Università degli Studi di Milano-Bicocca

FACOLTÀ DI SCIENZE STATISTICHE

Corso di Dottorato di Ricerca in Statistica XXIII ciclo

TESI DI DOTTORATO DI RICERCA

Population stratification in genome-wide association studies: comparison between multivariate analysis methods for dimensionality reduction

Relatore:	Candidato:
Prof Piergiorgio Lovaglio	Cristina Menni
Correlatore:	Matricola:
Dott.ssa Nadia Solaro	078763

Anno Accademico 2009-2010

Acknowledgements

I would like to thank several people for their contributions towards this thesis. First of all I would like to thank Sandosh Padmanabhan and the BHF Glasgow Cardiovascular Research Centre for introducing me to statistical genetics and for giving me the opportunity to work on genome-wide association data. I would also like to thank Sandosh for guiding me through most of the research I have undertaken during the last four years.

I am grateful to Nadia Solaro for supporting me from a statistical perspective and for her numerous suggestions, comments and insights towards the innovative part of this work.

I would also like to thank my colleagues Luisa Foco, Roberta Pastorino and Daniela Cianci for their many useful comments and discussions and for taking the time to proofread this thesis. Thanks also to Luisa Bernardinelli for her help and guidance throughout my PhD, to Vincenzo Bagnardi for his numerous suggestions, and to Giancarlo Cesana for his constant support. I am also grateful to Daniele Dellafiore and Roberto Zanotti for helping me solve practical software issues; and I thank Alessandra Bianchi and Pia Pozzi for their help to solve any administrational issue that has come up.

Further, I would like the thank all the PhD students in the Statistics group at the University of Milano-Bicocca, especially Viviana Amati and Isabella Romeo for many useful conversations; and Daniele Riggi for his help with the more practical issues.

On a more personal note, I am extremely grateful to my family, Nicolò, and my friends for their support and encouragement.

Contents

List of Abbreviations **viii**

Genetic Glossary **xi**

Introduction **1**

1 Background **7**

1.1 Basic Genetics 7

1.1.1 DNA, transcription and translation 8

1.1.2 Alleles, haplotype, genotype and phenotype 11

1.1.3 Meiosis and recombination 11

1.1.4 Polymorphisms and mutations 12

1.1.5 Mendelian genetics 15

1.1.6 Hardy-Weinberg equilibrium 15

1.2 Common complex diseases 16

1.3 Population genetics 18

1.3.1 Linkage versus association 18

1.4 Hypertension and genome-wide association studies 20

1.4.1 Evidence of a genetic component for hypertension 20

1.4.2 Linkage and candidate gene studies in hypertension . . . 21

1.4.3 The case for hypertension genome-wide association study 23

1.4.4 Designing a genome-wide association study for hypertension 25

1.4.5 Phenotyping of subjects 27

1.4.6 Analysis issues in genome-wide association studies 27

1.5 Genome-wide association study of blood pressure extremes . . . 29

1.5.1 Methods 30

1.5.1.1 Study design for the discovery cohort 30

1.5.1.2 Validation cohorts 30
1.5.1.3 Clinical functional studies 31
1.5.1.4 Urinary uromodulin measurements 32
1.5.1.5 Genotyping and quality control 32
1.5.1.6 Statistical analysis 33
1.5.1.7 Validation analysis 33
1.5.1.8 Continuous blood pressure trait modelling . . . 34
1.5.2 Results 34
1.5.2.1 Genome-wide association, replication and meta-analysis 34
1.5.2.2 Clinical functional studies 37
1.5.2.3 Cardiovascular outcomes and rs13333226 43
1.5.3 Discussion 46
1.6 Conclusions 48

2 Multivariate methods to correct for stratification 49
2.1 Genomic Control 51
2.2 Structured association 53
2.2.1 Phase 1: STRUCTURE 53
2.2.2 Phase II: STRAT 55
2.3 Principal components analysis 55
2.4 Propensity scores 59
2.5 Conclusions 60

3 Multivariate methods for non-metric or mixed data 62
3.1 The Gifi System 62
3.1.1 Homogeneity analysis 63
3.1.2 Nonlinear principal components analysis 70
3.1.3 Algorithmic implementation 74
3.1.3.1 SAS procedure for PRINCALS 75
3.2 Procrustes analysis 77
3.2.1 Algorithmic implementation in R 82

4 Comparison between linear and non-linear principal components analysis to correct for stratification 83
4.1 Description of the data matrix 84
4.1.1 Individual selection 86
4.1.2 SNPs selection 88

4.2 Application of PCA to the sample dataset 88
4.3 Application of the PRINCALS method 89
4.3.1 Phase 1: *proc prinqual* 89
4.3.2 Phase 2. Extracting the dimensions 90
4.4 Comparison of PCA and PRINCALS: graphical representation . 91
4.5 Comparison of PCA and PRINCALS: Procrustes and PROTEST 95
4.6 Comparison of PCA and PRINCALS: Genomic Control 103
4.7 Conclusions 105
4.8 R script to extract eigenvectors 106

Conclusion **111**

A Matrix Algebra **113**

B Java Script to recode SNPs and transpose the data matrix **120**

Bibliography **127**

List of Figures

1.1 Chromosome. 7
1.2 Human diploid chromosomes. 8
1.3 DNA. 9
1.4 Transcription and translation. 10
1.5 Crossing over. 12
1.6 Model of linkage disequilibrium mapping based on SNP genotyping. 13
1.7 Hypertension related loci across the genome 22
1.8 Dissecting the polygenic causation of hypertension using the genome-wide association approach. 24
1.9 Overview of association results in the discovery sample. 35
1.10 QQ plot of observed versus expected p-value distributions. 36
1.11 Association plot of the genomic region around rs13333226. 36
1.12 Forest Plots of association with rs13333226 and hypertension. . . 39
1.13 UMOD rs13333226 genotype and response to high salt intake. . . 45

2.1 Population Structure. 50

3.1 Bipartite graph of a toy example. 63

4.1 First and second axis of variation on 988 individuals and 405,402 SNPs. 86
4.2 Second and third axis of variation on 988 individuals and 405,402 SNPs. 87
4.3 First and third axis of variation on 988 individuals and 405,402 SNPs. 87
4.4 Scree plot. 91

4.5 First and second axis of variation on 90 individuals and 1,000 random unlinked SNPs. 92
4.6 Second and third axis of variation on 90 individuals and 1,000 random unlinked SNPs. 93
4.7 First and third axis of variation on 90 individuals and 1,000 random unlinked SNPs. 94
4.8 First principal component versus first dimension. 96
4.9 Second principal component versus second dimension. 97
4.10 Third principal component versus third dimension. 98
4.11 Fourth principal component versus fourth dimension. 99
4.12 Fifth principal component versus fifth dimension. 100
4.13 Procrustes superimposition plot. 101
4.14 Procrustes ordination plot of the residuals. 102

List of Tables

1.1 Demographic characteristics of the discovery case-control population. . . . 34
1.2 Results from the meta-analysis of rs13333226 and hypertension in discovery sample and after validation. . . . 38
1.3 Results from the meta-analysis of rs13333226 and hypertension before and after adjustment for eGFR. . . . 40
1.4 Univariate association analysis of rs13333226 in 256 hypertensive patients from the BRIGHT study. . . . 41
1.5 Univariate association analysis of rs13333226 in 110 participants from the HERCULES Study. . . . 42
1.6 Univariate association analysis of urinary uromodulin in relation to rs13333226 polymorphism and response to high and low salt intake (GRECO Study). . . . 44

2.1 Genotype distribution. . . . 52

3.1 Options for the PRINQUAL procedure. . . . 76

4.1 Ethnic distribution of the study sample. . . . 85
4.2 Scenarios analysis. . . . 104

List of Abbreviations

ALS: Alternating Least Squares.

ASW: African ancetry in Southwest USA.

BMI: Body Mass Index.

BP: Blood Pressure.

BRIGHT: British Genetics of Hypertension.

CDCV: Common Disease/ Common Variant.

CEU: Utah residents with Northern and Western European ancestry, CEPH collection.

CHB: Han Chinese in Beijing, China.

CHD: Chinese in Metropolitan Denver, Colorado.

CI: Confidence Interval.

CKD: Chronic Kidney Disease.

Cl: Chlorine.

CNV: Copy Number Variation.

CVD: Cardiovascular Diseases.

DBP: Diastolic Blood Pressure.

DNA: Deoxyribonucleic acid.

ECV: Extracellular Fluid Volume.

eGFR: Estimated Glomerular Filtration Rate.

FENa: Fractional Excretion of Sodium.

FTO: Fat mass and obesity associated.

GC: Genomic Control.

GCKD: Glomerulo Cystic Kidney Disease.

GIH: Gujarati Indians in Houston, Texas.

GPI: Glycosyphosphatidylinisitol.

GPS: Genomic Propensity Score.

GRECO: Groningen Renal Hemodynamic Cohort Study Group.

GWA: Genome-wide association.

GWAS: Genome-wide association studies.

HERCULES: Hypertension Elevation by Rembler and Calciuria Level Study.

HOMALS: Homogeneity Analysis by Means of Alternating Least Square.

HS: High Salt.

HWE: Hardy–Weinberg Equilibrium.

IBS: Identity By State.

JPT: Japanese in Tokyo, Japan.

K: Potassium.

LD: Linkage Disequilibrium.

LS: Low Salt.

LWK: Luhya in Webuye, Kenya.

MAF: Minor Allele Frequency.

MCKD2: Medullary Cystic Kidney Disease.

MCMC: Markov Chain Monte Carlo.

MDC: Malmö Diet and Cancer study.

MEX: Mexican ancestry in Los Angeles, California.

MKK: Maasai in Kinyawa, Kenya.

MDRD: Modification of Diet in Renal Disease study.

MDS: Multi-Dimensional Scaling.

MONICA: Monitoring Cardiovascular diseases.

MPP: Malmö Preventive Project.

mRNA: Messanger Ribonucleic acid.

Na: Sodium.

NESDA: The Netherlands Study of Depression and Anxiety.

NORDIL: Nordic Diltiazem study.

NLPCA: Non Linear Principal Components Analysis.

OR: Odds Ratio.

PAMELA: Pressioni Arteriose Monitorare E Loro Associazioni.

PC: Principal Component.

PCA: Principal Components Analysis.

PRA: Plasma Renin Activity.

PRINCALS: Principal Components Analysis by Means of Alternating Least Squares.

PS: Population Stratification.

RAAS: Renin Angiotensin Aldosterone System.

RNA: Ribonucleic acid.

SBP: Systolic Blood Pressure.
SLC12A1: solute carrier family 12 (sodium/potassium/chloride transporters).
SLC12A3: Solute carrier family 12 (sodium/chloride transporters).
SNP: Single Nucleotide Polymorphism.
STRAT: Structure population Association Test.
SSQ: Sum of Squares.
TAL: Ascending Limp of the Loop of Henle.
TDT: Transmission Disequilibrium Test.
THP: Tamm Horsfall Protein.
TSI: Toscani in Italia.
UMOD: Uromodulin.
YRI: Yoruba in Ibadan, Nigeria.
WTCCC: Wellcome Trust Case Control Consortium.

Genetic Glossary

Additive: A model for dominance in which the heterozygote is at intermediate risk between the two homozygotes.

Allele: Different version of the same gene at the same position on corresponding chromosomes.

Amino acids: The building blocks of proteins.

Candidate gene: A gene whose location and biological function are already known, which may be plausibly related to the phenotype under study.

Chromosome: Storage units of genes.

Coding region: A sequence of DNA that codes for a protein sequence.

Complex disease: A disorder that appears to have a genetic component with no simple Mendelian pattern of single-gene inheritance; multiple genes and environmental factors appear to be involved.

Diploid: Most body cells containing two copies of the genome (one from the father and one from the mother).

DNA: Deoxyribonucleic acid, which makes up genes and chromosomes, composed of a double- stranded helix of nucleotides.

Dominance: The joint effect of the two copies of a gene in influencing the phenotype probability.

Dominant: A model for dominance in which individuals with a single copy a disease susceptibility allele have the same risk of disease as those with two copies.

Exon: DNA coding sequence.

Gamete: Germ cell (sperm or egg).

Gene: The basic unit of genetic information.

Genome: Collection of genetic information.

Genome scan: The process of searching for a disease gene using widely space markers scattered throughout the entire genome.

Genotype: The set of alleles at a particular locus.

Haploid: Cells containing one copy of the genome (e.g. sperm and unfertilised eggs).

Haplotype: Allelic configuration along a single chromosome.

HapMap project: A global consortium mapping all common SNPs in different populations across the world.

Hardy-Weinberg Equilibrium: A tendency for the population allele frequencies to remain invariant across generations, with genotype probabilities that are a particular function of the population allele frequency.

Heterozygote: An individual who carries two different alleles for a particular gene.

Homozygote: An individual who carries two identical copies of an allele for a particular gene.

Intron: DNA non-coding sequence in a gene.

Linkage disequilibrium: A tendency for certain pairs of alleles at two linked loci to be associated with each other in the population more often than would be expected by chance.

Locus: Unique chromosomal location defining the position of an individual gene or DNA sequence.

Marker: A polymorphic genomic feature whose physical location is known.

Meiosis: Cell division process that leads to the creation of sex cells or gametes.

Mutation: Rare genetic change.

Nucleotide: A nitrogenous base linked to a sugar (ribose or deoxyribose) and a phosphate. Nucleotides are the building blocks for nucleic acids (DNA and RNA).

Penetrance: Probability of a phenotype given a genotype.

Phenotype: The observable trait or disease status that may be influenced by a genotype.

Polymorphism: A tendency for a gene to exist in more than one form (the variant should be relatively common, i.e. present in at least 1% of the population).

Recessive: A model for dominance in which individuals with only one copy of the susceptibility allele have the same risk as those with none.

Recombination: The phenomenon in which genes from two different homologous chromosomes are joined during meiosis.

RNA: A type of single stranded nucleic acid that serves as an intermediate between genomic information (DNA) and its phenotypic expression (protein).

Single Nucleotide Polymorphism (SNP): A DNA sequence variation consisting of a change in a single nucleotide.

Tag SNP: A representative SNP in a region of the genome with high linkage disequilibrium.

Transcription: The first step in the conversion of genomic information to protein. It is the process that copies genetic information from DNA to RNA.

Translation: The second step in conversion of genomic information to protein. It is the process that copies genetic information from mRNA to protein.

Introduction

Common complex diseases have a multifactorial etiology arising as the result of the interplay between many genetic factors and environmental exposure. Determination of the genetic variants involved in a particular disease should provide new insights into susceptibility to the disease and will also have a major impact on public health, improving prevention, diagnosis and treatment [1, 2, 3].

However, the enormous diversity within complex traits, not only in their environmental determinants, but also in their genetic components of risk, means specific genetic variants causally associated with common diseases will have small effects [4, 5], and their identification so far has been an uphill struggle [6, 7, 8]. To succeed in finding complex disease genes, a study must detect a relatively weak statistical signal, and potentially, the choices made in study design can have a dramatic impact on the probability of success [9].

The most common genetic variant studied in complex trait association is single nucleotide polymorphism (SNP). This is a DNA sequence variation, occurring when a single nucleotide, adenine (A), thymine (T), cytosine (C) or guanine (G), in the genome sequence is altered. A variation must occur in at least 1% of the population to be considered a SNP.

In the human genome, which consists of 3 billion nucleotide bases, SNPs occur every 100 to 300 bases. As the coding region comprises only 5% of the genome, most of the SNPs are found in regions that do not affect protein structure. If they occur in the coding regions, they may be synonymous, i.e. coding for the same protein without any change in amino acid sequence. Thus, most of the common SNPs might not be so informative. Amino acid altering coding region, non-synonymous SNPs, are rare and are harder to be found because of expected selection against them in human evolution.

Until recently, SNPs have been used in candidate gene association studies. This approach involves the selection of candidate genes based on a mechanistic understanding of the roles of the encoded proteins in disease regulation, or the location within a previously identified region. However, this will at best identify only a fraction of genetic risk factors even for diseases in which the pathophysiology is well understood. To date, roughly 50 genes and their allelic variants that contribute to complex diseases have been conclusively identified [10]. Now, emerging technologies are allowing researchers to study hundreds of thousands of genetic variants as risk factors for common complex diseases, with promising and exciting results [7, 11, 12, 13, 14, 15]. Genome-wide association studies (GWAS) are large-scale association mapping using SNPs, making no assumptions of the genomic location of the causal variant. They are a comprehensive approach to test the hypothesis that common alleles contribute to heritable phenotype variation. GWAS are mostly case-control studies, where SNP frequencies are compared between the two groups, and those that differ significantly are then validated in independent samples. It is not yet technically feasible to resequence every base on the genome or genotype the approximately 11 million currently known SNPs. However, the availability of genome-wide SNP arrays that genotype 650,000 to 1,000,000 SNPs per sample and the availability of linkage disequilibrium patterns on a genome-wide scale through the HapMap project (a global consortium mapping all common SNPs in different populations across the world) [16, 17], make these studies possible. Such studies typically measure sets of special DNA ‘tagging’ SNPs identified in the HapMap project, enriched with non-synonymous SNPs, as well as SNPs in evolutionary conserved regions of the genome. The genome-wide association approach thus represents an unbiased, yet fairly comprehensive option that can be attempted even in the absence of convincing evidence regarding the function or location of the causal genes [18]. There are major statistical issues in GWAS to be overcome including optimal study designs, population stratification, multiple testing, environment and gene interactions, and the effect of epigenetics and structural chromosomal variations.

Population stratification (PS) is a form of confounding that results by the presence of a systematic difference in allele frequencies between cases and controls due to different ancestries rather than association of genes with disease. Indeed, if the study population consists of subpopulations that differ genetically, and if the disease of interest is at high frequency in one subpopulation, we can expect to find such group over represented among the cases. Then any marker allele at higher frequency in that subpopulation compared to the others will appear to be associated with the disease, regardless of where it is in the genome [19]. PS is perhaps the most often cited reason for non replication of genetic association studies, since undetected stratification can mimic the signal of association and lead to more false positive findings or miss the real effects [20].

Although a well designed case-control study attempts to draw cases and controls from the same population, a hidden fine-scale genetic substructure within that single population cannot be ruled out. So, in the last few years, several statistical methods were developed to account for PS so that association studies could proceed even in the presence of structure. These methods use genotype information either from a set of random markers or from a set of selected ancestry informative markers and can be broadly classified into three classes [21, 22]: (i) genomic control [23]; (ii) structured association [24]; and (iii) principal component methods and multidimensional scaling [25]. Recently, a propensity score method has also been suggested [26]. No consensus, however, has been reached as to which method is the best, even though, for GWAS data, principal component analysis (PCA) represents a sort of "gold standard", being easy to apply and computationally feasible.

PCA uses genotype data to extract continuous (principal) axes of variation, which can be used to adjust for association attributable to ancestry along each axis. Under the assumption that markers are biallelic (e.g SNPs), each marker is quantified by fixing a reference and a variant allele and by counting the number of mutations. Hence, an individual who is homozygous wild type will have no variant alleles and will be assigned a value of zero; an individual who is heterozygous will have one variant and one reference allele and will be assigned a value of one; and an individual who is mutated homozygous will have two variant alleles and will be assigned a value of two. Data can then be seen as a large rectangular matrix of 0, 1 and 2 with rows indexed by individuals and columns indexed by polymorphic markers. The PCA algorithm, introduced for GWAS data by Price and collegues [25, 27] and implemented in the software EIGENSTRAT, involves the calculation of the eigenvalue decomposition of the data covariance matrix of the individuals, after mean centring and normalising the data for each attribute. It is important to notice that the covariance matrix is performed on the individuals, rather than on the markers, and hence its dimension will be equal to the number of samples in the dataset. The reason for this is that we are chiefly interested in the situation where the number of individuals is considerably less than the number of markers. Axes of variation (eigenvectors) are then used as covariates in the multilinear regression model.

Besides its computational merits, there are various pros to the use of PCA. It has been shown that its performance is very stable under various stratification levels in terms of power, type 1 error rates, accuracy, and positive prediction values [28]. However, PCA rests upon the assumption that the variables under study are continuous and SNPs are therefore quantified as previously explained. Counting the number of mutations for each marker and recoding genotypic data into 0, 1 and 2 assumes that the distance between homozygous wild type and heterozygous is the same as the distance between heterozygous and homozygous mutant and hence it assumes an additive model of inheritance. This model is the most conservative, it is very static and moreover it is not necessarily the correct one.

Our approach is to treat SNPs as ordinal qualitative variables. This means that we agree that there is an order between homozygous wild type, heterozygous and homozygous mutant, but that the distance between each pair is not necessarily the same. As we no longer assume a model of inheritance, we believe that our approach is more flexible and can potentially capture some information which linear PCA misses out.

We apply a multivariate technique to reduce dimensionality in the presence of non-metric data known as non linear principal components analysis (NLPCA, also known as PRINCALS: Principal components analysis by means of alternating least squares) and introduced by Gifi as derived from homogeneity analysis with restrictions [29, 30]. We apply both PCA and PRINCALS to a sample dataset of 90 individuals belonging to three very distinct subpopulations and 1,000 randomly chosen uncorrelated SNPs and compare the results graphically, using Procrustean superimposition approach and the test Protest and finally with a scenarios analysis.

The plan of the thesis can be summarised as follows. Chapter 1 begins with a brief excursus on the principles of human genetics which are required for a better comprehension of this work. It then contains a general overview of genetic epidemiology with particular focus on genome-wide association studies and hypertension as an example of complex disease. The chapter ends with the description, results and scientific issues which were encountered in a personal experience with genome-wide association studies. In the context of the InGenious HyperCare Network, the University of Glasgow (where I am an honorary research assistant) and the Istituto Auxologico Italiano carried out a case-control study of two large groups of extremely well phenotyped hypertensive cases and fully normotensive controls, using SNP-based genome-wide analyses (610K chip). This study evaluated 1,621 cases chosen among the hypertensive patients recruited in Sweden for the NORDIL trial, and 1,699 fully normotensive controls chosen among the subjects enrolled in the Malmo Diet and Cancer (MDC) study. We identified a locus on chromosome 16 in the $5'$ region of the uromodulin gene and were able to validate this result in an additional 19,845 cases and 16,541 controls (UMOD; rs13333226, combined p-value of 3.6 x 10^{-11}). When we first analysed the Swedish discovery sample, we encountered a lot of stratification and this was the starting point for this work.

Chapter 2 is dedicated to the description of the most used methods to correct for population stratification. Particular emphasis is given to PCA which represents a "gold standard" for GWAS.

Chapter 3 describes Gifi' s system for multivariate dimensionality reduction in the presence of qualitative or mixed data with particular attention to homogeneity analysis and PRINCALS. Also the goodness of fit measure Procrustes and the test Protest are presented. Procrustes rotation can be used to compare two data matrices and the test Protest tests matrix association using a correlation like statistic derived from the Procrustes sum of squares.

Chapter 4 contains the application of PCA and PRINCALS to a sample dataset consisting of 90 individuals drawn from three ethnically distinct HapMap populations (30 Caucasian Europeans, 30 Chinese from Beijing, China, 30 Yoruba from Ibadan, Nigeria) and 1,000 random SNPs not in LD to detect and correct for PS. PCA is applied using an R algorithm that mimics the EIGENSTRAT software, while we use SAS *proc prinqual* followed by the same R algorithm for PRINCALS. The sample dataset is described with particular emphasis on individuals and SNPs selection. Results obtained with the two different approaches are compared graphically, then by mean of the Procrustean superimposition approach and by the test Protest. Finally we perform a scenarios analysis. By randomly allocating the case/control label to generate structure, we calculate genomic control first on the raw data and then adjusting respectively for the first three axis of variation obtained with PCA and for the first three axis of variation obtained with PRINCALS. We compare the magnitude of genomic control in the three cases. We consider 10 different scenarios each with a different level of PS.

The thesis ends with a summary of some basic linear matrix algebra instrumental for a better understanding of Chapter 2 and 3 and a Java script to prepare the data matrix for the R algorithm. These are presented, respectively, in appendices A, and B.

Chapter 1

Background

1.1 Basic Genetics

A brief summary of the principle of molecular biology on which the phenomena of transmission of genes are based, is now given [31, 32, 33, 34]. Those familiar with human genetics may wish to skip this section.

A gene is the basic unit of genetic information, which determines the inherited characters. The genome is the collection of genetic information and chromosomes are storage units of genes (Figure 1.1). A locus is a unique chromosomal location defining the position of an individual gene or DNA sequence.

The human genome consists of 23 pairs of chromosomes: 22 pairs of autosomes (numbered 1-22) and 1 sex chromosome pair (XX or XY). Somatic cells are termed diploid as they contain two copies of the genome – one from the

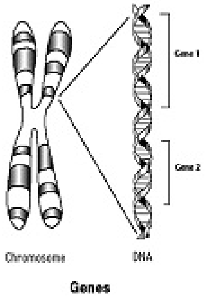

Figure 1.1: Chromosome.

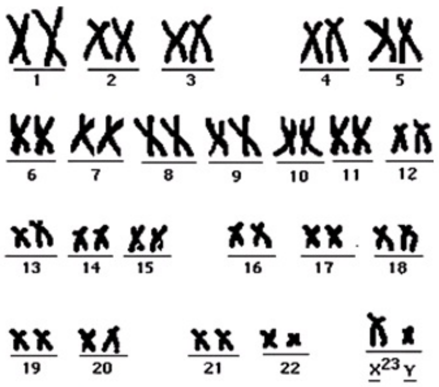

Figure 1.2: Human diploid chromosomes.

father and one from the mother. Germ cells such as sperm or unfertilised egg cells which, by contrast, contain only one copy of the genome, are said to be haploid (Figure 1.2).

One chromosome in each of the 22 homologues pairs is derived from the mother and one from the father, and the two homologues have the same sequence of genes in the same position, but usually exhibit sequence variation at several loci and can therefore be distinguished.

1.1.1 DNA, transcription and translation

The human genome is made up of DNA which consists of a long sequence of nucleotide bases of four types: adenine (A), cytosine (C), guanine (G), thymine (T). Strong covalent bonds bind bases together along a single strand, and weaker hydrogen bonds pair A with T and C with G between the two strands. Each single strand has two different ends, $5'$ and $3'$, oriented in opposite directions (Figure 1.3).

In the nucleus, DNA is double stranded. Double stranded DNA is replicated by the breakage of two strands and construction of a new complementary strand for each, resulting in two identical copies of the original. The code is embodied in the sequence of bases along the DNA strand of the gene: a set of three bases (known as a codon) specifies an amino acid. Some amino acids are specified by more than one codon, and some codons stand for ‘stop’ signals indicating the end of the protein. The protein-coding sequence of most genes is interrupted by non-coding sequences called introns; the protein-coding sections are called exons.

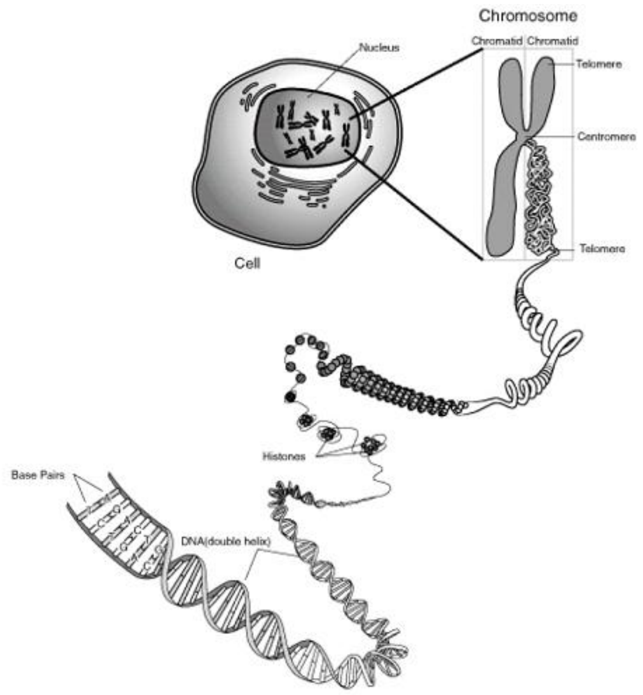

Figure 1.3: DNA.

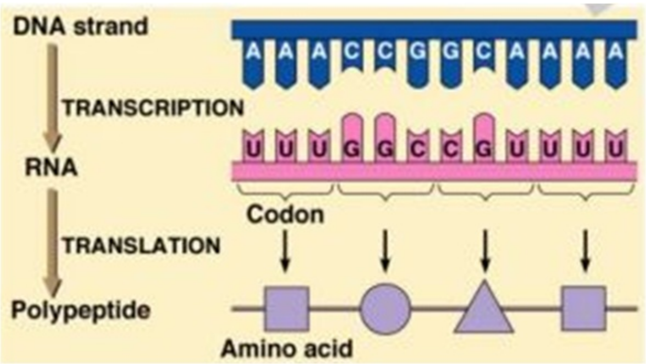

Figure 1.4: Transcription and translation.

Genes also contain regulatory sequences (usually located outside the coding region) that control whether and when that protein is made. Every cell in the body contains a complete set of DNA instructions for all the millions of different proteins the body needs. The human genome contains 3 billion base pairs and 22-25,000 genes, which represent no more than a few per cent of the total DNA sequence. The functions of the remaining 'junk' DNA are largely unknown. The discrepancy between the number of genes and the number of proteins is explained partly by mechanisms such as alternative splicing, where different combinations of exons within the same DNA sequence encode different proteins, and partly by production of proteins with different functional properties by variations in post-transcriptional and post-translational processing.

A single strand of DNA can also act as a template for a complementary strand of RNA. This transcription RNA is similar to DNA, but T is replaced by U (uracil) and deoxyribose is replaced by ribose. In certain regions of the DNA (genes), transcribed RNA encodes instructions that tell the cell how to assemble amino acids to make proteins. Most genes contain alternating regions, called exons and introns. The RNA that is transcribed is complementary to the whole gene (exons and introns). Mature messanger RNA, mRNA, is then created by post-transcriptional processing, which cuts out the introns and splices the exonic elements to produce mRNA which codes for proteins. The production of protein via mRNA is called translation (Figure 1.4). It is mainly through altered protein functions that changes in DNA affect health and disease.

1.1.2 Alleles, haplotype, genotype and phenotype

Different versions of the same gene at the same position on corresponding chromosomes (i.e. at the same genetic locus) are known as alleles. Everyone has two alleles of each gene in their autosomes, one inherited from their mother and the other from their father. Presence of two copies of the same allele at a locus means they are homozygous for that allele. Presence of different alleles at one locus means that they are heterozygous for that allele. The allelic configuration along a single chromosome is called a haplotype. The set of alleles at a particular locus is defined as the genotype at that locus. The genotype of an individual remains unchanged throughout its life, regardless of the environment surrounding and affecting it. On the other hand, the phenotype of an individual is the set of observable characteristics (e.g. its total physical appearance and constitution or a specific manifestation of a trait, such as size, eye colour, or behaviour). The phenotype is the result of the interaction between the genotype and environmental factors and varies between individuals.

1.1.3 Meiosis and recombination

When sex cells or gametes (sperm or eggs) are produced, by a special cell division mechanism called meiosis, the genetic complement must be halved. Two sequential divisions produce four gametes, each containing only one representative of each homologous pair plus one sex chromosome. Fertilisation then restores the full chromosome complement.

During the first division of meiosis, the members of each homologous pair separate (or 'segregate') to the two resulting cells independently of the members of any other pair, so that the 23 chromosomes in the different gametes produced by an individual represent different assortments of the original 46. This is an important source of variation that results from sexual reproduction. Variation also results from a process called crossing over and recombination (Figure 1.5). During the first division of meiosis, homologous chromosomes can swap portions of their DNA, which means that the chromosomes in the gametes may contain different sets of alleles from the chromosomes of the individual who produced them. There are typically one or two crossovers per chromosome during meiosis. It is possible to follow the inheritance of alleles down through generations. Because of recombination and the independent segregation of chromosomes in meiosis, most genes show inheritance patterns that are independent of the patterns shown by other genes. For example, the genes determining eye colour is

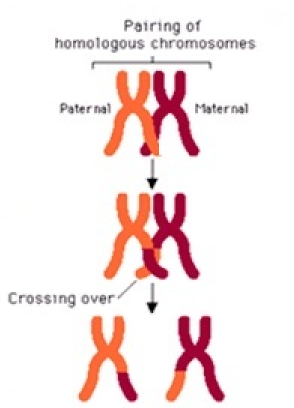

Figure 1.5: Crossing over.

inherited independently of the genes determining shape of the ear lobes. However, genes that are close together on the same chromosome tend to be inherited together from parent to offspring more often than expected under independent inheritance, as they are less frequently separated by recombination. The frequency with which two genes are inherited together, or linked, depends on the length of the DNA between the two genes: the longer the gap, the greater the chance that recombination will separate them. Linkage between genes is the basis of genetic mapping using patterns of inheritance to determine the relative positions of genes on the chromosomes.

Another key concept in genetic epidemiology is linkage disequilibrium (LD). LD is the non-random association of alleles at two or more loci on a chromosome and results in greater co-occurrence of two genetic markers on the same chromosome in a population than would be expected for independent markers (Figure 1.6). In short, linkage measures the co-segregation between a genetic marker and a disease affection status in a pedigree, due to meiotic recombination events in the last 2–3 generations, while LD measures co-segregation in a population (a very large pedigree extending back to the founders) resulting from much earlier ancestral meiotic recombination events. In general, two loci in LD are also linked, but the reverse is not always true [31, 35].

1.1.4 Polymorphisms and mutations

Despite the wide range of phenotypes observed in the human race, our DNA has little variability. More than 99% of the nucleotides in the DNA are the

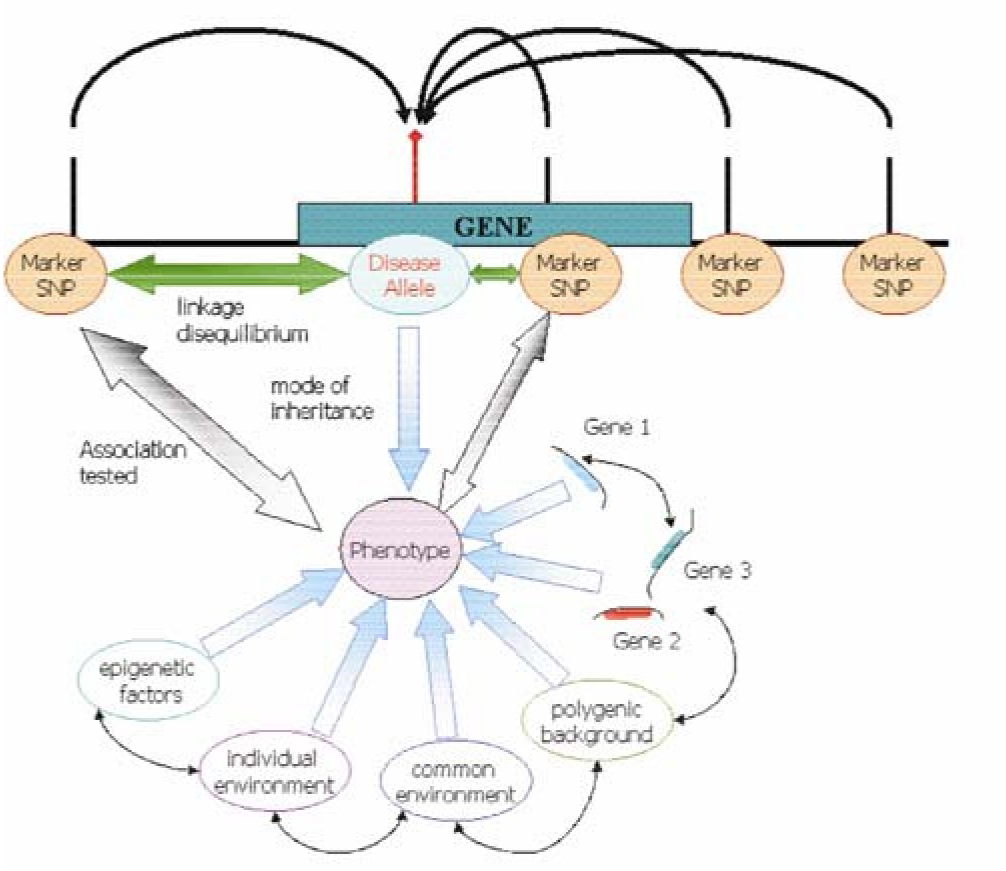

Figure 1.6: Model of linkage disequilibrium mapping based on SNP genotyping. Association is tested on a marker in LD with the disease allele. Correlation between phenotype and the marker allele should indicate the presence of the causative SNP in LD. However, the strength of the association is affected by the presence of other factors that influence phenotype. The success on LD mapping is determined by genotyping sufficient SNPs so as the capture all the LD blocks in the genomic region and by minimising the effect of environmental confounders and other causes of etiological heterogeneity (modified from Weiss and Terwilliger [36]).

same in all humans. Those DNA locations that vary from person to person are said to be polymorphic. Generally, the term 'mutation' is used to refer to a rare (present in less than 1% of the population) and deleterious genetic change, whereas 'polymorphism' refers to a normal variant (present in at least 1% of the population).

The most common and most useful for many purposes type of polymorphism is the single nucleotide polymorphism (SNP) which represents variation in a single nucleotide (e.g. C to T or A to G) [37]. Each SNP has a "major" allele and a "minor" allele based on their observed frequency in the general population. As humans are diploid, at a chosen SNP, a person can have one of several genotypes: homozygous for the major allele, heterozygous or homozygous for the minor allele.

Other very useful markers are the DNA microsatellites, traditionally employed in linkage analysis. Microsatellites are short sequences of DNA (usually one to six nucleotides) that repeat multiple times (usually 10 to 60 times).

Although SNPs are not as polymorphic as microsatellites, the abundance of SNPs in the human genome and their potential to be genotyped in a large scale, automated fashion makes them the best DNA-based markers for genetic case-control studies [38]. Their abundance, however, does not guarantee that the resulting genetic map is powerful for all disease association studies. Indeed, many SNPs are correlated with one another making it difficult to identify the SNP that impacts the phenotype from the several SNPs associated with it [39, 40].

SNPs occur on average once every 100 to 300 base pairs in the human genome [41, 42, 43]. It is expected that there are approximately 6 millions common SNPs in the human genome, i.e. SNPs with a minor allele frequency ranging from 5% to $>$ 20%. Most of these do not occur in the coding region of genes or even in genes [44], and of those that occur in the coding regions, even fewer change an amino acid within a protein ("non-synonymous" SNPs) [42]. Non-synonymous SNPs are the obvious suspects in causing a proportion of human disease, but they do not account for all SNPs that can cause disease or disease susceptibility. Indeed, other functional SNPs implicated in disease or disease susceptibility include SNPs located in promoters [45], introns [46], splice sites [47], intrageneic regions [46] and even synonymous SNPs [48]. These SNPs alter the DNA sequence, but do not change the protein coding sequence as interpreted at translation. Because of the redundancy in the code, several different codes specify the same amino acid.

1.1.5 Mendelian genetics

The first coherent description of the inheritance of genes was presented by Gregor Mendel in 1865, based on the breeding experiments with pea plants, which he summarised in two principles.

1. Law of segregation: each individual carries two copies of each gene, one inherited from each parent. Alleles at any given gene are transmitted randomly and with equal probability.

2. Law of independent assortment: alleles of different genes are transmitted independently.

A third concept, distinct from the two basic Mendelian principles, but generally considered part of the Mendelian framework, assumes that the expression of two genes is independent of which parent they come from, that is, heterozygotes Dd have the same penetrance (i.e. probability of a phenotype given the genotype) irrespective of whether the D allele comes from the father and the d allele comes from the mother or viceversa [33]. During gamete formation the segregation of the alleles of one allelic pair is independent of the segregation of the alleles of another allelic pair.

1.1.6 Hardy-Weinberg equilibrium

The Hardy–Weinberg principle states that both allele and genotype frequencies in a population remain constant, i.e.they are in equilibrium, from generation to generation under the following assumptions:

- Large population. The population must be large to minimise random sampling errors.

- Random mating. There is no mating preference. For example an AA male does not prefer an aa female.

- No mutation. The alleles must not change.

- No migration. Exchange of genes between the population and another population must not occur. Most populations are relatively isolated with some rare exchange of marriage partners between groups. It has been shown that an average of one immigrant per generation is enough to keep genetic drift partially at bay and to avoid complete fixation of alleles.

- No natural selection. Natural selection must not favour any particular individual.

Let's consider a diallelic locus with alleles a and A with population frequencies 1-q and q respectively. Under HWE, the probabilities of the three possible genotypes (aa, aA, AA) are ($(1-q)^2$, $2q$(1-q), q^2). Further, HW law states that in non-homogeneous but randomly mating populations, these frequencies are established in a single generation after mixing.

1.2 Common complex diseases

Common complex diseases such as cancer, diabetes and heart disease, arise as a result of the interplay between many genetic factors and environmental exposures, and impose a significant health burden worldwide. Determination of the genetic variants involved in a particular disease should provide new insights into the susceptibility to the disease, disease progression and severity, leading to novel pharmaceutical targets, with the ultimate goal of improving prevention, diagnosis and treatment [1, 2, 3]. There is empirical evidence that specific genetic variants causally associated with common diseases will have small effects (risk ratios < 2.0). While individually these effect sizes are minor, the combination of even a few small effects, caused by less than 20 common genetic variants could account [49] for a sizeable population attributable fraction of common diseases and shed important light on etiopathogenesis. A complex interplay of genetic and environmental factors likely accounts for the largest attributable fraction of common diseases. Genetic variants with large effect sizes manifest as Mendelian or single-gene disorders and account for only a small fraction of cases. Tracing patterns of genetic segregation in families have been used successfully in single-gene Mendelian diseases, but have provided little success in the identification of genes that underlie common complex traits.

To succeed in finding complex disease genes, a study must detect a relatively weak statistical signal, and choices in study design can potentially have a dramatic impact on the probability of success [9]. The key factors which determine which study design would be most successful are the allelic architecture, their frequencies and penetrances.

The common disease common variant hypothesis (CDCV) [50] holds that the genetic variants underlying complex traits occur with a relatively high frequency ($> 1\%$), have undergone little or no selection in earlier populations and are likely to date back to more than 100,000 years ago. In contrast, the common disease rare variant hypothesis argues that diseases are common because of highly prevalent environmental influences, not because of common disease alleles in the population. The fact that the human population rapidly expanded from a small founder pool over a short time [51, 52], and that disease-risk alleles that were common in the founder population take a long time to be diluted out by new alleles generated during population growth [49] support the CDCV hypothesis. However, it is clear that environmental factors have an important role in complex traits and the individual genetic variants associated with these traits have low attributable risk. The CDCV model depends on the survival of common risk alleles that are today capable of significantly influencing health. Mendelian disease (or monogenic disease) mutations on the other hand, are highly penetrant and usually under very strong selection, which keeps them at low frequencies with high levels of allelic heterogeneity (a form of genetic expression in which distinct mutant alleles at the same locus lead to the same disease phenotype). Susceptibility variants involved in complex diseases seem to have low or medium penetrance, and are probably not subject to such strong selection resulting in lower allelic heterogeneity. Nevertheless, previous selection can also be a factor in complex traits, since there is a case that several of the common variants underlying disease today have increased within the last 5,000 years as a result of selection. These variants may have exerted significant phenotypic effects in the past and hence are more likely to do so again today under changed environmental circumstances. Selectively neutral alleles are a random selection of variants arising throughout evolutionary history, the outcome of the evolutionary processes of mutation, selection and random drift. It seems unlikely that large numbers of these random, functionless events will significantly influence common disease traits. However, it is likely that some of these alleles will reach intermediate frequencies of more than 5–10% and are likely to represent only a small fraction of all the loci that are involved in disease susceptibility, but they will contribute disproportionately to the total population risk [53]. These are the loci amenable to mapping by association or linkage studies. The challenge of complex disease mapping is that the marginal increase in risk due to the at-risk genotype at a disease gene is quite small. The most common genetic variants studied in complex trait association are SNPs.

1.3 Population genetics

Population genetics is concerned with understanding patterns of genetic variation within and between populations. In particular, it seeks to delineate the relative roles of factors such as mutation, recombination, selection, geographic structure, migration and mate selection in contributing to genetic variation.

1.3.1 Linkage versus association

There are two broad classes of gene mapping methods: linkage and association. The underlying idea of both methods is that a disease-predisposing allele will pass from generation to generation together with variants at tightly-linked marker alleles. Therefore, if the transmission across generations of a marker allele is correlated with the transmission of the phenotype, it may be that the marker is tightly linked with a locus at which a disease-predisposing allele has arisen. If not, recombination should have broken down the correlation between the marker allele and the phenotype.

Most common methods and data types are:

- Linkage methods
    1. Parametric linkage: extended pedigree over multiple generations.
    2. Nonparametric linkage: affected relative pairs.
- Association methods
    1. Population-based: samples of unrelated cases and controls.
    2. Family- based: e.g. trios of two unaffected parents and one affected child

Linkage analysis is used to map genetic loci by use of observation of related individuals (pedigree/family tree) and it thus require family data. Many generations and/or families are needed for there to be sufficient recombination events. It can be difficult to obtain and verify this type of data.

The terms parametric and nonparametric refer to whether or not the penetrance functions (i.e. probability of a phenotype given a genotype) need to be specified.

Linkage analysis is often the first stage in genetic investigation of a trait since it can be used to identify broad genomic regions that may contain a disease gene, even in the absence of previous biologically driven hypothesis [35].

In **population association methods**, which are based on LD, comparisons are made between unrelated individuals and so recombinations over perhaps hundreds or thousands of generations can break down associations of phenotype with all but the most tightly-linked marker alleles, thus achieving "fine" mapping. Association differs from linkage in that the same allele (or alleles) is associated with the phenotype in a similar manner across the whole population, while linkage allows different alleles to be associated with the trait in different families. Nevertheless, genetic association arises only because human populations share common ancestry and association studies can thus be seen as a special form of linkage studies where the extended family is the wider population [54]. Since the pedigree is not available in association methods, the transmission of phenotype over generations cannot be traced and one must rely on correlations of current phenotype with current marker alleles. Population-association studies are mostly case-control studies, where SNP frequencies are compared between the two groups. Those that differ significantly are then validated in independent samples. With the advent of new technologies, it is now possible to genotype hundred of thousand of genetic markers and genome-wide association studies are emerging as they represent an unbiased, yet fairly comprehensive option that can be attempted in the absence of convincing evidence regarding the function or location of the causal genes [18]. They are large-scale association mapping using SNPs, making no assumptions of the genomic location of the causal variant. They measure sets (typically 650,000 to 1,000,000) of special DNA 'tagging' SNPs identified in the HapMap project [16, 17], enriched with non-synonymous SNPs, as well as SNPs in evolutionary conserved regions of the genome. Population association studies are the easiest and cheapest to perform, but there is the risk of spurious association. Indeed, association does not necessarily imply causation, but may reflect either linkage to a nearby gene (because of LD if a disease gene and a marker gene are closely linked, their alleles may not be independently distributed in the population; i.e. carrier of the disease allele may also be more likely to carry a particular marker allele) or simply a spurious result due to some underlying stratification or admixture in the population (see Chapter 2).

Family-based association methods exploit associations between the two parents as well as the single generation transmission from parents to child: therefore they are tests of both association and linkage making them insensitive to population structure. This type of studies, however, are more expensive and it can be difficult to collect family data.

As the present work will deal with genome-wide association studies (GWAS) and arises from a personal experience of GWAS in hypertension, the remaining part of this chapter focuses on hypertension and GWAS. First, we outline where genetic research in hypertension has arrived so far [55, 56]. Then, we present a personal and positive experience of GWAS and hypertension [15].

1.4 Hypertension and genome-wide association studies

The main determinants of blood pressure (BP), cardiac output and total peripheral resistance are controlled by a complex network of interacting pathways involving renal, neural, endocrine, vascular and other mechanisms. Multiple genes within each of these systems contribute to the specialised functions regulating BP, and hence it is likely that many genes will participate in the development of hypertension [6, 57, 58]. However, at present, it is not known whether the inheritance of hypertension susceptibility in humans is attributable to variation in a narrow or a large subset of these genes. Identifying the genes, and thus causative mechanisms leading to hypertension, should aid the early diagnosis and development of more specific targeted preventive measures. Knowledge of the disease pathway and/or drug response profile will also help tailoring of therapy to achieve optimal BP control and simultaneously reduce the financial and personal costs of ineffective or dangerous treatments.

1.4.1 Evidence of a genetic component for hypertension

The influence of genes on BP has been suggested by family studies demonstrating correlation of BP among siblings and between parents and children. BP variability attributable to all genetic factors varies from 25% in pedigree studies to 65% in twin studies [59, 60]. The heritability of systolic BP is around 15-40% and 15-30% for diastolic BP, whereas the sibling recurrent risk of hypertension is around 1.2-1.5, indicating a phenotype with modest genetic effect. From an evolutionary perspective, hypertension is a disease of civilisation and may be an undesirable pleiotropic effect of a preserved genotype that may have optimised fitness in the ancient environment [61]. The rates of hypertension and sodium sensitivity are generally higher in individuals carrying the ancestral alleles of sodium-conserving genes, which show strong latitudinal clines with the ancestral sodium-conserving alleles much more prevalent in African populations and less so in the northern regions [62, 63, 64]. It is also hypothesised that the renin–angiotensin–aldosterone (RAAS) system was initially adapted for sodium conservation with modern civilisations facing detrimental effects even with its normal state of activity and adapting by selection for lesser RAAS activity [65]. Genetic variants with large effect sizes manifest as Mendelian or single-gene disorders with mutations in at least 10 genes known to cause hypertension or hypotension, primarily by affecting renal tubular electrolyte transport functions [59, 66]. However, these rare alleles account for less than 1% of human hypertension and have not been associated with common forms. Studies are ongoing to determine whether other alleles of these genes, perhaps with smaller effects, are involved in essential hypertension, with the WNK1 [67, 68] gene showing suggestive association but, in general, there is an absence of unequivocal evidence that this is the case.

1.4.2 Linkage and candidate gene studies in hypertension

Genome-wide linkage studies [69, 70, 71, 72, 8] with microsatellites have shown evidence for existence of several chromosomal regions that are linked to BP or hypertension on almost all chromosomes (see Figure 1.7).

Most of these studies, however, have not been replicated, and the main reasons are the polygenic nature of hypertension involving possibly multiple pleiotropic variants of low penetrance, epistasis, ethnic diversity of human populations, phenotypic heterogeneity and the inability to control environmental factors. The multiple linkage loci identified especially in human chromosomes 1,2,3,17 and 18 with overlapping confidence intervals indicate a low likelihood of there being a single genomic region with a large effect on predisposition to hypertension. In addition, linkage analysis has poor power for detecting common alleles that have low penetrance.

The candidate gene association study approach involves the selection of candidate genes based on a mechanistic understanding of the roles of the encoded proteins in BP regulation, or the location within a previously determined region of linkage. However, this will at best identify only a fraction of genetic risk factors even for diseases in which the pathophysiology is relatively well understood. Candidate genes for hypertension fall into five broad categories: the RAAS, adrenergic system, metabolism related genes and novel genetic pathways including loci that encode growth factors, oxidative stress and inflammatory re-

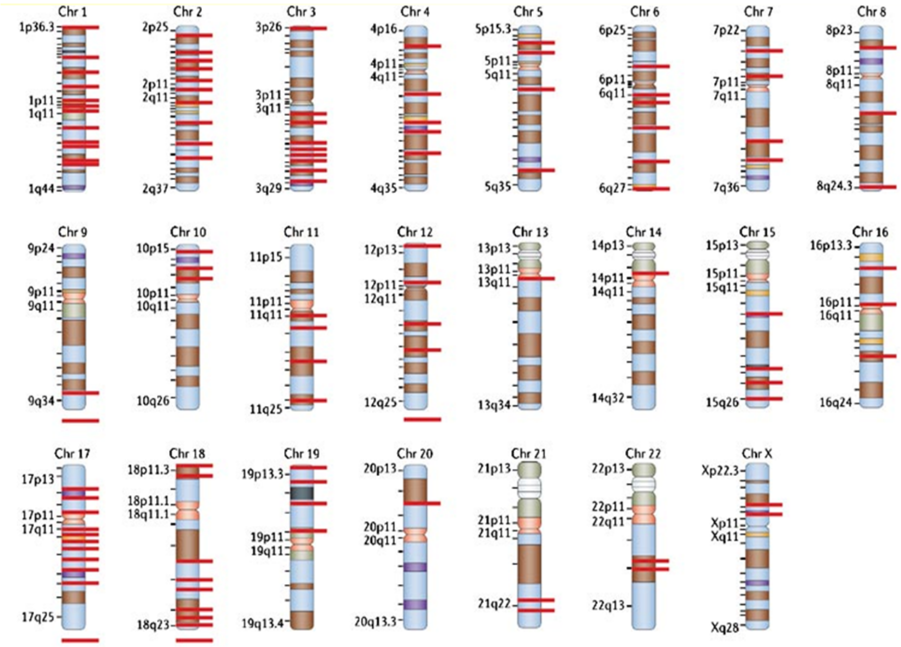

Figure 1.7: Hypertension related loci across the genome[6].

sponse [73]. Figure 1.8a shows the complex multifactorial interplay of genetic and environmental factors in the causation of essential hypertension. Monogenic forms of hypertension are completely genetically determined and are a rare cause of hypertension in the general population . None of the candidate gene studies have so far shown reproducible associations with hypertension and we also had a negative experience [74]. In collaboration with the BHF Glasgow Cardiovascular Research centre, we performed a gene-centric experiment with dense tag SNP coverage including the flanking 10 kb regions of major cardiovascular genes in the large accurately phenotyped Italian population Pressioni Arteriose Monitorate E Loro Associazioni (PAMELA) [75] with clinic, home and ambulatory BP measures in all the participants. This design enabled us to compare association signals in the same well powered sample for different BP measurement methods capturing all the information in the selected genes. However, not only among the SNPs analysed none reached experiment-wide significance threshold for replication, but also we found no consistency in SNPs association results between ambulatory, clinic and home BP readings. Moreover, evidence of significant sexual dimorphism emerged.

The failure of candidate gene association studies to identify the genetic basis of the common forms of hypertension suggests that there are several limitations to this approach. First, the choice of candidate genes may be inappropriate. Second, the causative genes might be either upstream of the points of action or in the downstream signalling pathways of the selected candidates. Third, the SNPs selected for association studies may provide incomplete coverage of all the variants within the genes studied. Fourth, most studies are underpowered and problems arise due to population stratification, phenotypic and locus heterogeneity. Finally, candidate gene studies rely on prior hypotheses about disease mechanisms, so that discovery of genetic variants in previously unknown pathways is precluded.

1.4.3 The case for hypertension genome-wide association study

There may be uncertainty about the extent of oligogenic versus polygenic influences on hypertension, but few doubt that non-genetic factors play a major role. There is evidence from modelling of complex diseases that neutral susceptibility alleles contribute little and that alleles under weak selection may constitute the bulk of the genetic variance of the underlying disease. This would indicate that hypertension may be attributable to loci in which susceptibility mutations are mildly deleterious and in which the overall mutation rate and hence allelic heterogeneity are relatively high [76]. As previously described, a large number of genes are involved in the regulation of BP, resulting in hundreds of genes that interact in different combinations in different individuals to influence hypertension. This high locus heterogeneity, in which more than one locus contributes to disease, amplifies the difficulty posed by allelic heterogeneity for successful hypertension gene mapping. The most comprehensive analysis of candidate genes (assuming that both common as well as rare alleles will contribute to hypertension) is obtained by resequencing the entire candidate gene regions in patients and controls, and searching for a variant or set of variants that is enriched or depleted in hypertensive patients. But these studies are laborious and expensive and there are challenges in properly interpreting the results particularly when considering rare non-coding variants. The recent surge in GWA studies for complex diseases with replicated results of SNP association would suggest that pursuing GWA approach for common variants is a valid strategy. Because no assumptions are made about the genomic location of the causal variants,

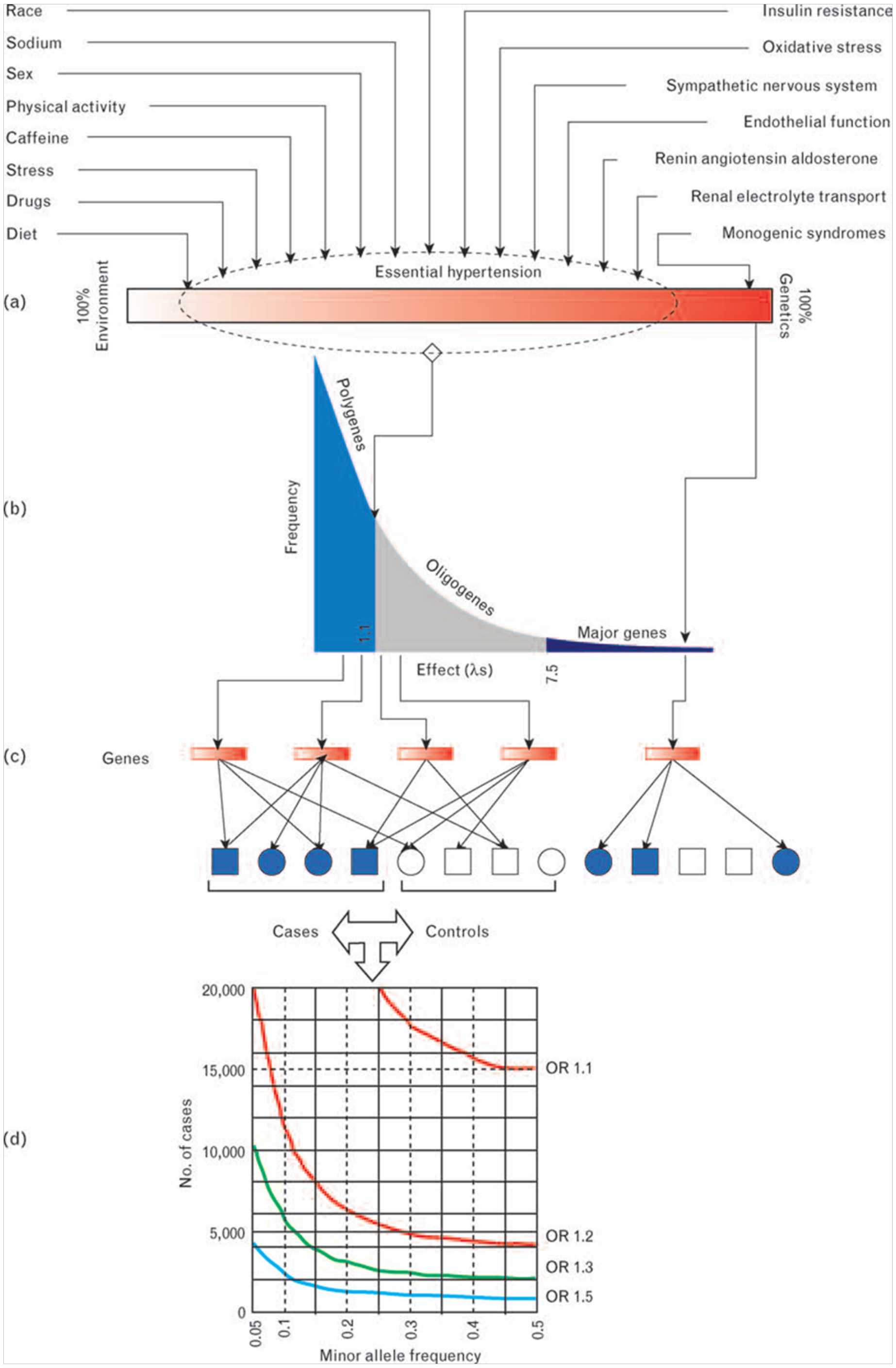

Figure 1.8: Dissecting the polygenic causation of hypertension using the genome-wide association approach.

this approach could exploit the strengths of association studies without having to guess the identity of the causal genes. The crucial factors that have made GWA studies possible are the availability of high throughput technologies, the decreasing cost of genotyping and the determination of linkage disequilibrium on a genome-wide scale through the HapMap project [77]. Such studies typically measure sets of special DNA tag SNPs selected from the catalogue of common human genetic variations provided by the HapMap project [16], enriched with non-synonymous SNPs, as well as SNPs in evolutionarily conserved regions of the genome. Most of these studies are case-control studies, in which SNP frequencies are compared between the two groups, and those that differ significantly are then validated in independent samples.

1.4.4 Designing a genome-wide association study for hypertension

There are two main approaches to a GWA study. The direct or sequence-based study design tests variants with known biological function located within all the genes, whereas the indirect or map-based approach takes advantage of linkage disequilibrium to genotype a set of tag SNPs as proxies for the entire set of SNPs. The International HapMap Project [77] has information on more than 3.8 million validated SNPs along with linkage disequilibrium between them, whereas gene-based SNP discovery projects like Seattle SNPs Program for Genomic Applications (http://pga.gs.washington.edu/) have identified SNPs in genes that can be used for gene-centric approaches to GWA studies. For indirect GWA, one can use quasi-random or anonymous SNPs that are spread across the genome (Affymetrix, Santa Clara, California, USA), or sets of linkage disequilibrium-based tag SNPs that are specifically chosen to saturate the genome (Illumina, San Diego, California, USA). For gene-centric GWA, one can use customised candidate gene chips that provide comprehensive coverage of genes by including both quasi-random or linkage disequilibrium tagging sets of SNPs and those from resequencing data. Based on the results of recent GWA studies, both the gene-centric and indirect map-based approaches can be successful.

Coverage is a measure of how well the SNPs that are part of a genotyping set capture all known variants and is an important factor not only in the ability to capture all causal variants, but also in terms of power of the study. The maximum r^2 (squared correlation coefficient for each SNP) can be used to translate coverage to the sample size that is required for an indirect association study. For a particular variant, the effective sample size of an indirect association study is simply the product of the actual sample size and the maximum r^2 value of the variant. So, as coverage decreases, a larger sample size will be needed to obtain the same power. In turn, the overall power of a GWA study can be estimated using the effective sample size for each variant [9, 78]. A gene-centric approach requires much less genotyping than an indirect approach and has increased power as a more liberal significance level can be used reflecting the smaller number of tests. Additionally, it has been shown that, when half of all the causal variants occur outside the genes, the gene-centric SNP approach is more efficient than whole-genome approaches [78], but, in general, it provides little coverage of SNPs that lie outside genes. Indirect GWA studies on the other hand can have higher overall power than gene-centric SNP studies, but have lower power and coverage for genic SNPs than non-genic SNPs. For a hypertension GWA, it would be reasonable to go for a combined approach by enriching the indirect approach with genic SNPs so as to maximise coverage of regions with a high prior probability of functional importance. A gene-centric approach would be useful when dealing with less frequent phenotypes like drug response, testing gene–gene interactions and for identification of cis enhancers in evolutionarily conserved regions [18, 9, 79].

As part of the Wellcome Trust Case Control Consortium (WTCCC) [7], a GWA study for hypertension was performed using 2,000 patients and 3,000 common population controls using the Affymetrix 500 K panel. This analysis revealed no SNPs with significance below 5 x 10^{-7}, the genome-wide significance level of this experiment. However, the number and distribution of association signals in the range 10^{-4}-10^{-7} was similar to other diseases studied. The study also did not detect any genes previously implicated by candidate gene association studies. Various reasons have been proposed for this, including poor coverage of many genes by the Affymetrix chip used in this study. Importantly, the use of common controls in this experiment may have caused significant attrition in power for hypertension (see Section1.4.5). Furthermore, it is highly plausible that hypertension may have fewer common risk alleles of larger effect sizes than other complex diseases, and genotype relative risks of less than 1.5 can only be detected by studying larger cohorts of more than 6,000 patients and controls (Figure 1.8d). Indeed, the thirteen loci very recently identified as responsible for BP control come from by two very large meta-analyses consortia [4, 80].

1.4.5 Phenotyping of subjects

Most of the successful replicated association study results are in diseases in which phenotyping specificity is increased with maximum attention to subject characterisation and selection eliminating much of the background noise from non-genetic factors. This reduces genetic heterogeneity and thus increases the chance of success. The Medical Research Council funded British Genetics of Hypertension (MRC BRIGHT) study [69] attempted to increase detectance (the probability of carrying any particular susceptibility genotype, given that the individual has a particular disease or trait phenotype) of the disease locus by ascertaining for disease severity and including only non-diabetic, non-obese hypertensive patients whose BP was in the top 5% of the BP distribution in the UK. Another method of reducing genetic and environmental heterogeneity is to use intermediate phenotypes [81] such as endothelial function or anti-hypertensive drug response [70] that may indicate shared underlying genetic mechanisms. Of note, patients and controls should undergo the same phenotyping protocol to establish both presence of disease in patients and also absence of disease in controls. The value of phenotyping controls is all the more evident in the case of negative results from the WTCCC GWA study of hypertension. It was estimated that a 5% misclassification of the common controls in this study would result in a loss of power equivalent to sample size reduction of 10%. However, the high prevalence of hypertension might have led to a much higher misclassification bias [7].

1.4.6 Analysis issues in genome-wide association studies

Standard association analysis considers each genetic marker individually and this can neglect information on their joint distribution. Haplotypes not only allow multiple potentially causal variants to be tested simultaneously for association, but also they may be a proxy for untyped causal markers. However, for haplotypes to be superior to the individual markers, at least at an initial screen, multiple functional markers must have a strong interaction when in cis (i.e. on the same chromosome) and yet have no detectable effect when considered individually. If multiple markers have detectable effects on their own, then it is appropriate to test haplotypes for yet stronger effects. The penalty for multiple tests required for adding haplotype-based tests is likely to be less severe when the use of haplotypes is restricted to blocks of strong linkage disequilibrium with low haplotype diversity [79, 82, 83, 84]. It is pertinent to point out here that the widespread adoption of tagging strategies diminishes the utility of haplotype analyses. Gene–gene and gene–environment interactions are important factors in hypertension causation. However, there are important power and multiple testing factors to be considered. A 500 K SNP study would result in about 10^{11} and 10^{16} two and three SNP combinations respectively, and various strategies to improve power and reduce the cost of computation and multiple testing have been published [85, 86, 87]. But there is no simple method, and the advantages and shortcomings of the analytic strategies should be considered in advance of developing the experimental design of any study of interaction effects.

The other major statistical problem to be overcome is identifying a true positive signal from a sea of false positives. Figure 1.8b shows that there is an inverse relationship between the allele frequency of a contributory locus and its phenotypic effect. Figure 1.8b is modified from [5]. Alleles with small, intermediate and large effects are polygenes, oligogenes and major genes, respectively. The sibling recurrent risk of hypertension is less than 1.5, indicating a phenotype with modest genetic effect. Modified from [23]. The common disease common variant hypothesis (Figure 1.8c) in hypertension holds that the genetic variants occur with a relatively high frequency ($>1\%$). Locus heterogeneity is a major problem in hypertension gene mapping. Mendelian or monogenic hypertension mutations are highly penetrant, and usually under very strong selection, which keeps them at low frequencies. Significant thresholds of p-value less than 10^{-6} have been proposed for GWA owing to the need to allow for the very small prior probability that any given locus or region is truly associated with disease. In the recent GWA studies, SNPs reaching a threshold of 5 x 10^{-7} have been successfully replicated [7, 9, 79, 84]. The positive associations should help future studies, as researchers can better understand statistical profiles of genuinely associated disease alleles, and this will enable the identification of loci of smaller effect that may contribute independently to disease. However, the gold standard for any genotype–phenotype association is replication in independent well-powered samples, and guidelines on this have recently been published [88]. Although replications are usually attempted in populations with the same ancestry as the initial study, extending this to populations of different ancestry will increase the confidence of the findings while failure to replicate will not invalidate the initial finding, but may indicate a population-specific risk. Owing to their robustness to population stratification, family-based studies can also serve as valuable replication studies. Several rounds of replication may be needed to establish a valid genotype–phenotype association [88].

Most of the current GWA studies are probably underpowered to detect odds ratios less than 1.5 and the requirement for larger sample sizes can be addressed efficiently by combining information from the different studies performed on the same disease. Figure 1.8d shows the needed sample size for a case–control GWAS using 500,000 SNPs with 80% power for various odds ratios assuming an equal number of patients and controls; prevalence=30%; p=5×10^{-7}. It would be worthwhile considering a priori meta-analyses when designing a GWA for hypertension, as the potential effect sizes are going to less than 1.3. Differences in genotyping platforms between studies, population stratification, phenotype misclassification and between-study heterogeneity may affect the gain in power offered by meta-analysis [89, 90]. A typical example is the association of fat mass-associated gene and obesity-associated gene (FTO) with diabetes through its effect on the correlated phenotype of obesity [7, 89]. The current trend to make publicly available GWA datasets can offset the problem of publication bias.

1.5 Genome-wide association study of blood pressure extremes

In this last section we outline a personal GWAS experience which takes place in the context of the Ingenious HyperCare Network, in collaboration with the BHF Glasgow Cardiovascular Research Centre and the Istituto Auxologico Italiano [15].

As described in Section 1.4, statistical power to detect a phenotype-genotype association is dependent upon the magnitude of effect, the frequency of causal alleles and the sample size. The hunt for genes responsible for BP control has so far identified thirteen loci from two large meta-analyses consortia, with each association explaining only a very small proportion of the total variation in systolic or diastolic blood pressure (SBP or DBP; roughly 0.05–0.10%, approximately 1 mmHg per allele SBP or 0.5 mmHg per allele DBP) [4, 80]. This suggests the existence of more undiscovered blood pressure related common variants. Cross-sectional studies of the general population have required extremely large sample sizes to detect such small effect sizes [7]. We explored an alternative strategy to increase power, using cases and controls drawn from the extremes of the BP distribution, and detected a novel locus associated with hypertension. We then validated this association using large-scale population and case-control studies,where similar extreme criteria for selection of cases and controls have been used. As the locus was related to uromodulin, a protein exclusively expressed intra-renally, we tested for dependency of the association on renal function (eGFR) and urinary excretion of uromodulin. Next, we tested the hypothesis that the locus affects sodium homoeostasis by studying the homoeostatic response to altered sodium intake. Finally, we tested associations with cardiovascular outcomes.

1.5.1 Methods

1.5.1.1 Study design for the discovery cohort

To identify novel susceptibility loci for hypertension, we used an extreme ends case-control design. Hypertensive cases had to have at least two consecutive BP measurements of $\geq$ 160 mmHg systolic and $\geq$ 100 mmHg diastolic, with the diagnosis made before age 63 years. We identified 2,000 cases in the Nordic Diltiazem study (NORDIL) [91]. These hypertensive subjects represent approximately the top 2% of the BP distribution in the Swedish population. 2,000 control subjects were drawn from the Malmö Diet and Cancer study (MDC) [92] who had a SBP $\leq$ 120 mmHg and DBP $\leq$ 80 mmHg. Controls had to be at least 50 years of age and free from cardiovascular events (coronary events and stroke) during 10 years of follow up [93] and not on any anti-hypertensive medication. Of the MDC population (n=27,000), 9.2% met these criteria and thus selected subjects were hyper-controls with low cardiovascular risk. In both NORDIL and MDC, BP was measured in the recumbent position after 5-10 minutes rest using a manual sphygmomanometer. Rigorously phenotyped samples minimise case/control misclassification, and the potential advantage of an extreme case/control design is greater power to detect variants associated with BP and hypertension, for a given total sample size and total genotyping cost.

1.5.1.2 Validation cohorts

For the validation we used phenotypic definitions (extreme SBP/DBP thresholds) to closely match our discovery samples. The BP criteria were slightly modified as most validation cohorts were general population cohorts. We identified as cases individuals less than 60 years of age with SBP $\geq$ 140 mmHg or DBP $\geq$ 90 mmHg or on current treatment with anti-hypertensive or on BP lowering medication commenced before age 60 years. We called controls those with SBP $\leq$ 120 mmHg and DBP $\leq$ 80 mmHg, at least 50 years of age, and free from any BP lowering medication. If age $\leq$ 50 years, then the criteria were slightly modified to SBP $\leq$ 115 mmHg and DBP $\leq$ 80 mmHg and free from BP lowering medications. The validation cohorts were the MONItoring trends and determinants of CArdiovascular diseases (MONICA) and the PAMELA study (894 cases/746 controls) from Northern Italy [94, 95], 1956 cases/1057 controls from the 2002-2006 follow-up exam of the Malmö Preventive Project (MPP) [96] and 6977 cases/6891 controls from the Malmö Diet and Cancer study [97] (MDC; non-overlapping with discovery samples), 509 cases/209 controls from The Netherlands Study of Depression and Anxiety study (NESDA) [98] and ten cohorts from a collaboration with the Global BPgen consortium [4]. Analyses reported here are distinct from those previously published [4], because they use phenotypic definitions to match our discovery samples. The combined sample size of the discovery and validation cohorts is 39,706 individuals (21,466 cases and 18,240 controls).

Estimated glomerular filtration rate (eGFR) was calculated using the Modification of Diet in Renal Disease (MDRD) Study equation [99].

1.5.1.3 Clinical functional studies

We studied functional associations of the top SNP in a hypertensive cohort and a population cohort with extensive urine phenotypes and one interventional study of low and high salt intake with extensive measurements of sodium balance.

- The British Genetics of Hypertension (BRIGHT) study [69] is a hypertension case-control study. Case inclusion criterion was a diagnosis of hypertension ($>$ 150/100 mmHg) prior to 50 years of age. Exclusion criteria included BMI $>$ 35, diabetes, secondary hypertension or co-existing illness. 24-hour urine collection was available for all the cases with measurements of urinary sodium, potassium, creatinine and microalbuminuria. We measured urinary uromodulin in 256 hypertensive subjects.

- Groningen Renal Hemodynamic Cohort Study Group (GRECO): The GRECO protocol comprises integrated measurement of renal hemodynamics and extracellular volume as applied in an ongoing series of studies in healthy subjects [100, 101]. For the current analysis 64 healthy adult males were included (mean age = 23 years), who had been studied after two seven-day periods: the first 7 days on a low sodium diet (LS, 50 mmol

Na+ per day, balance verified by repeated 24h urine), the second 7 days on a high-sodium diet (HS, 200 mmol Na+ per day).

- Hypertension Evaluation by Remler and CalciUria LEvel Study (HERCULES) is a substudy of the population-based CoLaus study from Lausanne Switzerland [102, 103]. A random sample of 411 CoLaus participants, aged 38-78 years, underwent ambulatory BP monitoring and 24 hour urine collection. The phenotypes available include 24-hour urine collection with measurement of creatinine clearance, endogenous lithium clearance, urinary sodium, potassium and uric acid excretion and microalbuminuria. We measured urinary uromodulin in 110 participants of this study.

1.5.1.4 Urinary uromodulin measurements

Urinary uromodulin was measured in duplicate in 24 hour urine samples using a commercially available ELISA (MD Biosciences, Zürich, Switzerland) as recommended by the manufacturer. The range of assay is 9.375 - 150 ng/mL and sensitivity <5.50 ng/mL. The inter-assay coefficient of variation was 11.9%. Urinary uromodulin levels were corrected for urine creatinine before analysis.

1.5.1.5 Genotyping and quality control

The GWAS samples were genotyped using Illumina 550K Single and Illumina 610 Quad V1 BeadChip (Illumina, Inc., San Diego, CA, USA). We included 551,629 SNPs common to both the Single and Quad chips, for analysis. SNPs with a minor allele frequency (MAF) <1% or in significant Hardy-Weinberg disequilibrium (P $<1x10^{-7}$) in pooled samples were removed leaving 521,220 SNPs for analysis. We assessed population structure within the data using principal components analysis as implemented in EIGENSTRAT [25] to infer continuous axes of genetic variation. After data quality control for unspecified sex (5 subjects removed), relatedness/duplicates (68 individuals removed), multidimensional scaling plot outliers (33 individuals removed), genetic outliers (i.e. individuals whose ancestry is at least 6 s.d. from the mean on one of the top ten axes of variation on principal component analysis- 388 individuals removed) and genotyping success of <95% (92 individuals removed), genotype information from 1,621 cases and 1,699 controls (1,510 males and 1,810 females) was available for analysis. Untyped SNPs were imputed using IMPUTE v1 24 with data from the August 2009 release of CEU phased haplotypes from Pilot 1 of the 1000 Genomes Project NCBI Build 36 (dbSNP b126) as the reference panel (from https://mathgen.stats.ox.ac.uk/impute/impute_v1.html). The probability threshold used for calling an imputed genotype was 0.9. Association analysis was performed using Plink [104] and SNPTEST [105] taking into account uncertainty in imputation.

1.5.1.6 Statistical analysis

In the GWAS samples, we tested each SNP for association using an additive genetic model. The model implies that a given allele at a given locus adds a constant to, or subtracts a constant from, the expected value of the trait. The amount added or subtracted varies in an unknown way from allele to allele and from locus to locus. The additive model will usually capture much of the etiological information tat can reasonably be explained by genes [31]) and logistic regression with adjustment for significant ancestry principal components [25] to correct for population stratification. There was still a slight overall inflation of test statistics, with a genomic control inflation factor (λ) of 1.07. All results are presented after application of genomic control to correct for this residual inflation [23]. Additionally two logistic regression analyses were performed, with adjustment for age, age^2, sex and BMI and with adjustment for age, age^2, sex, BMI and eGFR. Multiple linear regression was used to test association between genotype and urinary uromodulin levels, functional parameters like GFR, extracellular volume etc. with relevant covariates. In the GRECO study, as the numbers of GG genotypes were small, AG and GG were combined for analysis. Non-normally distributed traits were tested using the non-parametric Kruskal Wallis test.

1.5.1.7 Validation analysis

In validation samples, SNPs were tested for association using logistic regression, with adjustment for ancestry principal components where available to correct for population stratification. Meta-analysis of the combined discovery and validation results was conducted using an inverse-variance weighted (fixed-effects) meta-analysis.

A genome-wide significance threshold of 5x10^{-8} corresponding to a P value of 0.05 with a Bonferroni correction for 1 million independent tests was considered a priori as genome-wide significant [106].

1.5.1.8 Continuous blood pressure trait modelling

The associations between the validated SNP and SBP and DBP were analysed separately in the Stage 1 samples of the Global BPgen consortium (n=34,433) and in the overall MDC (n=27,000) and MPP (n=17,700) cohorts [4, 92, 96]. The results were combined using fixed-effect inverse variance weighted meta-analysis. Continuous SBP and DBP were adjusted for age, age^2, body mass index and any study-specific geographic covariates in sex-specific linear regression models. In individuals taking anti-hypertensive therapies, blood pressure was imputed by adding 15 mm Hg and 10 mm Hg for SBP and DBP, respectively [4, 107].

1.5.2 Results

1.5.2.1 Genome-wide association, replication and meta-analysis

The demographic characteristics of the discovery sample are presented in Table 1.1.

	Controls (n=1699)	Cases (n=1621)
Age at enrolment,years	57.4 (5.9)	55.4 (7.1)
BMI, kg/m^2	24.2 (3.5)	27.1 (7.8)
SBP, mmHg	115.8 (6.8)	175.8 (22.5)
DBP, mmHg	73.7 (5.7)	104.7 (11.8)

Table 1.1: Demographic characteristics of the discovery case-control population.

The results of the GWAS in the discovery sample are presented in Figure 1.9. The observed versus expected p-value distributions (quantile-quantile plots) are shown in Figure 1.10.

The top hit was rs13333226 (p=1.14x10^{-7}) . We selected the top 89 SNPs for validation analysis (corresponding to p-value≤5.6x10^{-4}) in the MONICA and PAMELA samples. Three SNPs crossed a p-value threshold of 5x10^{-7} in the combined analysis - rs13333226 (p=3.86x10^{-7}), rs4293393 (p=3.30x10^{-7}), rs13353058 (p=4.78x10^{-7}). The two SNPs, rs13333226 and rs4293393 are highly correlated with an r^2 = 0.996. We selected rs13333226, which is in close proximity to the uromodulin transcription start site at -1617 base pairs (Figure 1.11) for further analyses.

We found it to be associated with hypertension across the combined discovery and validation samples (Table 1.2, Figure 1.12A&B), with the minor G

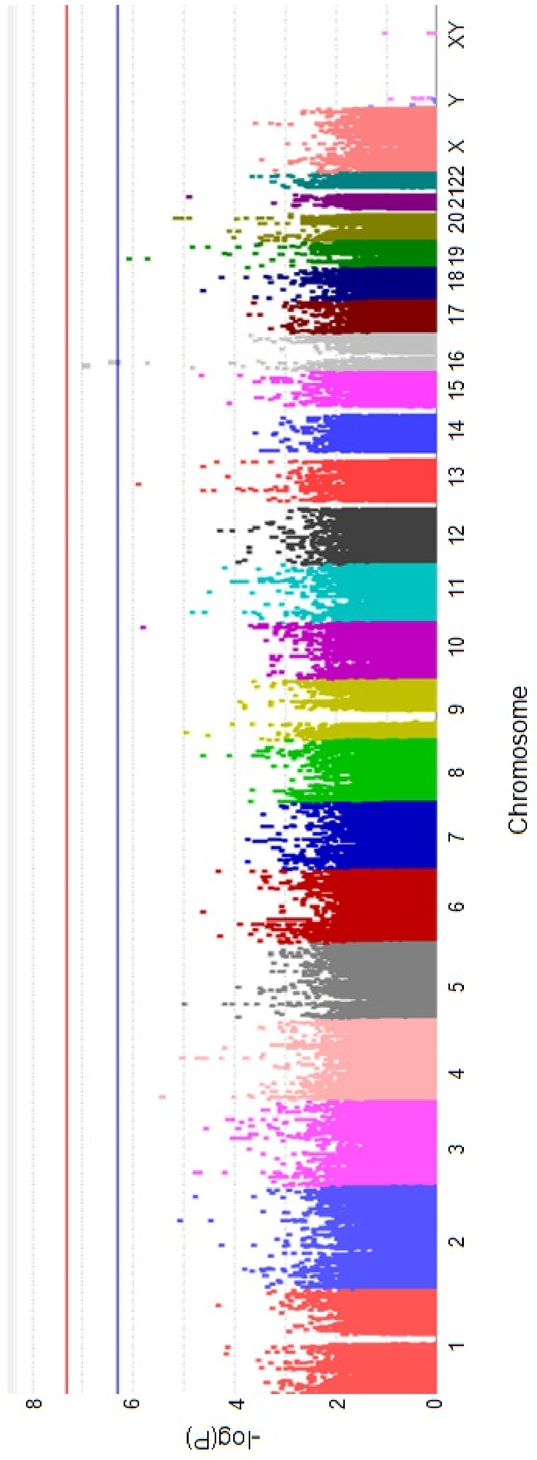

Figure 1.9: Overview of association results in the discovery sample. Association signal for each SNP is plotted as -log(p value) against genomic position. Red line indicates p=5x10^{-8} and blue line indicates p=5x10^{-7}.

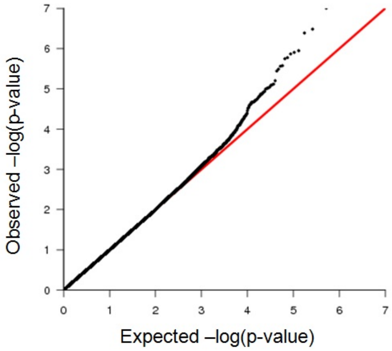

Figure 1.10: Quantile-Quantile plot of observed versus expected p-value distributions in the discovery sample.

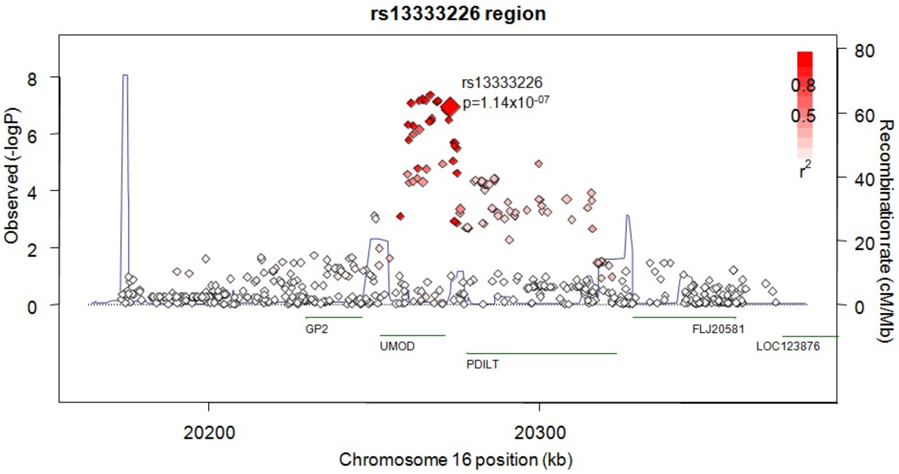

Figure 1.11: Association plot of the genomic region around rs13333226 showing both typed and imputed SNPs with location of genes and recombination rate.

allele associated with a lower risk of hypertension (OR [95%CI] = 0.87 [0.84; 0.91], p=3.6x10^{-11}).

The strong evidence of association remained when the discovery sample was excluded from the combined analysis (OR [95%CI] = 0.89 [0.86-0.93], p=7.36x10^{-8}). The association signal strengthened after adjustment for age, age^2, sex and BMI (OR [95%CI] = 0.85, p =1.5x10^{-13}; after excluding the discovery sample: OR [95%CI] = 0.86[0.83-0.90], p =1.61x10^{-10}). Heterogeneity across the study samples was assessed using the Q statistic (all samples p = 0.036, after excluding discovery samples p = 0.514). In the 13,446 individuals with eGFR measurements available, the strength of association of rs13333226 with hypertension was identical after correcting for eGFR and the effect sizes remained unchanged (unadjusted for eGFR: OR [95%CI] = 0.90[0.83;0.96], p = 0.004; after eGFR adjustment: OR [95%CI]=0.89[0.83;0.96], p=0.0030); (Table 1.3, Figure 1.12 C&D).

Association with SBP and DBP We find rs13333226 to be significantly associated with lower SBP (0.49 mmHg lower per copy of G allele, p = 2.6x10^{-5}) and DBP (0.30 mmHg lower per copy of G allele, p = 1.5x10^{-5}) on combined analysis of Global BPgen, MPP and MDC cohorts (n = 79,133).

1.5.2.2 Clinical functional studies

We studied the association between rs13333226 genotypes and different phenotypes including urinary uromodulin, in 256 hypertensive individuals from the BRIGHT cohort. The average age was 63 years and univariate analysis showed the G allele was significantly associated with higher eGFR (4.6 ml/min/1.73m^2 per copy of G allele; p = 0.005) and decreased uromodulin levels corrected for urine creatinine (0.2mg/mmol lower per G allele; p = 0.007) (Table 1.4).

The association with urinary uromodulin levels persisted after adjusting for sex, urine sodium and eGFR on multiple regression analysis (r^2 = 0.14, p<0.001).

Urinary uromodulin was also measured in 110 participants from the HERCULES study. Univariate analysis showed a significant association between rs13333226 and urinary excretion of uromodulin but not with creatinine clearance (Table 1.5). In the HERCULES Study, hypertension is defined based on 24 hours ambulatory blood pressure >135/85 or on anti-hypertensive treatment.

On multiple regression analyses, uromodulin remained significantly associated with rs13333226, independently of creatinine clearance. Urinary uro-

					Unadjusted Analysis		Adjusted for age, age^2, sex, BMI	
	Origin	Cases	Controls	MAF	OR [95%CI]	p	OR [95%CI]	p
Discovery sample	Swedish	1621	1699	0.17	0.65 [0.56-0.76]	1.10x10^{-7}	0.60 [0.50-0.73]	3.3x10^{-7}
BRIGHT/ASCOT	British/Irish	3069	1787	0.18	0.94 [0.84-1.04]	0.229	0.90 [0.80-1.02]	0.103
MPP	Swedish	1956	1057	0.18	0.91 [0.78-1.05]	0.193	0.91 [0.78-1.05]	0.186
MDC	Swedish	6977	6891	0.18	0.86 [0.80-0.92]	0.001	0.86 [0.80-0.92]	3.0x10^{-5}
PREVEND	Dutch	2411	1613	0.18	0.90 [0.80-1.02]	0.091	0.89 [0.77-1.03]	0.113
CoLaus	Swiss	1300	1375	0.19	0.97 [0.84-1.11]	0.634	0.93 [0.79-1.10]	0.375
KORA	German	457	300	0.16	0.80 [0.61-1.06]	0.128	0.70 [0.51-0.97]	0.030
SHIP	German	656	240	0.18	1.07 [0.81-1.41]	0.627	0.74 [0.50-1.10]	0.137
58BC	British	514	529	0.19	0.82 [0.66-1.02]	0.077	0.77 [0.61-0.97]	0.026
TwinsUK	British	245	845	0.19	0.88 [0.68-1.14]	0.332	0.84 [0.63-1.12]	0.236
MIGen	European Ancestry	316	278	0.21	0.68 [0.51-0.90]	0.004	0.61 [0.44-0.84]	0.002
DGI	Swedish/Finnish	277	161	0.23	1.11 [0.77-1.62]	0.572	1.15 [0.78-1.68]	0.483
Fenland	British	264	510	0.19	0.91 [0.69-1.19]	0.478	0.80 [0.58-1.09]	0.158
MONICA/PAMELA	Italian	894	746	0.19	0.91 [0.76-1.08]	0.282	0.87 [0.72-1.05]	0.145
NESDA	Dutch	509	209	0.18	0.98 [0.73-1.31]	0.898	0.93 [0.63-1.35]	0.689
Combined Analysis		21466	18240	0.187(*mean*)	0.87 [0.84-0.91]	3.60x10^{-11}	0.85 [0.81-0.89]	1.50x10^{-13}
Combined Analysis-excluding discovery		19845	16541	0.187 (*mean*)	0.89 [0.86-0.93]	7.36x10^{-8}	0.86 [0.83-0.90]	1.61x10^{-10}

Table 1.2: Results from the meta-analysis of rs13333226 and hypertension in discovery sample and after validation.

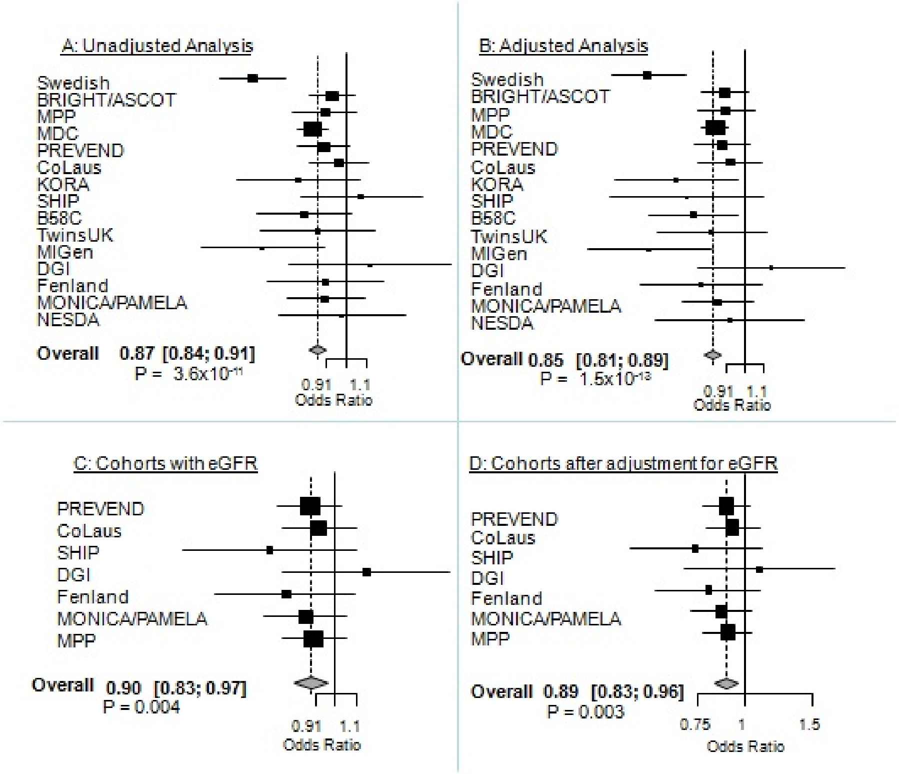

Figure 1.12: Forest Plots of association with rs13333226 and hypertension. **A**: Forest plot of association analysis unadjusted for any covariates - 21,466 cases and 18,240 controls. **B**: Forest plot of association analysis adjusted for age, age^2, sex and BMI - 21,466 cases and 18,240 controls. **C**: Forest plot of association analysis in the cohorts where eGFR was available and age, age^2, sex and BMI - 7427 controls and 5739 cases. **D**: Forest plot of association analysis in the cohorts where eGFR was available and age, age^2, sex, BMI and eGFR - 7427 controls and 5739 cases.

					Adjusted for age, age^2, sex, BMI		Adjusted for age, age^2, sex, BMI; eGFR	
	Controls	Cases	eGFR mean	eGFR SD	OR [95% CI]	p	OR [95% CI]	p
PREVEND	2404	1606	80.36	14.39	0.90 [0.77;1.03]	0.1130	0.89 [0.77;1.03]	0.1350
CoLaus	1375	1298	83.28	16.35	0.93 [0.79;1.10]	0.3747	0.93 [0.79;1.10]	0.3769
SHIP	240	656	87.62	19.78	0.74 [0.50;1.11]	0.1374	0.74 [0.50;1.10]	0.1437
DGI	120	141	72.69	11.67	1.09 [0.69;1.72]	0.4827	1.09 [0.78;1.68]	0.6988
Fenland	508	262	98.92	52.96	0.80 [0.58;1.10]	0.1580	0.80 [0.58;1.09]	0.1739
MONICA/PAMELA	824	719	84.30	16.59	0.87 [0.72;1.05]	0.1445	0.90 [0.74;1.09]	0.2777
MPP	1956	1057	88.20	15.10	0.91 [0.78;1.05]	0.1860	0.90 [0.78;1.05]	0.1790
Combined analysis	7427	5739			0.90 [0.83;0.97]	0.0036	0.89 [0.83;0.96]	0.0030

Table 1.3: Results from the meta-analysis of rs13333226 and hypertension before and after adjustment for eGFR.

	AA (n=141)	AG (n=93)	GG (n=22)	p
Male:Female	0.7	0.8	0.6	0.763
Age (years)	64.7(8.4)	63.9(7.8)	59.5(9.5)	0.036
SBP (mm HG)	156.0(19.5)	151.5(18.9)	153.3(14.5)	0.205
DBP (mm HG)	93.1(10.0)	90.9(10.7)	93.3(10.3)	0.2666
BMI (Kg/m^2)	26.8(4.6)	26.8(5.4)	27.2(3.9)	0.927
BSA (m^2)	1.8(0.2)	1.9(0.2)	1.8(0.2)	0.494
Sodium (mmol/L)	138.6(3.1)	138.9(3.0)	137.8(2.9)	0.341
Potassium (mmol/L)	4.4(0.9)	4.2(0.8)	4.4(1.0)	0.429
Urea (μmol/L)	6.3(1.6)	5.7(1.6)	6.0(1.6)	0.025
Creatinine (μmol/L)	92.2(21.7)	88.4(18.7)	82.9(20.0)	0.096
Urate (μmol/L)	0.3(0.1)	0.3(0.1)	0.3(0.1)	0.726
eGFR(ml/min/1.73m^2)	67.6(16.2)	70.3(12.3)	79.5(15.2)	0.005
Creatinine Clearance (ml/min)	70.6(20.3)	76.2(20.0)	86.6(26.6)	0.004
Urine Sodium (mmol/24h)	139.1(61.9)	158.9(70.6)	142.4(58.3)	0.073
Urine Potassium (mmol/24h)	66.4(24.1)	78.8(54.0)	69.2(18.8)	0.050
Creatinine excretion (mmol/24h)	10.2(3.6)	10.8(4.6)	10.7(3.1)	0.520
Uromodulin (mg/L)	5.3(5.3)	5.2(5.5)	3.2(3.4)	0.234
Fractional Excretion Sodium (%)	0.92(0.37)	0.95(0.36)	0.73(0.19)	0.032

Table 1.4: Univariate association analysis of rs13333226 in 256 hypertensive patients from the BRIGHT study.

	AA (n=52)	AG (n=46)	GG (n=12)	p
Male:Female	1.2	0.6	1.4	0.258
Age (years)	58(49-67)	56(49-66)	59(49-66)	0.889
24h SBP (mm HG)	115.4(107.7-123.0)	113.2(105.9-124.8)	118.4(111.4-130.7)	0.555
24h DBP (mm HG)	76.3(69.8-81.1)	77.1(71.5-85.2)	77.7(71.1-87.7)	0.547
Hypertension (%)	33	30	25	0.846
BMI (Kg/m^2)	26.1(23.6-29.3)	24.4(21.8-29.0)	24.7(24.0-28.0)	0.175
BSA (m^2)	1.84(1.72-1.98)	1.76(1.62-1.92)	1.87(1.77-2.00)	0.045
Sodium (mmol/L)	139.7(138.1-141.8)	139.9(138.2-141.5)	140.3(138.4-142.7)	0.708
Potassium (mmol/L)	4.0(3.8-4.3)	4.0(3.7-4.1)	3.8(3.6-4.0)	0.041
Urea (μmol/L)	5.3(4.4-6.1)	4.8(4.4-6.0)	4.4(4.1-5.0)	0.141
Creatinine (μmol/L)	82.0(73.5-91.5)	81.0(73.0-88.0)	76.5(72.5-80.5)	0.225
Urate (μmol/L)	318.0(290.0-378.0)	321.0(262.0-373.0)	294.0(268.0-318.0)	0.163
Creatinine Clearance (ml/min)	98.9(70.5-123.0)	93.7(75.4-123.3)	109.4(86.2-125.1)	0.607
Urine Sodium (mmol/24h)	138.0(86.0-176.0)	134.0(92.0-175.0)	109.0(84.0-161.0)	0.785
Urine Potassium (mmol/24h)	59.0(37.0-84.0)	61.0(45.0-74.0)	48.0(39.0-80.0)	0.865
Creatinine Excretion (mmol/Kg/24h)	0.15(0.12-0.19)	0.16(0.13-0.19)	0.16(0.12-0.19)	0.745
Uromodulin (mg/L)	30.8(15.6-51.7)	24.5(14.2-42.5)	14.0(10.6-16.5)	0.005
Uromodulin(mg/24h)	53.0(25.0-76.0)	40.0(28.0-68.0)	17.0(14.0-33.0)	0.005
Urine volume(mL)	1725.0(1200.0-2375.0)	1665.0(1150.0-2100.0)	1773.0(1125.0-2300.0)	0.864
Fractional Excretion Sodium (%)	1.2(0.6-1.8)	1.2(0.8-1.7)	0.7(0.6-1.7)	0.696
Fractional Excretion Lithium (%)	0.10(0.07-0.17)	0.15(0.09-0.21)	0.11(0.06-0.15)	0.031

Table 1.5: Univariate association analysis of rs13333226 in 110 participants from the HERCULES Study. *Median(Interquartile range)*

modulin (mg/24h) was also positively associated with urinary sodium excretion (mmol/24h) ($p = 0.025$) and with endogenous lithium clearance ($p = 0.038$), independently of creatinine clearance. Urinary uromodulin was significantly and positively associated with fractional excretion of endogenous lithium in individuals ($r^2 = 0.19$, $p = 0.045$). This suggests that low urinary uromodulin is associated with higher proximal tubular sodium reabsorption.

Furthermore, in the GRECO study urinary uromodulin concentration ($p = 0.004$) and 24 hour uromodulin excretion ($p = 0.002$; Wilcoxon's signed ranks test) were found to be significantly increased after a high salt intake (Table 1.6).

Moreover, there was a significant decrease in uromodulin levels with each additional copy of the G allele on the LS diet, and this was not apparent on the HS diet. The homoeostatic adaptation to altered sodium intake was significantly different between the genotypes: the G allele was associated with a significantly greater increase in measured GFR ($p = 0.015$), extracellular fluid volume (ECV, $p = 0.047$) and a greater suppression of plasma renin activity (PRA, $p = 0.055$) during HS compared to LS diet (Figure 4). During LS fractional excretion of sodium (FENa) was similar across the genotypes. During HS overall FENa was higher, as appropriate. The rise in FENa elicited by HS was however less in subjects with one or two G-alleles than in AA subjects, resulting in significant differences in FENa during HS, with lower FENa in presence of one or two G-alleles (Figure 1.13).

1.5.2.3 Cardiovascular outcomes and rs13333226

Finally, we evaluated the clinical significance of our findings by testing whether the low BP associated allele may protect against development of cardiovascular events during long-term follow-up at the population level. Among 26,654 subjects from the entire population based MDC study 15 who were free from prior cardiovascular events at baseline, 2,750 individuals developed cardiovascular events (CVD) during 12 years of follow-up. We found each copy of the G allele to be associated with a 7.7% reduction in risk of CVD events after adjusting for age, sex, BMI and smoking status (H.R. = 0.923, 95% CI 0.860-0.991; $p = 0.027$). When SBP was added to the Cox regression model, the directionality and risk remained almost identical (H.R. = 0.936, 95% CI 0.872-1.005; $p = 0.067$).

	AA (n=40)	AG (n=19)	GG (n=5)	p
Age(years)	26.0(8.0)	23.0(6.0)	23.0(4.0)	0.105
SBP LS (mm Hg)	120.0(10.0)	122.0(9.0)	118.0(11.0)	0.670
DBP LS (mm Hg)	68.0(9.0)	70.0(7.0)	69.0(8.0)	0.453
SBP HS (mm Hg)	123.0(10.0)	124.0(10.0)	122.0(11.0)	0.805
DBP HS (mm Hg)	69.0(8.0)	70.0(7.0)	68.0(8.0)	0.661
Weight (Kg)	81.0(10.0)	81.0(10.0)	77.0(10.0)	0.715
BSA (m^2)	2.05(0.14)	2.03(0.15)	2.02(0.15)	0.590
FLNa LS (mmol/min)	18.0(2.5)	16.9(2.6)	16.3(2.6)	0.060
FNLa HS (mmol/min)	18.8(2.5)	18.8(3.2)	18.6(3.0)	0.562
GFR LS (ml/min/1.73 m^2)	109.0(13.0)	104.0(14.0)	102.0(15.0)	0.127
GFR HS (ml/min/1.73 m^2)	114.0(14.0)	115.0(15.0)	116.0(15.0)	0.719
ERPF LS (ml/min/1.73 m^2)	472.0(74.0)	457.0(70.0)	419.0(54.9)	0.209
ERPF HS (ml/min/1.73 m^2)	502.0(90.0)	490.0(61.0)	484.0(98.0)	0.529
Δ ERPF (ml/min/1.73 m^2)	30.0(51.0)	33.0(40.0)	65.0(74.0)	0.445
ECV LS (1/1.73 m^2)	16.5(1.9)	16.2(1.7)	16.7(1.3)	0.657
ECV HS (1/1.73 m^2)	17.2(1.7)	18.1(1.8)	17.6(2.1)	0.093
FENa LS (%)	0.19(0.18)	0.23(0.28)	0.17(0.16)	0.342
FENa HS (%)	0.99(0.35)	0.81(0.34)	0.84(0.22)	0.001
PRA LS (nmol/L/h)	6.3(3.7)	6.7(3.4)	6.4(2.3)	0.723
PRA HS (nmol/L/h)	2.5(1.5)	2.1(0.9)	1.8(1.1)	0.155
UMOD LS (median (IQR) mg/L)	10.3(6.9-15.6)	9.9(7.4-15.7)	7.0(2.5-13.3)	0.002
UMOD HS (median (IQR) mg/L)	11.9(7.5-27.9)	11.6(11.4-23.0)	12.2(6.9-23.6)	0.513

Table 1.6: Univariate association analysis of urinary uromodulin in relation to rs13333226 polymorphism and response to high and low salt intake (GRECO Study).

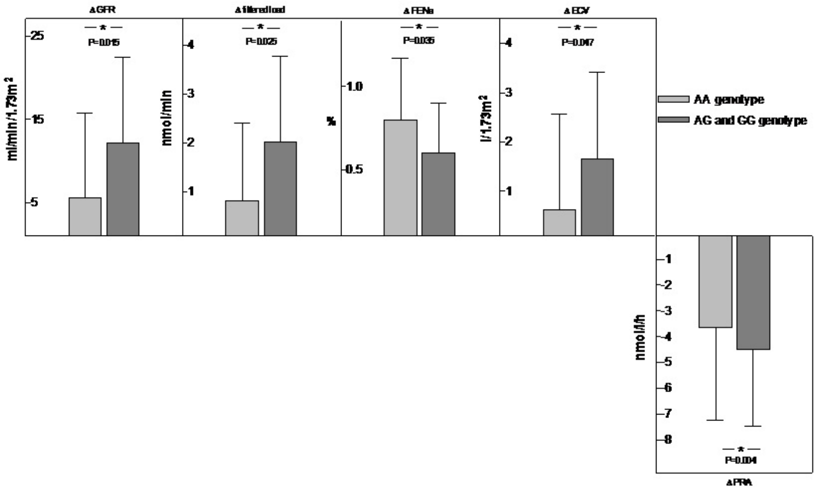

Figure 1.13: UMOD rs13333226 genotype and response to high salt intake. Filtered load: Filtered load sodium; ECV: Extracellular volume; FENa: Fractional excretion sodium; PRA: Plasma renin activity. The bars represent the difference in measurement between high salt and low salt intake for each measurement. In the GRECO subjects, we see that in response to a change in sodium intake the change in GFR is larger in G-allele carriers whereas the change in FENa is less. Hence, in G-allele carriers the restoration of sodium excretion towards the level that matches the altered intake is more dependent on the change in filtered load of sodium than in non-carriers. This is associated with a larger change in ECV in response to the change in sodium intake. The latter may be the driving force behind the larger change in GFR, and in plasma renin activity in G-allele carriers.

1.5.3 Discussion

We identified and validated a SNP upstream of the uromodulin (UMOD) gene whose minor allele is associated with a lower risk of hypertension, and a corresponding per allele reduction of 0.5 mmHg SBP and 0.3 mmHg DBP. The associated SNP (rs13333226) is in close proximity to the uromodulin transcription start site at -1617 base pairs. There is only one previous candidate gene study of UMOD and hypertension. This study tested rs6497476, located in the 5' region of the UMOD gene (-744 bp from UMOD transcriptional start point) and showed nominal association (p=0.04) of the minor allele with a lower risk of hypertension in a Japanese population [108]. This SNP is correlated with rs13333226 in the Japanese HapMap population (r^2 =0.91) and shows the same directionality of effect. A recent genome scan for chronic kidney disease (CKD) [109] has found the minor T allele at rs12917707, -3653 bp upstream from the UMOD transcription start site to be associated with a 20% reduction in risk of CKD. This association was consistent after adjusting for major CKD risk factors including SBP and hypertension. This SNP -rs12917707 is perfectly correlated (r^2 = 1 in HapMap CEU) with rs13333226. Our data shows the minor allele of rs13333226 is associated with increased eGFR (β=3.6, p=0.012), but adjustment for eGFR in our meta-analyses did not alter its association with lower risk for hypertension. Our findings indicate that the UMOD locus is independently associated with hypertension. We also show an association of this SNP with long term cardiovascular outcomes and despite the relatively small attenuation of the relationship after SBP adjustment, the association between rs13333226 and CVD could be mediated through a BP effect as baseline BP may not accurately enough reflect the differences in long term BP exposure, which are mediated through genetic UMOD variance.

The UMOD gene encodes the Tamm Horsfall protein (THP)/uromodulin, a glycosylphosphatidylinisitol (GPI) anchored glycoprotein. It is the most abundant tubular protein in the urine, which is expressed primarily in the thick ascending limb of the loop of Henle (TAL) with negligible expression elsewhere [110, 111]. Our findings in the BRIGHT and HERCULES studies demonstrate a direct relationship between urinary uromodulin and urinary sodium excretion in hypertensive patients and the general population, respectively. This is in line with the rise in urinary uromodulin elicited by the increase in sodium intake in the GRECO subjects and suggests a gene-environment interaction on urinary uromodulin excretion. In addition we also consistently show in three separate populations that the minor G allele of rs13333226 (associated with a lower risk of hypertension) is also associated with lower urinary uromodulin excretion. This effect was lost during the HS in the GRECO subjects. Together, these data would suggest that UMOD may be involved in regulating BP and facilitating onset of hypertension, possibly by an effect on sodium homoeostasis. In agreement with this hypothesis, our data in HERCULES show an increase in proximal sodium reabsorption in G allele subjects, a compensatory reaction to be expected when distal sodium loss in increased. Furthermore, in GRECO, when HS diet increased but equalized uromodulin excretion across rs13333226 genotypes, we see greater changes in GFR, filtered load of sodium, ECV and PRA in G allele carriers. This supports a unifying hypothesis which connects genetic variants in the UMOD gene and BP through a link between the gene product, uromodulin and volume homeostasis. Accordingly, a lower uromodulin excretion (either genetically determined or acquired) is associated with less sodium reabsorption in TAL, which could lead to lower ECV and lower BP.

In the context of our findings it is of interest to note that UMOD mutations (in exons 4 and 5) are implicated in monogenic syndromes such as familial juvenile hyperuricemic nephropathy, autosomal-dominant medullary cystic kidney disease [MCKD2] and glomerulocystic kidney disease (GCKD) (MIM603860, MIM162000, MIM609886) [112, 113, 114]. In previous small studies, urinary uromodulin levels were found to be decreased in older subjects and in subjects with renal impairment [115, 116]. In renal disease patients, uromodulin excretion was reduced in proportion to the extent of renal damage, and was a marker of distal tubular sodium reabsorption, but in these studies, the effects of BP on uromodulin were inconsistent [117, 118]. The TAL, where UMOD is selectively expressed is also the site where mutations of tubular transporters have resulted in rare Mendelian high or low BP syndromes [66]. Furthermore, recent data from Lifton's group demonstrated that heterozygous mutations in SLC12A3 (encoding the thiazide-sensitive Na-Cl cotransporter), SLC12A1 (encoding the Na-K-Cl cotransporter NKCC2), and KCNJ1 (encoding the K+ channel ROMK) discovered in the general population have been associated with lower BP and a 60% reduction in the development of hypertension [119].

Our strategy of using extremes of BP distribution has led to the discovery of a gene variant that could not be discovered when a less stringent case-control definition was used [4]. In addition to functional evidence, we show an association of this SNP with continuous BP and long term outcomes (though it is likely the risk associated with rs13333226 could be mediated partially through a BP effect). The effect size of the risk allele is comparable to the effect sizes of the previous robust association signals for blood pressure [120, 80]. This would suggest that using an extreme case-control strategy successfully enabled the discovery of a locus that previous GWAS meta-analysis failed to detect possibly due to the cost imposed by multiple testing correction. The main limitation of our study is that the functional studies are short-term studies, while the genotype-phenotype effects occur over prolonged time periods. The newly discovered UMOD locus for hypertension has the potential to give unique insights into pathways of renal sodium transport, and identify novel drugable targets.

1.6 Conclusions

The field of common complex disease genetics has moved from linkage to association study design, mainly GWAS, because association analysis has far greater power to detect variants of modest effects and of lower frequency [121]. However, there are still major statistical issues to be overcome including optimal study designs, population stratification, multiple testing, environment and gene interactions, and the effect of epigenetics and structural chromosomal variations.

When we analysed the Swedish data, our biggest problem was the presence of population stratification. Stratification is a form of confounding that can result in artefactual evidence of association. It occurs when there is a systematic difference in allele frequency between cases and controls and it may appear that the risk of disease is related to the marker allele when in fact it is not [20]. In this work, we develop an alternative approach to correct for the presence of stratification in GWAS. Hence, in the next chapter the most used methods to account for the presence of stratification will be presented.

Chapter 2

Multivariate methods to correct for population stratification

Population based case-control studies detect a non-random association between an allele and a trait and provide a powerful tool to identify multiple variants of small effect that modulate susceptibility to common, complex diseases [9]. By using a population sample, however, a significant association between disease and SNP allele can arise from three different sources: by chance, by tight linkage to a causal polymorphism, or, spuriously, by the impact of population stratification [122].

Population stratification, also referred to as population structure, (PS) is a form of confounding that arises when cases and controls are sampled from genetically distinct populations and it is perhaps the most often cited reason for non-replicability [20]. It refers to the presence of a systematic difference in allele frequencies between subpopulations in the study population possibly due to different ancestries. The most obvious cause of PS is migration where individuals from one population migrates into another population. After generations PS will become less due to admixture. Admixture occurs when individuals from two or more previously separated populations begin interbreeding and it results in the introduction of new genetic lineages into a population. Another form of PS is spurious relatedness where non-random mating causes a certain subpopulation to be more related with each other compared to the rest of the population.

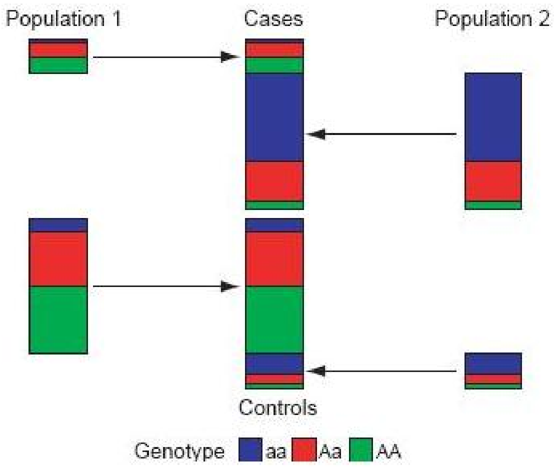

Figure 2.1: Population Structure.

In such cases, non-random associations can occur even at markers completely unlinked to a disease locus simply because of the underlying structure [20, 123]. Also the real disease causing locus might not be found in the study if the locus is less prevalent in the population where the case subjects are chosen. For instance, in a population that is a mixture of African Americans and Caucasians, cases of hypertension will occur disproportionately among African Americans, who are well known to have a higher prevalence of this disease [124]. Any allele that occur more commonly in African Americans will tend to be associated to the disease, even if it is completely unlinked to disease causing loci [125].

For a case-control sample to be stratified, both the following must be true: (i) the frequency of the marker genotype of interest varies significantly by race/ethnicity, and (ii) the background disease prevalence varies significantly by race/ethnicity [126, 127]. This means that the level of stratification is affected by the disease that is being studied, in addition to being affected by the ethnic mixture scenario in the study population. Figure 2.1 shows an example of this scenario with two populations in which the cases have an excess of individuals from population 2 and population 2 has a lower frequency of allele A than population 1. In this example, the structure mimics the signal of association in that there is a significant difference in allele and genotype frequencies between cases and controls [128].

Concerns about the effects of PS led to recommendation of using familial data and to the development of the transmission disequilibrium test (TDT)

[129]. The TDT is a family-based association test design to detect the presence of genetic linkage between a genetic marker and a trait by measuring the over-transmission of an allele from heterozygous parents to affected offsprings. It is an application of McNemar's test. Certainly effective at eliminating false positives due to PS and genetic admixture, the TDT design may result substantially lower in power than other association tests as it utilises only individuals who are informative for allelic transmission and exclude all others. Also, collecting DNA from relatives of affected individuals is generally harder than is collecting DNA from unrelated controls, especially for late-onset diseases. As a result, case-control studies tend to be cheaper than family-based studies of the same sample size. Further, the ability to use unrelated controls suggests the possibility of independent studies reusing database of control genotype data, thus reducing genotyping costs.

Hence, in the last few years, several statistical methods were developed to account for PS so that association studies could proceed even in the presence of structure. It has been argued that the effects of stratification can be eliminated by carefully matching cases and controls according to self reported ancestry and geographical origin [127], however a considerable amount of "cryptic stratification" may remain [125].

Methods for testing and/or adjusting for PS can be broadly classified into three classes [21, 22]:

- genomic control [23, 130, 131];
- structured association [132, 19, 24, 133];
- principal component methods and multidimensional scaling [25] .

Recently, a propensity score method has also been suggested [26].

2.1 Genomic Control

One of the most used method to control for stratification is genomic control (GC) proposed by Devlin and Roeder in 1999 [23]. It uses both a frequentist and a Bayesian approach, the latter being appropriate when dealing with a large number of candidate genes. The frequentist way of correcting for PS works by using markers that are not linked with the trait in question to correct for any inflation of the statistic caused by population stratification. The method was first developed for binary traits but has since been generalised for quantitative ones [122]. For binary traits and n biallelic markers, let N denote the number of subjects genotyped. The data for each marker are given in a standard 2 x 3 table of genotypes by case and control (see Table 2.1).

	A allele			
	0	1	2	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Table 2.1: Genotype distribution.

To test the lack of independence using an additive genetic model (which is the model normally used when no prior information about the mode of inheritance is known), Devlin and Roeder use Armitage's trend test

$$Y^2 \quad = \quad \frac{N[N(r_1+2r_2) - R(n_1+2n_2]^2}{R(N-R)[N(n_1+4n_2) - (n_1+2n_2)^2]} \,. \tag{2.1}$$

This test is equivalent to the score test in the logistic regression model. All is done under the assumption that the n loci under study consists of c biallelic polymorphisms in candidate genes and (n - c) null SNP dispersed throughout the genome. For the disorder of interest, it is also assumed that the null loci have no impact on liability and that they are not in linkage disequilibrium with polymorphisms affecting liability. Although the test statistic is computed for all n loci, only the candidate gene polymorphisms are tested for association. For each marker locus l a statistic Y^2 is obtained using the trend test, l = 1,..,n. When the marker is in linkage equilibrium with the disorder and there is no population substructure or cryptic relatedness, Y^2 is distributed as χ^2_1 . The GC model allows for extra variance by assuming that the test statistic is inflated by a factor λ [23, 125]; consequently,.

$$Y^2 \quad \approx \quad \lambda\chi^2_1, \tag{2.2}$$

where λ depends on the effect of stratification.

The above method rests upon the assumption that the inflation factor λ is constant, which means that the loci should have roughly equal mutation rates, should not be under different selection in the two populations, and the amount of Hardy-Weinberg disequilibrium measured in Wright's coefficient of inbreeding F should not differ between the different loci.

Genomic control, however, has some limits. First, it was shown by simulation that genomic control can lead to an anti-conservative p-value leading to false positives if the two populations (cases and controls) are extremely distinct and the number of unlinked markers is in the order 50 - 100 [128]. Also, some markers differ in their allele frequencies across ancestral populations more than others. This implies that the uniform adjustment applied by genomic control may be insufficient at markers having usually strong differentiation across ancestral populations and may be superfluous at markers devoid of such differentiation, leading to a loss of power.

2.2 Structured association

Structured association methods are more sophisticated than genomic control, and are computationally more demanding. These methods aim to allocate an individual's genome to one or more subpopulations, and to test for association conditional on this allocation. To cluster, structured association uses model-based methods (i.e. methods that assume that observations from each cluster are random draws from some parametric model. Inference for the parameters corresponding to each cluster is done together with inference for the cluster membership of each individual using standard statistical methods; e.g. maximum likelihood or Bayesian methods [134]) and it is based on a two-phase approach. First, using Bayesian Markov Chain Monte Carlo (MCMC) technology, genotype data from unlinked genetic markers are used to learn about population structure and to infer ancestry in the sample [134]. Second, association of candidate loci is tested taking into account the ancestry of cases and controls [24]. Pritchard and colleagues implemented the first phase in the software STRUCTURE and the second in the software STRAT (STRucture population Association Test).

2.2.1 Phase 1: STRUCTURE

The first challenge when applying model-based methods is to specify a suitable model for observations from each cluster. Consider a sample of unrelated individuals typed at many unlinked markers, yielding genotypes G. Assume that the sampled individuals have been ascertained from K discrete subpopulations and that each population is modelled by a characteristic set of allele frequencies. Under the assumption of HWE, the aim is to estimate the allele frequencies, p_k, within each subpopulation and the assignments Z, of each individual to each subpopulation. Under these assumptions each allele at each locus in each genotype is an independent draw from the appropriate frequency distribution. The idea is that the model accounts for the presence of HWE or linkage disequilibrium by introducing PS and attempts to find population groupings that are not in disequilibrium.

Using Bayes' theorem and conditioning on allele frequencies p_k, the posterior probability of assignment for the ith individual is:

$$\Pr(z_i = J|G_i, p_k) = \frac{\Pr(G_i|z_i = J, p_k)\Pr(z_i = J)}{\sum_{j=1}^{K}\Pr(G_i|z_i = j, p_k)\Pr(z_i = j)}, \tag{2.3}$$

where $\Pr(z_i = J)$ is the prior probability of assignment to subpopulation J, typically taken to be $1/K$

Similarly, conditioning on the subpopulation assignments of each individual Z, given a pre-specified prior density $\Pr(p_k)$, the posterior distribution of allele frequency is

$$\Pr(p_k|Z, G) = \frac{\Pr(Z|p_k, G)\Pr(p_k)}{\sum_{j=1}^{K}\Pr(Z = j|p_k, G)\Pr(p_k)}. \tag{2.4}$$

Structure uses Bayesian MCMC technique to sample, in turn, from the densities $\Pr(Z|G, p_k)$ and $\Pr(p_k|Z, G)$.

The algorithm runs for an initial burn-in period to allow for convergence. In the subsequent sampling period, sampled values of the population allele frequencies and subpopulation assignments for each individual are recorded to approximate the marginal posterior distribution of p_k and Z.

The basic structure model assumes that an individual belongs to just one discrete subpopulation. STRUCTURE also allows for admixed population, for which the subpopulations assignments, Z, are replaced by vectors of ancestry, Q, and $q_{i,k}$ denotes the proportion of the ancestry of the ith individual that comes from subpopulation k [134].

2.2.2 Phase II: STRAT

In order to test for association in the presence of PS, the standard null hypothesis of no overall association between allele frequencies at the candidate locus and phenotype, is replaced with a null hypothesis of no such association within each subpopulation. The alternative hypothesis then becomes that genotype frequencies vary with subpopulation and disease phenotype. Conditioning on the ancestries Q estimated by STRUCTURE, a test statistics is constructed by computing the likelihood ration under the two hypotheses:

$$\Lambda = \frac{\mathrm{Pr}_1(G, |\hat{p}_1, \hat{Q})}{\mathrm{Pr}_0(G, \hat{p}_0, \hat{Q})}, \tag{2.5}$$

where p_0 and p_1 denote subpopulation allele frequencies under H_0 and H_1 respectively. Large values of Λ indicate that the alternative model (in which allele frequencies at the candidate locus depend on the phenotype) is substantially better than the null model. The significance of a particular value of Λ is assessed by simulation [24].

Structure association has several advantages. First, it allows for discrete or admixed subpopulations, and can use both SNPs and microsatellites. Second, it provides a general framework for allowing for structure in association tests and can thus be extended to multi-locus or haplotype tests. Third, it provides detailed information about PS.

However, none of the structured association analyses is straightforward to implement, the most important difficulty being the selection of the number of subpopulations. Since the notion of subpopulation is a theoretical construct that only imperfectly reflects reality, it is clear that the problem of estimating the number of subpopulations will never satisfactorily be resolved and it is preferable, if feasible, to implement a method that does not rely on the number of subpopulations being correctly assessed [135]. Another limit of structured association is that is too computationally intensive for whole genome data. In such cases it requires a selection of unlinked markers.

2.3 Principal components analysis

Before describing how principal components analysis (PCA) applies to genetic data, a brief excursus on linear PCA will be given.

Suppose we have measurements of N individuals on M variables collected in a N x M matrix X. PCA is a variable reduction procedure which transforms the M possibly correlated variables into a smaller number p of uncorrelated variables called principal components, by projecting the M variables into a subset of $\mathbb{R}^p$. A principal component (PC) can be defined as a linear combination of optimally-weighted observed variables.

There are many different, but mathematically equivalent ways to define PCA. Simply speaking, PCA involves the calculation of the eigenvalue decomposition of the data covariance matrix, after mean centering the data for each attribute. PCA can also be formulated by means of a loss function (Eckart-Young theorem, see Appendix A). This involves finding N vectors a_i corresponding to the individuals and M vectors b_j corresponding to the variables such that $x_{ij} \approx a_i' b_j$ and it is defined as finding the scores of A and the loadings of B that minimise the loss function.

$$\sigma(A, B) = \sum_{i=1}^{N} \sum_{j=1}^{M} (x_{ij} - a_i' b_j)^2. \tag{2.6}$$

The N x p matrix A is the matrix of component scores and an M x p matrix B is the matrix of components loadings. Using the singular value decomposition theorem (see Appendix A) , the minimum value of the loss function (2.6) is given by:

$$\sigma(\hat{A}, \hat{B}) = \sum_{s=p+1}^{M} \lambda_s^2(X), \tag{2.7}$$

where $\lambda_s(X)$ are the ordered singular values of X.

The first component extracted in PCA accounts for a maximal amount of total variance in the observed variables. This means that the first component is correlated with at least some of the observed variables. The second component accounts for a maximum amount of variance in the dataset that was not accounted for by the first component; i.e. it is correlated with some of the observed variables that did not display strong correlations with the first component. Also, it is uncorrelated with the first component. The remaining components extracted display the same two characteristics: each component accounts for a maximal amount of variance in the observed variables that was not accounted for by the preceding components, and is uncorrelated with all of the preceding components.

PCA was first introduced to study genetic data over thirty years ago by Cavalli-Sforza and colleagues [136] and it has now become a standard tool in genetics. It was first applied at population level to obtain worldwide axes of human variation from the allele frequencies of various populations [136, 137]. It has recently been used at individual level especially now that data with hundreds or thousands of individuals and hundred of thousands of markers become available. Patterson and collegues [25, 27] use PCA to explicitly model ancestry differences between cases and controls. The algorithm, implemented in the software EIGENSTRAT, works as follows.

Suppose markers are biallelic (e.g. SNPs) and consider the **transposed** M x N data matrix G, with rows indexed by the polymorphic markers and columns indexed by individuals. Then, for each marker, choose a reference and a variant allele. Let g_{ij} be the number of variant alleles for marker i, individual j (so for autosomal loci, g_{ij} can take values 0,1,2).

The algorithm works in three steps.

1. From each entry in row i, subtract the row mean

$$\mu_i = \frac{\sum_{j=1}^{N} g_{ij}}{N} . \tag{2.8}$$

Missing entries are excluded from the computation of μ_i and set to zero.

2. Normalise row i by dividing each entry by $\sqrt{p_i(1-p_i)}$, where

$$p_i = \frac{(1 + \sum_{j=1}^{N} g_{ij})}{2 + 2N} \tag{2.9}$$

is a posterior estimate of the unobserved underlying frequency of SNP i. This normalisation step is motivated by the fact that the frequency change of a SNP due to genetic drift occurs at a rate proportional to $\sqrt{p_i(1-p_i)}$ per generation. Moreover, if the population is in HWE, it also normalises each row to have the same variance. If we denote by X the resulting matrix, this amounts to

$$x_{ij} = \frac{g_{ij} - \mu_i}{\sqrt{p_i(1-p_i)}} . \tag{2.10}$$

3. Compute the N x N covariance matrix of the individuals $\Psi = X'X$ and perform an eigenvalue decomposition.

The same eigenvalue decomposition can be obtained starting from the original data matrix, with rows indexed by the individuals and columns indexed by the polymorphic markers. In this case, columns are mean centred and normalised and the covariance matrix is computed on the individuals. Hence, this translates into computing the covariance matrix on the rows, rather than on the columns.

The kth axis of variation is then defined to be kth eigenvector of Ψ (i.e. the eigenvector with the kth largest eigenvalue). Eigenvectors corresponding to large eigenvalues are exposing non-random PS. The ancestry a_{jk} of individual j along the kth axis of variation equals to coordinate j of the kth eigenvector.

Finally, let

$$g_{ij,adjusted} = g_{ij} - \gamma_i a_j, \tag{2.11}$$

where

$$\gamma_i = \frac{\sum_j a_j g_{ij}}{\sum_j a_j^2} \tag{2.12}$$

is a regression coefficient for ancestry predicting genotype across individuals j with valid genotypes at SNP i. A similar adjustment is performed for each axis of variation. This is equivalent to use axis of variation as covariates in multilinear regression.

Correcting for stratification using continuous axis of variation has several advantages. First, they provide the most useful description of within-continent genetic variation [138]. Second, the continuous axis are constructed to be orthogonal and so results are insensitive to the number of axes inferred empirically. Third, EIGENSTRAT runs extremely quickly on large datasets making it computationally feasible for GWAS data. Fourth, the first formal tests for the presence of genetic data is provided. Finally, the PCA method does not attempt to classify all individuals into discrete populations or into linear combinations of populations. Rather, PCA outputs each individual's coordinates along axis of variation.

A method that is closely related to PCA and that yields similar results when applied to markers completely unlinked to disease causing loci, is multidimensional scaling (MDS). MDS aims to detect axes of variation between individuals that maximise the dissimilarities between them. The similarity between pairs of individuals is measured by mean of their identity by state (IBS) across the genome, i.e. by the number of genes with the same allele value. Over M markers, the IBS between the ith and jth individuals is given by:

$$\mathrm{IBS_{ij}} = 1 - \frac{1}{2M}\sum_{k}|G_{ik} - G_{jk}|, \tag{2.13}$$

where G_{ij} denotes the number of minor alleles (0, 1 or 2) carried by the ith individual at SNP k.

MDS is implemented in the software Plink [104]. As it requires the construction of a pairwise IBS matrix is more computationally complex than PCA

2.4 Propensity scores

Propensity score was initially introduced in design of experiments to provide an alternative method to estimate treatment effects when treatment assignment was not random, but could be assumed to be unconfounded [139]. It has now become widely accepted in observational studies [140, 141].

Propensity score is defined as the conditional probability of assignment to a particular group given a vector of covariates. If subjects are assigned to strata based on their propensity score, then the comparison groups within the strata are balanced with respect to the observed potential confounders. Exact adjustment using the propensity score will on average remove all of the bias in group effect estimates [142]. Recently, a genomic propensity score (GPS) framework has been developed to optimally estimate and correct for bias due to PS using both genetic and non-genetic factors [26]. GPS is defined as the likelihood of an individual having a particular test-locus genotype based on that individual covariate makeup, i.e.

$$\mathrm{GPS}_i(z_i, x_i) = P(G_i|z_i, x_i), \tag{2.14}$$

where $\mathrm{GPS}_i(z_i, x_i)$ is the genomic propensity score for subject i calculated from that subject's z_i and x_i, which represent that individual' s vector of genetic and non-genetic covariates and where G_i is that subject's test-locus genotype. It is assumed that, given the observed covariates, the $G_i(i = 1,..,N)$ are independent and identically distributed where N is the total size of the study cohort. Suppressing the i, it is also assumed that given GPS(z, x), (Z, X) and G are conditionally independent. A general class of models that specify the potential relation among disease, test-locus genotype, genetic and non-genetic covariates is then defined to be:

$$f(E(D|Z, \quad X)) = \quad \eta, \tag{2.15}$$

where *f(.)* is a link function, such as the logit function, that determines the relationship between the outcome variable D and predictor variables Z and X. $E(D|Z,X)$ denotes the conditional mean of D given Z and/or X; η is a linear function of covariates.

This method is able to correct for confounding both for genetic and non-genetic factors using a single variable. A simulation shows that GPS consistently outperforms the PCA and MDS methods in terms of standard error and coverage probability under moderate PS, while PCA consistently outperforms both in terms of power [26]. Under severe PS the GPS method consistently outperformed the PCA method in terms of estimation, bias, standard error and coverage probability. Both GPS and PCA methods always outperformed the MDS in terms of standard error under severe PS.

However, GPS is not easily applicable to GWAS studies or other large-scale studies using hundreds or thousands of SNPs because each test requires the estimation of a propensity score first, and then requires the fitting of a logistic regression model.

2.5 Conclusions

In this chapter, we have described methods for testing and adjusting for PS. Each has its pros and cons, simulation studies [26] have been done to access which method is the best, but no consensus has yet been reached. Also, methods such as STRUCTURE or GPS are computationaly demanding and can not be easily applied to GWAS data, where PCA still represents a sort of gold standard. However, the assumption underlying PCA is that the variable under study should be continuous and this is achieved by transforming the qualitative values AA, AB and BB of an individual at a given SNP into 0, 1, 2. While this transformation is easy to apply, it assumes that the distance between AA and AB is the same as the distance between AB and BB and hence it assumes an additive model of inheritance. This model is the most conservative, but not necessarily the correct one.

Our approach is to treat SNPs as ordinal qualitative variables. This implies that we agree that there is an order between AA, AB and BB, but that the distances AA-AB and AB-BB are not necessarily the same. We try to detect and correct for PS by applying a multivariate analysis method to reduce dimensionality in the presence of non-metric data. The next chapter, is dedicated mainly to the description of such a technique.

Chapter 3

Multivariate analysis methods for non-metric or mixed-type data

In this chapter we outline Gifi System to reduce dimensionality in the presence of qualitative or mixed-type variables. In particular, we describe homogeneity analysis (HOMALS) and non-linear principal components analysis (NLPCA or PRINCALS: principal components analysis by means of alternating least squares). We also present the goodness of fit measure Procrustes which can be used to compare two data matrices and the Procrustes's test Protest which tests the significance of the Procrustes statistics. PRINCALS is then applied to genetic data in order to correct for PS assuming that SNPs are ordinal qualitative variables. Procrustes and Protest are used to compare the performances of PCA and PRINCALS. A compendium of basic linear and matrix algebra, which can serve as a convenient source of reference for this chapter, is found in Appendix A.

3.1 The Gifi System

The Gifi System [30] represents a unified theoretical framework under which many well known descriptive multivariate techniques are organised. The common properties shared by all Gifi models are the specification of a loss function

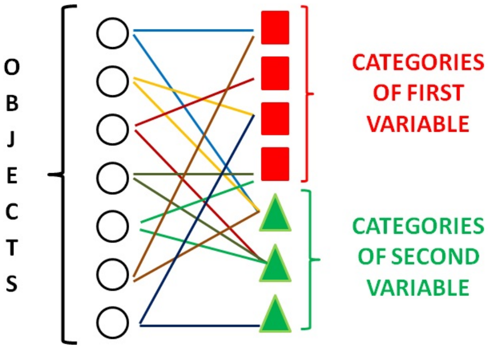

Figure 3.1: Bipartite graph of a toy example.

iteratively optimised by the alternating least squares (ALS) algorithm and transformation of the variables which lead to quantifications of the categories. This latter issue implies the concept of optimal scaling, a procedure that transforms the observed response categories according to some specific criterion, and allows to account for the scaling level of the variables. Most importantly, in all Gifi System, variables reduction, in the sense of PCA, and variables quantification are obtained simultaneously.

In this section, the basic model of homogeneity analysis is presented along with its extensions and generalisation to non-linear principal components analysis (NLPCA or PRINCALS) [29, 143] . The terms NLPCA and PRINCALS are used interchangeably in the remaining of this chapter.

3.1.1 Homogeneity analysis

Suppose that for N objects, data on J categorical variables are collected. Each of the corresponding J variables can take l_j possible levels or categories. Given such a data matrix, all the available information can be represented by a bipartite graph (i.e. a graph whose vertexes can be divided into two disjoint sets U and V such that every edge connects a vertex in U to one in V) where the first set of N vertexes corresponds to the objects and the second set of $\sum_{j \in J} l_j$ vertexes to the categories of the J variables.

Each object is connected to the categories of the variable it belongs to. See Figure 3.1 for the bipartite graph of a toy example of 7 objects and 2 variables with 4 and 3 categories respectively.

The set of all possible $N \sum_{j \in J} l_j$ edges provides information about which categories an object belongs to and about which object belongs to a specific category. Now the degree of a vertex is the size of its neighbourhood (i.e. the number of edges incident to it). So the N vertexes corresponding to the objects all have degree J as every object can take one value/category per variable, while the $\sum_{j\in J} l_j$ vertexes have varying degrees depending on the number of objects in the categories since more than one individual can take the same value/category for one variable.

The bipartite graph, however, is not very helpful for big datasets and so the aim is to construct a low-dimensional joint map of objects and categories in the Euclidean space $\mathbb{R}^p$. The problem is formulated by means of a loss function and it is solved by the ALS algorithm.

Let p denote the number of dimensions that we want to keep in the analysis, let X be the N x p object score matrix containing the coordinates of the object vertexes in $\mathbb{R}^p$ and let Y_j, $j \in J$ be the l_j x p category quantifications matrix containing the coordinates of the l_j category vertexes of variable j, then the aim of homogeneity analysis is to make a graph plot that minimises the total squared length on the edges.

Homogeneity analysis computes object scores and category quantifications on p dimensions or, in other words, solves a projection problem:

$$\mathbb{R}^J \to \mathbb{R}^p \text{ with } p \ll J.$$

To this purpose, data are coded using N x l_j binary indicator or dummy matrices G_j such that

$$G_j(i,t) = \begin{cases} 1 & \text{if object}\, i \in \text{category}\, t \\ 0 & \text{if object}\, i \notin \text{category}\, t, \end{cases} \tag{3.1}$$

where G_j is the adjacency matrix which contains, for the jth variable, as many dummy columns as the number of its categories; $i = 1, ..., N$ and $t = 1, ..., l_j$.

The whole set of indicator matrices can be collected in a block matrix

$$G = [G_1, ..., G_J].$$

We can think of G as the adjacency matrix of a graph in which object i is adjacent to category t of variable j if i is in category t of variable j. The average squared edge length over all variables is then given by the loss function:

$$\begin{aligned} \sigma(X, Y_1, ..., Y_J) &= \frac{1}{J}\sum_{j=1}^{J} SSQ(X - G_jY_j) && (3.2)\\ &= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(X - G_jY_j)'(X - G_jY_j), && (3.3)\end{aligned}$$

where SSQ(H) denotes the sum of squares of the its argument H. This function is the heart of the Gifi System [30].

In equations (3.2) and (3.3), G_j only is known. X and Y_j are unknown and have to be determined during optimisation. In order to avoid the trivial solution $X = 0$, and $Y_j = 0$, $\forall\ j \in J$, optimisation is done under the following normalisation constraints:

$$\begin{aligned} X'X &= NI_p, && (3.4)\\ u'_N X &= 0. && (3.5)\end{aligned}$$

The first constraint (equation (3.4)) standardises the squared length of the object scores to be equal to N, and in two or more dimensions, it also requires the columns of X to be orthogonal. This means that the p dimensions should be uncorrelated and standardised. The second constraint (equation (3.5)) requires the graph plot to be centred around the origin.

So, under the above constraints (i.e. the p dimensions are uncorrelated, with mean zero, unit variance and the solutions should be non trivial), homogeneity analysis has two **simultaneous** goals:

1. to reduce the number of variables from J to p, in the sense of PCA (matrix X);
2. to quantify the qualitative variables (matrix G_jY_j).

The ALS algorithm for solving this problem is a numerical iteration process. At iteration $t = 0$, ALS begins with a starting solution, conveniently fixed, $X^{(0)}$ for the object scores. Consequently, it is possible to update the category scores $Y_j^{(1)}$. In the next step, the object scores $X^{(1)}$ are updated and normalised. Based on these normalised scores, ALS updates the category in the next iteration and so forth. The algorithm stops when the loss function (3.2) does not decrease significantly any more (i.e. the loss difference between two iterations is below a specified threshold ϵ).

Formally, in the **first step**, function (3.2) is minimised with respect to Y_j for fixed X. X needs to be initialised. For each $j \in J$, the multivariate linear model

$$X = G_j Y_j + \ error, \tag{3.6}$$

is fitted and the solution is given by

$$\hat{Y}_j = \frac{1}{D_j} G'_j X, \qquad j \in J, \tag{3.7}$$

where

$$D_j = G'_j G_j \tag{3.8}$$

is the $l_j \, \mathrm{x} \, l_j$ diagonal matrix containing on its diagonal the relative frequencies of the categories of variable j. Equation (3.7), which quantifies categories, represents the so called first centroid principle: a category quantification is in the centroid of the object scores that belong to it [144]. Thus, a category point is the centroid of objects belonging to that category.

In the **second step**, function (3.2) is minimised with respect to X for fixed Y_j' s. The optimal $\hat{X}$ is given by

$$\hat{X} = \frac{1}{J} \sum_{j=1}^{J} G_j Y_j. \tag{3.9}$$

This is known as the second centroid principle. The object score is the average of the quantifications of the categories it belongs to. This implies that objects with the same response pattern receive identical scores.

In the **third step**, the object scores X are columned centred by setting

$$W = \hat{X} - u_N \left(u'_N \frac{\hat{X}}{N} \right), \tag{3.10}$$

and then orthonormalised by the modified Gram-Schmidt procedure [145]

$$X = \sqrt{N} GRAM(W), \tag{3.11}$$

so that constraints (3.4) and (3.5) are both satisfied.

The ALS algorithm cycles through these three steps until it converges. This solution is known as the HOMALS solution (Homogeneity Analysis by Means of Alternating Least Square) and it is implemented in various platforms (see Section 3.1.3 below). It is important to notice that with HOMALS, multiple category quantifications are possible.

Once the ALS algorithm converges,

$$\begin{aligned}\hat{Y}_j' D_j \hat{Y}_j &= \hat{Y}_j' D_j \left(\frac{1}{D_j} G_j' \hat{Y}_j \right) \\ &= \hat{Y}_j' G_j' \hat{X}. \end{aligned} \tag{3.12}$$

We can then decompose the loss function (3.2) as follows:

$$\begin{aligned}
\sigma(X, Y_1, ..., Y_J) &= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(\hat{X} - G_j\hat{Y}_j)'(\hat{X} - G_j\hat{Y}_j) \qquad (3.13)\\
&= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}\{[(\hat{X}' - (G_j\hat{Y}_j)'](\hat{X} - G_j\hat{Y}_j)\}\\
&= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}[(\hat{X}' - \hat{Y}_j'G_j')(\hat{X} - G_j\hat{Y}_j)]\\
&= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(\hat{X}'\hat{X} - \hat{X}'G_j\hat{Y}_j - \hat{Y}_j'G_j'\hat{X} + \hat{Y}_j'G_j'G_j\hat{Y}_j)\\
&= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(\hat{X}'\hat{X} - \hat{X}'G_j\hat{Y}_j - \hat{Y}_j'G_j'\hat{X} + \hat{Y}_j'D_j\hat{Y}_j)\\
&= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(\hat{X}'\hat{X} - \hat{X}'G_j\hat{Y}_j - \hat{Y}_j'G_j'\hat{X} + \hat{Y}_j'G_j'\hat{X})\\
&= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(\hat{X}'\hat{X} - \hat{X}'G_j\hat{Y}_j)\\
&= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}[\hat{X}^{\prime}\hat{X} - (\hat{Y}_j'G_j'\hat{X})']\\
&= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}[\hat{X}'\hat{X} - (\hat{Y}_j'D_j\hat{Y}_j)']\\
&= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(\hat{X}^{\prime}\hat{X} - \hat{Y}_j'D_j'\hat{Y}_j)\\
&= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(\hat{X}^{\prime}\hat{X} - \hat{Y}_j'D_j\hat{Y}_j)\\
&= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(NI_p - \hat{Y}_j'D_j\hat{Y}_j)\\
&= \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(NI_p) - \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(\hat{Y}_j'D_j\hat{Y}_j)\\
&= \frac{1}{J}JNp - \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(\hat{Y}_j'D_j\hat{Y}_j)\\
&= Np - \frac{1}{J}\sum_{j=1}^{J} \mathrm{tr}(\hat{Y}_j'D_j\hat{Y}_j).
\end{aligned}$$

This loss function decomposition mimics variance decomposition of linear analysis. Np can be seen as a sort of total inertia, while $\frac{1}{J}\sum_{j=1}^{J}\operatorname{tr}(\hat{Y}_j' D_j \hat{Y}_j)$ a sort of explained variation. Hence, the loss function can be considered as a sort of residual variance.

The sum of the diagonal elements of the matrices $\hat{Y}_j' D_j \hat{Y}_j$ is called *fit of the solution*.

Moreover, the discriminant measures, which give the average squared distance (weighted by the marginal frequencies) of the category quantifications to the origin of the p-dimensional space, of variable j in dimension s are given by:

$$\eta_{js}^2 \quad \equiv \quad \frac{\hat{Y}_j'(\cdot,s) D_j \hat{Y}_j(\cdot,s)}{N}, \quad j \in J, \quad s = 1, ..., p, \tag{3.14}$$

where $(\cdot, s)$ denotes the sth column of matrix $\hat{Y}_j'$. Equation (3.14) measures how well the p variables represent the J starting variables. A variable discriminates better to the extent that its category points are further apart.

Now, assuming that there are no missing values, the discrimination measures are equal to the squared correlation between an optimally quantifies variable $G_j\hat{Y}_j(\cdot, s)$ in dimension s, and the corresponding column of object scores $\hat{X}(\cdot, s)$. Hence, the loss function (3.2) can be expressed as:

$$\begin{aligned} \sigma(X, Y_1, ..., Y_J) &= Np - \frac{1}{J}\sum_{j=1}^{J}\operatorname{tr}(\hat{Y}_j' D_j \hat{Y}_j) \qquad (3.15)\\ &= Np - \frac{1}{J}\sum_{j=1}^{J}\sum_{s=1}^{p} N\eta_{js}^2 \\ &= N\left(p - \frac{1}{J}\sum_{j=1}^{J}\sum_{s=1}^{p}\eta_{js}^2\right) \\ &= N\left(p - \sum_{s=1}^{p}\gamma_s\right), \end{aligned}$$

where

$$\gamma_s = \frac{1}{J}\sum_{j=1}^{J}\eta_{js}^2, \quad s = 1, ..., p$$

are the eigenvalues. They correspond to the average of the discrimination measures and give a measure of the fit of the HOMALS solution in the sth dimension.

3.1.2 Nonlinear principal components analysis

Given an N x M matrix with metric variables, PCA is a common technique to reduce the dataset dimensionality by projecting the variables into a subset of $\mathbb{R}^p$. The Eckart-Young theorem states that this classical form of PCA can be formulated by means of a loss function. Its minimisation leads to an N x p matrix of component scores and an M x p matrix of components loadings. In the case of non-metric variables, NLPCA is appropriate [143]. Note that in Gifi terminology, the term non-linear refers to non-linear transformations of the observed categories. NLPCA (or PRINCALS) is derived as homogeneity analysis with restrictions [146]. The crucial difference to homogeneity analysis concerns the category score matrix Y_j. In classical HOMALS, Y_j is unrestricted; while in NLPCA is expressed by a linear combination

$$Y_j = Q_j B_j', \quad j \in J, \tag{3.16}$$

where Q_j is the restricted $l_j \text{x} r_j$ quantification matrix (r_j represents the lower rank) and B_j is the p x r_j weight matrix. From a practical point of view, the most important special case [29] is the rank-1 restricted formulation

$$Y_j = q_j \beta_j', \quad j \in J, \tag{3.17}$$

imposed on the multiple category quantifications. q_j is the l_j-column vector of single category quantifications for variable j, and β_j is a p-column vector of weights (component loadings). Hence, each quantification matrix Y_j is restricted to be of rank-1, implying that the quantifications in p-dimensional space become proportional to each other. Introducing the rank-1 restrictions allows the existence of multidimensional solutions for object scores with a single quantification (optimal scaling) for the categories of the variables, and makes it straightforward to incorporate different scale levels (ordinal, numerical) on the variable.

In NLPCA, the aim is still to minimise the loss function but this should be done also under restriction (3.17). The algorithm starts as before, by computing the $\hat{Y}_j$'s as in equation (3.7). Then, recalling equation (3.8) and that if A is a squared matrix, $\text{tr}(A) = \text{tr}(A')$, the Gifi loss function (3.2) can be partitioned as follows:

$$
\begin{aligned}
J\sigma(X, Y_1, ..., Y_J) &= \sum_j \operatorname{tr}(X - G_jY_j)'(X - G_jY_j) \qquad (3.18)\\
&= \sum_j \operatorname{tr}[X - G_j(Y_j + \hat{Y}_j - \hat{Y}_j)]'[X - G_j(Y_j + \hat{Y}_j - \hat{Y}_j)]\\
&= \sum_j \operatorname{tr}\{X - G_j[\hat{Y}_j + (Y_j - \hat{Y}_j)]\}'\{X - G_j[\hat{Y}_j + (Y_j - \hat{Y}_j)]\}\\
&= \sum_j \operatorname{tr}[(X - G_j\hat{Y}_j) - G_j(Y_j - \hat{Y}_j)]'[(X - G_j\hat{Y}_j) - G_j(Y_j - \hat{Y}_j)]\\
&= \sum_j \operatorname{tr}\{(X - G_j\hat{Y}_j)' - [G_j(Y_j - \hat{Y}_j)]'\}\\
&\cdot [(X - G_j\hat{Y}_j) - G_j(Y_j - \hat{Y}_j)]\\
&= \sum_j \operatorname{tr}\{(X - G_j\hat{Y}_j)'(X - G_j\hat{Y}_j) - (X - G_j\hat{Y}_j)'G_j(Y_j - \hat{Y}_j)\\
&- [G_j(Y_j - \hat{Y}_j)]'(X - G_j\hat{Y}_j) + [G_j(Y_j - \hat{Y}_j)]'G_j(Y_j - \hat{Y}_j)\}\\
&= \sum_j \operatorname{tr}\{(X - G_j\hat{Y}_j)'(X - G_j\hat{Y}_j)\\
&+ (Y_j - \hat{Y}_j)'D_j(Y_j - \hat{Y}_j) - (X - G_j\hat{Y}_j)'G_j(Y_j - \hat{Y}_j)\\
&- [(X - G_j\hat{Y}_j)'G_j(Y_j - \hat{Y}_j)]'\}\\
&= \sum_j \operatorname{tr}\{(X - G_j\hat{Y}_j)'(X - G_j\hat{Y}_j)\\
&+ (Y_j - \hat{Y}_j)'D_j(Y_j - \hat{Y}_j) - P_j - P_j'\},
\end{aligned}
$$

where P_j is the cross product $(X - G_j\hat{Y}_j)'G_j(Y_j - \hat{Y}_j)$ and $\hat{Y}_j$ is given by equation (3.7).

Now, using equation (3.7),

$$
\begin{aligned}
P_j &= (X - G_j\hat{Y}_j)'G_j(Y_j - \hat{Y}_j) \qquad (3.19)\\
&= (X - G_j\hat{Y}_j)'G_jY_j - (X - G_j\hat{Y}_j)'G_j\hat{Y}_j\\
&= X'G_jY_j - \hat{Y}_j'G_j'G_jY_j - X'G_j\hat{Y}_j + \hat{Y}_j'G_j'G_j\hat{Y}_j\\
&= X'G_jY_j - X'G_j\frac{1}{D_j}D_jY_j - X'G_j\frac{1}{D_j}G_j'X + X'G_j\frac{1}{D_j}D_j\frac{1}{D_j}G_j'X\\
&= X'G_jY_j - X'G_jY_j - X'G_j\frac{1}{D_j}G_j'X + X'G_j\frac{1}{D_j}G_j'X\\
&= 0.
\end{aligned}
$$

Hence,

$$\begin{aligned} J\sigma(X, Y_1, ..., Y_J) &= \sum_j \mathrm{tr}\{(X - G_j\hat{Y}_j)'(X - G_j\hat{Y}_j) \\ &\quad +(Y_j - \hat{Y}_j)'D_j(Y_j - \hat{Y}_j) - P_j - P_j'\} \\ &= \sum_j \mathrm{tr}\{(X - G_j\hat{Y}_j)'(X - G_j\hat{Y}_j) + (Y_j - \hat{Y}_j)'D_j(Y_j - \hat{Y}_j)\} \\ &= \sum_j \mathrm{tr}(X - G_j\hat{Y}_j)'(X - G_j\hat{Y}_j) \\ &\quad + \sum_j \mathrm{tr}(Y_j - \hat{Y}_j)'D_j(Y_j - \hat{Y}_j). \end{aligned} \tag{3.20}$$

At this point, imposing the rank-1 restrictions (3.17) on the Y_j's leaves to minimise

$$\sum_j \mathrm{tr}(Y_j - \hat{Y}_j)'D_j(Y_j - \hat{Y}_j) = \sum_j \mathrm{tr}(q_j\beta_j' - \hat{Y}_j)'D_j(q_j\beta_j' - \hat{Y}_j), \tag{3.21}$$

with respect to q_j and β_j. This is done with another ALS loop alternating over q_j and β_j which gives, for fixed q_j's,

$$\hat{\beta}_j = \frac{\hat{Y}_j'D_jq_j}{q_j'D_jq_j}, \quad j \in J, \tag{3.22}$$

$$\hat{q}_j = \frac{\hat{Y}_j'\beta_j}{\beta_j'\beta_j}, \quad j \in J. \tag{3.23}$$

The restrictions imposed by the measurement levels of the variables need to be considered. This amounts to project the estimated vector $\hat{q}_j$ on some cone .

For *ordinal data*, the cone C_j is the cone of monotone transformations given by

$$C_j = \{q_j | q_j(1) \leq q_j(2) \leq ... \leq q_j(l_j)\}, \tag{3.24}$$

and the projection to this cone is solved by a weighted ordinal regression in the metric D_j (the weights).

For *numerical data*, the corresponding cone is a ray given by

$$C_j = \{q_j | q_j = \gamma_j + \delta_j s_j\}, \tag{3.25}$$

where s_j is a given vector. The projection to this cone is a linear regression problem.

For *nominal data*, the cone is the $\mathbb{R}^{l_j}$ space and the projection is done by setting

$$q_j = \hat{q}_j. \tag{3.26}$$

Then, for given $\hat{q}_j$ and $\hat{\beta}_j$, $\hat{Y}_j$ can be updated as follows:

$$\hat{Y}_j = \hat{q}_j \hat{\beta}'{}_j. \tag{3.27}$$

and object scores are computed. This solution, which takes into consideration the measurement level of the variables, is known as the PRINCALS solution (Principal Components Analysis by Means of Alternating Least Squares).

Formally, the PRINCALS algorithm works as follows:

0. Initialise X under the two normalisation constraints (3.4) and (3.5).

1. As with HOMALS, estimate the multiple category quantifications by equation (3.7)

$$\hat{Y}'_j = \frac{1}{D_j} G'_j X, \; j \in J.$$

2. Estimate the component loadings by equation (3.22)

$$\hat{\beta}_j = \frac{\hat{Y}'_j D_j q_j}{q'_j D_j q_j},$$

where q_j is suitably fixed and $j \in J$.

3. Estimate the single category quantifications by equation (3.23)

$$\hat{q}_j = \frac{\hat{Y}'_j \beta_j}{\beta'_j \beta_j}, \quad j \in J.$$

4. Account for the measurement level of the jth variable by performing a monotone or linear regression.

5. Update the multiple category quantifications by setting

$$\hat{Y}_j = \hat{q}_j \hat{\beta}_j', \quad j \in J.$$

6. Estimate the object scores by

$$\hat{X} = \frac{1}{J} \sum_j G_j Y_j.$$

7. Column centre and orthonormalise the matrix of object scores.

8. Check the convergence criterion.

Step 2-5 minimise the second term of the loss function (3.20). In step 6, $\hat{X}$ is updated. In step 8, the convergence criterions are checked. If those are not met, the algorithm is repeated from step 1. Step 0, on the other hand, is never repeated. It is simply an initialisation step.

3.1.3 Algorithmic implementation

There are packages in R, SPSS and SAS to work with PRINCALS.

The *homals* package in R does principle component analysis, correspondence analysis, multiple correspondence analysis, regression, canonical correlation analysis and multiset canonical correlation analysis. It allows for treating variables nominal, ordinal, numerical; as well as single and multiple.

A similar set of options is available in *SPSS Categories*, except that they are distributed over various programmes.

The packages in both R and SPSS have been implemented by de Leeuw and colleagues and they follow the algorithm just described (Section 3.1), simultaneously obtaining the two goals of homogeneity analysis. The algorithms written in R are still very tentative, especially compared to the least squared methods which are very well tested and understood.

In SAS, the two goals of homogeneity analysis are not achieved simultaneously. Two different procedures are applied one after the other. The first, *proc prinqual*, quantifies the qualitative variables; the second, *proc princomp*, applies PCA to the quantified variables hence reducing their number from J to p.

3.1.3.1 SAS procedure for PRINCALS

The *prinqual* procedure in SAS provides methods to find transformations that decrease the rank of the covariance matrix of the transformed variables or maximise the variance accounted for by a few linear combinations. As such, it generalises PCA to ordinal or qualitative data.

The data for the *prinqual* procedures can contain variables with nominal, ordinal, interval, or ratio scales. Any mix is allowed: nominal variables can be transformed by scoring the categories to optimise the covariance matrix; ordinal variables can be transformed monotonically or transformed to ranks; interval and ratio variables can be transformed linearly, or non linearly with spline, or monotone spline transformations. In addition, the procedure provides methods for estimating missing data.

Prinqual is primarily a scoring procedure, which produces very little printed output. It creates an output data set that contains both the original and transformed variables. The original variables are named X1, X2, and X3, and the transformed (optimally scaled) variables are named AX1, AX2, and AX3. *Prinqual* also produces an iteration history table that displays (for each iteration) the iteration number, the maximum and average absolute change in standardised variable scores computed over the iteratively transformed variables, the criterion being optimised, and the criterion change.

Table 3.1 summarises options available in the *proc prinqual* statement.

TASK	**Option**
Identify input data set	
specifies input SAS data set	DATA=
Specify details for output data set	
outputs approximations to transformed variables	APPROXIMATIONS
specifies prefix for approximation variables	APREFIX=
outputs correlations and component structure matrix	CORRELATIONS
specifies a multidimensional preference analysis	MDPREF
specifies output data set	OUT=
specifies prefix for principal component scores variables	PREFIX=
replaces raw data with transformed data	REPLACE
outputs principal component scores	SCORES
standardises principal component scores	STANDARD
specifies transformation standardisation	TSTANDARD=
specifies prefix for transformed variables	TPREFIX=
Control iterative algorithm	
analyses covariances	COVARIANCE
initialises using dummy variables	DUMMY
specifies iterative algorithm	METHOD=
specifies number of principal components	N=
suppresses numerical error checking	NOCHECK
specifies number of MGV models before refreshing	REFRESH=
restart iterations	REITERATE
specifies singularity criterion	SINGULAR=
specifies input observation type	TYPE=
Control the number of iterations	
specifies minimum criterion change	CCONVERGE=
specifies number of first iteration to be displayed	CHANGE=
specifies minimum data change	CONVERGE=
specifies number of MAC initialisation iterations	INITITER=
specifies maximum number of iterations	MAXITER=
Specify details for handling missing values	
includes monotone special missing values	MONOTONE=
excludes observations with missing values	NOMISS
unties special missing values	UNTIE=
Suppress displayes output	
suppresses displayed output	NOPRINT

Table 3.1: Options for the PRINQUAL procedure.

The dataset containing the transformed (optimally scaled) variables can be used as the input to other SAS procedures.

3.2 Procrustes analysis

Procrustes transformations permit to compare two configurations (i.e. data matrices) in order to measure their degree of concordance [147, 148]. One data matrix, referred to as the target matrix, is kept fixed; the other is scaled and rotated to find an optimal superimposition that maximises their fit. The sum of the squared residuals between configurations in their optimal superimposition can be then used as a metric of association. A permutation procedure (Protest), implemented by Jackson [149], can be used to assess the statistical significance of the Procrustean fit.

Let X be a n x p matrix of the coordinates of n points obtained from matrix G by one technique and let Y be a n x q matrix obtained with a different technique. Since the superimposition process requires matrices to have the same dimensionality, the matrix with the smaller number of variables, say Y (supposing $q \leq p$), can be filled with columns of zeros until it matches the dimensionality of the larger matrix. In order to find the optimal superimposition one configuration is kept fixed as a reference (X) while the other (Y) is moved successively until the residual sum of squares

$$\sum_{r=1}^{n}(x_r - y_r)'(x_r - y_r) \tag{3.28}$$

is minimised. y_r can be moved relative to x_r through rotation, reflection and translation, i.e. by

$$A'y_r + b, \quad r = 1, ..., n, \tag{3.29}$$

where A is a p x p orthogonal matrix.

Hence, the aim is to solve:

$$R^2 = \min_{A,b} \sum_{r=1}^{n}(x_r - A'y_r - b)'(x_r - A'y_r - b). \tag{3.30}$$

A and b are found by least squares using the following result.

Theorem. Let $Z = Y'X$ and let's assume that the column means of X and Y are zero, i.e. $\bar{x} = \bar{y} = 0$. Using the singular value decomposition theorem,

$$Z = V\Gamma U' \tag{3.31}$$

where V and U are orthogonal p x p matrices and Γ is a diagonal matrix of non-negative values. Then the minimising values of A and b are given by

$$\hat{b} = 0, \qquad \hat{A} = VU', \tag{3.32}$$

and

$$R^2 = \mathrm{tr}(XX') + \mathrm{tr}(YY') - 2\mathrm{tr}(\Gamma). \tag{3.33}$$

Proof.

Equation (3.30) can be rewritten as

$$\begin{aligned} R^2 &= \min_{A,b} \sum_{r=1}^{n} (x_r' - y_r'A - b')(x_r - A'y_r - b) \qquad (3.34) \\ &= \min_{A,b} \sum_{r=1}^{n} (x_r'x_r - x_r'A'y_r - x_r'b - y_r'Ax_r \\ &\quad +y_r'AA'y_r + y_r'Ab - b'x_r + b'A'y_r + b'b). \end{aligned}$$

To find the minimum we need to differentiate equation (3.30) with respect to$A$ and b and equate the partial derivatives to zero.

Differentiating equation (3.30) with respect to b gives,

$$\begin{aligned} \frac{\partial R^2}{\partial b} &= -\sum_{r=1}^{n} x_r' + \sum_{r=1}^{n} y_r'A - \sum_{r=1}^{n} x_r' + \sum_{r=1}^{n} y_r'A + 2nb' \qquad (3.35) \\ &= -2n\bar{x}' + 2n\bar{y}'A + 2nb', \end{aligned}$$

where $\bar{y} = \frac{\sum y_r}{n}$ and $\bar{x} = \frac{\sum x_r}{n}$.

Equating equation (3.35) to zero gives

$$b' = \bar{x}' - \bar{y}'A, \tag{3.36}$$

and taking the transposed on both sides gives

$$\hat{b} = \overline{x} - A'\overline{y}. \tag{3.37}$$

Since both configurations are centred,

$$\hat{b} = 0.$$

□

Hence, equation (3.30) can be rewritten as:

$$\begin{aligned}
R^2 &= \min_A \sum_{r=1}^{n} (x_r - A'y_r)'(x_r - A'y_r) \qquad (3.38)\\
&= \min_A \operatorname{tr}(X - YA)'(X - YA)\\
&= \min_A \operatorname{tr}(X - YA)(X - YA)'\\
&= \min_A \operatorname{tr}(X - YA)(X' - A'Y')\\
&= \min_A \operatorname{tr}(XX' - XA'Y' - YAX' + YAA'Y')\\
&= \min_A \operatorname{tr}(XX' - XA'Y' - YAX' - YY')\\
&= \operatorname{tr}(XX') + \operatorname{tr}(YY') - 2\max_A \operatorname{tr}(X'YA).
\end{aligned}$$

A is such that that $AA' = I$, i.e, $a_i'a_i = 1,\ a_i'a_j = 0$, where a_i' is the ith row of A. Hence, there are $p(p+1)/2$ constraints.

Let $\frac{1}{2}\Lambda$ be the p x p symmetric matrix of Lagrange multipliers for these constraints. We need to maximise

$$\operatorname{tr}\{Z'A - \frac{1}{2}\Lambda(AA' - I)\}, \qquad (3.39)$$

where $Z' = X'Y$.

Differentiating equation (3.39) with respect to A gives:

$$\frac{\partial}{\partial A} \quad \operatorname{tr}\{Z'A - \tfrac{1}{2}\Lambda(AA' - I)\} \quad = \frac{\partial}{\partial A}\operatorname{tr}(Z'A) - \frac{1}{2}\frac{\partial}{\partial A}\Lambda(AA' - I). \quad (3.40)$$

Now,

$$\frac{\partial}{\partial A}\operatorname{tr}(Z'A) = Z, \qquad (3.41)$$

and

$$\frac{\partial}{\partial A}\operatorname{tr}(\Lambda AA') = 2\Lambda A. \qquad (3.42)$$

So, differentiating equation (3.40) and equating the derivatives to zero gives

$$\begin{aligned} \frac{\partial}{\partial A}\mathrm{tr}\{Z'A - \frac{1}{2}\Lambda(AA' - I)\} \quad &= \quad 0 \qquad (3.43) \\ &\Longleftrightarrow \quad Z - \Lambda A = 0 \\ &\Longleftrightarrow \quad Z = \Lambda A \\ &\Longleftrightarrow \quad \Lambda = ZA'. \end{aligned}$$

Recall that A is orthogonal and that Λ is symmetric. Then,

$$\begin{aligned} \Lambda^2 &= ZA'AZ' \qquad (3.44) \\ &= ZZ' \\ &= (V\Gamma U')(U\Gamma V') \\ &= V\Gamma^2 V'. \qquad (3.45) \end{aligned}$$

So we can take $\Lambda = V\Gamma V'$. Thus,

$$\begin{aligned} Z = \Lambda A \quad &\Longleftrightarrow \quad V\Gamma U' = V\Gamma V'A \qquad (3.46) \\ &\Longleftrightarrow \quad \Gamma U' = \Gamma V'A \\ &\Longleftrightarrow \quad U' = V'A \\ &\Longleftrightarrow \quad A = VU'. \end{aligned}$$

Hence,

$$\hat{A} = VU'. \qquad (3.47)$$

□

Note that $\hat{A}$ is orthogonal. Substituting equation (3.47) into equation (3.38) gives

$$\begin{aligned} R^2 &= \operatorname{tr}(XX') + \operatorname{tr}(YY') - 2\max_A \operatorname{tr}(X'YA) \qquad (3.48) \\ &= \operatorname{tr}(XX') + \operatorname{tr}(YY') - 2\operatorname{tr}(X'YVU') \\ &= \operatorname{tr}(XX') + \operatorname{tr}(YY') - 2\operatorname{tr}(U\Gamma V'VU') \\ &= \operatorname{tr}(XX') + \operatorname{tr}(YY') - 2\operatorname{tr}(U\Gamma U') \\ &= \operatorname{tr}(XX') + \operatorname{tr}(YY') - 2\operatorname{tr}(\Gamma UU') \\ &= \operatorname{tr}(XX') + \operatorname{tr}(YY') - 2\operatorname{tr}(\Gamma). \end{aligned}$$

□

Now, we only have to verify that $\hat{A}$ maximises equation (3.39) and it is not just a stationary point. This can be achieved by differentiating equation (3.39) again with respect to A.

For this purpose it is is convenient to write A as a vector $a = (a'_{(1)}, ..., a'_{(p)})'$. Then equation (3.39) is a quadratic form of the elements of a and its second derivative can be expressed as the matrix $-I_p \otimes \Lambda$. Since $\Lambda = V\Gamma V'$ and the diagonal elements of Γ are non-negative, the second derivative matrix is negative semi-definite. Hence $\hat{A}$ maximises equation (3.39).

■

Since we assumed that X and Y are centred at the origin (i.e. their column means are zero), the best rotation of Y relative to X is $Y\hat{A}$ where $\hat{A}$ is given by equation (3.47), and $\hat{A}$ is called the Procrustes rotation of Y relative to X.

Now, as

$$X'YY'X = U\Gamma^2U', \qquad (3.49)$$

equation (3.38) can be rewritten as:

$$R^2 = \operatorname{tr}(XX') + \operatorname{tr}(YY') - 2\operatorname{tr}(X'YY'X)^{1/2}. \qquad (3.50)$$

It can be seen that equation (3.38) is zero if and only if the y_r can be rotated to the x_r exactly.

Scale factor: If the scales of the two configurations are different, equation (3.29) becomes

$$cA'y_r + b, \qquad (3.51)$$

where $c > 0$.

Using the above procedure,

$$\hat{c} = \frac{\mathrm{tr}(\Gamma)}{\mathrm{tr}(YY')}. \tag{3.52}$$

$\hat{b}$ and $\hat{A}$ remain as before and the new minimum residual sum of squares becomes:

$$R^2 = \mathrm{tr}(XX') + \hat{c}^2\mathrm{tr}(YY') - 2\hat{c}\mathrm{tr}(X'YY'X)^{1/2}. \tag{3.53}$$

This transformation is called the Procrustes rotation with scaling of Y relative to X.

This procedure is not symmetrical with respect to X and Y. However, symmetry can be obtained by selecting scaling so that

$$\mathrm{tr}(XX') = \mathrm{tr}(YY'). \tag{3.54}$$

3.2.1 Algorithmic implementation in R

In R, Procrustes and PROTEST are implemented in the package “vegan”. Function `procrustes` rotates a configuration to maximum similarity with another configuration. The main things that need to be specified are: the target matrix X, the matrix to be rotated Y, whether the scaling of axes of Y should be allowed and whether the symmetric Procrustes statistic should be used.

If `scale=FALSE`, the function only rotates matrix Y. If `scale=TRUE`, it scales linearly configuration Y for maximum similarity. Since Y is scaled to fit X, the scaling is non-symmetric. However, with `symmetric=TRUE`, the configurations are scaled to equal dispersions and a symmetric version of the Procrustes statistic is computed.

Function `protest` calls `procrustes(..., symmetric = TRUE)` repeatedly to estimate the significance of the Procrustes statistic. It tests the non-randomness (i.e. the significance) between two configurations using a correlation-like statistic derived from the symmetric Procrustes sum of squares.

Chapter 4

Comparison between linear and non-linear principal components analysis to correct for stratification

In this final chapter, we describe and compare results obtained using linear PCA and PRINCALS to correct for PS. Linear PCA assumes that variables under study are continuous, and so, SNPs are quantified by fixing a reference and a variant allele, and by counting the number of mutations. This implies that: (i) SNPs can take values 0, 1 and 2; (ii) the distance between homozygous wild type and heterozygous is the same as the distance between heterozygous and homozygous mutant; and (iii) the model of inheritance is additive. PRINCALS, on the other hand, treats SNPs as ordinal qualitative variables. This means that there is an order between homozygous wild type, heterozygous and homozygous mutant, but that the distance is not the same. Hence, it does not assume a model of inheritance, it is more flexible and it can potentially capture some information which linear PCA misses out.

The chapter is organised as follows. First, there is a description of the sample dataset including how individuals and SNPs were selected. Second, the application of linear PCA to the sample dataset is presented. PCA was applied using an R algorithm written *ad hoc* which follows precisely the EIGENSTRAT procedure described in Section 2.3. Third, the application of Gifi's PRINCALS to the same sample dataset is outlined. Results obtained with the two different approaches are then compared graphically, by mean of the Procrustean superimposition approach and by the test Protest which tests matrices association, and finally with a scenarios analysis.

For clarity, we refer to the components extracted with linear PCA as principal components (PC), and to those obtained with PRINCALS as dimensions.

4.1 Description of the data matrix

The study sample was drawn from HapMap [17] and so we knew the ethnicity of all individuals. Clearly, ethnicity is typically unknown in advance.We had 988 individuals genotyped at 404,502 SNPs. The ethnic distribution of the study sample is reported in Table 4.1.

We organised data in a large 988 x 404,502 matrix G, with rows indexed by individuals and columns indexed by SNPs.

G is in pedigree format, i.e. its first six columns are respectively:

1. Family ID
2. Individual ID
3. Paternal ID
4. Maternal ID
5. Sex (1=male; 2=female; other=unknown)
6. Phenotype

and from column 7 onwards there are the genotypes.

It has been shown that effective stratification correction is insensitive to the number of samples and that, though EIGENSTRAT has difficulty inferring a perfectly accurate axis of variation when there are less than 5,000 markers (M), stratification correction at random candidate SNPs is effective for M $\geq$200 [25]. Hence, for the purpose of this thesis, we decided to reduce the number of samples to include 90 individuals belonging to three very distinct ethnic groups and to 1,000 randomly chosen SNPs. Individuals and SNPs were selected as described in Sections 4.1.1 and 4.1.2.

Ethnic Group	**Frequency (n)**	**Percent (%)**
African ancestry in Southwest USA (ASW)	49	4.96
Utah residents with Northern and Western European ancestry, CEPH collection (CEU)	112	11.34
Han Chinese in Beijing, China (CHB)	84	8.5
Chinese in Metropolitan Denver, Colorado (CHD)	85	8.6
Gujarati Indians in Houston, Texas (GIH)	88	8.91
Japanese in Tokyo, Japan (JPT)	86	8.7
Luhya in Webuye, Kenya (LWK)	90	9.11
Mexican ancestry in Los Angeles, California (MEX)	50	5.06
Maasai in Kinyawa, Kenya (MKK)	143	14.47
Toscani in Italia (TSI)	88	8.91
Yoruba in Ibadan, Nigeria (YRI)	113	11.44
Total	988	100

Table 4.1: Ethnic distribution of the study sample.

4.1.1 Individual selection

We first ran EIGENSTRAT on the entire study sample to see how individuals clustered. Figures 4.1, 4.2 and 4.3 show respectively the plots of the first and second, second and third, and first and third axis of variation (i.e. eigenvectors).

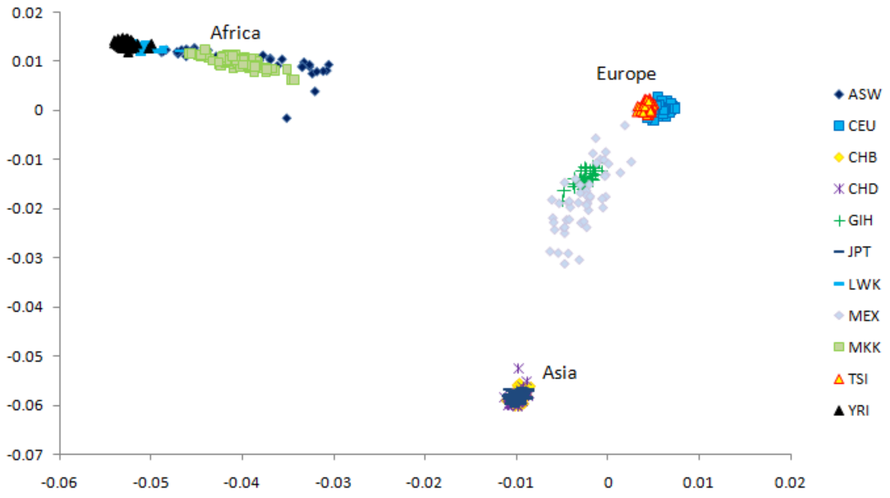

Figure 4.1: First and second axis of variation on 988 HapMap individuals and 405,402 SNPs.

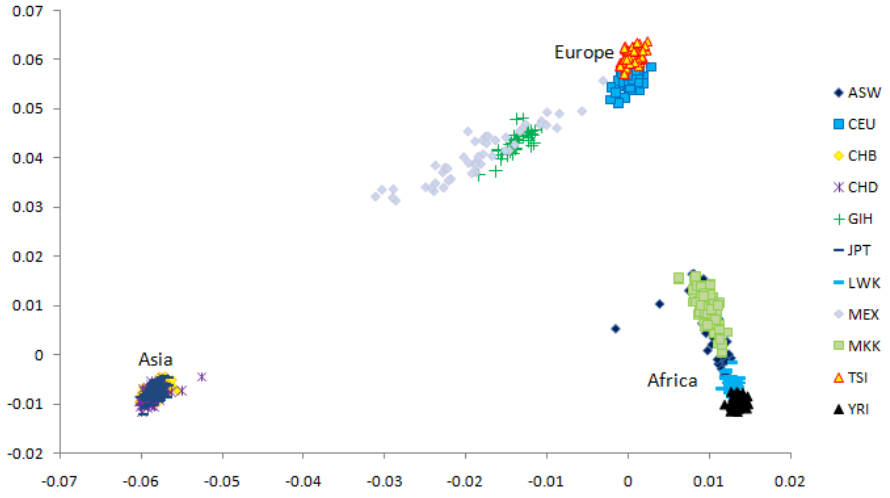

Figure 4.2: Second and third axis of variation on 988 HapMap individuals and 405,402 SNPs.

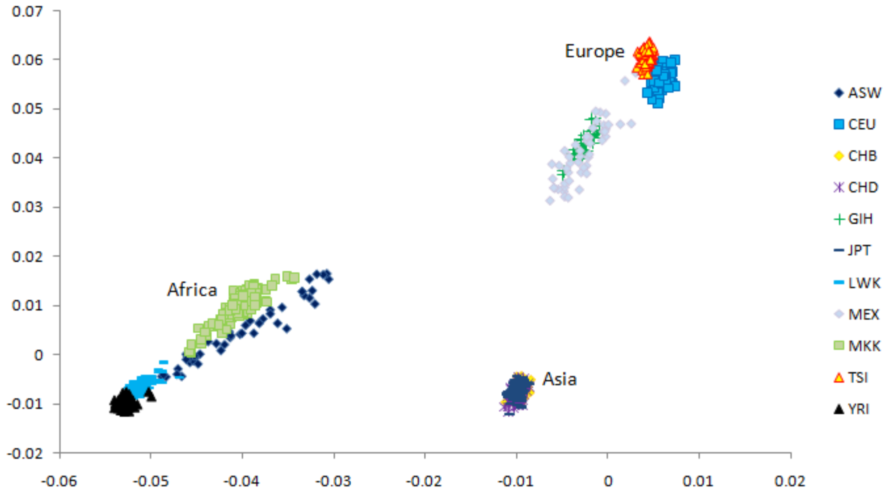

Figure 4.3: First and third axis of variation on 988 HapMap individuals and 405,402 SNPs.

As we can observe from the figures, CEU, CHB and YRI always have a very distinct cluster. We thus decided to restrict our attention to 90 individuals, randomly chosen from those ethnicities. We fixed the prior probability of being

selected from one of the three subpopulations to be 1/3. The reduced study sample thus included 30 CEU, 30 CHB and 30 YRI.
CEU: These samples were collected from people living in Utah with ancestry from Northern and Western Europe. Collection was organised by the Centre d'Etude du Polymorphisme Humain (CEPH) in 1980.
CHB: These samples were collected from individuals living in the residential community at Beijing Normal University who were self-identified as having at least three out of four Han Chinese grandparents.
YRI: These samples were collected in a particular community in Ibadan, Nigeria, from individuals who identified themselves as having four Yoruba grandparents.

4.1.2 SNPs selection

SNPs were chosen as follows. First, SNPs were pruned to exclude markers in LD with each other. We generated a subset of SNPs in approximate linkage equilibrium, i.e. uncorrelated, by calculating the pairwise genotypic correlation and recursively removing SNPs within a sliding window. We chose an r^2 threshold of 0.5 and performed SNPs pruning using Plink [104] with the option "`--indep-pairwise 50 5 0.5`" . With this command we: (i) considered a window of 50 SNPs; (ii) calculated LD between each pair of SNPs in the window and removed one of a pair of SNPs if the LD was greater than 0.5; and (iii) shifted the window 5 SNPs forward; then repeated the procedure. With pruning we reduced the number of SNPs from 412,354 to 142,792.

Second, SNPs with missing values on any of the randomly chosen individuals were removed. This further reduced the number of SNPs to 113,521.

Then, we applied the R function `[sample(nrowSNP,1000,replace=FALSE),]` to randomly select 1,000 SNPs.

4.2 Application of PCA to the sample dataset

As previously mentioned, we applied linear PCA to the study sample using an R script which follows precisely the EIGENSTRAT procedure described in Section 2.3. The R script is reported at the end of this chapter (see Section 4.8). We also wrote a script in Java (see Appendix B) to recode SNPs as 0, 1 and 2 by choosing for each marker a reference and a variant allele and counting the number of mutations. We then transposed the 90 x 1000 sample data matrix and imported it into R. We mean centred its rows (i.e. the SNPs), normalised them and computed the covariance matrix on the individuals. Finally, we performed an eigenvalue decomposition on the covariance matrix.

As it is not yet clear how normalisation can be applied to PRINCALS, and as it has been shown that results are mathematically valid without the normalisation [27], we also performed an eigenvalue decomposition omitting the normalisation step and merely subtracting the row mean. In the remainder of this chapter, we present results obtained without normalisation.

4.3 Application of the PRINCALS method

We applied PRINCALS to the reduced dataset consisting of 90 individuals (30 CEU, 30 CHB and 30 YRI) and 1,000 randomly chosen SNPs not in LD.

Since the algorithms implemented in the R homals package are still very tentative, we decided to use SAS *proc prinqual*. However, as mentioned in section 3.1.3, in SAS, the two goals of homogeneity analysis (i.e. (i) to reduce the number of variables in the sense of PCA and (ii) to quantify the qualitative variables) are not achieved simultaneously. Rather, SAS *proc prinqual* quantifies the qualitative variables. A data reduction procedure (e.g. SAS *proc princomp* or the PCA R script of section 4.8) can then be applied to the quantified variables, thus reducing their number. So, we used a two phased method:

Phase 1. Quantification of the qualitative variables using SAS *proc prinqual*

Phase 2. Extraction of the dimensions using the R script on the quantified variables.

4.3.1 Phase 1: *proc prinqual*

Step 1. We recoded the data matrix of 0, 1 and 2 into 1, 2 and 3, as SAS treats 0 as missing. We then imported the matrix into SAS. The first column of the matrix contained the subject id, the remaining 1,000 columns contained the SNPs. We removed the column headings, so that columns were named by SAS *VAR1-VAR1001*. SNPs were therefore named *VAR2-VAR1001*.

```
PROC IMPORT out= WORK.source
datafile= "D:\G_90ids_1000snps.txt"
dbms=DLM replace;
delimite=’20’x;
getnames=NO;
datarow=1;
RUN;
```

Step 2. We performed the proc prinqual procedure on the data matrix to quantify variables.

```
PROC PRINQUAL data=source maxiter=2000 out=quantified_variables
replace approximations scores;
TRANSFORM monotone (VAR2-VAR1001);
run;
```

Option maxiter = 2000 told SAS the maximum number of iterations. Replace asked to replace raw data with transformed data. Approximations and scores respectively demanded as output approximations to transformed variables and principle components scores. Finally, under our assumption that SNPs should be treated as ordinal qualitative variables,with TRANSFORM monotone, we requested a monotone transformation of the markers.

The algorithm converged in 40 iterations.

Step 3. We created a dataset containing only the quantified, optimally scaled, variables (which by default, in SAS, are called AVAR). This dataset was then exported and imported into R.

```
data nlpca;set quantified_variables ;
keep AVAR2-AVAR1001;
run;
PROC EXPORT data = work.nlpca outfile = ”D:\quantified
variables.txt”;
dbms = TAB replace;
run;
```

4.3.2 Phase 2. Extracting the dimensions

We imported the matrix containing the quantified optimally scaled variables into R, we transposed the matrix and we ran the R script. As mentioned, we omitted the normalisation step as no longer applicable. Indeed, with the normalisation step, each row entry is divided by $\sqrt{p_i(1-p_i)}$ where

$$p_i = \frac{(1 + \sum_{j=1}^{N} g_{ij})}{2 + 2N}$$

and $\sum_j g_{ij}$ is the sum of the row of the values on each row. While with PCA, this sum could take as maximum value $2N$ (as g_{ij} could take values 0, 1 and 2) and so $p_i \leq 1$, with PRINCALS the upper limit of the sum could be > 2. Therefore, p_i could be greater than 1.

4.4 Comparison of PCA and PRINCALS: graphical representation

Figure 4.4 shows the scree plot obtained both with PCA (in blue) and PRINCALS (in red). As the elbow is after the third eigenvalue in both cases, we considered the first three PC and the first three dimensions.

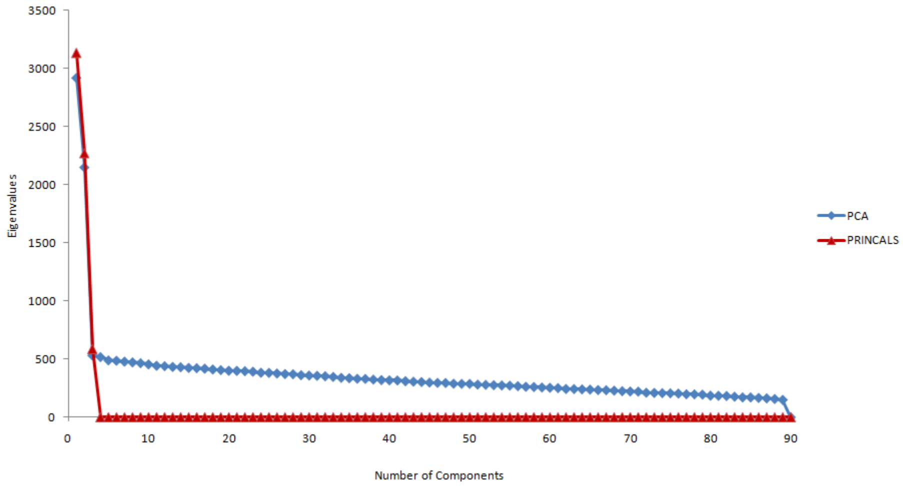

Figure 4.4: Scree plot.

Figures 4.5, 4.6 and 4.7 respectively show the scatter plot of the first and second, second and third, and first and third axes of variation (i.e. eigenvectors) obtained with PCA and PRINCALS. Individuals of CEU origin are indicated in blue, of CHB origin in yellow and of YRI origin in black.

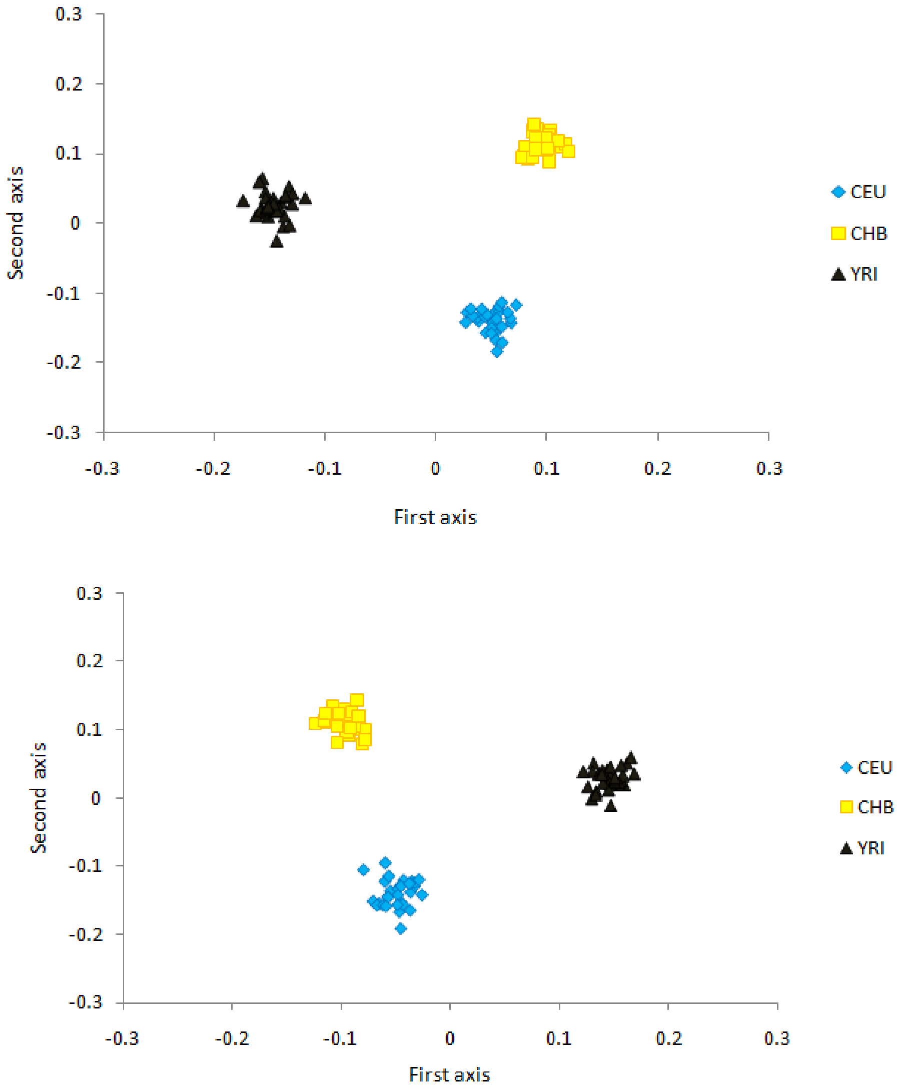

Figure 4.5: First and second axis of variation on 90 individuals and 1,000 random unlinked SNPs. The graph above is obtained from linear PCA, the one below from PRINCALS.

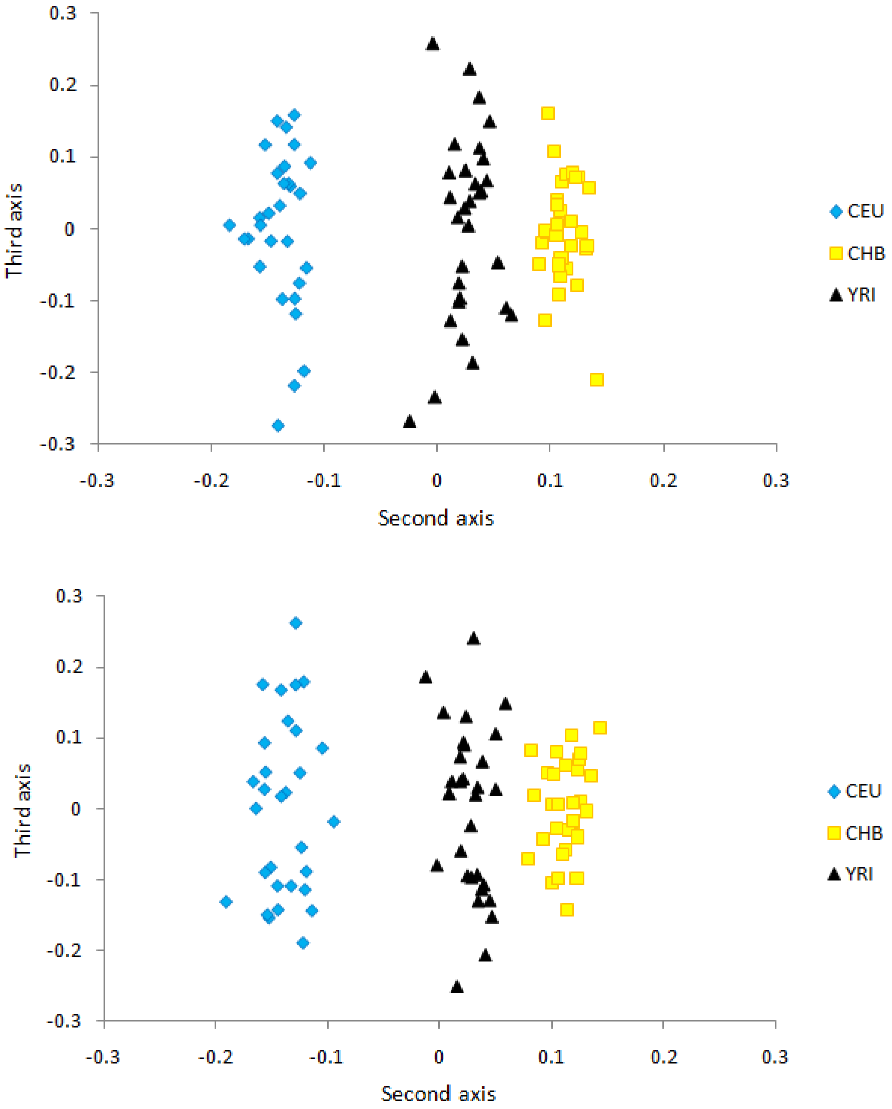

Figure 4.6: Second and third axis of variation on 90 individuals and 1,000 random unlinked SNPs. The graph above is obtained from linear PCA, the one below from PRINCALS.

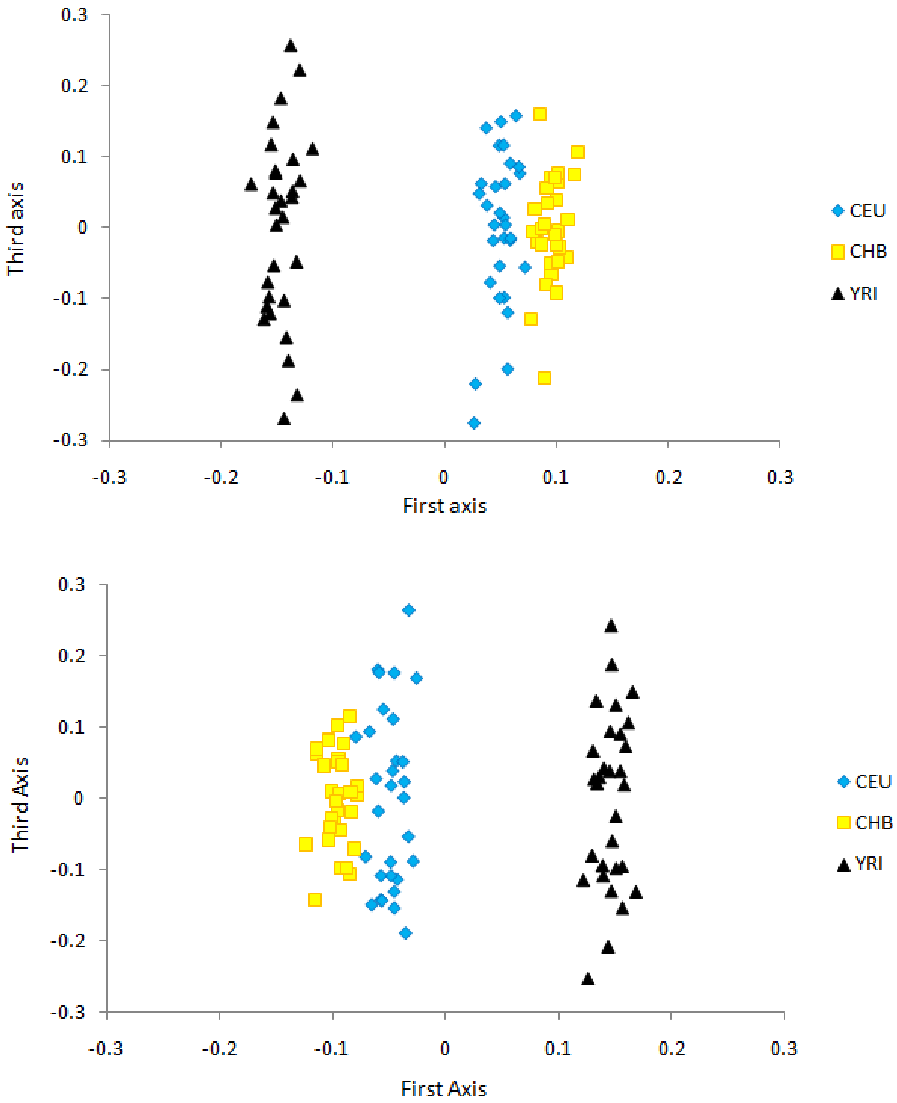

Figure 4.7: First and third axis of variation on 90 individuals and 1,000 random unlinked SNPs. The graph above is obtained from linear PCA, the one below from PRINCALS.

From these graphs, we can see that PCA and PRINCALS give opposite scores for the first axis of variation. CEU and CHB positive scores are translated to negative values, while YRI negative values become positive. Also, the third axis of variation has a bigger range for PCA than PRINCALS.

However, from a graphical perspective, we can see that for both PCA and PRINCALS the first axis of variation identifies macro ethnic differences. In particular, it reflects genetic variation between Eurasia and Africa. Notably, it is the second axis of variation that separates CEU and CHB.

Figure 4.8 shows the scatter plot produced by plotting the first PC versus the first dimension. The points are mainly on a line with a negative gradient, but there are several exceptions. We looked for patterns by studying the genotypic distribution of the outliers, but nothing emerged. We thus went on to plot the second PC vs the second dimension (Figure 4.9) and the third PC versus the third dimension (Figure 4.10). A similar pattern emerged, though the line had positive gradient in both cases. However, plots of PC versus dimension from the fourth component onwards were different, as the points no longer lay on a line (Figures 4.11 and 4.12). We decided to proceed with Procrustes rotation to better compare the two matrices.

4.5 Comparison of PCA and PRINCALS: Procrustes and PROTEST

We compared the matrix of PC and the matrix containing the dimensions using the Procrustean superimposition approach. As described in section 3.2, Procrustes analysis tests the overall degree of association between two matrices by scaling and rotating them in order to find an optimal superimposition that maximises their fit. The sum of the squared residuals between matrices in their superimposition can then be used as a metric of association. There are several strategies for Procrustes analysis. We adopted the simplest approach of the least squares superimposition of one matrix to a reference matrix.

Let C be the matrix containing all PC extracted from the sample dataset and let D be the matrix containing all the dimensions extracted from the same sample dataset. Clearly, both C and D are 1000 x 1000 matrices. Indeed, we extracted eigenvectors from the 90 x 90 covariance matrix constructed on the individuals, thus producing 90 eigenvectors and 1,000 PC/dimensions.

We set the matrix of principle components C to be the reference matrix and we moved the matrix of dimensions, D, successively until the sum of squared residuals $\sum_{j=1}^{1000}(c_j - d_j)'(c_j - d_j)$ was minimised.

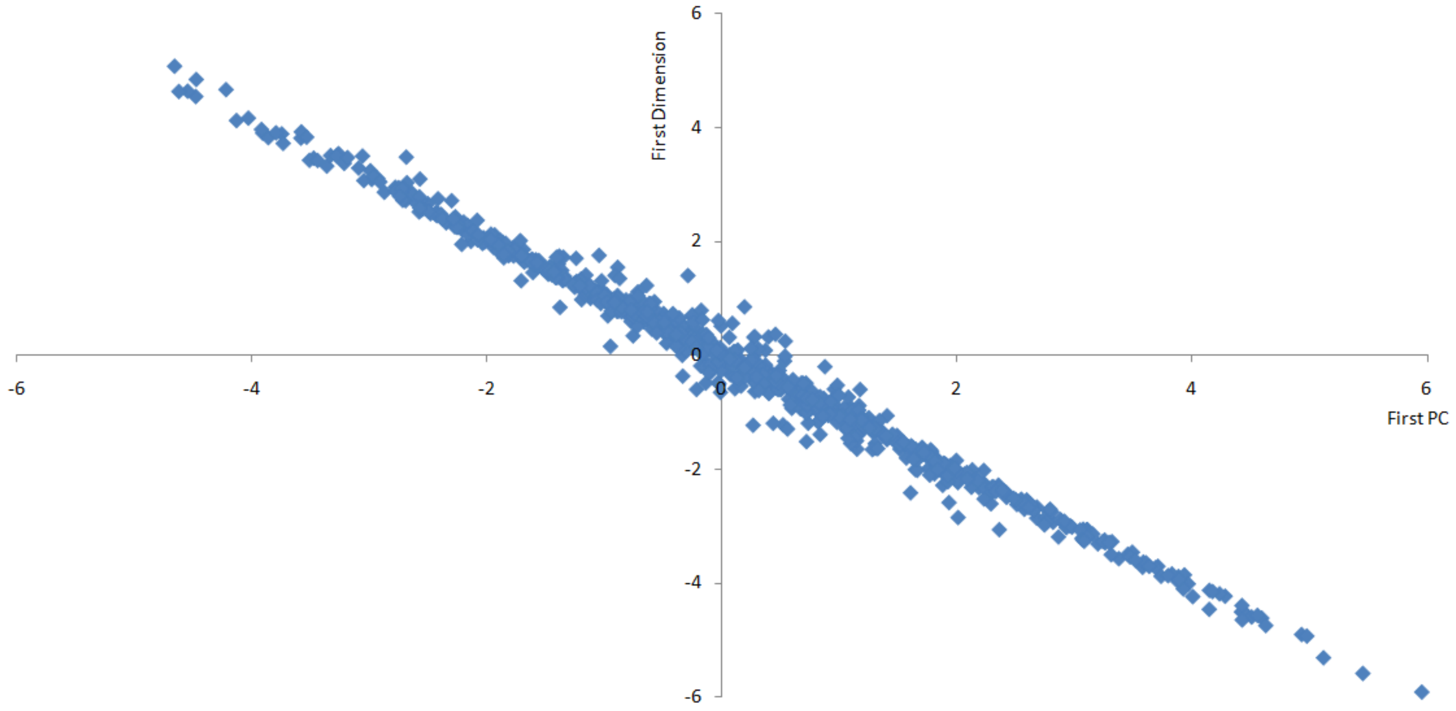

Figure 4.8: First principal component versus first dimension.

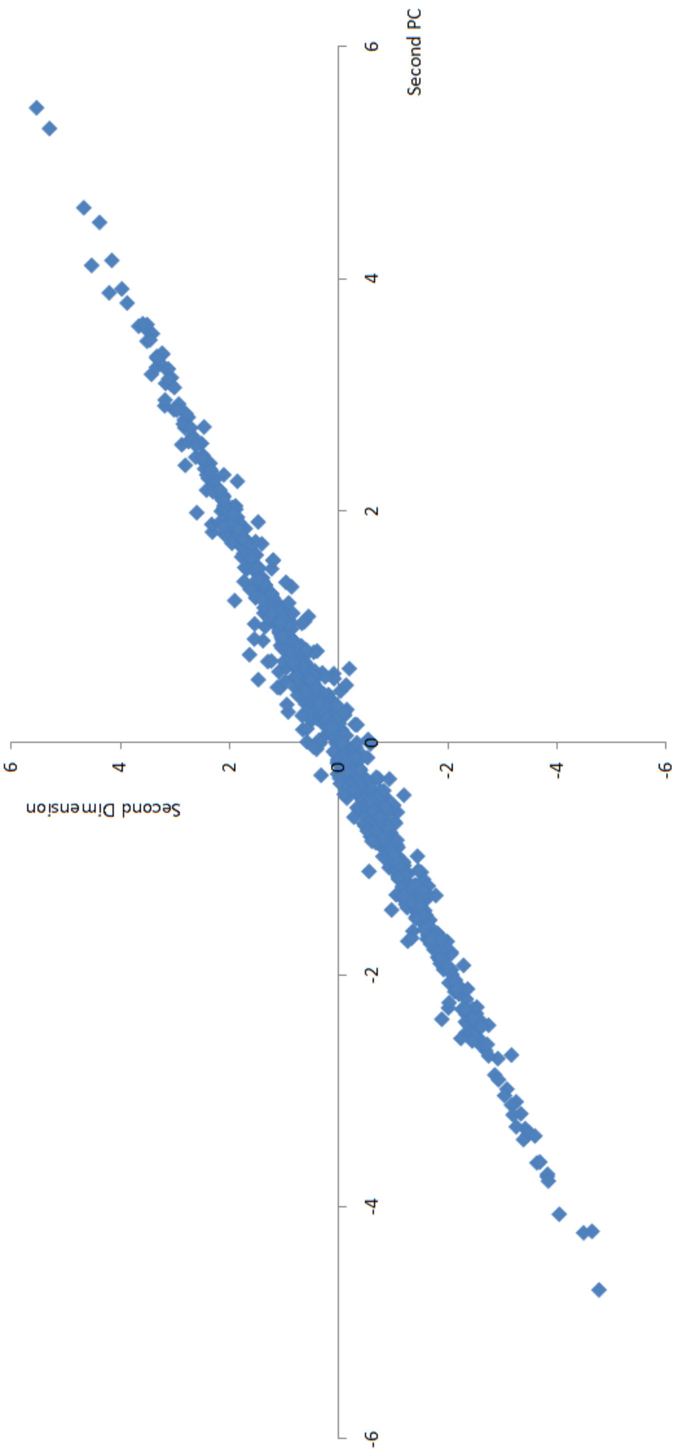

Figure 4.9: Second principal component versus second dimension.

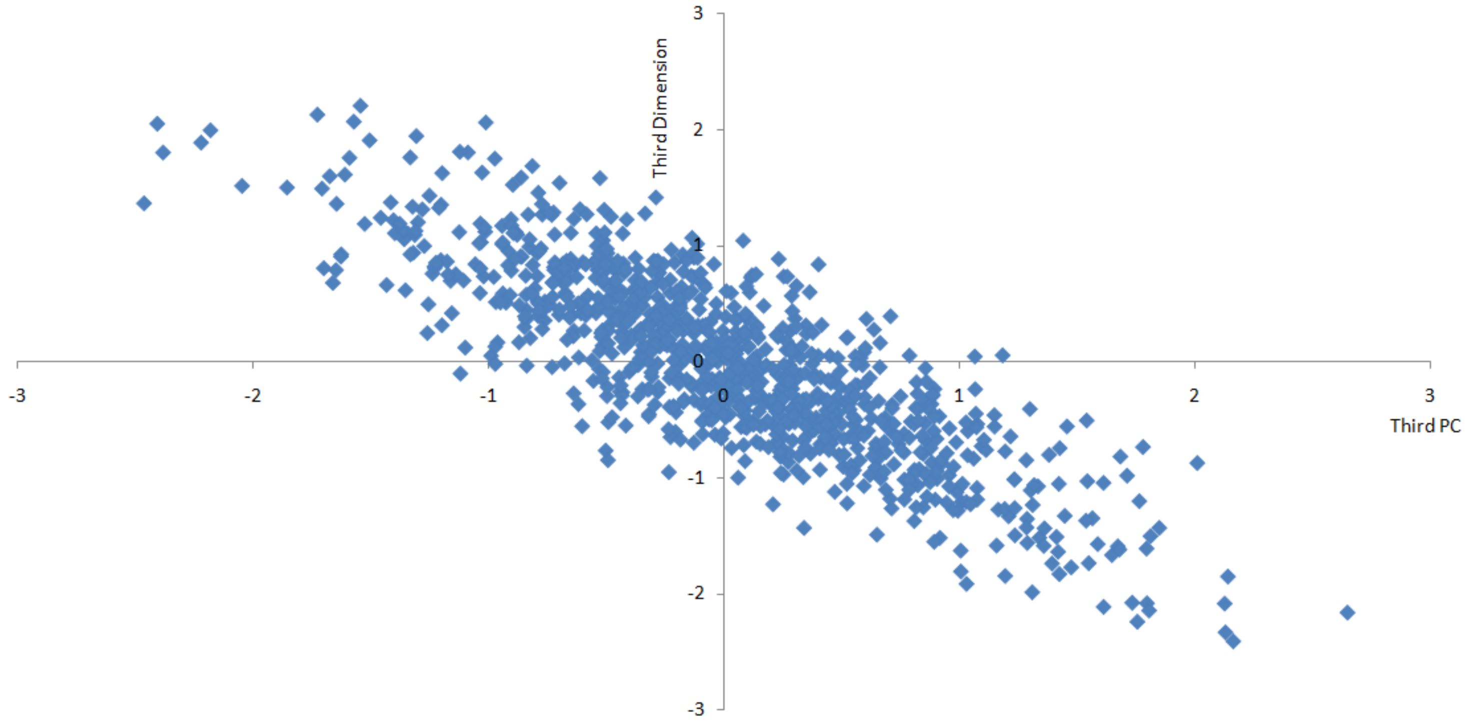

Figure 4.10: Third principal component versus third dimension.

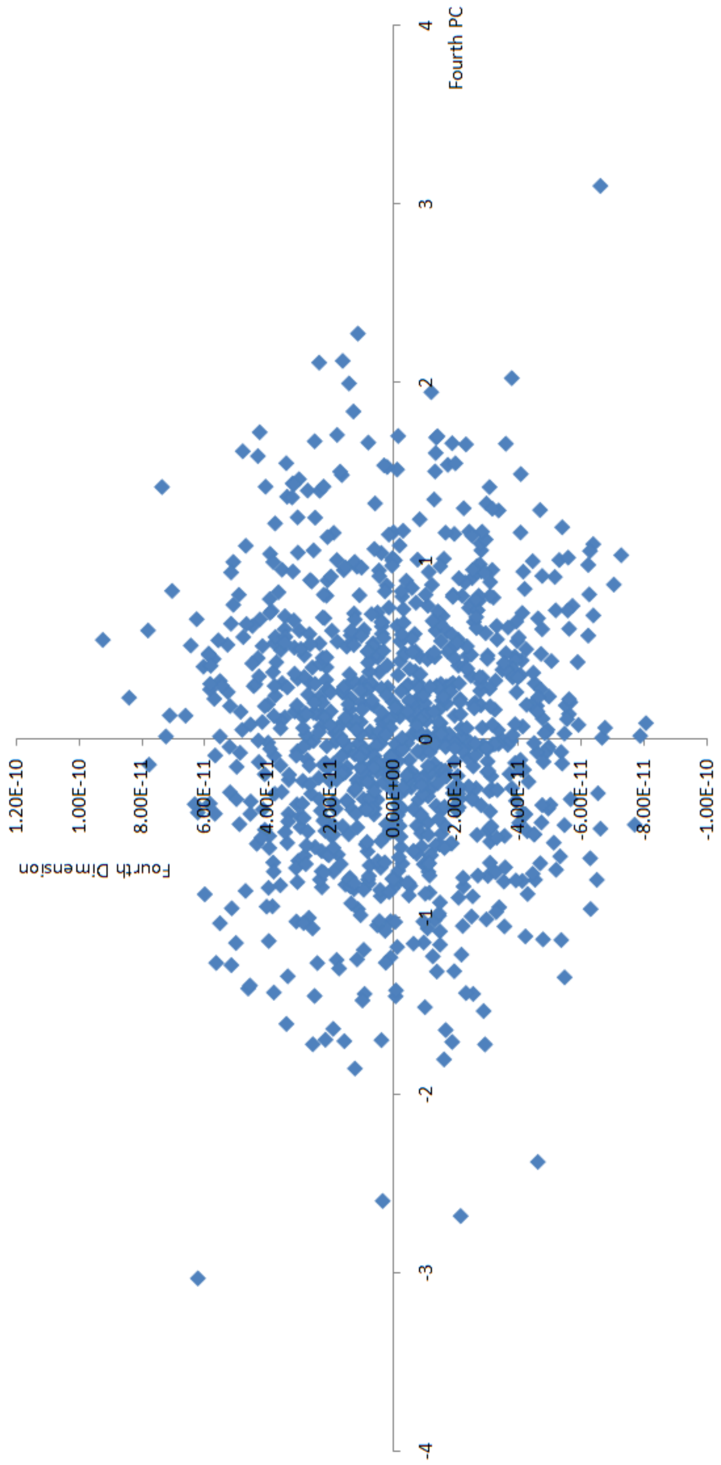

Figure 4.11: Fourth principal component versus fourth dimension.

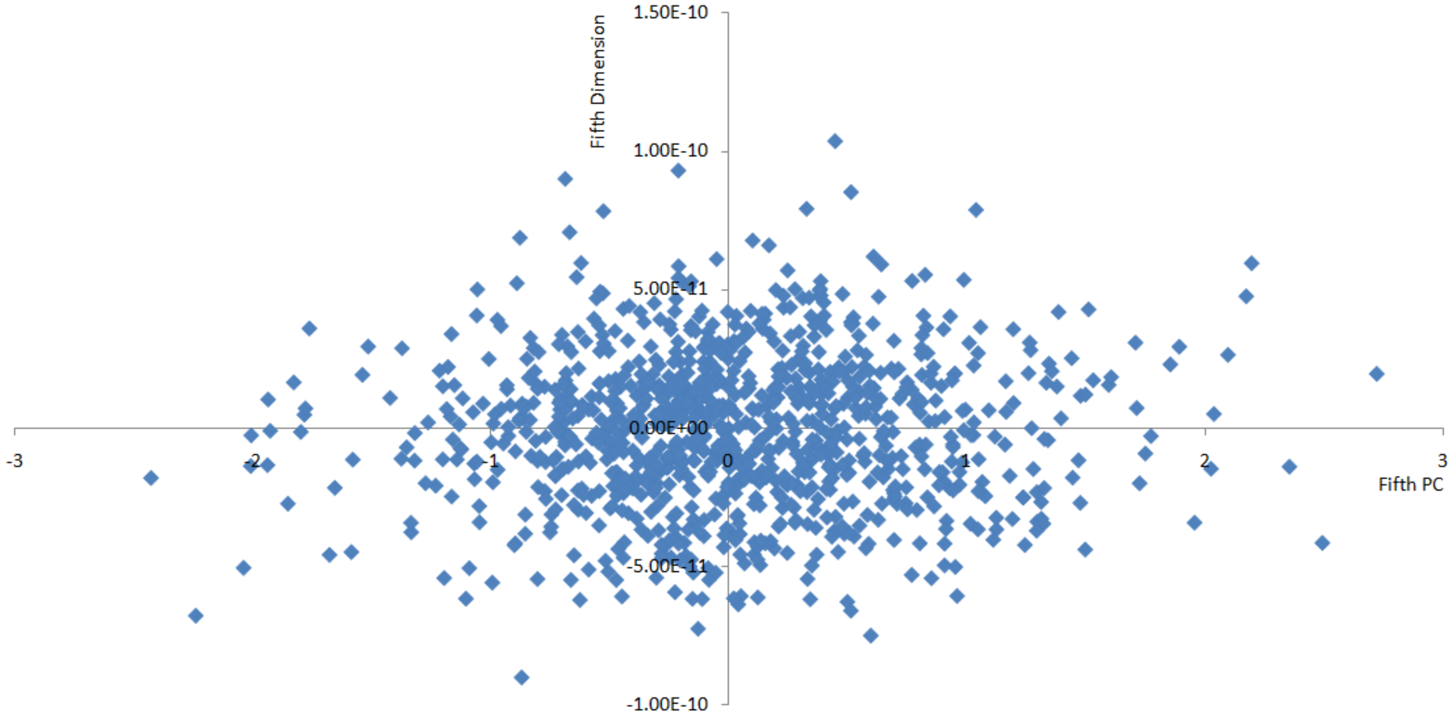

Figure 4.12: Fifth principal component versus fifth dimension.

We performed Procrustes analysis with the "vegan" package in R. Figure 4.13 is the superimposition plot. The target matrix (i.e. the matrix C of principle components) is represented by the solid circles. The matrix of dimensions is represented by the arrow heads. Solid lines represent Procrustes residuals.The longer the line, the greater is the residual. We can observe that an almost perfect super imposition is obtained for some markers, whilst high residuals are seen for others.

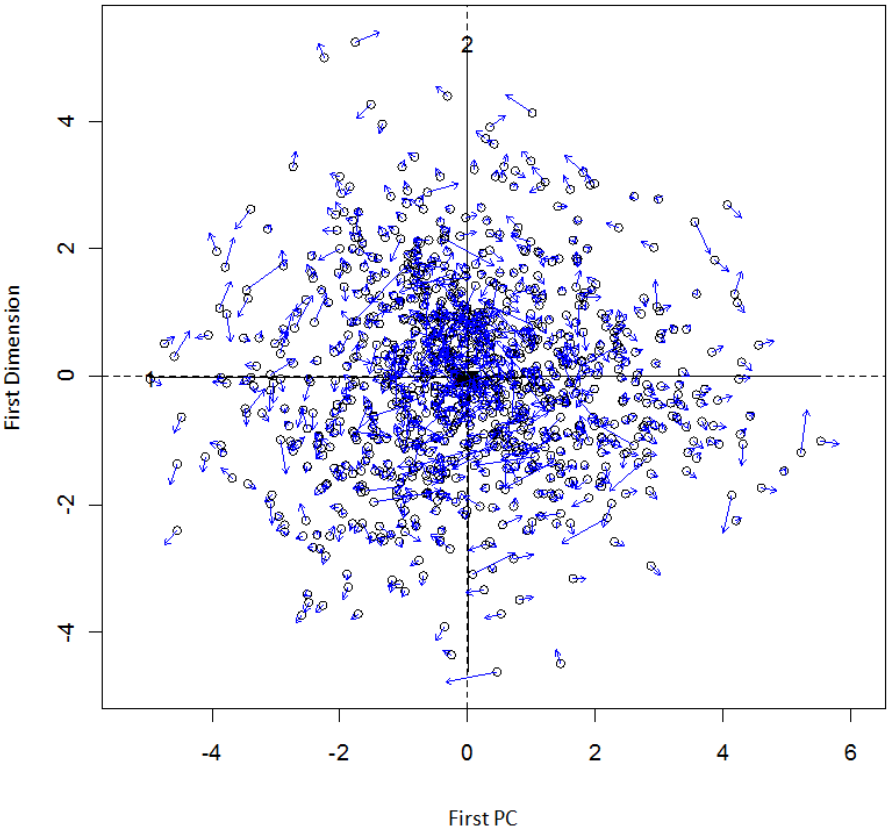

Figure 4.13: Procrustes superimposition plot.

Figure 4.14 is the ordination plot of the residuals. SNPs are ordered depending on the magnitude of the residuals. Again, we can observe that residuals magnitude ranges from 1.7 to 7.3.

By studying the residuals, we see a trend. PC and PRINCALS yield similar

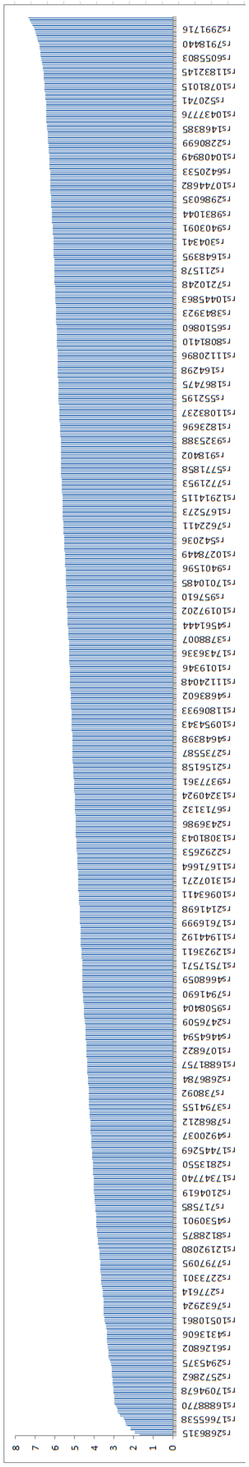

Figure 4.14: Procrustes ordination plot of the residuals.

results (low residuals) for markers that are homogeneous in the study sample. As the genotypic variability increases across the population, so do the scores produced by the two methods.

Finally, we tested matrices association using Protest, a permutation procedure which tests the significance of the Procrustean fit. The function Protest uses a correlation-like statistic derived from the symmetric Procrustes sum of squares ss as $r = \sqrt{1 - ss}$. Based on 1,000 permutations, this gave a correlation in a symmetric Procrustes rotation of 0.4137 with p = 0.001 thus indicating that the two matrices are significantly different.

4.6 Comparison of PCA and PRINCALS: Genomic Control

For the purpose of this study, we focused on the situation where a single set of loci (the 1,000 randomly chosen SNPs), was used both to test for association, and to learn about structure (as in a genome-wide study for association). We present the results from the following model. Samples cases and controls were drawn as previously described so that the study population of 90 individuals consisted of three subpopulations. Case/control label was randomly assigned to produce association due to PS. The first three axis of variation (eigenvectors) obtained with PCA and the first three axis of variation obtained with PRINCALS, were then separately put in a logistic regression multivariate model as covariates. Genomic control was calculated in the three situations. The scenarios analysis was done using Plink [104] with the options "`--logistic --adjust`". Table 4.2 reports the results from 10 different scenarios each with different level of stratification. For each scenario, we report: (i) the ratio cases : controls by ethnicity; (ii) the unadjusted genomic control; (iii) the genomic control obtained correcting for the first three eigenvectors from linear PCA; and (iv) the genomic control obtained correcting for the first three eigenvectors obtained with PRINCALS.

As we can observe from Table 4.2, if the genomic control on the study sample is < 1.5, then adjusting for the first three eigenvectors from PCA gives a lower genomic control than adjusting for the first three eigenvectors from PRINCALS. On the other hand if the original genomic control is > 1.5, then correcting for PRINCALS yields a lower genomic control than correcting for PC. Hence, it looks like PCA performs better under mild PS; while under moderate/severe

Scenario	Case : Control distribution			Genomic control		
	CEU	CHB	YRI	Unadjusted	Adjusted PC	Adjusted PRINCALS
1	15:15	15:15	20:10	1.21825	1.09508	1.11958
2	15:15	15:15	25:5	1.63573	1.12909	1.11401
3	10:20	10:20	25:5	2.23657	1.08550	1.05369
4	5:25	16:14	17:13	1.45822	1.04396	1.02821
5	10:20	10:20	30:0	3.34007	1	1
6	15:15	10:20	20:10	1.41165	1	1.02751
7	10:20	10:20	20:10	1.53488	1.06174	1.0559
8	5:25	10:20	20:10	2.15665	1.09331	1.07075
9	16:14	14:16	20:10	1.41657	1.1199	1.12096
10	10:20	20:10	15:15	1.53609	1.07948	1.03279

Table 4.2: Scenarios analysis.

PS, PRINCALS outperforms PCA.

4.7 Conclusions

In this chapter, we have compared the results obtained using PCA and PRINCALS to detect and correct for PS. It emerges that the two methods yield similar scores (PC and dimensions) for markers that have a low/null genotypic variability across the study sample, whilst scores are different for markers with a high genotypic variability, thus suggesting that the two methods capture different intra-subject variability. This is confirmed by the Protest analysis where we find that the matrix of PC and the matrix of dimensions are statistically different and, further, by the scenarios analysis. In particular, we see that, as the level of PS increases, PRINCALS outperforms PCA. The reason for this can be explained by the fact that PCA provides quite a rigid framework by imposing an additive model of inheritance for each marker. Biology is full of non-linear phenomena and so is genetics. PRINCALS does not assume any model of inheritance and it weighs every marker differently depending on the distribution in the study population. This relaxing of the boundaries can perhaps better capture the real situation and hence produce a better correction.

Clearly, these results should be validated in multiple independent samples and by using a greater number of SNPs.

4.8 R script to extract eigenvectors

The following R script replicates entirely the PCA implemented in EIGENSTRAT and described in Section 2.3. We wrote the script in R in order to better control the algorithm (e.g. to avoid normalisation and to run the algorithm on the transformed, optimally scaled variables).

R script

```
#Import genotype data
#NB:the matrix is already transposed.  SNPs are coded as 0, 1, 2
#There are 1000 SNPs not in LD and 30 IDS: 30 CEU, 30 CHB, 30 YRI
G1<-read.table("matrix_1000SNPs_90ids_pruned.txt",header=T)
a<-G1[,1]
G<-G1[,-1]
row.names(G)<-a
G<-as.matrix(G)
rm(G1)
#Generate the variables number of row (nrowG)
nrowG <- nrow(G)
#Generate the variable number of columns (ncolG)
ncolG <- ncol(G)
#Find the vector of row means
#(row means as we are working on the SNPs)
m<-rowMeans(G)
m<-as.matrix(m)
#Centre the data matrix by subtracting the row mean from each
#entry in row i
one<-matrix(1,1,ncolG)
tmp<-m %*% one
G_centred<-G-tmp
#Normalise the data matrix
M<-ncolG*m
one_tmp <-matrix(1,nrowG ,1)
p_tmp<-(one_tmp +M)/(2*ncolG+2)
p<-sqrt(p_tmp*(1-p_tmp))
p<-as.matrix(p)
P<-p%*%one
#Divide every cell of the centred matrix for the corrisponding p_i
```

```
X<-G_centred/P
X<-as.matrix(X)
#Find the variance and covariance matrix S
S<-(t(X))%*%X
#Extract eigenvalues and eigenvectors
eigen(S)
eigenvalue<-eigen(S)$values
eigenvector<-eigen(S)$vectors
#Scree Plot (to determine the significant principal components)
plot(eigenvalue, type="b", main="Scree Diagram",
xlab="Number of components", ylab="Eigenvalues")
#Plot the significant axis of variation (eigenvectors)
plot(eigenvector[,1], eigenvector[,2], main="eigen1 vs.  eigen2",
xlab="eigenvactor 1", ylab="eigenvector 2", pch=18, col="black")
text(eigenvector[,1],eigenvector[,2],row.names(eigenvector),
cex=0.6, pos=4, col="black")
plot(eigenvector[,1],eigenvector[,3],main="eigen1 vs.  eigen3",
xlab="eigenvector 1", ylab="eigenvector 3", pch=18, col="black")
text(eigenvector[,1],eigenvector[,3],row.names(eigenvector),
cex=0.6, pos=4, col="black")
plot(eigenvector[,2],eigenvector[,3], main="eigen2 vs.  eigen3",
xlab="eigenvector 2", ylab="eigenvector 3", pch=18, col="black")
text(eigenvector[,2],eigenvector[,3],row.names(eigenvector),
cex=0.6, pos=4, col="black")
eigenvector<-as.matrix(eigenvector)
#Compute the matrix of PC
C=X%*%eigenvector

#Eigenvalue decomposition omitting the normalisation step
#Find the variance and covariance matrix (S_centred) from the
#centred matrix
S_centred<-as.matrix(G_centred)
S_centred<-(t(G_centred))%*%G_centred
#Extract eigenvalues and eigenvectors from S_centred
eigen(S_centred)
eigenvalue_centred<-eigen(S_centred)$values
eigenvector_centred<-eigen(S_centred)$vectors
```

```
eigenvector_centred<-as.data.frame (eigenvector_centred)
#Scree Plot (to determine the significant principal components)
plot(eigenvalue_centred,type="b",main="Scree Diagram",
xlab="Number of components",ylab="Eigenvalues")
#Plot the significant axis of variation
plot(eigenvector_centred[,1],eigenvector_centred[,2],
main="eigenvector 1 vs.  eigenvector 2",
xlab="eigenvector 1", ylab="eigenvector 2", pch=18, col="black")
text(eigenvector_centred[,1],eigenvector_centred[,2],
row.names(eigenvector_centred),cex=0.6, pos=4, col="black")
plot(eigenvector_centred[,1],eigenvector_centred[,3],
main="eigenvector 1 vs.  eigenvector 3",
xlab="eigenvector 1", ylab="eigenvector 3", pch=18, col="black")
text(eigenvector_centred[,1],eigenvector_centred[,3],
row.names(eigenvector_centred), cex=0.6, pos=4, col="black")
plot(eigenvector_centred[,2],eigenvector_centred[,3],
main="eigenvector 2 vs.  eigenvector 3",
xlab="eigenvector 2", ylab="eigenvector 3", pch=18, col="black")
text(eigenvector_centrato[,2],eigenvector_centred[,3],
row.names(eigenvector_centred), cex=0.6, pos=4, col="black")
eigenvector_centred<-as.matrix(eigenvector_centred)
#Find the matrix of PC
C_centred=G_centred%*%eigenvector_centred

#Eigenvalue decomposition on the transformed variables
#Import the transformed optimally scaled variables obtained from SAS
proc prinqual
A<-as.matrix(read.table("princals_1000snps_90ids.txt",header=T))
#Transpose matrix A
A_t<-t(A)
#Generate the variables number of row (nrowA_t)
nrowA_t <- nrow(A_t)
#Generate the variables number of columns (ncolA_t)
ncolA_t <- ncol(A_t)
#Find the vector of row means
m<-rowMeans(A_t)
m<-as.matrix(m)
```

```
#Centre A_t by subtracting the row mean from each
#entry in row i
one<-matrix(1,1,ncolA_t)
tmp<-m %*% one
A_t_centred<-A_t-tmp
#Find the variance and covariance matrix from A_t_centred
S_ordinal<-(t(A_t_centred))%*%A_t_centred
#Extract eigenvalues and eigenvectors from S_ordinal
eigen(S_ordinal)
eigenval_ordinal<-eigen(S_ordinal)$values
eigenvec_ordinal<-eigen(S_ordinal)$vectors
eigenvec_ordinal<-as.data.frame(eigenvec_ordinal)
#Scree Plot (to determine the significant dimensions)
plot(eigenvalue_ordinal,type="b",main="Scree Diagram",
xlab="Number of components",ylab="Eigenvalues")
#Plot the significant axis of variation
plot(eigenvec_ordinal[,1],eigenvec_ordinal[,2],
main="eigenvector 1 vs eigenvector 2",
xlab="eigenvector 1", ylab="eigenvector 2", pch=18, col="black")
text(eigenvec_ordinal[,1],eigenvec_ordinal[,2],
row.names(eigenvec_ordinal), cex=0.6, pos=4, col="black")
plot(eigenvec_ordinal[,1],eigenvec_ordinal[,3] ,
main="eigenvector 1 vs eigenvector 3",
xlab="eigenvector 1", ylab="eigenvector 3", pch=18, col="black")
text(eigenvec_ordinal[,1],eigenvec_ordinal[,3],
row.names(eigenvec_ordinal), cex=0.6, pos=4, col="black")
plot(eigenvec_ordinal[,2],eigenvec_ordinal[,3],
main="eigenvector 2 vs eigenvector 3",
xlab="eigenvector 2", ylab="eigenvector 3", pch=18, col="black")
text(eigenvec_ordinal[,2],eigenvec_ordinal[,3],
row.names(eigenvec_ordinal), cex=0.6, pos=4, col="black")
eigenvec_ordinal<-as.matrix(eigenvec_ordinal)
#Find the matrix of dimensions
D=A_t_centred%*%eigenvec_ordinal
#Plot the first PC (centred) vs first dimension
plot(C_centred[,1],D[,1],
main="First PC vs First dimension", xlab="First PC ",
ylab="First Dimension", pch=18, col="black")
text(C_centred[,1],D[,1], row.names(D), cex=0.6, pos=4, col="blue")
```

Conclusion

In this thesis we have applied non-linear principal components analysis (PRINCALS) to correct for the presence of population stratification in genome-wide association studies for complex diseases. The approach adopted differs from previous ones in that SNPs are not considered as quantitative variables, but they are instead treated as ordinal qualitative variables. This implies that there is an order between homozygous wild type, heterozygous and homozygous mutant, but that the distance between each pair is not necessarily the same. As a consequence, we no longer assume an additive model of inheritance. Instead, different models of inheritance are allowed. Our approach is thus more flexible and can potentially capture some information that linear PCA, a "gold standard" in this field, misses out. Indeed, with PRINCALS, each marker is weighted differently depending on its distribution in the study population. As biology is full of non-linear phenomena, this relaxing of the boundaries can perhaps better capture the real situation and hence produce a better correction. When we compare the performances of PRINCALS and PCA, we find that the two methods yield similar scores for markers with a low/null genotypic variability across the study sample, while scores differ as the level of genotypic variability increases. This suggests that the two methods capture intra-subject variability differently. Procrustes analysis and scenarios analysis confirm this. Indeed, the matrix of PC and the matrix of dimensions are shown to be statistically different by the test PROTEST and, in the scenarios analysis, we find that, as the level of PS increases, PRINCALS appears to outperform PCA.

As previously mentioned, our results should be validated, perhaps also with a simulation, in multiple independent samples, increasing the number of markers and of individuals.

There are interesting areas of research that can be further investigated. In particular, to our knowledge, there is not yet a validated methodology, even for PCA, to choose the optimum number of components (axes of variation) to keep in the model when adjusting for PS. The rule of thumb is to blindly use the first 10 axes of variation as covariates in the regression model. However, this can over correct and produce many false negatives. The other criterion of keeping in the model those axes of variation that explain a certain fraction of the total variance, can miss out the “intermediate” components. When designing a GWAS, much attention is put towards the homogeneity of the study sample with respect to ethnicity. The first axes of variation show macro-ethnic differences. If the study sample is homogeneous, there should not be many individuals that display macro-ethnic differences and typically those are removed from the analysis. Intermediate components are then responsible of PS as, event though they do not explain very much of the total variance, they are crucial in identifying minor ethnic variability.

We are thinking to tackle the problem using cross-validation methodologies. This is our current work in progress.

Appendix A

Matrix Algebra

This appendix gives a summary of some of the basic definitions and results in matrix algebra. It is designed to be a convenient source of reference for Chapters 2 and 3.

Matrix

A matrix is a rectangular array of numbers. We say a matrix A is n x m if it has n rows and m columns.

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1m} \\ a_{21} & a_{22} & ... & a_{2m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ a_{n1} & a_{n2} & ... & a_{nm} \end{pmatrix} = (a_{ij}).$$

Operations on matrices

Given two n x m matrices A= (a_{ij}) and $B = (b_{ij})$, and and m x p matrix C, then

1. $A + B = (a_{ij} + b_{ij})$;

2. $A - B = (a_{ij} - b_{ij})$;

3. $\lambda A = (\lambda a_{ij})$, where λ is a scalar;

4. the matrix product AC is the n x p matrix D, given by $d_{ik} = \sum_{j=1}^{m} a_{ij} c_{jk}$.

Transposed Matrix

The transposed of a matrix A, A' is obtained by interchanging rows and columns

$$A' = \begin{pmatrix} a_{11} & a_{21} & ... & a_{n1} \\ a_{12} & a_{22} & ... & a_{n2} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ a_{1m} & a_{2m} & ... & a_{nm} \end{pmatrix}.$$

Properties:

1. $(A')' = A$.
2. $(A+B)' = A' + B'$.
3. $(AB)' = B'A'$.

Identity matrix

The n x n identity matrix I has entries

$$s_{ij} = \begin{cases} 1 & \text{if} \quad i = j; \\ \\ 0 & \text{if} \quad i \neq j. \end{cases}$$

Inverse matrix

Given an n x n matrix A, the n x n matrix B is an inverse of A if

$$AB = I = BA.$$

In this case we write

$$B = A^{-1}.$$

If A has an inverse we say that it is invertible. If not, we say it is singular.

Diagonal matrix

A square n x n matrix D is a diagonal matrix if $d_{ij} = 0 \; \forall \; i \neq j$.

Properties:

1. $A' = A$.

Trace

Given an n x n matrix A, the trace of A is defined as

$$\text{tr}(A) = \sum_{i=1}^{n} a_{ii}.$$

Some Properties:

1. $\text{tr}(A + B) = \text{tr}(A) + tr(B)$;

2. $\text{tr}(cA) = c\,\text{tr}(A)$, where c is a scalar;

3. tr (AB) = tr (BA).

Rank

The rank of an n x n matrix A is the maximum number of linearly independent rows (columns) in A.

Determinant

Given a 2 x 2 matrix A, the determinat of A is given by

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}.$$

Given an n x n matrix A,the determinat of A is given by

$$\det(A) = \sum_{j=1}^{n} (-1)^{1+j} a_{1j} \det(A_{ij}).$$

Some properties:

1. An n x n matrix A is non-singular if $\det(A) \neq 0$. Otherwise it is singular.

2. $\det(AB) = \det(A)\det(B)$.

3. $\det(A') = \det(A)$.

4. $\det(A)^{-1} = (\det(A))^{-1}$.

Adjoint Matrix

Let A be a square matrix. Then the adjoint of A, Adj(A), is the transpose of the matrix obtained from A by replacing each entry by its cofactor, i.e. letting C_{ij} be the ij-cofactor of A, the ij-entry of Adj(A) is C_{ij}.

Eigenvector

Given an n x n matrix A, an eigenvector of A is a non-zero vector $\mathbf{u} \in \mathbb{R}^n$ such that $A\mathbf{u} = \lambda\mathbf{u}$ for some $\lambda \in \mathbb{R}$. We say that λ is the **eigenvalue** of the eigenvector $\mathbf{u}$.

Computation The eigenvalue equation can be expressed as

$$A\mathbf{u} - \lambda I\mathbf{u} = 0,$$

where I is the identity matrix.
This can be rearranged to

$$(A - \lambda I)\mathbf{u} = 0.$$

If $(A-\lambda I)$ were invertible, then both sides of the equation coul be left multiplied by the inverse to obtain the trivial solution $\mathbf{u} = 0$. Thus we require $(A - \lambda I)$ to be singular. This amount to put the determinant equal to zero:

$$\det(A - \lambda I) = 0.$$

The determinant requirement is called the **characteristic equation** of A, and the left-hand side is called the **characteristic polynomial**. When expanded, this gives a polynomial equation for λ.

Some properties of eigenvalues Given an n x n matrix A with eigenvalues λ_i. Then

1. $\mathrm{tr}(A) = \lambda_1 + \lambda_2 + ... + \lambda_n$;
2. $\det(A) = \lambda_1\lambda_2 \cdots \lambda_n$.

Singular value decomposition theorem

Given an n x m matrix A of rank p, A can be written as

$$A = ULV',$$

where the n x p matrix U and the p x m matrix V are column orthonormal matrices ($U'U = V'V = I$) and L is a diagonal matrix with positive elements.

Derivatives of linear functions

$$\frac{\partial}{\partial x}(a'x) = \frac{\partial}{\partial x}(x'a) = a'.$$

$$\frac{\partial}{\partial x}(Ax) = A.$$

$$\frac{\partial}{\partial x'}(Ax) = A'.$$

Derivatives of quadratic functions

$$\frac{\partial}{\partial x}(x'Ax) = x'(A' + A).$$

Derivatives of matrix trace

$$\frac{\partial}{\partial X}tr(AXB) = \frac{\partial}{\partial X}tr(B'X'A') = BA.$$

Vector Space

A vector space consists of a given set of objects and two given operations (addition and scalar multiplication) defined on that set and satisfying certain special properties.

Linearly Dependent

A set of vectors $\{\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_n}\}$ in a vector space $\mathbf{V}$ is linearly dependent if either

- The vector equation $c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{0}$ has a solution with at least one $c_i \neq 0$.

or equivalently

- One of its vectors can be written as a linear combination of the others.

A set of vectors is linearly independent if and only if it is not linearly dependent.

Spans

A set of vectors $\Im$ in a vector space $\mathbf{V}$ spans $\mathbf{V}$ if every vector in $\mathbf{V}$ can be written as a finite linear combination of those in $\Im$.

Basis

A set of vectors $\Im$ in a vector space $\mathbf{V}$ is a basis for $\mathbf{V}$ if $\Im$ is linearly independent and spans $\mathbf{V}$.

Inner product space

Let $\mathbf{V}$ be a vector space over the real numbers. An inner product on $\mathbf{V}$ is a function that associates a real number, denoted (u, v), with every pair of vectors u and v in V such that:

1. $(\mathbf{u},\mathbf{v}) = (\mathbf{v},\mathbf{u})$;
2. $(\mathbf{u}, \mathbf{v} + \mathbf{w}) = (\mathbf{u}, \mathbf{v}) + (\mathbf{u}, \mathbf{w})$;
3. $c(\mathbf{u}, \mathbf{v}) = (c\mathbf{u}, \mathbf{v})$;
4. $(\mathbf{u}, \mathbf{v}) \geq 0$; $(\mathbf{u}, \mathbf{u}) = 0$ if and only if $\mathbf{u} = 0$.

The vector space $\mathbf{V}$, together with the inner product is called an inner product space.

Norm

Given an inner product space, the norm of $\mathbf{u}$, denoted $||\mathbf{u}||$, is given by $||\mathbf{u}|| = \sqrt{(\mathbf{u}, \mathbf{u})}$.

Orthogonal and orthonormal

Vectors $\mathbf{u}$ and $\mathbf{v}$ in an inner product space are orthogonal if $(\mathbf{u}, \mathbf{v}) = 0$. A set of vectors $\Im$ in an inner product space is orthogonal if every pair of vectors in $\Im$ is orthogonal; an orthogonal set $\Im$ is orthonormal if every vector in $\Im$ has norm equal to 1.

Gram-Schmidt Process

Let V be an inner product space with basis $\Im = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$. Then by the Gram-Schmidt process we can construct an orthonormal basis from the $\mathbf{u}_i$
Let $\Im = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ be a basis for an inner product space $\mathbf{V}$. Let $\Im' = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be defined as follows:

$$\begin{aligned}
\mathbf{v}_1 &= \mathbf{u}_1; \\
\mathbf{v}_2 &= \mathbf{u}_2 - \frac{(\mathbf{u}_2, \mathbf{v}_1)}{(\mathbf{v}_1, \mathbf{v}_1)}\mathbf{v}_1; \\
\mathbf{v}_3 &= \mathbf{u}_3 - \frac{(\mathbf{u}_3, \mathbf{v}_1)}{(\mathbf{v}_1, \mathbf{v}_1)}\mathbf{v}_1 - \frac{(\mathbf{u}_2, \mathbf{v}_2)}{(\mathbf{v}_2, \mathbf{v}_2)}\mathbf{v}_2; \\
&\vdots \\
\mathbf{v}_n &= \mathbf{u}_n - \frac{(\mathbf{u}_n, \mathbf{v}_1)}{(\mathbf{v}_1, \mathbf{v}_1)}\mathbf{v}_1 - \frac{(\mathbf{u}_n, \mathbf{v}_2)}{(\mathbf{v}_2, \mathbf{v}_2)}\mathbf{v}_2 - \cdots - \frac{(\mathbf{u}_n, \mathbf{v}_{n-1})}{(\mathbf{v}_{n-1}, \mathbf{v}_{n-1})}\mathbf{v}_{n-1}.
\end{aligned}$$

Then the set $\Im'$ is an orthogonal basis for $\mathbf{V}$. An orthonormal basis for $\mathbf{V}$ is given by $\Im'' = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n\}$, where

$$\mathbf{w}_i = \frac{1}{||\mathbf{v}_i||}\mathbf{v}_i \ \ \forall i = 1, \ldots, n.$$

Appendix B

Java Script to recode SNPs and transpose the data matrix

We now report the Java Script which we designed, with the help of a computer expert, to prepare data to run an eigenvalue decomposition on the data covariance matrix.

The main script (Ped Transformer) calls all the others. The input files are the PED and the MAP file which come out from Illumina. The output file is the transposed matrix which can then be imported into R.

PED file

The PED file is a white-space (space or tab) delimited file: the first six columns are mandatory: Family ID, Individual ID, Paternal ID, Maternal ID, Sex (1=male; 2=female; other=unknown), Phenotype. A PED file must have 1 and only 1 phenotype in the sixth column. The phenotype can be either a quantitative trait or an affection status column. Affection status, by default, should be coded: 0 if missing, 1 if unaffected and 2 if affected. Genotypes (column 7 onwards) should also be white-space delimited; they can be any character (e.g. 1,2,3,4 or A,C,G,T or 1,2 or anything else) except 0 which is, by default, the missing genotype character. Our Java script is written for genotypes coded as 11, 12 and 22. If genotypes are in other format, they can be recoded using the `--recode12` Plink [104] option. All markers should be biallelic. All SNPs (whether haploid or not) must have two alleles specified. Either both alleles should be missing (i.e. 0) or neither. No header row should be given.

MAP file

The MAP file contains exactly 4 columns: chromosome (1-22, X, Y or 0 if unplaced), rs# or snp identifier, genetic distance (morgans) and base-pair position (bp units). By default, each line of the MAP file describes a single marker.

Ped Transposer

```
import genotype.AlleleToGenotypeFileTransformer;
import genotype.MapTrasposer;
import java.io.BufferedReader;
import java.io.BufferedWriter;
import java.io.File;
import java.io.FileReader;
import java.io.FileWriter;
import matrix.TextTraspose;
public class PedTransposer {
public static void main(final String[] args) throws Exception {
// String inputFilename = "nameoffile";
String inputFilename = "pca.recode";

/*
* 1.  Code genotype, row by row, on a new file:  out.gen
* 2.  Insert labels as first row of out.gen FID, IID, FA, MO, SEX,

STATUS and then the second transposed column of the MAP file
* 3.  Transpose in the file out.txt
* (4.  Replace MISSING values with a space (from row 7 onward))
*/


final BufferedReader pedInput = new BufferedReader(new FileReader
(new File("input", inputFilename + ".ped")));
final File output = new File("output", inputFilename + ".gen");
output.createNewFile();
```

```
final BufferedWriter genotypeWriter = new BufferedWriter
(new FileWriter(output));
BufferedReader genotypeReader = null;
BufferedWriter srcWriter = null;
BufferedReader inputMapReader = null;
try
{
// 1 new AlleleToGenotypeFileTransformer(pedInput,
genotypeWriter).run();
genotypeWriter.flush();
System.out.println("STEP 1 - COMPLETE");

// 2 genotypeReader = new BufferedReader(new FileReader(output));
inputMapReader = new BufferedReader(new FileReader
(new File("input", inputFilename + ".map")));
final File src = new File("output", inputFilename + ".src");
srcWriter = new BufferedWriter(new FileWriter(src));
final String columns = new MapTrasposer
(inputMapReader).extract();
final String header = new StringBuilder().append("FID
IID FA MO SEX STATUS ").append(columns).toString();
srcWriter.write(header);
String line = genotypeReader.readLine();
int counter = 0;
while (line != null) { srcWriter.write(line);
srcWriter.write("\n");
line = genotypeReader.readLine();
System.out.println("scrivo il src (gen + header)"
+ counter++);
}
srcWriter.flush();
System.out.println("STEP 2 - COMPLETE");

// First column has as header:  FIDIID,
// 3
String resultFilename = inputFilename + ".result";
TextTraspose textTraspose = new TextTraspose(src,
resultFilename);
textTraspose.run();
} catch (final Exception e) { e.printStackTrace(); }
finally {
srcWriter.close();
genotypeReader.close();
inputMapReader.close();
pedInput.close();
genotypeWriter.close(); }
}
public static final String FILENAME = "input/nameoffile.ped";
public static final String FILEMAP = "input/nameoffile.map"; }
```

Allele to genotype

```
import java.io.BufferedReader;
import java.io.BufferedWriter;
import java.io.IOException;
public class AlleleToGenotypeFileTransformer {
private final BufferedReader reader;
private final BufferedWriter writer;
public AlleleToGenotypeFileTransformer(final BufferedReader reader,

final BufferedWriter writer) {
this.reader = reader;
this.writer = writer;
}
public void run() throws IOException {
String personaInAllele = reader.readLine();
int counter = 0;
while (personInAllele != null) {
final CodeGenotype CodeGenotype = new
CodeGenotype(personInAllele);
writer.write(codificaGenotipo.transform());
writer.write("\n");
personaInAllele = reader.readLine();
System.out.println("code genotype line:  " +
counter++); } } }
```

Code genotype

```
package genotype;
public class CodeGenoype {
private final String input;
public CodeGenoype(final String input) {
this.input = input;
}
public String transform() {
final String[] split = input.split(" ");
final StringBuilder builder = new StringBuilder();
for (int i = 0; i < 6; i++)
{ builder.append(split[i]);
builder.append(" ");
}
for (int j = 6; j < split.length; j++) {
final Snp snp = new Snp(split[j], split[j + 1]);
builder.append(snp.getGenotype());
builder.append(" ");
j++;
}
return builder.toString();
}
}
```

Map transposer

```
package genotype;
import java.io.BufferedReader;
import java.io.IOException;
public class MapTrasposer {
private final BufferedReader reader;
public MapTrasposer(final BufferedReader reader) {
this.reader = reader;
}
public String extract() throws IOException {
String line = reader.readLine();
final StringBuilder sb = new StringBuilder();
while (line != null) {
final String[] columns = line.split("\t");
sb.append(columns[1]); sb.append(" ");
line = reader.readLine(); }
return sb.toString(); }
}
```

SNP

```
package genotype;
public class Snp {
private static final String _0 = "0";
private static final String _1 = "1";
private static final String _2 = "2";
private static final String MISSING = ".";
private int firstAllele; private int secondAllele;
public Snp(final String firstAllele, final String
secondAllele) {
this.firstAllele = Integer.parseInt(firstAllele);
this.secondAllele = Integer.parseInt(secondAllele);
if (this.firstAllele > 2) {
this.firstAllele = 0;
}
if (this.secondAllele > 2) {
this.secondAllele = 0;
}
}
public String getGenotipo() {
switch (firstAllele) {
case 0:
```

```
return MISSING;
case 1:
if (secondAllele == 0) {
return MISSING;
} else if (secondAllele == 1) {
return _0;
} else { return _1; }
case 2:
if (secondAllele == 0) {
return MISSING;
} else if (secondAllele == 1) {
return _1;
} else {
return _2;
}
}
return
MISSING;
}
}
```

Bibliography

[1] M. J. Khoury, R. Davis, M. Gwinn, M. L. Lindegren, and P. Yoon. Do we need genomic research for the prevention of common diseases with environmental causes? *Am. J. Epidemiol.*, 161:799–805, May 2005.

[2] K. R. Merikangas and N. Risch. Genomic priorities and public health. *Science*, 302:599–601, Oct 2003.

[3] L. J. Palmer and L. R. Cardon. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet*, 366:1223–1234, Oct 2005.

[4] C. Newton-Cheh, T. Johnson, V. Gateva, M. D. Tobin, M. Bochud, L. Coin, S. S. Najjar, J. H. Zhao, S. C. Heath, S. Eyheramendy, K. Papadakis, B. F. Voight, L. J. Scott, F. Zhang, M. Farrall, T. Tanaka, C. Wallace, J. C. Chambers, K. T. Khaw, P. Nilsson, P. van der Harst, S. Polidoro, D. E. Grobbee, N. C. Onland-Moret, M. L. Bots, L. V. Wain, K. S. Elliott, A. Teumer, J. Luan, G. Lucas, J. Kuusisto, P. R. Burton, D. Hadley, W. L. McArdle, M. Brown, A. Dominiczak, S. J. Newhouse, N. J. Samani, J. Webster, E. Zeggini, J. S. Beckmann, S. Bergmann, N. Lim, K. Song, P. Vollenweider, G. Waeber, D. M. Waterworth, X. Yuan, L. Groop, M. Orho-Melander, A. Allione, A. Di Gregorio, S. Guarrera, S. Panico, F. Ricceri, V. Romanazzi, C. Sacerdote, P. Vineis, I. Barroso, M. S. Sandhu, R. N. Luben, G. J. Crawford, P. Jousilahti, M. Perola, M. Boehnke, L. L. Bonnycastle, F. S. Collins, A. U. Jackson, K. L. Mohlke, H. M. Stringham, T. T. Valle, C. J. Willer, R. N. Bergman, M. A. Morken, A. Doring, C. Gieger, T. Illig, T. Meitinger, E. Org, A. Pfeufer, H. E. Wichmann, S. Kathiresan, J. Marrugat, C. J. O'Donnell, S. M. Schwartz, D. S. Siscovick, I. Subirana, N. B. Freimer, A. L. Hartikainen, M. I. McCarthy, P. F. O'Reilly, L. Peltonen, A. Pouta, P. E. de Jong,

H. Snieder, W. H. van Gilst, R. Clarke, A. Goel, A. Hamsten, J. F. Peden, U. Seedorf, A. C. Syvanen, G. Tognoni, E. G. Lakatta, S. Sanna, P. Scheet, D. Schlessinger, A. Scuteri, M. Dorr, F. Ernst, S. B. Felix, G. Homuth, R. Lorbeer, T. Reffelmann, R. Rettig, U. Volker, P. Galan, I. G. Gut, S. Hercberg, G. M. Lathrop, D. Zelenika, P. Deloukas, N. Soranzo, F. M. Williams, G. Zhai, V. Salomaa, M. Laakso, R. Elosua, N. G. Forouhi, H. Volzke, C. S. Uiterwaal, Y. T. van der Schouw, M. E. Numans, G. Matullo, G. Navis, G. Berglund, S. A. Bingham, J. S. Kooner, J. M. Connell, S. Bandinelli, L. Ferrucci, H. Watkins, T. D. Spector, J. Tuomilehto, D. Altshuler, D. P. Strachan, M. Laan, P. Meneton, N. J. Wareham, M. Uda, M. R. Jarvelin, V. Mooser, O. Melander, R. J. Loos, P. Elliott, G. R. Abecasis, M. Caulfield, and P. B. Munroe. Genome-wide association study identifies eight loci associated with blood pressure. *Nat. Genet.*, May 2009.

[5] N. E. Morton. Significance levels in complex inheritance. *Am. J. Hum. Genet.*, 62:690–697, Mar 1998.

[6] A.W. Cowley. The genetic dissection of essential hypertension. *Nature Reviews Genetics*, 7(11):829–840, 2006.

[7] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani, J. A. Todd, P. Donnelly, J. C. Barrett, P. R. Burton, D. Davison, P. Donnelly, D. Easton, D. Evans, H. T. Leung, J. L. Marchini, A. P. Morris, C. C. Spencer, M. D. Tobin, L. R. Cardon, D. G. Clayton, A. P. Attwood, J. P. Boorman, B. Cant, U. Everson, J. M. Hussey, J. D. Jolley, A. S. Knight, K. Koch, E. Meech, S. Nutland, C. V. Prowse, H. E. Stevens, N. C. Taylor, G. R. Walters, N. M. Walker, N. A. Watkins, T. Winzer, J. A. Todd, W. H. Ouwehand, R. W. Jones, W. L. McArdle, S. M. Ring, D. P. Strachan, M. Pembrey, G. Breen, D. St Clair, S. Caesar, K. Gordon-Smith, L. Jones, C. Fraser, E. K. Green, D. Grozeva, M. L. Hamshere, P. A. Holmans, I. R. Jones, G. Kirov, V. Moskvina, I. Nikolov, M. C. O'Donovan, M. J. Owen, N. Craddock, D. A. Collier, A. Elkin, A. Farmer, R. Williamson, P. McGuffin, A. H. Young, I. N. Ferrier, S. G. Ball, A. J. Balmforth, J. H. Barrett, D. T. Bishop, M. M. Iles, A. Maqbool, N. Yuldasheva, A. S. Hall, P. S. Braund, P. R. Burton, R. J. Dixon, M. Mangino, S. Suzanne, M. D. Tobin, J. R. Thomp-

son, N. J. Samani, F. Bredin, M. Tremelling, M. Parkes, H. Drummond, C. W. Lees, E. R. Nimmo, J. Satsangi, S. A. Fisher, A. Forbes, C. M. Lewis, C. M. Onnie, N. J. Prescott, J. Sanderson, C. G. Mathew, J. Barbour, M. K. Mohiuddin, C. E. Todhunter, J. C. Mansfield, T. Ahmad, F. R. Cummings, D. P. Jewell, J. Webster, M. J. Brown, D. G. Clayton, G. M. Lathrop, J. Connell, A. Dominczak, N. J. Samani, C. A. Marcano, B. Burke, R. Dobson, J. Gungadoo, K. L. Lee, P. B. Munroe, S. J. Newhouse, A. Onipinla, C. Wallace, M. Xue, M. Caulfield, M. Farrall, A. Barton, I. N. Bruce, H. Donovan, S. Eyre, P. D. Gilbert, S. L. Hider, A. M. Hinks, S. L. John, C. Potter, A. J. Silman, D. P. Symmmons, W. Thomson, J. Worthington, D. G. Clayton, D. B. Dunger, S. Nutland, H. E. Stevens, N. M. Walker, B. Widmer, J. A. Todd, T. A. Frayling, R. M. Freathy, H. Lango, J. R. Perry, B. M. Shields, M. N. Weedon, A. T. Hattersley, G. A. Hitman, M. Walker, K. S. Elliott, C. J. Groves, C. M. Lindgren, N. W. Rayner, N. J. Timpson, E. Zeggini, M. I. McCarthy, M. Newport, G. Sirugo, E. Lyons, F. Vanberg, A. V. Hill, L. A. Bradbury, C. Farrar, J. J. Pointon, P. Wordsworth, M. A. Brown, J. A. Franklyn, J. M. Heward, M. J. Simmonds, S. C. Gough, S. Seal, M. R. Stratton, N. Rahman, M. Ban, A. Goris, S. J. Sawcer, A. Compston, D. Conway, M. Jallow, M. Newport, G. Sirugo, K. A. Rockett, D. P. Kwiatowski, S. J. Bumpstead, A. Chaney, K. Downes, M. J. Ghori, R. Gwilliam, S. E. Hunt, M. Inouye, A. Keniry, E. King, R. McGinnis, S. Potter, R. Ravindrarajah, P. Whittaker, C. Widden, D. Withers, P. Deloukas, H. T. Leung, S. Nutland, H. E. Stevens, N. M. Walker, J. A. Todd, D. Easton, D. G. Clayton, P. R. Burton, M. D. Tobin, J. C. Barrett, D. Evans, A. P. Morris, L. R. Cardon, N. J. Cardin, D. Davison, T. Ferreira, J. Pereira-Gale, I. B. Hallgrimsdottir, B. N. Howie, J. L. Marchini, C. C. Spencer, Z. Su, Y. Y. Teo, D. Vukcevic, P. Donnelly, D. Bentley, M. A. Brown, L. R. Gordon, M. Caulfield, D. G. Clayton, A. Compston, N. Craddock, P. Deloukas, P. Donnelly, M. Farrall, S. C. Gough, A. S. Hall, A. T. Hattersley, A. V. Hill, D. P. Kwiatkowski, C. Mathew, M. I. McCarthy, W. H. Ouwehand, M. Parkes, M. Pembrey, N. Rahman, N. J. Samani, M. R. Stratton, J. A. Todd, and J. Worthington. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, Jun 2007.

[8] X. Wu, D. Kan, M. Province, T. Quertermous, D. C. Rao, C. Chang,

T. H. Mosley, D. Curb, E. Boerwinkle, and R. S. Cooper. An updated meta-analysis of genome scans for hypertension and blood pressure in the NHLBI Family Blood Pressure Program (FBPP). *Am. J. Hypertens.*, 19:122–127, Jan 2006.

[9] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273:1516–1517, Sep 1996.

[10] K. E. Lohmueller, C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.*, 33:177–182, Feb 2003.

[11] A. Helgadottir, G. Thorleifsson, A. Manolescu, S. Gretarsdottir, T. Blondal, A. Jonasdottir, A. Jonasdottir, A. Sigurdsson, A. Baker, A. Palsson, G. Masson, D. F. Gudbjartsson, K. P. Magnusson, K. Andersen, A. I. Levey, V. M. Backman, S. Matthiasdottir, T. Jonsdottir, S. Palsson, H. Einarsdottir, S. Gunnarsdottir, A. Gylfason, V. Vaccarino, W. C. Hooper, M. P. Reilly, C. B. Granger, H. Austin, D. J. Rader, S. H. Shah, A. A. Quyyumi, J. R. Gulcher, G. Thorgeirsson, U. Thorsteinsdottir, A. Kong, and K. Stefansson. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science*, 316:1491–1493, Jun 2007.

[12] T. M. Frayling, N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, J. R. Perry, K. S. Elliott, H. Lango, N. W. Rayner, B. Shields, L. W. Harries, J. C. Barrett, S. Ellard, C. J. Groves, B. Knight, A. M. Patch, A. R. Ness, S. Ebrahim, D. A. Lawlor, S. M. Ring, Y. Ben-Shlomo, M. R. Jarvelin, U. Sovio, A. J. Bennett, D. Melzer, L. Ferrucci, R. J. Loos, I. Barroso, N. J. Wareham, F. Karpe, K. R. Owen, L. R. Cardon, M. Walker, G. A. Hitman, C. N. Palmer, A. S. Doney, A. D. Morris, G. D. Smith, A. T. Hattersley, and M. I. McCarthy. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316:889–894, May 2007.

[13] E. Zeggini, M. N. Weedon, C. M. Lindgren, T. M. Frayling, K. S. Elliott, H. Lango, N. J. Timpson, J. R. Perry, N. W. Rayner, R. M. Freathy, J. C. Barrett, B. Shields, A. P. Morris, S. Ellard, C. J. Groves, L. W. Harries, J. L. Marchini, K. R. Owen, B. Knight, L. R. Cardon, M. Walker, G. A.

Hitman, A. D. Morris, A. S. Doney, M. I. McCarthy, and A. T. Hattersley. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316:1336–1341, Jun 2007.

[14] L. J. Scott, K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, Y. Li, W. L. Duren, M. R. Erdos, H. M. Stringham, P. S. Chines, A. U. Jackson, L. Prokunina-Olsson, C. J. Ding, A. J. Swift, N. Narisu, T. Hu, R. Pruim, R. Xiao, X. Y. Li, K. N. Conneely, N. L. Riebow, A. G. Sprau, M. Tong, P. P. White, K. N. Hetrick, M. W. Barnhart, C. W. Bark, J. L. Goldstein, L. Watkins, F. Xiang, J. Saramies, T. A. Buchanan, R. M. Watanabe, T. T. Valle, L. Kinnunen, G. R. Abecasis, E. W. Pugh, K. F. Doheny, R. N. Bergman, J. Tuomilehto, F. S. Collins, and M. Boehnke. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, 316:1341–1345, Jun 2007.

[15] S. Padmanabhan, O. Melander, T. Johnson, A. M. Di Blasio, W. K. Lee, D. Gentilini, C. Hastie, C. Menni, M. C. Monti, C. Delles, S. Laing, B. Corso, G. Navis, A. Kwakernaak, P. van der Harst, M. Bochud, M. Maillard, M. Burnier, T. Hedner, S. Kjeldsen, B. Wahlstrand, M. Sjogren, C. Fava, M. Montagnana, E. Danese, O. Torffvit, B. Hedblad, H. Snieder, J. M. Connell, M. Brown, N. J. Samani, M. Farrall, G. Cesana, G. Mancia, S. Signorini, G. Grassi, S. Eyheramendy, H. E. Wichmann, M. Laan, D. P. Strachan, P. Sever, D. Shields, A. Stanton, P. Vollenweider, A. Teumer, H. Volzke, R. Rettig, C. Newton-Cheh, P. Arora, F. Zhang, N. Soranzo, T. Spector, G. Lucas, S. Kathiresan, D. Siscovick, J. Luan, R. J.F Loos, N.J Wareham, B. W Penninx, I. M Nolte, M. McBride, W. H. Miller, S. A. Nicklin, A. H Baker, D. Graham, A. McDonald, J. Pell, N. Sattar, P. Welsh, P. B. Munroe, M. Caulfield, A. Zanchetti, and A. F. Dominiczak. Genomewide association study of blood pressure extremes identifies variant in UMOD associated with hypertension. *Plos Genetics*, 6:e1001177, 2010.

[16] No authors listed. The International HapMap Project. *Nature*, 426:789–796, Dec 2003.

[17] No authors listed. A haplotype map of the human genome. *Nature*, 437:1299–1320, Oct 2005.

[18] J. N. Hirschhorn and M. J. Daly. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, 6:95–108, Feb 2005.

[19] J. K. Pritchard and N. A. Rosenberg. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.*, 65:220–228, Jul 1999.

[20] E. S. Lander and N. J. Schork. Genetic dissection of complex traits. *Science*, 265:2037–2048, Sep 1994.

[21] D. J. Balding. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.*, 7:781–791, Oct 2006.

[22] J. S. Barnholtz-Sloan, B. McEvoy, M. D. Shriver, and T. R. Rebbeck. Ancestry estimation and correction for population stratification in molecular epidemiologic association studies. *Cancer Epidemiol. Biomarkers Prev.*, 17:471–477, Mar 2008.

[23] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55:997–1004, Dec 1999.

[24] J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *Am. J. Hum. Genet.*, 67:170–181, Jul 2000.

[25] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38:904–909, Aug 2006.

[26] H. Zhao, T. R. Rebbeck, and N. Mitra. A propensity score approach to correction for bias due to population stratification using genetic and non-genetic factors. *Genet. Epidemiol.*, 33:679–690, Dec 2009.

[27] N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet.*, 2:e190, Dec 2006.

[28] F. Zhang, Y. Wang, and H. W. Deng. Comparison of population-based association study methods correcting for population stratification. *PLoS ONE*, 3:e3392, 2008.

[29] G. Michailidis and J. de Leeuw. The Gifi system of descriptive multivariate analysis. *Statistical Science*, 13(4):307–336, 1998.

[30] A. Gifi. *Non linear multivariate analysis.* Wiley, New York, 1990.

[31] P. R. Burton, M. D. Tobin, and J. L. Hopper. Key concepts in genetic epidemiology. *Lancet*, 366:941–951, 2005.

[32] M.C. Speer. Basic concepts in genetics. In J.L. Haines and M.A. Pericak-Vance, editors, *Approaches to gene mapping in complex human diseases*, pages 17–52. Wiley-Liss, 1998.

[33] C.T. Thomas. *Statistical methods in genetic epidemiology.* Oxford Univeristy Press, New York, 2004.

[34] L.H. Hartwell, L. Hood, M.L. Goldberg, A.E. Reynolds, R.C. Silver, and R.C. Veres. *Genetics - from gene to genome.* McGraw-Hill, Milan, 2004.

[35] M. Dawn Teare and J. H. Barrett. Genetic linkage studies. *Lancet*, 366:1036–1044, 2005.

[36] K. M. Weiss and J. D. Terwilliger. How many diseases does it take to map a gene with SNPs? *Nat. Genet.*, 26:151–157, Oct 2000.

[37] L. L. Cavalli-Sforza and M. W. Feldman. The application of molecular genetic approaches to the study of human evolution. *Nat. Genet.*, 33 Suppl:266–275, Mar 2003.

[38] D. C. Crawford and D. A. Nickerson. Definition and clinical importance of haplotypes. *Annu. Rev. Med.*, 56:303–320, 2005.

[39] M. J. Rieder, S. L. Taylor, A. G. Clark, and D. A. Nickerson. Sequence variation in the human angiotensin converting enzyme. *Nat. Genet.*, 22:59–62, May 1999.

[40] X. Zhu, C. A. McKenzie, T. Forrester, D. A. Nickerson, U. Broeckel, H. Schunkert, A. Doering, H. J. Jacob, R. S. Cooper, and M. J. Rieder. Localization of a small genomic region associated with elevated ACE. *Am. J. Hum. Genet.*, 67:1144–1153, Nov 2000.

[41] J. C. Stephens, J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya, S. E. Stanley, R. Jiang, C. J. Messer, A. Chew, J. H. Han, J. Duan, J. L. Carr, M. S. Lee, B. Koshy, A. M. Kumar, G. Zhang, W. R. Newell, A. Windemuth, C. Xu, T. S. Kalbfleisch, S. L. Shaner, K. Arnold, V. Schulz, C. M.

Drysdale, K. Nandabalan, R. S. Judson, G. Ruano, and G. F. Vovis. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293:489–493, Jul 2001.

[42] M. K. Halushka, J. B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.*, 22:239–247, Jul 1999.

[43] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, 74:106–120, Jan 2004.

[44] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner,

S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291:1304–1351, Feb 2001.

[45] A. H. Klerkx, M. W. Tanck, J. J. Kastelein, H. O. Molhuizen, J. W. Jukema, A. H. Zwinderman, and J. A. Kuivenhoven. A polymorphism and a novel promoter variant are independently associated with CETP concentration. *Hum. Mol. Genet.*, 12:111–123, Jan 2003.

[46] S. Tokuhiro, R. Yamada, X. Chang, A. Suzuki, Y. Kochi, T. Sawada, M. Suzuki, M. Nagasaki, M. Ohtsuki, M. Ono, H. Furukawa, M. Nagashima, S. Yoshino, A. Mabuchi, A. Sekine, S. Saito, A. Takahashi, T. Tsunoda, Y. Nakamura, and K. Yamamoto. An intronic SNP in a RUNX1 binding site of SLC22A4, encoding an organic cation transporter, is associated with rheumatoid arthritis. *Nat. Genet.*, 35:341–348, Dec 2003.

[47] D. C. Betticher, N. Thatcher, H. J. Altermatt, P. Hoban, W. D. Ryder, and J. Heighway. Alternate splicing produces a novel cyclin D1 transcript.

Oncogene, 11:1005–1011, Sep 1995.

[48] J. Duan, M. S. Wainwright, J. M. Comeron, N. Saitou, A. R. Sanders, J. Gelernter, and P. V. Gejman. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum. Mol. Genet.*, 12:205–216, Feb 2003.

[49] A. F. Wright and N. D. Hastie. Complex genetic diseases: controversy over the Croesus code. *Genome Biol.*, 2:Comment2007, 2001.

[50] E. S. Lander. The new genomics: global views of biology. *Science*, 274:536–539, Oct 1996.

[51] D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *Trends Genet.*, 17:502–510, Sep 2001.

[52] G. Marth, G. Schuler, R. Yeh, R. Davenport, R. Agarwala, D. Church, S. Wheelan, J. Baker, M. Ward, M. Kholodov, L. Phan, E. Czabarka, J. Murvai, D. Cutler, S. Wooding, A. Rogers, A. Chakravarti, H. C. Harpending, P. Y. Kwok, and S. T. Sherry. Sequence variations in the public human genome data reflect a bottlenecked population history. *Proc. Natl. Acad. Sci. U.S.A.*, 100:376–381, Jan 2003.

[53] J. K. Pritchard and N. J. Cox. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum. Mol. Genet.*, 11:2417–2423, Oct 2002.

[54] H. J. Cordell and D. G. Clayton. Genetic association studies. *Lancet*, 366:1121–1131, 2005.

[55] S. Padmanabhan, O. Melander, C. Hastie, C. Menni, C. Delles, J. M. Connell, and A. F. Dominiczak. Hypertension and genome-wide association studies: combining high fidelity phenotyping and hypercontrols. *J. Hypertens.*, 26:1275–1281, Jul 2008.

[56] S. Padmanabhan, C. Menni, D. Prabhakaran, and A. F. Dominiczak. Discovering the genetic determinants of complex diseases. *Current science*, 97:385–391, Aug 2009.

[57] A.C. Guyton, T.G. Coleman, A.W. Cowley, J.F. Liard, R.A. Norman, and R.D. Manning. Systems analysis of arterial pressure regulation and hypertension. *Annals of Biomedical Engineering*, 1(2):254–281, 1972.

[58] O. A. Carretero and S. Oparil. Essential hypertension : part II: treatment. *Circulation*, 101:446–453, Feb 2000.

[59] J. G. Mongeau, P. Biron, and C. F. Sing. The influence of genetics and household environment upon the variability of normal blood pressure: the Montreal Adoption Survey. *Clin Exp Hypertens A*, 8:653–660, 1986.

[60] M. Feinleib, R. J. Garrison, R. Fabsitz, J. C. Christian, Z. Hrubec, N. O. Borhani, W. B. Kannel, R. Rosenman, J. T. Schwartz, and J. O. Wagner. The NHLBI twin study of cardiovascular disease risk factors: methodology and summary of results. *Am. J. Epidemiol.*, 106:284–285, Oct 1977.

[61] J.V. Neel. Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress" . *Am J Hum Genet*, 14:353–362, 1962.

[62] A. B. Weder. Evolution and hypertension. *Hypertension*, 49:260–265, Feb 2007.

[63] T. Nakajima, S. Wooding, T. Sakagami, M. Emi, K. Tokunaga, G. Tamiya, T. Ishigami, S. Umemura, B. Munkhbat, F. Jin, J. Guan-Jun, I. Hayasaka, T. Ishida, N. Saitou, K. Pavelka, J. M. Lalouel, L. B. Jorde, and I. Inoue. Natural selection and population history in the human angiotensinogen gene (AGT): 736 complete AGT sequences in chromosomes from around the world. *Am. J. Hum. Genet.*, 74:898–916, May 2004.

[64] J. H. Young, Y. P. Chang, J. D. Kim, J. P. Chretien, M. J. Klag, M. A. Levine, C. B. Ruff, N. Y. Wang, and A. Chakravarti. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS Genet.*, 1:e82, Dec 2005.

[65] H. R. Brunner and H. Gavras. Is the renin system necessary? *Am. J. Med.*, 69:739–745, Nov 1980.

[66] R. P. Lifton. Genetic dissection of human blood pressure variation: common pathways from rare phenotypes. *Harvey Lect.*, 100:71–101, 2004.

[67] S. J. Newhouse, C. Wallace, R. Dobson, C. Mein, J. Pembroke, M. Farrall, D. Clayton, M. Brown, N. Samani, A. Dominiczak, J. M. Connell, J. Webster, G. M. Lathrop, M. Caulfield, and P. B. Munroe. Haplotypes of the WNK1 gene associate with blood pressure variation in a severely hypertensive population from the British Genetics of Hypertension study. *Hum. Mol. Genet.*, 14:1805–1814, Jul 2005.

[68] F. H. Wilson, S. Disse-Nicodeme, K. A. Choate, K. Ishikawa, C. Nelson-Williams, I. Desitter, M. Gunel, D. V. Milford, G. W. Lipkin, J. M. Achard, M. P. Feely, B. Dussol, Y. Berland, R. J. Unwin, H. Mayan, D. B. Simon, Z. Farfel, X. Jeunemaitre, and R. P. Lifton. Human hypertension caused by mutations in WNK kinases. *Science*, 293:1107–1112, Aug 2001.

[69] M. Caulfield, P. Munroe, J. Pembroke, N. Samani, A. Dominiczak, M. Brown, N. Benjamin, J. Webster, P. Ratcliffe, S. O'Shea, J. Papp, E. Taylor, R. Dobson, J. Knight, S. Newhouse, J. Hooper, W. Lee, N. Brain, D. Clayton, G. M. Lathrop, M. Farrall, and J. Connell. Genome-wide mapping of human loci for essential hypertension. *Lancet*, 361:2118–2123, Jun 2003.

[70] S. Padmanabhan, C. Wallace, P. B. Munroe, R. Dobson, M. Brown, N. Samani, D. Clayton, M. Farrall, J. Webster, M. Lathrop, M. Caulfield, A. F. Dominiczak, and J. M. Connell. Chromosome 2p shows significant linkage to antihypertensive response in the British Genetics of Hypertension Study. *Hypertension*, 47:603–608, Mar 2006.

[71] L. Koivukoski, S. A. Fisher, T. Kanninen, C. M. Lewis, F. von Wowern, S. Hunt, S. L. Kardia, D. Levy, M. Perola, T. Rankinen, D. C. Rao, T. Rice, B. A. Thiel, and O. Melander. Meta-analysis of genome-wide scans for hypertension and blood pressure in Caucasians shows evidence of susceptibility regions on chromosomes 2 and 3. *Hum. Mol. Genet.*, 13:2325–2332, Oct 2004.

[72] C. C. Gu, S. C. Hunt, S. Kardia, S. T. Turner, A. Chakravarti, N. Schork, R. Olshen, D. Curb, C. Jaquish, E. Boerwinkle, and D. C. Rao. An investigation of genome-wide associations of hypertension with microsatellite markers in the family blood pressure program (FBPP). *Hum. Genet.*, 121:577–590, Jun 2007.

[73] A. F. Dominiczak, D. Graham, M. W. McBride, N. J. Brain, W. K. Lee, F. J. Charchar, M. Tomaszewski, C. Delles, and C. A. Hamilton. Corcoran Lecture. Cardiovascular genomics and oxidative stress. *Hypertension*, 45:636–642, Apr 2005.

[74] S. Padmanabhan, C. Menni, W. K. Lee, S. Laing, P. Brambilla, R. Sega, R. Perego, G. Grassi, G. Cesana, C. Delles, G. Mancia, and A. F. Do-

miniczak. The effects of sex and method of blood pressure measurement on genetic associations with blood pressure in the PAMELA study. *J. Hypertens.*, 28:465–477, Mar 2010.

[75] G. Mancia, R. Sega, C. Bravi, G. De Vito, F. Valagussa, G. Cesana, and A. Zanchetti. Ambulatory blood pressure normality: results from the PAMELA study. *J. Hypertens.*, 13:1377–1390, Dec 1995.

[76] J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, 69:124–137, Jul 2001.

[77] K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbol, S. M. Leal, S. Pasternak, D. A. Wheeler, T. D. Willis, F. Yu, H. Yang, C. Zeng, Y. Gao, H. Hu, W. Hu, C. Li, W. Lin, S. Liu, H. Pan, X. Tang, J. Wang, W. Wang, J. Yu, B. Zhang, Q. Zhang, H. Zhao, H. Zhao, J. Zhou, S. B. Gabriel, R. Barry, B. Blumenstiel, A. Camargo, M. Defelice, M. Faggart, M. Goyette, S. Gupta, J. Moore, H. Nguyen, R. C. Onofrio, M. Parkin, J. Roy, E. Stahl, E. Winchester, L. Ziaugra, D. Altshuler, Y. Shen, Z. Yao, W. Huang, X. Chu, Y. He, L. Jin, Y. Liu, Y. Shen, W. Sun, H. Wang, Y. Wang, Y. Wang, X. Xiong, L. Xu, M. M. Waye, S. K. Tsui, H. Xue, J. T. Wong, L. M. Galver, J. B. Fan, K. Gunderson, S. S. Murray, A. R. Oliphant, M. S. Chee, A. Montpetit, F. Chagnon, V. Ferretti, M. Leboeuf, J. F. Olivier, M. S. Phillips, S. Roumy, C. Sallee, A. Verner, T. J. Hudson, P. Y. Kwok, D. Cai, D. C. Koboldt, R. D. Miller, L. Pawlikowska, P. Taillon-Miller, M. Xiao, L. C. Tsui, W. Mak, Y. Q. Song, P. K. Tam, Y. Nakamura, T. Kawaguchi, T. Kitamoto, T. Morizono, A. Nagashima, Y. Ohnishi, A. Sekine, T. Tanaka, T. Tsunoda, P. Deloukas, C. P. Bird, M. Delgado, E. T. Dermitzakis, R. Gwilliam, S. Hunt, J. Morrison, D. Powell, B. E. Stranger, P. Whittaker, D. R. Bentley, M. J. Daly, P. I. de Bakker, J. Barrett, Y. R. Chretien, J. Maller, S. McCarroll, N. Patterson, I. Pe'er, A. Price, S. Purcell, D. J. Richter, P. Sabeti, R. Saxena, S. F. Schaffner, P. C. Sham, P. Varilly, D. Altshuler, L. D. Stein, L. Krishnan, A. V. Smith, M. K. Tello-Ruiz, G. A. Thorisson, A. Chakravarti, P. E. Chen, D. J. Cutler, C. S. Kashuk, S. Lin, G. R. Abecasis, W. Guan, Y. Li, H. M. Munro, Z. S. Qin, D. J. Thomas, G. McVean, A. Auton, L. Bottolo, N. Cardin, S. Eyheramendy, C. Freeman, J. Marchini, S. Myers, C. Spencer, M. Stephens, P. Donnelly, L. R. Cardon, G. Clarke, D. M.

Evans, A. P. Morris, B. S. Weir, T. Tsunoda, J. C. Mullikin, S. T. Sherry, M. Feolo, A. Skol, H. Zhang, C. Zeng, H. Zhao, I. Matsuda, Y. Fukushima, D. R. Macer, E. Suda, C. N. Rotimi, C. A. Adebamowo, I. Ajayi, T. Aniagwu, P. A. Marshall, C. Nkwodimmah, C. D. Royal, M. F. Leppert, M. Dixon, A. Peiffer, R. Qiu, A. Kent, K. Kato, N. Niikawa, I. F. Adewole, B. M. Knoppers, M. W. Foster, E. W. Clayton, J. Watkin, R. A. Gibbs, J. W. Belmont, D. Muzny, L. Nazareth, E. Sodergren, G. M. Weinstock, D. A. Wheeler, I. Yakub, S. B. Gabriel, R. C. Onofrio, D. J. Richter, L. Ziaugra, B. W. Birren, M. J. Daly, D. Altshuler, R. K. Wilson, L. L. Fulton, J. Rogers, J. Burton, N. P. Carter, C. M. Clee, M. Griffiths, M. C. Jones, K. McLay, R. W. Plumb, M. T. Ross, S. K. Sims, D. L. Willey, Z. Chen, H. Han, L. Kang, M. Godbout, J. C. Wallenburg, P. L'Archeveque, G. Bellemare, K. Saeki, H. Wang, D. An, H. Fu, Q. Li, Z. Wang, R. Wang, A. L. Holden, L. D. Brooks, J. E. McEwen, M. S. Guyer, V. O. Wang, J. L. Peterson, M. Shi, J. Spiegel, L. M. Sung, L. F. Zacharia, F. S. Collins, K. Kennedy, R. Jamieson, and J. Stewart. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861, Oct 2007.

[78] E. Jorgenson and J. S. Witte. A gene-centric approach to genome-wide association studies. *Nat. Rev. Genet.*, 7:885–891, Nov 2006.

[79] W. Y. Wang, B. J. Barratt, D. G. Clayton, and J. A. Todd. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.*, 6:109–118, Feb 2005.

[80] D. Levy, G. B. Ehret, K. Rice, G. C. Verwoert, L. J. Launer, A. Dehghan, N. L. Glazer, A. C. Morrison, A. D. Johnson, T. Aspelund, Y. Aulchenko, T. Lumley, A. Kottgen, R. S. Vasan, F. Rivadeneira, G. Eiriksdottir, X. Guo, D. E. Arking, G. F. Mitchell, F. U. Mattace-Raso, A. V. Smith, K. Taylor, R. B. Scharpf, S. J. Hwang, E. J. Sijbrands, J. Bis, T. B. Harris, S. K. Ganesh, C. J. O'Donnell, A. Hofman, J. I. Rotter, J. Coresh, E. J. Benjamin, A. G. Uitterlinden, G. Heiss, C. S. Fox, J. C. Witteman, E. Boerwinkle, T. J. Wang, V. Gudnason, M. G. Larson, A. Chakravarti, B. M. Psaty, and C. M. van Duijn. Genome-wide association study of blood pressure and hypertension. *Nat. Genet.*, May 2009.

[81] P. A. Doris. Hypertension genetics, single nucleotide polymorphisms, and the common disease:common variant hypothesis. *Hypertension*, 39:323–331, Feb 2002.

[82] C. Newton-Cheh and J. N. Hirschhorn. Genetic association studies of complex traits: design and analysis issues. *Mutat. Res.*, 573:54–69, Jun 2005.

[83] K. T. Zondervan and L. R. Cardon. The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.*, 5:89–100, Feb 2004.

[84] S. Wacholder, S. Chanock, M. Garcia-Closas, L. El Ghormli, and N. Rothman. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl. Cancer Inst.*, 96:434–442, Mar 2004.

[85] R. C. Culverhouse, B. K. Suarez, L. Beckmann, P. Chen, Y. S. Chen, Y. F. Chiu, J. Chang-Claude, A. Dempfle, R. Hein, R. Kazma, J. J. Lebrec, S. Lee, S. Lim, B. S. Maher, T. Park, H. Perdry, K. S. Wang, P. P. Wolkow, and W. Xu. Gene by environment interactions. *Genet. Epidemiol.*, 31 Suppl 1:68–74, 2007.

[86] R. Hein, L. Beckmann, and J. Chang-Claude. Sample size requirements for indirect association studies of gene-environment interactions (G x E). *Genet. Epidemiol.*, 32:235–245, Apr 2008.

[87] C. Kooperberg and M. Leblanc. Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet. Epidemiol.*, 32:255–263, Apr 2008.

[88] S. J. Chanock, T. Manolio, M. Boehnke, E. Boerwinkle, D. J. Hunter, G. Thomas, J. N. Hirschhorn, G. Abecasis, D. Altshuler, J. E. Bailey-Wilson, L. D. Brooks, L. R. Cardon, M. Daly, P. Donnelly, J. F. Fraumeni, N. B. Freimer, D. S. Gerhard, C. Gunter, A. E. Guttmacher, M. S. Guyer, E. L. Harris, J. Hoh, R. Hoover, C. A. Kong, K. R. Merikangas, C. C. Morton, L. J. Palmer, E. G. Phimister, J. P. Rice, J. Roberts, C. Rotimi, M. A. Tucker, K. J. Vogan, S. Wacholder, E. M. Wijsman, D. M. Winn, and F. S. Collins. Replicating genotype-phenotype associations. *Nature*, 447:655–660, Jun 2007.

[89] R. Saxena, B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. de Bakker, H. Chen, J. J. Roix, S. Kathiresan, J. N. Hirschhorn, M. J. Daly, T. E. Hughes, L. Groop, D. Altshuler, P. Almgren, J. C. Florez, J. Meyer, K. Ardlie, K. Bengtsson Bostrom, B. Isomaa, G. Lettre, U. Lindblad, H. N. Lyon, O. Melander, C. Newton-Cheh, P. Nilsson, M. Orho-Melander, L. Rastam, E. K. Speliotes, M. R. Taskinen, T. Tuomi, C. Guiducci, A. Berglund, J. Carlson, L. Gianniny, R. Hackett, L. Hall, J. Holmkvist, E. Laurila, M. Sjogren, M. Sterner, A. Surti, M. Svensson, M. Svensson, R. Tewhey, B. Blumenstiel, M. Parkin, M. Defelice, R. Barry, W. Brodeur, J. Camarata, N. Chia, M. Fava, J. Gibbons, B. Handsaker, C. Healy, K. Nguyen, C. Gates, C. Sougnez, D. Gage, M. Nizzari, S. B. Gabriel, G. W. Chirn, Q. Ma, H. Parikh, D. Richardson, D. Ricke, and S. Purcell. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316:1331–1336, Jun 2007.

[90] J. P. Ioannidis, N. A. Patsopoulos, and E. Evangelou. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS ONE*, 2:e841, 2007.

[91] L. Hansson, T. Hedner, P. Lund-Johansen, S. E. Kjeldsen, L. H. Lindholm, J. O. Syvertsen, J. Lanke, U. de Faire, B. Dahlof, and B. E. Karlberg. Randomised trial of effects of calcium antagonists compared with diuretics and beta-blockers on cardiovascular morbidity and mortality in hypertension: the Nordic Diltiazem (NORDIL) study. *Lancet*, 356:359–365, Jul 2000.

[92] G. Berglund, S. Elmstahl, L. Janzon, and S. A. Larsson. The Malmo Diet and Cancer Study. Design and feasibility. *J. Intern. Med.*, 233:45–51, Jan 1993.

[93] S. Kathiresan, O. Melander, D. Anevski, C. Guiducci, N. P. Burtt, C. Roos, J. N. Hirschhorn, G. Berglund, B. Hedblad, L. Groop, D. M. Altshuler, C. Newton-Cheh, and M. Orho-Melander. Polymorphisms associated with cholesterol and risk of cardiovascular events. *N. Engl. J. Med.*, 358:1240–1249, Mar 2008.

[94] R. Sega, R. Facchetti, M. Bombelli, G. Cesana, G. Corrao, G. Grassi, and G. Mancia. Prognostic value of ambulatory and home blood pressures compared with office blood pressure in the general population: follow-up results from the Pressioni Arteriose Monitorate e Loro Associazioni (PAMELA) study. *Circulation*, 111:1777–1783, Apr 2005.

[95] M. Ferrario, R. Sega, and G. Cesana. Lessons from the MONICA study in northern Italy. *J Hypertens Suppl*, 9:7–14, Dec 1991.

[96] V. Lyssenko, A. Jonsson, P. Almgren, N. Pulizzi, B. Isomaa, T. Tuomi, G. Berglund, D. Altshuler, P. Nilsson, and L. Groop. Clinical risk factors, DNA variants, and the development of type 2 diabetes. *N. Engl. J. Med.*, 359:2220–2232, Nov 2008.

[97] J. G. Smith, P. G. Platonov, B. Hedblad, G. Engstrom, and O. Melander. Atrial fibrillation in the Malmo Diet and Cancer study: a study of occurrence, risk factors and diagnostic validity. *Eur. J. Epidemiol.*, 25:95–102, Feb 2010.

[98] C. M. Licht, E. J. de Geus, A. Seldenrijk, H. P. van Hout, F. G. Zitman, R. van Dyck, and B. W. Penninx. Depression is associated with decreased blood pressure, but antidepressant use increases the risk for hypertension. *Hypertension*, 53:631–638, Apr 2009.

[99] A. S. Levey, J. P. Bosch, J. B. Lewis, T. Greene, N. Rogers, and D. Roth. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Ann. Intern. Med.*, 130:461–470, Mar 1999.

[100] F. W. Visser, A. H. Boonstra, A. Titia Lely, F. Boomsma, and G. Navis. Renal response to angiotensin II is blunted in sodium-sensitive normotensive men. *Am. J. Hypertens.*, 21:323–328, Mar 2008.

[101] F. W. Visser, J. H. Muntinga, R. A. Dierckx, and G. Navis. Feasibility and impact of the measurement of extracellular fluid volume simultaneous with GFR by 125I-iothalamate. *Clin J Am Soc Nephrol*, 3:1308–1315, Sep 2008.

[102] M. Bochud, J. A. Staessen, M. Maillard, M. J. Mazeko, T. Kuznetsova, A. Woodiwiss, T. Richart, G. Norton, L. Thijs, R. Elston, and M. Burnier. Ethnic differences in proximal and distal tubular sodium reabsorption are heritable in black and white populations. *J. Hypertens.*, 27:606–612, Mar 2009.

[103] M. Bochud, P. Bovet, P. Vollenweider, M. Maillard, F. Paccaud, G. Wandeler, A. Gabriel, and M. Burnier. Association between white-coat effect

and blunted dipping of nocturnal blood pressure. *Am. J. Hypertens.*, 22:1054–1061, Oct 2009.

[104] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81:559–575, Sep 2007.

[105] J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, 39:906–913, Jul 2007.

[106] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, 9:356–369, May 2008.

[107] M. D. Tobin, N. A. Sheehan, K. J. Scurrah, and P. R. Burton. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Stat Med*, 24:2911–2935, Oct 2005.

[108] N. Iwai, K. Kajimoto, Y. Kokubo, and H. Tomoike. Extensive genetic analysis of 10 candidate genes for hypertension in Japanese. *Hypertension*, 48:901–907, Nov 2006.

[109] A. Kottgen, N. L. Glazer, A. Dehghan, S. J. Hwang, R. Katz, M. Li, Q. Yang, V. Gudnason, L. J. Launer, T. B. Harris, A. V. Smith, D. E. Arking, B. C. Astor, E. Boerwinkle, G. B. Ehret, I. Ruczinski, R. B. Scharpf, Y. D. Ida Chen, I. H. de Boer, T. Haritunians, T. Lumley, M. Sarnak, D. Siscovick, E. J. Benjamin, D. Levy, A. Upadhyay, Y. S. Aulchenko, A. Hofman, F. Rivadeneira, A. G. Uitterlinden, C. M. van Duijn, D. I. Chasman, G. Pare, P. M. Ridker, W. H. Kao, J. C. Witteman, J. Coresh, M. G. Shlipak, and C. S. Fox. Multiple loci associated with indices of renal function and chronic kidney disease. *Nat. Genet.*, May 2009.

[110] S. Bachmann, R. Metzger, and B. Bunnemann. Tamm-Horsfall protein-mRNA synthesis is localized to the thick ascending limb of Henle's loop in rat kidney. *Histochemistry*, 94:517–523, 1990.

[111] N. Malagolini, D. Cavallone, and F. Serafini-Cessi. Intracellular transport, cell-surface exposure and release of recombinant Tamm-Horsfall glycoprotein. *Kidney Int.*, 52:1340–1350, Nov 1997.

[112] T. C. Hart, M. C. Gorry, P. S. Hart, A. S. Woodard, Z. Shihabi, J. Sandhu, B. Shirts, L. Xu, H. Zhu, M. M. Barmada, and A. J. Bleyer. Mutations of the UMOD gene are responsible for medullary cystic kidney disease 2 and familial juvenile hyperuricaemic nephropathy. *J. Med. Genet.*, 39:882–892, Dec 2002.

[113] L. Rampoldi, G. Caridi, D. Santon, F. Boaretto, I. Bernascone, G. Lamorte, R. Tardanico, M. Dagnino, G. Colussi, F. Scolari, G. M. Ghiggeri, A. Amoroso, and G. Casari. Allelism of MCKD, FJHN and GCKD caused by impairment of uromodulin export dynamics. *Hum. Mol. Genet.*, 12:3369–3384, Dec 2003.

[114] P. Vylet'al, M. Kublova, M. Kalbacova, K. Hodanova, V. Baresova, B. Stibarkova, J. Sikora, H. Halkova, J. Zivny, J. Majewski, A. Simmonds, J. P. Fryns, G. Venkat-Raman, M. Elleder, and S. Kmoch. Alterations of uromodulin biology: a common denominator of the genetically heterogeneous FJHN/MCKD syndrome. *Kidney Int.*, 70:1155–1169, Sep 2006.

[115] O. Torffvit, P. E. Jorgensen, A. L. Kamper, N. H. Holstein-Rathlou, P. P. Leyssac, S. S. Poulsen, and S. Strandgaard. Urinary excretion of Tamm-Horsfall protein and epidermal growth factor in chronic nephropathy. *Nephron*, 79:167–172, 1998.

[116] P. Zurbig, S. Decramer, M. Dakna, J. Jantos, D. M. Good, J. J. Coon, F. Bandin, H. Mischak, J. L. Bascands, and J. P. Schanstra. The human urinary proteome reveals high similarity between kidney aging and chronic kidney disease. *Proteomics*, 9:2108–2117, Apr 2009.

[117] J. Dulawa, F. Kokot, M. Kokot, and H. Pander. Urinary excretion of Tamm-Horsfall protein in normotensive and hypertensive elderly patients. *J Hum Hypertens*, 12:635–637, Sep 1998.

[118] O. Torffvit, C. D. Agardh, and T. Thulin. A study of Tamm-Horsfall protein excretion in hypertensive patients and type 1 diabetic patients. *Scand. J. Urol. Nephrol.*, 33:187–191, Jun 1999.

[119] W. Ji, J. N. Foo, B. J. O'Roak, H. Zhao, M. G. Larson, D. B. Simon, C. Newton-Cheh, M. W. State, D. Levy, and R. P. Lifton. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.*, 40:592–599, May 2008.

[120] J. A. Cutler, P. D. Sorlie, M. Wolz, T. Thom, L. E. Fields, and E. J. Roccella. Trends in hypertension prevalence, awareness, treatment, and control rates in United States adults between 1988-1994 and 1999-2004. *Hypertension*, 52:818–827, Nov 2008.

[121] C. Hastie, S. Padmanabhan, and A. F. Dominiczak. Genome-wide association studies of hypertension: light at the end of the tunnel. *International Journal of Hypertension*, 2010:doi:10.4061/2010/509581, 2010.

[122] S. A. Bacanu, B. Devlin, and K. Roeder. Association studies for quantitative traits in structured populations. *Genet. Epidemiol.*, 22:78–93, Jan 2002.

[123] R. Chakraborty and K. M. Weiss. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Acad. Sci. U.S.A.*, 85:9119–9123, Dec 1988.

[124] N.M. Kaplan. *Clinical Hypertension*. Williams and Wilkins, Sixth edition, 1994.

[125] D. E. Reich and D. B. Goldstein. Detecting association in a case-control study while correcting for population stratification. *Genet. Epidemiol.*, 20:4–16, Jan 2001.

[126] S. Wacholder, N. Rothman, and N. Caporaso. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J. Natl. Cancer Inst.*, 92:1151–1158, Jul 2000.

[127] S. Wacholder, N. Rothman, and N. Caporaso. 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, Jun 2002.

[128] J. Marchini, L. R. Cardon, M. S. Phillips, and P. Donnelly. The effects of human population structure on large genetic association studies. *Nat. Genet.*, 36:512–517, May 2004.

[129] R. S. Spielman, R. E. McGinnis, and W. J. Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, 52:506–516, Mar 1993.

[130] S. A. Bacanu, B. Devlin, and K. Roeder. The power of genomic control. *Am. J. Hum. Genet.*, 66:1933–1944, Jun 2000.

[131] B. Devlin, K. Roeder, and L. Wasserman. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol*, 60:155–166, Nov 2001.

[132] J. K. Pritchard and P. Donnelly. Case-control studies of association in structured or admixed populations. *Theor Popul Biol*, 60:227–237, Nov 2001.

[133] C. J. Hoggart, E. J. Parra, M. D. Shriver, C. Bonilla, R. A. Kittles, D. G. Clayton, and P. M. McKeigue. Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet.*, 72:1492–1504, Jun 2003.

[134] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, Jun 2000.

[135] E. Setakis, H. Stirnadel, and D. J. Balding. Logistic regression protects against population structure in genetic association studies. *Genome Res.*, 16:290–296, Feb 2006.

[136] P. Menozzi, A. Piazza, and L. Cavalli-Sforza. Synthetic maps of human gene frequencies in Europeans. *Science*, 201:786–792, Sep 1978.

[137] L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza. Demic expansions and human evolution. *Science*, 259:639–646, Jan 1993.

[138] N. A. Rosenberg, S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet.*, 1:e70, Dec 2005.

[139] P.R. Rosenbaum and D.B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[140] P.R. Rosenbaum. *Observational studies.* Springer, 1995.

[141] J. K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med*, 23:2937–2960, Oct 2004.

[142] P.R. Rosenbaum and D.B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Stat*, 39:33–38, 1985.

[143] P. Mair and J. de Leeuw. Rank and set restrictions for homogeneity analysis in R: the "homals" package. 2008.

[144] J.P. Benzecri. *Analyse des Donnees*. Dunod, 1973.

[145] G.H. Golub and C.F. Van Loan. *Matrix computations*. John Hopkins University Press, 1989.

[146] J. De Leeuw and J. Van Rijckevorsel. Homals and princals. Some generalizations of principal components analyis. In E. Diday, L.and Tomassone Lebart, L. Escoufier, J. Pages, and Y. Schektman, editors, *Data analysis and informatics*, pages 231–242. North-Holland, 1980.

[147] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate analysis*, chapter 14.7. Academic press, 1979.

[148] P.R. Peres-Neto and D.A. Jackson. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologica*, 129:169–178, 2001.

[149] D.A. Jackson. PROTEST: a Procrustean randomization test of community environment concordance. *Ecoscience*, 2:297–303, 1995.