL, Heaton C, Ripley BD (2013) fastICA: FastICA Algorithms to perform ICA and Projection Pursuit. R package
[Online:] URL: http://CRAN.R-project.org/package=fastICA [Stand: version 1.2-0, 21.05.2013, 10:15]
Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. *Nature* 461:747-753
McMurdie PJ, Holmes S (2014) Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol* **10**:e1003531
McVean G (2009) A Genealogical Interpretation of Principal Components Analysis. *PLoS Genet* **5**:e1000686
Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform **17***:628:641
Mersha TB, Ding L, He H, Alexander ES, Zhang X, Kurowski BG, Pilipenko V, Kottyan L, Martin LJ, Fardo DW (2015) Impact of population stratification on family-based association in an admixed population. *Int J Genomics* **2015**:501617
Miller CL, Pjanic M, Wang T, Nguyen T, Cohain A, Lee JD, Perisic L, Hedin U, Kundu RK, Majmudar D, Kim JB, Wang O, Betsholtz C, Ruusalepp A, Franzén O, Assimes TL, Montgomery SB, Schadt EE, Björkegren JL, Quertermous T (2016) Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. *Nat Commun* **7**:12092
Montassier E, Al-Ghalith GA, Ward T, Corvec S, Gastinne T, Potel G, Moreau P, de la Cochetiere MF, Batard E, Knights D (2016) Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. *Genome Med* **8**:49
Moors JJA (1986) The meaning of kurtosis: Darlington reexamined. *Am Stat* **40**:283-284
Moors JJA (1988) A quantile alternative for kurtosis. *Statistician* **37**:25-32
Müller B, Wilcke A, Boulesteix AL, Brauer J, Passarge E, Boltze J, Kirsten H (2016) Improved prediction of complex diseases by common genetic markers: state of the art and further perspectives. *Hum Genet* **135**:259-272
Nasidze I, Li J, Quinque D, Tang K, Stoneking M (2009) Global diversity in the human salivary microbiome. *Genome Res* **19**:636-643
Nilsson D, Andiappan AK, Halldn C, Tim CF, Säll T, Wang DY, Cardell L-O (2013) Poor Reproducibility of Allergic Rhinitis SNP Associations. *PLoS One* **8**:e53975
Okbay A, Baselmans BM, De Neve JE, Turley P, Nivard MG, Fontana MA, Meddens SF, Linnér RK, Rietveld CA, Derringer J, Gratten J, Lee JJ, Liu JZ, de Vlaming R, Ahluwalia TS, Buchwald J, Cavadino A, Frazier-Wood AC, Furlotte NA, Garfield V, Geisel MH, Gonzalez JR, Haitjema S, Karlsson R, van der Laan SW, Ladwig KH, Lahti J, van der Lee SJ, Lind PA, Liu T, Matteson L, Mihailov E, Miller MB, Minica CC, Nolte IM, Mook-Kanamori D, van der Most PJ, Oldmeadow C, Qian Y, Raitakari O, Rawal R, Realo A, Rueedi R, Schmidt B, Smith AV, Stergiakouli E, Tanaka T, Taylor K, Wedenoja J, Wellmann J, Westra HJ, Willems SM, Zhao W; LifeLines Cohort Study, Amin N, Bakshi A, Boyle PA, Cherney S, Cox SR, Davies G, Davis OS, Ding J, Direk N, Eibich P, Emeny RT, Fatemifar G, Faul JD, Ferrucci L, Forstner A, Gieger C, Gupta R, Harris TB, Harris JM, Holliday EG, Hottenga JJ, De Jager PL, Kaakinen MA, Kajantie E, Karhunen V, Kolcic I, Kumari M, Launer LJ, Franke L, Li-Gao R, Koini M, Loukola A, Marques-Vidal P, Montgomery GW, Mosing MA, Paternoster L, Pattie A, Petrovic KE, Pulkki-Råback L, Quaye L, Räikkönen K, Rudan I, Scott RJ, Smith JA, Sutin AR, Trzaskowski M, Vinkhuyzen AE, Yu L, Zabaneh D, Attia JR, Bennett DA, Berger K, Bertram L, Boomsma DI, Snieder H, Chang SC, Cucca F, Deary IJ, van Duijn CM, Eriksson JG, Bültmann U, de Geus EJ, Groenen PJ, Gudnason V, Hansen T, Hartman CA, Haworth CM, Hayward C,

Heath AC, Hinds DA, Hyppönen E, Iacono WG, Järvelin MR, Jöckel KH, Kaprio J, Kardia SL, Keltikangas-Järvinen L, Kraft P, Kubzansky LD, Lehtimäki T, Magnusson PK, Martin NG, McGue M, Metspalu A, Mills M, de Mutsert R, Oldehinkel AJ, Pasterkamp G, Pedersen NL, Plomin R, Polasek O, Power C, Rich SS, Rosendaal FR, den Ruijter HM, Schlessinger D, Schmidt H, Svento R, Schmidt R, Alizadeh BZ, Sørensen TI, Spector TD, Steptoe A, Terracciano A, Thurik AR, Timpson NJ, Tiemeier H, Uitterlinden AG, Vollenweider P, Wagner GG, Weir DR, Yang J, Conley DC, Smith GD, Hofman A, Johannesson M, Laibson DI, Medland SE, Meyer MN, Pickrell JK, Esko T, Krueger RF, Beauchamp JP, Koellinger PD, Benjamin DJ, Bartels M, Cesarini D (2016) Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet* **48**:624-633

Oliehoek PA, Windig JJ, van Arendonk JAM, Bijma P (2006) Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics* **173**:483-496

Ott J, Kamatani Y, Lathrop M (2011) Family-based designs for genome-wide association studies. *Nat Rev Genet* **12**:465-474

Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, Tanaka T (2002) Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature Genet* **32**:650-654

Ozkavruk Eliyatkin N, Aktas S, Ozgur H, Ercetin P, Kupelioglu A (2016) The role of p95HER2 in trastuzumab resistance in breast cancer. *J BUON* **21**:382-389

Paulson JN, Stine OC, Bravo HC, Pop M (2013) Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* **10**:1200-1202

Paulson JN, Talukder H, Pop M, Bravo HC (2015) metagenomeSeq: Statistical analysis for sparse high-throughput sequencing. Bioconductor package: 1.14.2.
[Online:] URL: http://cbcb.umd.edu/software/metagenomeSeq [Stand: 25.03.2014, v1.4.2]

Pearson K (1901) On lines and planes of closest fit to systems of points in space. *Philos mag* **2**:559-572

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, CournapeauD, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**:2825-2830

Peterson P (2009) F2PY: a tool for connecting Fortran and Python programs. *Int J Comput Sci Eng* **4**:296-305

Price A, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**:904-909

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**:459-463

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* **65**:220-228

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**:945-959

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* **81**:559-575
[Online:] URL: http://pngu.mgh.harvard.edu/purcell/plink/ [Stand: version v1.07]

R Core Team, Chambers J, Gentleman R, Ihaka R (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
[Online:] URL: https://www.r-project.org/ [Stand: version 3.2.1 (18.06.2015)]

Ray B (2014) Outliers Dominate Signaling at Cell Membrane. *Sci Signal* **7**:ec189

Reich DE, Goldstein DB (2001) Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* **20**:4-16

Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, Charrondière UR, Hémon B, Casagrande C, Vignat J, Overvad K, Tjønneland A, Clavel-Chapelon F, Thiébaut A, Wahrendorf J, Boeing H, Trichopoulos D, Trichopoulou A, Vineis P, Palli D, Bueno-De-Mesquita HB, Peeters PH, Lund E, Engeset D, González CA, Barricarte A, Berglund G, Hallmans G, Day NE, Key TJ, Kaaks R, Saracci R (2002) European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* **5**:1113-1124

Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D (2015) Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* **16**:85-97

Robinson MR, Hemani G, Medina-Gomez C, Mezzavilla M, Esko T, Shakhbazov K, Powell JE, Vinkhuyzen A, Berndt SI, Gustafsson S, Justice AE, Kahali B, Locke AE, Pers TH, Vedantam S, Wood AR, van Rheenen W, Andreassen OA, Gasparini P, Metspalu A, Berg LH, Veldink JH, Rivadeneira F, Werge TM, Abecasis GR, Boomsma DI, Chasman DI, de Geus EJ, Frayling TM, Hirschhorn JN, Hottenga JJ, Ingelsson E, Loos RJ, Magnusson PK, Martin NG, Montgomery GW, North KE, Pedersen NL, Spector TD, Speliotes EK, Goddard ME, Yang J, Visscher PM (2015) Population genetic differentiation of height and body mass index across Europe. *Nat Genet* **47**:1357-1362

Roman T, Nayyeri A, Fasy BT, Schwartz R (2015) A simplicial complex-based approach to unmixing tumor progression data. *BMC Bioinformatics* **16**:254

Rousseeuw PJ (1984) Least median of squares regression. *J Am Stat Assoc* **79**:871-880

Rousseeuw PJ, van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**:212-223

Rousseeuw PJ, van Zomeren BC (1990) Unmasking Multivariate Outliers and Leverage Points. *J Am Stat Assoc* **85**:633-639

Saad MN, Mabrouk MS, Eldeib AM, Shaker OG (2016) Identification of rheumatoid arthritis biomarkers based on single nucleotide polymorphisms and haplotype blocks: A systematic review and meta-analysis. *J Adv Res* **7**:1-16

Sadee W, Hartmann K, Seweryn M, Pietrzak M, Handelman SK, Rempala GA (2014) Missing heritability of common diseases and treatments outside the protein-coding exome. *Hum Genet* **133**:1199-1215

Saito YA, Talley NJ, de Andrade M, Petersen GM (2006) Case-control genetic association studies in gastrointestinal disease: review and recommendations. *Am J Gastroenterol* **101**:1379-1389

Salanti G, Sanderson S, Higgins JPT (2005) Obstacles and opportunities in meta-analysis of genetic association studies. *Genet Med* **7**:13-20

Sampson DL, Parker TJ, Upton Z, Hurst CP (2011) A Comparison of Methods for Classifying Clinical Samples Based on Proteomics Data: A Case Study for Statistical and Machine Learning Approaches. *PLoS One* **6**:e24973

Satten GA, Flanders WD, Yang Q (2001) Accounting for Unmeasured Population Substructure in Case-Control Studies of Genetic Association Using a Novel Latent-Class Model. *Am J Hum Genet* 68:466-477

Schommer NN, Gallo RL (2013) Structure and function of the human skin microbiome. *Trends in Microbiology* **21**:660-668

Scott A, Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, Varambally S, Cao X, Tchinda J, Kuefer R, Lee C, Montie JE, Shah RB, Pienta KJ, Rubin MA, Chinnaiyan AM (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**:644-648

Sillanpää MJ (2011). Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses. *Heredity* **106**:511-519

Sinha R, Abnet C2, White 3, Knight 4, Huttenhower C (2015) The microbiome quality control project: baseline study design and future directions. *Genome Biol* **16**:276

Shen Y, Liu Z, Ott J (2010) Systematic removal of outliers to reduce heterogeneity in case-control association studies. *Hum Hered* **70**:227-231

Shibata K, Hozawa A, Tamiya G, Ueki M, Nakamura T, Narimatsu H, Kubota I, Ueno Y, Kato T, Yamashita H, Fukao A, Kayama T; Yamagata University Genomic Cohort Consortiuma (2013) The confounding effect of cryptic relatedness for environmental risks of systolic blood pressure on cohort studies. *Mol Genet Genomic Med* **1**:45-53

Shinkareva SV, Wang J, Wedell DH (2013) Examining similarity structure: multidimensional scaling and related approaches in neuroimaging. *Comput Math Methods Med* **2013**:796183

Slattery ML, Lundgreen A, Mullany LE, Penney RB, Wolff RK (2014) Influence of CHIEF pathway genes on gene expression: a pathway approach to functionality. *Int J Mol Epidemiol Genet* **5**:100-111

Sommerville DMY (1929) Chapter VIII Mensuration: Content, 123-125
In: Sommerville DMY: An introduction to the geometry of n dimensions. Methuen, London

Song X, Wu M, Jermaine C, Ranka S (2007) Conditional Anomaly Detection. *IEEE Trans Knowl Data Eng* **19**:631-645

Spence I, Lewandowsky S (1989) Robust Multidimensional Scaling. *Psychometrika* **54**:501-513

Suvrit S, Tropp J, Dhillon S (2005) Triangle Fixing Algorithms for the Metric Nearness Problem. *Adv Neural Inf Process Syst* **17**:361-368

The Human Microbiome Project Consortium (2012) Structure, Function and Diversity of the Healthy Human Microbiome. *Nature* **486**:207-214

Thomas DC, Witte JS (2002) Point: population stratification: a problem for case-control
studies of candidate-gene associations? *Cancer Epidemiol. Biomarkers Prev* **11**:505-512

Thomas F and Ros L (2005) Revisiting trilateration for robot localization. *IEEE Trans Robot* **21**:93-101

Thornton T, Mc Peek MS (2010) ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet* **86**:172-184

Tibshirani R (1996) Regression Shrinkage and Selection via the lasso. *J Roy Stat Soc B Met* **58**:267-288

Torgerson WS (1952) Multidimensional scaling: I. Theory and method. *Psychometrika* **17**:401-419

Tropsha A1, Singh RK, Vaisman II, Zheng W (1996) Statistical geometry analysis of proteins: implications for inverted structure prediction. *Pac Symp Biocomput* **1996**:614-623.

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**:1027-1031

Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI (2007) The Human Microbiome Project. *Nature* **449**:804-810
[Online:] http://hmpdacc.org/HMMCP/ [Stand: 30.11.2015]

Turner S, Armstrong LL, Bradford Y, Carlson CS, Crawford DC, Crenshaw AT, de Andrade M, Doheny KF, Haines JL, Hayes G, Jarvik G, Jiang L, Kullo IJ, Li R, Ling H, Manolio TA, Matsumoto M, McCarty CA, McDavid AN, Mirel DB, Paschall JE, Pugh EW, Rasmussen LV, Wilke RA, Zuvich RL, Ritchie MD (2011) Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet* **1**:Unit1.19

Tveite MD (1985) Statistical aspects of L1 regression. Retrospective Theses and Dissertations. Paper 8751. [Online:] URL: http://lib.dr.iastate.edu/rtd/8751 [Stand: 02.08.2016, 21:35]

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Insights into community structure and metabolism by reconstruction of microbial genomes from the environment *Nature* **428**:37-43

Van den Broeck J, Cunningham SA, Eeckels R, Herbst K. (2005) Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med* **2**:e267

Van Rossum G (1995) Python tutorial. Technical Report CS-R9526 Amsterdam: Centrum voor Wiskunde en Informatica (CWI).
[Online:] https://www.python.org/ [Stand: version 2.7.6, 22.06.2015]

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90:7-24

Vogtmann E, Goedert JJ (2016) Epidemiologic studies of the human microbiome and cancer. *B J Cancer* **114**:237-242

Voight B, Pritchard JK (2005) Confounding from Cryptic Relatedness in Case-Control Association Studies. *PLoS Genet* **1**:e32

Wacholder S, Rothman N, Caporas N (2002) Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol Biomarkers Prev* **11**:513-520

Wächter A, Biegler LT (2006) On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Math Program* **106**:25-57

Wallace BD, Wang H, Lane KT, Scott JE, Orans J, Koo JS, Venkatesh M, Jobin C, Yeh LA, Mani S, Redinbo MR (2010) Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. *Science* **330**: 831-835

Wang D, Sun Y, Stang P, Berlin JA, Wilcox MA, Li Q (2009) Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling. *BMC Proc* **3**:S109

Wang IX, Ramrattan G, Cheung VG (2015) Genetic variation in insulin-induced kinas signaling. *Mol Syst Biol* **11**:820

Wang J, Jia H (2016) Metagenome-wide association studies: fine-mining the microbiome. *Microbiome* **14**:508-522

Widmer C, Lippert C, Weissbrod O, Fusi N, Kadie C, Davidson R, Listgarten J, Heckerman D (2014) Further improvements to linear mixed models for genome-wide association studies. *Sci Rep* **4**:6874

Wu C, DeWan A, Hoh J, Wang Z (2011) A comparison of association methods correcting for population stratification in case-control studies. *Ann Hum Genet* **75**:418-427

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of heritability for human height. *Nat Genet* **42**:565-569

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014) Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**:100-106

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). *Science* **296**:79-92

Zhang M, Mu H, Lv H, Duan L, Shang Z, Li J, Jiang Y, Zhang R (2016) Integrative analysis of genome-wide association studies and gene expression analysis identifies pathways associated with rheumatoid arthritis. *Oncotarget* **7**:8580-8589

Zhang S, Pakstis AJ, Kidd KK, Zhao H (2001) Comparisons of Two Methods for Haplotype Reconstruction and Haplotype Frequency Estimation from Population Data. *Am J Hum Genet* **69**:906-912

Zhang Y, Pan W (2014) Adjusting for population stratification and relatedness with sequencing data *BMC Proc* **8**:S42

Zmora N, Zeevi D, Korem T, Segal E, Elinav E (2016) Taking it personally: personalized utilization of the human microbiome in health and disease. *Cell Host Microbe* **19**:12-20

Zou W, Wolchok JD, Chen L (2016) PD-L1 (B7-H1) and PD-1 pathway blockade for cancer therapy: Mechanisms, response biomarkers, and combinations. *Sci Transl Med* **8**:328

7 LIST OF PUBLICATIONS

Legrand C, Hartmann GH, Karger CP (2012) Experimental determination of the effective point of measurement for cylindrical ionization chambers in 60Co gamma radiation. Phys Med Biol 57:3463-3475

Legrand C, Hartmann GH, Karger CP (2012) Experimental determination of the effective point of measurement and the displacement correction factor for cylindrical ionization chambers in a 6 MV photon beam. Phys Med Biol 57:6869-6880

Kesselmeier M, Legrand C, Peil B, Kabisch M, Fischer C, Hamann U, Lorenzo Bermejo J (2014) Practical Investigation of the Performance of Robust Logistic Regression to Predict the Genetic Risk of Hypertension. BMC Proc 8:S65

Pardini B, Lorenzo Bermejo J, Naccarati A, Di Gaetano C, Rosa F, Legrand C, Novotny J, Vodicka P, Kumar R (2014) Inherited variability in a master regulator polymorphism (rs4846126) associates with survival in 5-FU treated colorectal cancer patients. Mutat Res 766:7-13

González Silos R, Karadag Ö, Peil B, Fischer C, Kabisch M, Legrand C, Lorenzo Bermejo J (2015) Using next generation DNA sequence data for genetic association tests based on allele counts with and without consideration of zero-inflation. BMC Proc 9:S56

Legrand C, Tuorto F, Hartmann M, Liebers R, Lyko F. Statistically robust methylation calling for whole-transcriptome bisulfite sequencing reveals distinct methylation patterns for mouse RNAs. (In preparation.)

8 APPENDIX

Table S1. Adjusted R^2 for the robustness to synthetic outliers, EPIC dataset.

Table S2. Adjusted R^2 for the robustness to real outliers, EPIC dataset.

Table S3. Adjusted R^2 for the robustness to outliers, HMP dataset.

Table S1. Adjusted R^2 for robustness to synthetic outliers, EPIC dataset.

%outliers	axis 1	axis 2	axis 3	axis 4	axis 5	axis 6	axis 7	axis 8	axis 9	axis 10
					EIG					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-0.01	0.00
0.5	1.00	1.00	1.00	1.00	1.00	0.93	0.06	-0.01	0.00	0.00
0.75	1.00	1.00	1.00	1.00	0.03	0.00	0.00	0.01	0.00	0.00
1	1.00	1.00	0.08	0.05	0.03	0.01	0.00	0.02	0.00	0.01
2	0.94	0.17	0.06	0.06	0.05	0.01	-0.01	0.00	0.03	0.01
5	0.89	0.14	0.09	0.03	0.07	0.01	0.00	0.00	0.03	0.03
7	0.86	0.11	0.09	0.04	0.07	0.01	0.00	0.01	0.03	0.02
10	0.85	0.23	0.07	0.04	0.05	0.01	0.00	0.01	0.03	0.02
					SPH					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	-0.01
0.5	1.00	1.00	1.00	1.00	1.00	1.00	0.02	-0.01	0.00	-0.01
0.75	1.00	1.00	1.00	1.00	0.20	0.33	0.01	0.01	0.00	0.00
1	1.00	1.00	0.34	0.18	0.17	0.26	0.02	0.01	0.01	0.01
2	1.00	0.92	0.27	0.16	0.14	0.24	0.01	0.00	0.00	0.01
5	1.00	0.86	0.26	0.17	0.16	0.24	0.01	-0.01	0.00	0.00
7	1.00	0.85	0.41	0.20	0.17	0.21	0.02	-0.01	0.00	0.00
10	1.00	0.82	0.41	0.23	0.19	0.21	0.01	0.02	0.00	0.01
					MCD					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	0.95	0.85	0.76	0.56	0.38	0.16	0.02	0.10	0.05	0.04
0.5	0.95	0.86	0.73	0.52	0.37	0.15	0.01	0.04	0.05	0.02
0.75	0.95	0.85	0.75	0.52	0.00	0.03	0.00	0.02	0.01	0.02
1	0.94	0.85	0.04	0.01	0.02	0.01	0.00	0.04	0.01	0.00
2	0.84	0.11	0.05	0.01	0.03	0.01	0.02	0.03	0.02	0.00
5	0.80	0.12	0.10	0.03	0.02	0.01	0.02	0.01	0.01	0.01
7	0.77	0.13	0.09	0.03	0.03	0.02	0.01	0.02	0.02	0.01
10	0.69	0.15	0.11	0.03	0.02	0.02	0.02	0.01	0.02	0.01
					IBS					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-0.01	0.00
0.5	1.00	1.00	1.00	1.00	1.00	1.00	0.16	-0.01	0.00	0.01
0.75	1.00	1.00	1.00	1.00	0.22	0.18	0.09	0.00	0.00	0.01
1	1.00	1.00	0.47	0.12	0.20	0.13	0.07	0.01	0.01	0.01
2	1.00	0.98	0.43	0.11	0.20	0.14	0.07	0.00	0.01	0.03
5	1.00	0.96	0.37	0.12	0.16	0.14	0.08	0.01	0.02	0.02
7	1.00	0.95	0.34	0.16	0.17	0.14	0.08	0.01	0.03	0.02
10	1.00	0.93	0.41	0.14	0.15	0.13	0.09	0.01	0.03	0.02

gMCD										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	0.96	1.00	0.99	0.98	0.97	0.96	0.82	0.63	0.00	0.12
0.5	0.96	1.00	0.99	0.98	0.96	0.96	0.07	0.02	0.00	0.00
0.75	0.96	1.00	0.99	0.98	0.34	0.16	0.03	0.02	0.01	0.00
1	0.96	1.00	0.36	0.28	0.21	0.10	0.01	0.01	0.01	0.00
2	0.96	0.92	0.36	0.26	0.17	0.10	0.02	0.02	0.00	0.00
5	0.96	0.85	0.39	0.24	0.20	0.09	0.04	0.02	-0.01	0.01
7	0.96	0.83	0.33	0.26	0.19	0.08	0.03	0.02	-0.01	0.01
10	0.96	0.84	0.35	0.23	0.16	0.08	0.02	0.04	0.00	0.00
LAR										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	0.98	0.00	0.00	0.00	-0.01	0.00	0.00	-0.01	0.00	0.01
0.5	0.99	0.01	0.02	0.01	0.00	0.00	0.01	0.00	-0.01	-0.01
0.75	0.99	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01
1	0.99	0.00	-0.01	0.01	0.00	0.00	0.01	-0.01	0.01	0.01
2	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.99	0.00	-0.01	0.00	0.00	0.00	0.01	-0.01	-0.01	0.00
7	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.99	0.00	0.01	0.00	-0.01	0.00	0.01	-0.01	-0.01	-0.01
RMDS										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	0.75	-0.01	-0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00
0.5	0.75	0.00	0.00	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00
0.75	0.77	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
1	0.75	0.00	0.01	0.00	0.00	0.01	0.00	0.00	-0.01	0.00
2	0.74	0.00	-0.01	-0.01	0.00	0.00	-0.01	0.01	-0.01	0.00
5	0.78	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
7	0.78	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
10	0.77	0.00	0.01	0.00	0.01	0.00	0.00	-0.01	0.00	-0.01
RSMDS										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1.00	0.03	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.01
0.5	1.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.02	0.00
0.75	1.00	0.01	0.03	0.01	0.01	0.02	0.02	0.00	0.01	0.00
1	1.00	0.03	0.02	0.00	0.02	0.02	0.01	0.01	0.01	0.00
2	1.00	0.01	0.00	0.00	0.01	0.01	0.00	0.02	0.01	0.04
5	1.00	0.02	0.01	0.01	0.00	0.01	0.03	0.01	0.01	0.00
7	1.00	0.00	0.01	0.01	0.02	0.00	0.00	0.01	0.01	0.00
10	0.97	0.01	0.02	0.00	0.01	0.02	0.00	0.00	0.03	0.00
nmMDS										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	0.00	0.11	0.05	0.04	0.05	0.01	0.02	0.00	-0.01	0.06

0.5	0.00	0.11	0.04	0.00	0.05	0.01	0.02	0.00	0.00	0.11
0.75	0.00	0.12	0.04	0.01	0.02	0.01	0.01	0.01	0.00	0.06
1	0.01	0.10	0.06	0.01	0.02	0.03	0.01	0.00	-0.01	0.04
2	0.00	0.09	0.05	0.01	0.03	0.02	0.02	0.00	0.00	0.07
5	0.00	0.10	0.03	0.01	0.03	0.03	0.01	0.01	0.00	0.09
7	0.00	0.14	0.03	0.02	0.03	0.01	0.02	0.00	0.01	0.07
10	0.00	0.11	0.05	0.01	0.03	0.01	0.03	0.00	0.00	0.06
					gMDS					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	-0.01
0.5	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.01	0.00
0.75	1.00	1.00	1.00	1.00	0.01	0.00	0.01	0.00	0.01	0.00
1	1.00	1.00	0.04	0.01	0.02	0.00	0.01	0.01	0.01	0.01
2	0.89	0.17	0.08	0.03	0.02	0.01	0.02	0.00	0.00	0.01
5	0.85	0.10	0.07	0.06	0.01	0.00	0.02	0.00	0.01	0.00
7	0.86	0.12	0.07	0.07	0.00	0.01	0.01	0.02	0.02	0.00
10	0.81	0.10	0.04	0.04	0.00	0.01	0.03	0.01	0.01	0.01
					nSimplices					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1.00	0.99	0.98	0.96	0.53	0.36	0.73	0.49	0.23	0.17
0.5	1.00	0.99	0.98	0.96	0.52	0.35	0.70	0.00	0.00	0.02
0.75	1.00	0.99	0.98	0.96	0.02	0.00	-0.01	0.00	0.00	0.00
1	1.00	0.99	0.03	0.04	0.03	0.01	0.00	0.00	0.01	0.02
2	0.98	0.34	0.06	0.02	0.03	0.01	0.00	0.00	0.00	0.01
5	0.91	0.20	0.10	0.03	0.03	0.02	0.03	0.00	0.00	0.01
7	0.90	0.40	0.03	0.03	0.03	0.01	0.01	0.03	0.00	0.01
10	0.88	0.41	0.05	0.03	0.03	0.03	0.02	0.01	0.00	0.01

Table S2. Adjusted R^2 for robustness to real outliers, EPIC dataset.

%outliers	axis 1	axis 2	axis 3	axis 4	axis 5	axis 6	axis 7	axis 8	axis 9	axis 10
					EIG					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1.00	1.00	1.00	1.00	0.97	0.99	0.86	0.99	0.95	0.00
0.5	1.00	1.00	1.00	1.00	0.97	0.99	0.82	0.98	0.89	0.01
0.75	1.00	1.00	1.00	1.00	0.97	0.99	0.79	0.97	0.86	0.00
1	1.00	1.00	1.00	1.00	0.97	0.99	0.79	0.97	0.81	0.03
2	1.00	1.00	1.00	1.00	0.95	0.99	0.63	0.95	0.14	0.01
5	1.00	1.00	1.00	0.99	0.93	0.02	0.51	0.04	0.13	0.12
7	1.00	1.00	1.00	0.99	0.93	0.02	0.51	0.05	0.13	0.13
10	1.00	1.00	0.99	0.99	0.92	0.67	0.51	0.05	0.10	0.04
					SPH					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1.00	1.00	1.00	1.00	1.00	0.93	1.00	0.92	0.91	0.66
0.5	1.00	1.00	1.00	1.00	1.00	0.93	1.00	0.87	0.87	0.63
0.75	1.00	1.00	1.00	1.00	1.00	0.93	0.99	0.87	0.87	0.26
1	1.00	1.00	1.00	1.00	0.99	0.93	0.99	0.83	0.67	0.13
2	0.99	1.00	1.00	1.00	0.99	0.91	0.98	0.81	0.52	0.13
5	0.98	1.00	1.00	1.00	0.99	0.90	0.98	0.23	0.18	0.07
7	0.98	1.00	1.00	1.00	0.99	0.90	0.97	0.22	0.18	0.07
10	0.98	1.00	1.00	1.00	0.99	0.89	0.96	0.23	0.18	0.06
					MCD					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	0.95	0.85	0.76	0.58	0.46	0.24	0.13	0.10	0.06	0.05
0.5	0.95	0.86	0.77	0.61	0.61	0.22	0.02	0.11	0.10	0.01
0.75	0.95	0.86	0.77	0.62	0.59	0.29	0.03	0.10	0.04	0.00
1	0.95	0.86	0.77	0.64	0.58	0.32	0.03	0.11	0.03	0.03
2	0.95	0.86	0.77	0.65	0.58	0.22	0.01	0.12	0.05	0.02
5	0.95	0.87	0.76	0.65	0.62	0.28	0.02	0.16	0.05	0.01
7	0.95	0.86	0.76	0.65	0.61	0.28	0.01	0.17	0.04	0.02
10	0.95	0.87	0.75	0.63	0.63	0.11	0.02	0.18	0.04	0.03
					IBS					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	0.97	1.00	0.99	0.98	0.97	0.84	0.79	0.68	0.61	0.00
0.5	0.97	1.00	0.99	0.98	0.97	0.83	0.76	0.67	0.46	0.01
0.75	0.97	1.00	0.99	0.98	0.97	0.85	0.78	0.68	0.20	0.06
1	0.97	1.00	0.99	0.98	0.97	0.85	0.79	0.68	0.14	0.05
2	0.97	1.00	0.99	0.98	0.97	0.84	0.74	0.56	0.07	0.01
5	0.97	1.00	0.99	0.98	0.97	0.47	0.30	0.02	0.02	0.04
7	0.97	1.00	0.99	0.98	0.97	0.44	0.29	0.02	0.02	0.04
10	0.97	1.00	0.99	0.98	0.97	0.44	0.28	0.02	0.02	0.02

%outliers	axis 1	axis 2	axis 3	axis 4	axis 5	axis 6	axis 7	axis 8	axis 9	axis 10
					gMCD					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1.00	1.00	1.00	1.00	1.00	1.00	0.92	0.97	0.51	0.33
0.5	1.00	1.00	1.00	1.00	0.99	0.99	0.76	0.92	0.05	0.08
0.75	1.00	1.00	1.00	1.00	0.99	0.99	0.75	0.93	0.06	0.10
1	1.00	1.00	1.00	1.00	0.99	0.97	0.56	0.43	0.09	0.05
2	1.00	1.00	1.00	0.99	0.97	0.94	0.63	0.09	0.04	-0.01
5	1.00	1.00	1.00	0.99	0.97	0.93	0.64	0.18	0.01	0.02
7	1.00	1.00	1.00	0.99	0.97	0.93	0.64	0.19	0.02	0.01
10	1.00	1.00	1.00	0.99	0.97	0.93	0.61	0.09	0.01	0.01
					LAR					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	0.98	0.00	0.00	0.00	-0.01	0.00	0.00	-0.01	0.00	0.01
0.5	0.98	0.01	0.02	0.01	0.00	0.00	0.01	0.00	-0.01	-0.01
0.75	0.97	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
1	0.98	0.00	-0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.01
2	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.98	0.00	-0.01	0.00	0.00	0.00	0.01	-0.01	-0.01	0.00
7	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.99	0.00	0.01	0.00	-0.01	0.00	0.01	-0.01	-0.01	-0.01
					RMDS					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	0.76	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
0.5	0.75	0.00	0.00	-0.01	-0.01	-0.01	0.00	0.00	0.00	0.00
0.75	0.76	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
1	0.75	0.00	0.01	0.00	0.00	0.01	-0.01	0.00	-0.01	0.00
2	0.75	0.00	-0.01	-0.01	0.00	0.00	-0.01	0.01	0.00	0.00
5	0.73	0.00	0.00	0.00	0.00	-0.01	0.01	0.00	0.00	0.00
7	0.74	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
10	0.72	0.00	0.01	0.00	0.00	0.00	0.00	-0.01	0.00	0.00
					RSMDS					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1.00	0.03	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.01
0.5	1.00	0.00	0.00	0.01	0.01	0.01	0.01	0.01	0.02	0.00
0.75	0.99	0.01	0.03	0.01	0.01	0.02	0.01	0.00	0.01	0.00
1	0.99	0.02	0.02	0.00	0.01	0.02	0.01	0.01	0.01	0.00
2	0.99	0.01	0.00	0.00	0.01	0.01	0.00	0.02	0.01	0.04
5	0.98	0.01	0.00	0.00	0.00	0.00	0.03	0.01	0.01	0.00
7	0.98	0.00	0.01	0.01	0.02	0.00	0.00	0.00	0.01	0.00
10	0.97	0.01	0.02	0.00	0.01	0.02	0.00	0.00	0.03	0.00

%outliers	axis 1	axis 2	axis 3	axis 4	axis 5	axis 6	axis 7	axis 8	axis 9	axis 10
					nmMDS					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	0.00	0.11	0.03	0.00	0.05	0.01	0.01	0.01	0.00	0.09
0.5	0.01	0.09	0.06	0.00	0.02	0.01	0.01	0.00	0.01	0.06
0.75	-0.01	0.11	0.04	0.00	0.03	0.02	0.01	0.01	0.00	0.07
1	0.00	0.11	0.05	0.00	0.04	0.01	0.03	0.00	0.00	0.08
2	-0.01	0.10	0.03	0.00	0.03	0.00	0.00	0.01	-0.01	0.03
5	0.01	0.12	0.03	0.01	0.05	0.00	0.01	0.00	0.00	0.08
7	0.00	0.13	0.04	0.01	0.03	0.01	0.01	0.00	0.00	0.05
10	0.00	0.13	0.04	0.01	0.04	0.00	0.02	0.00	0.01	0.08
					gMDS					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1.00	0.99	0.98	0.96	0.93	0.77	0.83	0.64	0.35	0.35
0.5	1.00	0.99	0.98	0.96	0.92	0.78	0.83	0.61	0.04	0.08
0.75	1.00	0.99	0.98	0.97	0.89	0.77	0.75	0.09	0.03	0.02
1	1.00	0.99	0.98	0.97	0.89	0.57	0.25	0.04	0.03	0.02
2	1.00	0.99	0.97	0.90	0.82	0.34	0.16	0.02	0.03	0.01
5	1.00	0.98	0.94	0.86	0.78	0.33	0.20	0.03	0.00	0.00
7	0.99	0.98	0.94	0.84	0.79	0.37	0.18	0.03	0.01	0.00
10	0.99	0.98	0.94	0.84	0.81	0.41	0.15	0.03	0.02	0.03
					nSimplices					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.25	1.00	1.00	1.00	0.99	0.92	0.92	1.00	0.98	1.00	0.02
0.5	1.00	1.00	1.00	0.99	0.90	0.90	1.00	0.98	1.00	0.00
0.75	1.00	1.00	1.00	0.99	0.90	0.90	1.00	0.98	1.00	0.00
1	1.00	1.00	1.00	0.99	0.90	0.89	1.00	0.98	1.00	-0.01
2	1.00	1.00	1.00	0.99	0.90	0.89	0.99	0.04	0.06	-0.01
5	1.00	1.00	1.00	0.99	0.89	0.88	0.99	0.03	0.01	0.00
7	1.00	1.00	1.00	0.99	0.88	0.88	0.99	0.03	0.00	-0.01
10	1.00	1.00	1.00	0.99	0.87	0.87	0.98	0.02	0.00	-0.01

Table S3. Adjusted R^2 for the robustness to outliers, HMP dataset.

%outliers	axis 1	axis 2	axis 3	axis 4	axis 5	axis 6	axis 7	axis 8	axis 9	axis 10
					MDSm					
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.26
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.17
7.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.15
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.14

%outliers	axis 1	axis 2	axis 3	axis 4	axis 5	axis 6	axis 7	axis 8	axis 9	axis 10
MDSe										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.26
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.17
7.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.15
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.14
LAR										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.5	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7.5	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RMDS										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.5	1.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
5	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7.5	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
RSMDS										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.5	0.19	0.16	0.10	0.13	0.15	0.06	0.17	0.15	0.19	0.16
5	0.02	0.02	0.04	0.03	0.02	0.05	0.02	0.02	0.01	0.02
7.5	0.18	0.12	0.08	0.11	0.11	0.02	0.16	0.11	0.17	0.13
10	0.08	0.04	0.07	0.08	0.03	0.04	0.08	0.03	0.05	0.04
wMDS										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.5	0.47	0.03	0.19	0.32	0.10	0.07	0.17	0.01	0.02	0.38
5	0.43	0.03	0.16	0.29	0.10	0.06	0.15	0.01	0.02	0.28
7.5	0.41	0.03	0.15	0.29	0.10	0.06	0.14	0.00	0.02	0.25
10	0.40	0.03	0.13	0.29	0.09	0.06	0.13	0.00	0.02	0.24
q_{NB}-MDS										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
q_{E}-MDS										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99
7.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.97

%outliers	axis 1	axis 2	axis 3	axis 4	axis 5	axis 6	axis 7	axis 8	axis 9	axis 10
CSS										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.5	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.99	1.00	0.05
5	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.99	1.00	0.07
7.5	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.99	1.00	0.08
10	1.00	1.00	0.99	1.00	1.00	1.00	0.96	0.98	1.00	0.08
nSimplices q_{NB}-MDS										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.93
7.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.90
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.98	0.87
nSimplices q_{E}-MDS										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7.5	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
10	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00
nSimplicesCSS										
0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.85
5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.95	0.48
7.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.33
10	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.24

9 CURRICULUM VITAE

Carine LEGRAND, M.Sc., Dipl.-Ing.

Personal Information

Date of birth:	October 5th, 1978
Place of birth:	La Garenne-Colombes (Hauts-de-Seine, France)
Father and mother:	Jean-Louis and Marie-France Legrand

Education

1989 — 1993	Collège Sainte Anne, Montesson. Final examination: *Brevet*, June 1993
1993 — 1996	Lycée Saint Érembert, St-Germain-en-Laye
June 1996	*Baccalauréat* in natural sciences, major mathematics (*magna cum laude*)

Upper Education

1996 — 1998	Preparatory classes for admission in a *Grande École*, Lycée Pasteur, Neuilly-sur-Seine Final examination: Admission to École Centrale Paris.
1998 — 2001	École Centrale Paris. Obtained degree: Dipl.-Ing.
2010 — 2011	Master II Medical Physics, University of Nantes. Obtained degree: M.Sc.

Professional Background

2001	Diploma thesis, Airbus Space, Les Mureaux Fast dynamics modeling and application to environmental survey satellite ENVISAT
2001 — 2009	Space mission analysis and Flight software Engineer, Airbus Space, Les Mureaux - Robust Estimation, trajectory and orbit calculations (Rosetta, Ariane 5 launchers), - Pilot studies and trajectories calculations (suborbital reentry, Ariane 6, etc.) - Flight prediction, compliance and quality for Ariane 5 vehicle flight software.
2010	Intern Medical Physicist, Brainlab, Feldkirchen bei München High-accuracy dosimetric measurements
2011 — 2012	Master Thesis, DKFZ, Angewandte Medizinische Physik, Heidelberg Practical investigation of displacement effect in dosimetry, for Co-60 and 6MV radiation The PTW award for dosimetry was granted for this work
2012 — 2015	PhD candidate and Biostatistician, IMBI, Ruprecht-Karl-Universität, Heidelberg - Robust PCA for Genetic Association Studies - Differential methylation, miRNA or gene expression
2015 — 2016	Post-doctoral student, DKFZ, Abteilung Epigenetik, Heidelberg - Whole Transcriptome Bisulfite Sequencing - Ribosome profiling

Award

PTW-Dosimetriepreis 2012 für "eine hervorragende wissenschaftliche Arbeit auf dem Gebiet der Dosimetrie in der Medizin verliehen" http://www.dgmp.de/de-DE/160/ptw-dosimetriepreis

10 ACKNOWLEDGMENTS

My strongest thanks go to my supervisor Prof. Dr. rer. nat. Frank Lyko. Demanding but fair and generous, sharp and constructive, Prof. Dr. rer. nat. Frank Lyko has a talent to encourage people in a fast forward and inspiring manner. This will remain a tremendously positive and inspiring time for me.

I am deeply grateful to Prof. Dr. rer. nat. Ekaterina Kostina who gave a new impetus to my thesis and provided decisive ideas and guidance. I am profoundly sorry that I wasn't up to all the promises that stemmed from her supervision.

Thanks to my previous doctoral supervisor apl. Prof. Dr. sc. agr. Justo Lorenzo Bermejo who gave me the opportunity to do a thesis and to learn at great pace in statistical genetics.

Many thanks to Dr. Federico Canzian who kindly made data from the EPIC cohort available to me, thanks also for the information and for the open and insightful discussions.

Many thanks also to Prof. Dr. John Paul Brooks who kindly provided indispensable information on the microbiome data.

I would also like to express special thanks to Miriam Kesselmeier for insightful and critical discussion, as well as to Dr. sc. hum. Maria Kabisch and to Dr. sc. hum. Barbara Peil.

My grateful thanks to Prof. Dr. Christian Karger who kindly agreed to meet and advise me to help in the progress of the thesis.

To a broader extent, thanks to all the wonderful people whom I got to know and who made these years in Heidelberg colourful, lively and friendly.

Thanks to my companion in life Frédéric Pain, who stayed at my side and was a strong support as well as a stimulating presence at all times.