Of Long-Tails, Microarrays, and Marker Sets: Integrative Approaches in Functional Genomics, Population Genetics and Genetic Epidemiology

Dissertation

zur Erlangung des Doktorgrades
der Mathematische-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität
zu Kiel

vorgelegt von
Timothy Te Hua Lu

盧德華

Kiel

2009

Referent: Prof. Dr. Manuela Dittmar

Konferent: Prof. Dr. Michael Krawczak

Tag der mündlichen Prüfung: 21.07.2009

Zum Druck genehmigt: Kiel, _____.

der Dekan

Table of Contents

Summary

The work in the three presented articles provides several demonstrations of how an integrative approach to scientific research has led to a better understand of biological phenomena. The first article incorporates research from the overlapping fields of biotechnology, functional genomics, and bioinformatics. The study's objective is to describe the nature of the distribution of gene expression levels measured with microarrays with the aim of developing an inter-array normalization method. The normalization method is compared to other existing normalization methods and is found to be especially suited to so-called *boutique* microarrays. The second article uses genotyping data generated by microarrays with the goal of examining the population genetic structure of the European human population. This study combines aspects of the fields of biotechnology, bioinformatics, and population genetics and sheds light on the genetic differences between Europeans by characterizing a strong correlation between geographic and genetic distance. In the final article, focus switches from genetic differences to genetic similarities in the same European individuals by examining the relationship structure of genetic nearest neighbors. Observations about these relationships lead to the proposal of a genetic matched-pair study design that contributes a methodological improvement to the field of genetic epidemiology. The proposed study design has the potential to increase the power of analysis of genome-wide association studies which are used to discover disease-causing genes. A presentation of previously unpublished research which was generated during the course of the work is also included. Finally, a discussion of long-tail data distributions initially observed in the first article leads to conclusions on the fundamental properties of the informational content of genetic marker sets ascertained in the last two articles.

Zusammenfassung

Die Arbeiten, die in den drei vorliegenden Artikeln präsentiert werden, zeigen, wie ein integrativer wissenschaftlicher Ansatz zu einem besseren Verständnis biologischer Phänomene führt. Der erste Artikel verknüpft Forschung aus den sich überlappenden Fachgebieten Biotechnologie, funktionelle Genomik und Bioinformatik. Das Ziel der Studie war es, mittels Mikroarrays die Verteilungsform der Genexpressionsniveaus zu bestimmen, um eine Normalisierungsmethode zu entwickeln. Diese Normalisierungsmethode wurde mit anderen bereits bekannten Normalisierungsmethoden verglichen und sie erwies sich als besonders geeignet für sogenannte Boutique-Mikroarrays. Der zweite Artikel verfolgt das Ziel, mit Hilfe von humanen Genotypisierungsdaten aus Mikroarrays die populationsgenetische Struktur der europäischen Population zu charakterisieren. Diese Studie verbindet Aspekte der Forschungsgebiete Biotechnologie, Bioinformatik und Populationsgenetik und gibt damit Aufschluss über die Muster genetischer Unterschiede zwischen Europäern: Es konnte eine hohe Korrelation zwischen geographischen und genetischen Distanzen gezeigt werden. Der letzte Artikel richtet den Blick auf die genetischen Gemeinsamkeiten der selben europäischen Individuen, indem er die Verwandtschaftsstruktur mittels eines genetischen „nearest neighbors“-Algorithmus untersucht. Die beobachteten Verwandtschaftsstrukturen führen zum Vorschlag eines genetischen Matched-Pair-Studiendesigns, das auf dem Gebiet der genetischen Epidemiologie eine erhebliche methodische Verbesserung darstellt. Das vorgeschlagene Studiendesign kann die Aussagekraft der statistischen Analysen bei Genom-weiten Assoziationsstudien erhöhen, also bei Studien, die durchgeführt werden, um krankheitsverursachende Gene zu identifizieren. Darüber hinaus werden bisher unveröffentlichte Forschungsergebnisse vorgestellt, die im Zusammenhang mit den obigen Studien gewonnen wurden. Eine abschließende Diskussion der Long-Tailed-Verteilung der Daten, die zunächst in der ersten Studie beobachtet wurde, führt zu Schlussfolgerungen über die grundlegenden Eigenschaften des Informationsgehaltes genetischer Markersets, welche nachfolgend in den letzten beiden Studien bestätigt wurden.

Introduction

Overview

The work presented in this thesis spans five broad fields of biological study, which are biotechnology, population genetics, genetic epidemiology, functional genomics, and bioinformatics (Table 1). These diverse fields and the overlap between them provide the foundation upon which the research presented in this thesis is based. Because this thesis draws from narrower areas of research from within each of these broad fields, a general overview of each field will be followed by more specific background information required to place the research presented in the three articles that follow in the necessary context. In the first article, a novel normalization method for expression microarrays is proposed and evaluated. Concepts introduced in this article draw from the fields of functional genomics, biotechnology, and bioinformatics. Article two examines the genetic structure present in the European human population, making use of genotypes generated using microarray technology to examine a population genetic question. The thesis culminates in the evaluation of a novel method for controlling for population stratification in genome wide association studies, drawing from the diverse fields population genetics, genetic epidemiology, and biotechnology. Because each of the general topics are very broad, only the background information relevant to each of the categories will be discussed; however, the interested reader will find more detailed material provided in the introduction to each of the articles.

Table 1. Integrated Areas of Research

	Thesis Article		
Area of Research	1	2	3
Biotechnology	•	•	•
Population Genetics		•	•
Genetic Epidemiology			•
Functional Genomics	•		
Bioinformatics	•	•	
Publication reference number (see CV)	14	2	1

Table putting the articles presented in this thesis into the context of broader fields of biological research. All thesis articles have been previously published under primary authorship. Other publications as supporting author are listed along with publication reference numbers in the *Curriculum vitae*.

The articles presented in this thesis have been previously published elsewhere. The references to the three publications corresponding to the articles are as follows:

1. Lu, et. al. Can Zipf's law be adapted to normalize microarrays? BMC Bioinformatics. 2005 ;6:37.

2. Lao & Lu*, et. al. Correlation between genetic and geographic structure in Europe. Curr Biol. 2008 Aug 26;18(16):1241-8.

3. Lu & Lao*, et. al.. An evaluation of the genetic-matched pair study design using genome-wide SNP data from the European population. Eur J Hum Genet. 2009 Jan 21;[online advanced publication] doi:10.1038/ejhg.2008.266

* shared first authorship.

Biotechnology - Microarray Technology

Biotechnology differs from the other fields of study previously mentioned in the sense that it is not a classical academic subject of study but instead a collection of tools, techniques and methods used to facilitate the acquisition of knowledge from various other fields of biological research. Even so, fundamental research in biotechnology is an important endeavor wherein scientists develop novel methods of observing physical phenomena as well as increase the efficiency and sensitivity of existing analytical techniques. The overlap between biotechnology and other fields of study, such as bioinformatics, functional genomics, genetic epidemiology, and population genetics is an area where knowledge can be combined to spur scientific discovery. Innovations in these areas often allow research to proceed in directions not previously anticipated under the previous scientific framework. An example of one such revolutionary advance in biotechnology is DNA microarray technology, which makes a recurring appearance throughout this thesis and is one of the foundation technologies upon which this work is based.

Microarray technology has been at the leading edge of advancements in both the speed and scale of data collection in many areas of molecular and cell biology since its development in the mid-1990's as a high-throughput method for measuring gene expression[1], it came into widespread use in the late-1990's[2] and was soon adapted to be used for analyzing genotypes[3], sequencing DNA[4] and detecting mRNA splice variants[5]. It continues to be a field of rapid technological development. As microarray research advanced, the technology moved out of the

academic research environment to be developed as application oriented high-throughput commercial systems. The continued success of microarray technology is attributable to three main factors. First, experiments can be performed with very small quantities of test material, typically in the range of nanograms to picograms, presenting a significant improvement over previous technologies. Second, microarrays further increase experimental efficiency by miniaturizing the physical format upon which an individual experiment takes place. The number of targets which can be screened simultaneously ranges from the thousands to the millions depending on the type of array. Third, microarrays are flexible. Expanding on the basic idea of the expression microarray by varying the type of targets, probes, or chemistry of the system is a powerful way of developing microarrays with completely novel purposes.

The description of microarrays presented here focuses on the expression type of array as this was the first to be developed. At the most fundamental level, a microarray experiment begins with single stranded DNA of a known sequence (target) covalently bound to a solid support material. A mixture of unknown single stranded DNAs (probe) is applied to the support with the purpose of detecting a match, which occurs when the target and one of the probes hybridize. Detection is facilitated by a label (i.e. radioactive, fluorescent or chemiluminescent) that tags the target-probe DNA hybrid. A scanner is used to quantify the amount of label present at the target-probe hybrid. Many different target DNAs are arranged in a grid pattern on the solid support which is then referred to as an “array”. Early microarrays were manufactured by physically spotting target solutions onto the solid support, thus target DNA locations are sometimes referred to as “spots”. With all microarray systems, the location of the spot and thereby the mapping of the target sequence to a physical position is critical to the final interpretation of target activity. One distinctive feature of microarrays that makes them especially powerful is the large number of spots (thousands to a million) spaced very closely together on the solid support. The spots are typically on the scale of tens of micrometers in diameter and the whole array a few millimeters in size, thus the descriptive name “microarray”. By decreasing the size of spots and/or increasing their physical proximity it is possible to increase the number of measurements collected in a single experiment sometimes gaining exponential increases in efficiency. In fact, it has even been proposed that the process of miniaturization and corresponding increase in throughput for microarrays may have parallels with

Moore's Law, which describes price decreases corresponding to performance increases in computer chips[6]. Microarrays have found application in many fields of genetics, molecular biology, and cell biology. The use of microarrays in this thesis is limited to the specialized fields of population genetics, genetic epidemiology and functional genomics where two distinct types of microarray systems were used, one for quantifying the level of gene expression, and one for genotyping genetic markers.

Gene expression measurement was the original purpose for which microarrays were developed. The term “gene expression” is used to describe the quantified level of activity of a gene. An estimated thirty thousand genes are present in the human genome[7], each with its own pattern of regulation and a level of expression that varies with the cells' environment. To understand the concept of quantifying gene expression, it is useful to recall the textbook central dogma of molecular biology[8] coined by Frances Crick in 1958 which states that DNA is transcribed to RNA which is, in turn, translated to protein. In microarrays, measurement of gene expression is accomplished by extracting messenger RNA (mRNA) from a sample, converting it to complimentary DNA (cDNA) which is then labeled and used to probe the microarray, which is itself spotted cDNA of known sequence from a library of expressed genes. Genes expressed at a high level will have correspondingly high levels of messenger RNA present in the sample extract and therefore high levels of cDNA probe that hybridize to the target. Similarly, genes with low expression levels, or genes that are not expressed at all, will have correspondingly low target binding activity.

Genotyping is the other purpose to which microarrays are tasked in the work presented here. Genotyping is done using microarrays designed to detect single base-pair changes in the DNA sequence. The genetic variants occurring at this most fundamental level of the genome are called single nucleotide polymorphisms (SNPs). The human genome contains approximately fifteen million SNPs according to one estimate[9], making them ideal genetic markers for the study of human genetics diseases, for example. Microarrays designed to detect SNPs differ from those used to quantify gene expression in many respects making them a good example for demonstrating the flexibility of microarray technology by showing how the same basic concept can be used in a different context by varying the selection of probes, targets, and hybridization chemistry[10]. For example, the probes used on expression microarrays are cDNA synthesized from extracted

mRNA. In contrast, for genotyping microarrays, probes are derived from whole genomic DNA. In a similar vein, the targets on expression microarrays are gene transcripts, usually from cDNA libraries, representing only coding regions of the genome, whereas the targets on genotyping microarrays are the allele specific oligonucleotides, which may occur almost anywhere in the genome. Furthermore, in expression microarrays, the detection chemistry is optimized to quantify the amount of probe hybridized to the target. In contrast, typical genotyping microarrays have a more complex chemistry that allows differentiation between two alternate genetic variants (alleles) using two florescent markers of different color. The expression microarrays presented in Article One of this thesis were designed to evaluate approximately thirty thousand target genes per experiment while the genotyping microarrays presented in Articles Two and Three can simultaneously measure approximately five hundred thousand target genetic markers per experiment.

Despite the differences between expression and genotyping microarrays, they share several notable similarities with most other types of microarrays. The first similarity is the requirement for data normalization. If the values measured on one microarray are to be compared to the values measured on another microarray, the microarrays must be normalized to each other[11,12]. This is accomplished by a systematic adjustment of the magnitude of measured values with the intention of removing unwanted variation introduced by experimental conditions (covered in more detail in the bioinformatics section of the Introduction). Another feature that all microarrays have in common is the large number of observations generated per experiment. This is both an advantage and a disadvantage. It is a disadvantage because the problem of false discovery becomes much more prominent. High numbers of false discoveries are unavoidable when large numbers of statistical tests are applied. However, because microarrays are so efficient in collecting many observations simultaneously, this reduces cost and increases the speed at which experiments can be performed, opening new possibilities for data analysis (such as the dimension reduction techniques presented in the bioinformatics section of the Introduction).

Population Genetics – The European Human Population

All of the articles in this thesis deal with the European human population either directly or indirectly. Article two in particular focuses on the population genetic structure of Europe, while the other articles use data derived from the European population. A brief overview of population genetic studies in general and the chronology of studies on the Europeans specifically will assist the reader in placing the work presented here into the broader context of previous knowledge. Genetic structure is the term used to describe the uneven geographical distribution of alleles from one or more genes within a population. Genetic structure can be observed at the level of individual genetic markers as a dynamic phenomenon, changing over time under the influence of four driving factors: migration, selection, mutation, and genetic drift[13]. Genes may be located in close physical proximity to each other on the genome. In this case, changes in genetic structure observed at one genetic locus is not independent of changes observed at the other locus because the loci are physically *linked*. The linkage observed between markers is directly correlated with their physical proximity. All of these factors combine to produce intricate and dynamic patterns of genetic variation over a geographic range. Documenting and interpreting these patterns is the core focus of this research.

The first comprehensive study of European human population genetic structure was done in the late 1970's by a group lead by Luigi Luca Cavalli-Sforza[14]. The study's methodology was groundbreaking on two accounts. First, it was the first Europe-wide study that examined a large number of genetic loci (n=38) and second, it used multivariate principal components analysis (PCA) to generate maps depicting genetic gradients across the continent. Most of the loci in this early study were blood groups and HLA (human lymphocyte antigens). These loci are very informative because of their high levels of polymorphism. The study reported a primary southeast-northwest genetic gradient across Europe, which was hypothesized to correspond with the neolithic expansion of agriculture from the Middle East. This primary gradient has since been repeatedly verified[15,16]. Beside the primary principal component, other principal components revealed genetic gradients across Europe in various directions, although explanations for these gradients were more tenuous. By the early 1990's, technological advances and the discovery of new genetic markers brought the number of genetic loci

examined in Europe to over one hundred, including more HLA alleles as well as allozymes (enzyme variants), immunoglobulin variants, and DNA polymorphisms. In addition to the previously used multivariate analysis methods such as PCA and multidimensional scaling (MDS), data were analyzed using spatial autocorrelation and clustering methods, recognized as being useful on hierarchical data. A comprehensive review showed inconsistent clustering of European populations and genetic gradients in a variety of directions across the continent, although predominantly southeast-northwest[17]. The main conclusions drawn by these studies were the existence of genetic associations between geography and language and a complex migration history evidenced by inconsistent patterns of clusters and gradients. Little evidence for selection was noted although there was speculation that future methods may discover variation related to north-south differences in physical traits. (An example of selection is presented in the Additional Results portion of this thesis, *Evidence of Selection in OCA2/HERC2 and LCT Genes*). Throughout the 1990's and into the new century the advancement of DNA analysis techniques brought with it an increase in the number of studies using mitochondrial DNA (mtDNA) and Y-chromosomal markers which shed further light on the population structure in Europe[18]. Data derived from Y-chromosome markers reaffirmed the genetic gradient indicative of neolithic expansion from the Middle East, but interestingly did not show significant association with language. In contrast, mtDNA markers were more evenly distributed and showed a weaker neolithic genetic gradient than was observed in previous studies. This and further evidence lead to the hypothesis that human males are more sedentary than human females[19].

In the late 1990's, commercial microarray technology became widely available enabling cost effective SNP genotyping of unprecedented numbers of markers[10] and population geneticist began using these autosomal genetic markers in studies. In 2006, a major development in the analysis of autosomal SNP marker data was made which allowed PCA analysis to be done using individuals as the unit of measurement[20]. This was a significant improvement over the previous method where genetic distance measures were based on allele frequencies estimated from populations because 1) the accuracy of such allele frequency estimations are very sensitive to sample size and 2) the definition of *a priori* populations can be problematic. These two developments led to several studies (one example is included as Article Two of this thesis) which combined aspects of

the fields of biotechnology and population genetics to add further detail to the understanding of European population genetic structure[16,21-23]. Results of these population genetic studies have an obvious application in genetic epidemiology where population structure is referred to as *stratification* and has been a major concern of researchers during the past quarter century.

Genetic Epidemiology - Genetic Association Studies and Genetic Matching

Discovering the genetic causes of human diseases is a major focus of biomedical research. This endeavor is complicated by many factors which, taken together, obscure the causal link between genes and disease. The use of statistical analysis and the development of designs to maximize the experimental power have therefore become an important area of research in the biological and medical sciences.

With the formulation of the modern theory of genetics in the late 19th century, Gregor Mendel provided a mechanism for explaining how heritable diseases are transmitted[24]. Diseases which can be accurately explained by this model are called simple monogenic diseases (controlled by a single gene). Because they are easily identified by their characteristic pattern of transmission, more than fifteen thousand of these simple diseases have so far been characterized[25], many of them associated with the disruption of metabolic pathways. However, although simple diseases are numerous they are the exception among diseases, occurring in very few individuals in the population because the alleles causing these diseases are rare. This observation is a direct result of their straightforward mechanism of transmission which allows natural selection to act as a strong force to reduce the frequency of the disease in the population. The simple model of a single genetic marker displaying either dominant or recessive inheritance presented in biology textbooks belies the complexity faced by researchers when examining genetic diseases. Most common diseases (i.e. coronary heart disease[26], Alzheimer's disease[27] and Parkinson's disease[28]) occur at high frequencies in the human population despite the large financial and scientific investment by both academic researchers and pharmaceutical companies in combating these diseases. These diseases do not display the characteristic patterns of transmission that allow explanation by the simple Mendelian model of monogenetic inheritance. For this reason, these

diseases are called complex, however, it must be kept in mind that diseases do not necessarily have to have a genetic component; they may be caused entirely by environmental factors.

Complex diseases are defined as those which are not easily explained by the simple Mendelian model of monogenic inheritance but still have some genetic component. Many factors can obscure the patterns of transmission expected under the simple model, thus leading them to be classified as complex diseases[29]. 1) A disease may be caused by several genes (epistasis). In this case, disease inheritance is no longer a straightforward matter but depends upon the sum total contribution of many alleles at many genetic loci. Because of genetic recombination that takes place during meiosis, the exact genetic composition causing the disease is unlikely to be passed on in the same state to the offspring. This makes observation of the inheritance pattern of the disease difficult. 2) The presence of a disease allele may result in varying degrees of manifestation of a disease in different individuals (penetrance). It may even be the case that some individuals carrying a disease gene are completely unaffected by the disease. 3) In some cases, several different genes may cause the same or very similar diseases (multigenic causes and genetic heterogeneity) 4) There are some diseases which, although autosomal, are inherited only maternally or paternally (imprinting). Finally, 5) Many diseases are also strongly influenced by environmental effects that obscure patterns of inheritance in genetic studies. Compounding the matter, each of the confounding factors mentioned above may be acting in the presence of one or more of the others. It may be of interest here to mention a class of simple diseases which is expressed with a phenotype very similar to particular complex diseases. These rare monogenic forms have been very important in helping understand the disease process than have their more common complex disease counterparts because the genetics of the monogenic forms are more easily understood.

Table2. Simple vs Complex Diseases

	Simple Diseases	Complex Diseases
Genetic Model	Monogenic	Polygenic
Influence of Selection	Strong	Weak
Frequency in Population	Rare	Common
Confounding Factors	Few or none	Many
Gene Detection	Easier	Very Difficult

A comparison summarizing the main features of each disease category.

Different methods have been developed to identify disease causing (or disease associated) genes despite the confounding factors found in epidemiological data. Traditionally, genetic epidemiological studies have employed well-ordered series of analyses which were applied in sequence. However, analyses have become less structured today. The first analysis which should be done for any disease being studied is a confirmation of the diseases genetic component. This can be done for example by comparing the recurrence risk of disease in affected families versus the general population risk, or by examining disease co-occurrence in monozygotic over dizygotic twins to estimate the heritable component of the phenotype in question. Once the disease has been shown to have a genetic component, segregation analysis[30] can determine the inheritance pattern of the disease by examining co-transmission of phenotypic and genotypic traits within pedigrees. In studies of human diseases, segregation analysis is a data intensive process that requires analysis of pedigree and phenotypic data from as many families affected by the disease as possible. The analysis described up to this point is done without having to collect any genotype data. After the mode of inheritance has been determined by segregation analysis, linkage analysis is used to identify the location of the disease gene on the genome. Linkage studies[31] are usually performed on familial data, but can also be performed on population-based data. In familial linkage analysis, the transmission of the disease phenotype together with alleles of markers of known genomic positions are observed to identify markers which are physically close, or 'linked', to the disease causing gene. Once linkage has been identified, a focused genotyping effort is made using additional genetic markers in the region of interest and an association analysis[32] is performed to identify alleles co-occurring with the disease at the population level. Association analyses therefore typically use population-based data, but they can also be performed using familial data. Because microarray technology has drastically lowered the cost of genotyping, the investment in segregation analysis and even linkage analysis is often dispensed with in favor of simply performing association testing on genome-wide scans of population-based samples. It should also be noted that in the context of genome-wide scans, family-based linkage studies and population-based association studies have differing detection sensitivities, the former on rare alleles with weak effects and the latter on common alleles with strong effects.

One weakness in population-based case-control studies, is the possible

presence of either population genetic structure (see Introduction, Section Two), or stratification of the sampled population, which can both lead to spurious results[33,34]. The most dramatic example would be a situation where the 'cases' and 'controls' come from two different populations. If this were the case, it is possible that the genetic differences seen between cases and controls are due entirely to population genetic differences, and not at all related to the disease state. When this problem became a matter of concern in the genetic epidemiology community in the 1980's, researchers developed two strategies to respond to it. One was to focus on family-based association analysis, using core families as the experimental unit[35]. The analysis examines the transmission of marker alleles and looks at the correlation between these transmission events and the transmission of the disease. Family-based association analysis is not affected by population structure because the observation of transmission events does not depend on population allele frequencies in the families tested. Another strategy to deal with population stratification in association studies was to measure its magnitude and to adjust for it appropriately when conducting the statistical test for association using two methods: 1) the genomic controls method[36], which adjusts the test statistic according to an inflation factor estimated from control markers located throughout the genome, and 2) the structured association method[37], which first tests for the presence of structure in the sampled population and then stratifies the analysis according to the observed structure. Both methods were developed before microarray technology spawned the massive genome-wide screening projects sponsored by the British Wellcome Trust[38], the American National Institutes of Health[39] (NIH), and the German Nationales Genomforschungsnetz[40,41] (NGFN) in a wide range of complex diseases such as diabetes[42], arthritis[43,44], Crohn's disease[45], Sarciodosis[46], gallstone disorders[47], Parkinson's disease[48], periodontitis[49], coronary heart disease[50], bipolar disorder[51] and hypertension[52].

One of the criticisms raised against genome-wide association studies is that the number of false positives is high, which is an unavoidable side effect of the massive numbers of genetic markers screened by microarrays. This casts some doubt on the validity and replicability of studies using this method[53]. Partly in response to this line of criticism, the main topic of the third article is an evaluation of genetic-matched pair study design. This study design is an alternative to the association tests usually performed in genome-wide association studies and reduces the potential for false positive results by restricting the statistical

comparisons to genetically similar pairs of cases and controls.

Functional Genomics – Gene Expression

The development of high-throughput genotyping technology at the end of the 20th century led soon after to the complete sequencing of the human genome and stimulated further interest in the field of functional genomics. Simply defined, functional genomics is the study of how genes work[54]. It encompasses all aspects of molecular biology's central dogma (see Introduction, Biotechnology) and includes study of the entire spectrum of mechanisms and processes that control and influence transcription and translation. Functional genomics covers, among other areas, comparative sequence analysis both between genes and between species, protein analysis including protein sequencing and protein structure analyses, aspects of molecular phylogeny and evolution, and research into mRNA splice variants. Although functional genomics is a broad field of study, the focus for purposes of this introduction will be on gene expression[55]. Gene expression is measured from mRNA extracted from cells or tissue of interest and reverse transcribed to cDNA to facilitate quantification using microarrays (see Introduction, Section One). The principle advantage of using expression microarrays is the large number of gene expression levels that can be measured simultaneously in a single experiment. Typical gene expression studies involve comparing gene expression levels from two or more different sources with the intention of measuring differences of biological interest. The scientific question dictates whether the comparisons made are between different tissue types, between treatment and control samples, diseased or healthy organs, or from a developmental time series, for example. However, even before comparisons are considered, interesting observations can be made about the mathematical distribution of expression levels, which is a topic of scientific interest in its own right. It was observed early on that the expression levels did not follow a Poisson distribution; however studies on the distribution of expression levels in microarrays gave conflicting descriptions of the underlying mathematical function. Some researchers observed a power-law distribution[56], and others a log-normal[57] or a skewed binomial differential distribution[58]. In Article One of this thesis, the distribution of gene expression levels in the microarray system we examined is shown to follow Zipf's Law and this observation is further developed into a

normalization method for between-array comparisons and functionally specific microarrays. These microarrays are also known as *boutique* microarrays because they contain a selected set of functionally specific expression targets making them unsuitable for global normalization or the standard Quantile method. Additionally, the normalization method presented here found practical implementation in three gene expression studies[59-61].

Bioinformatics – Data Normalization

Bioinformatics, the use of computer science or informatics to develop techniques and procedures for handling and analyzing biological data, includes topics ranging from mundane data management software to the development of complex algorithms for sequence alignment or protein conformation prediction. Technological advances in both the fields of biology and computer science have increased the scale and speed of experimental data collection to a heretofore unprecedented level. Historically, extremely large data sets have been available to biologists only through long, labor intensive data collection. Currently, it is not uncommon for biologists to collect millions of experimental observations in a single day making it necessary for biologists to manage datasets as large or larger than those previously dealt with in such disciplines as physics, chemistry, astronomy, and economics.

While it may seem that biologists are following in the footsteps of others by borrowing analytical methods from different fields to analyze large data sets, history shows that many of these methods were originally developed for use with biological data. For example, principle components analysis is one such method featured in Article Two of this work. This mathematical technique, which reduces the number of variables while minimizing the loss of information in a dataset, has for many years been put to extensive use in fields outside biology, although it was developed in 1901 by Karl Pearson for use in biology among other applications[62]. Another analytical method often used in population genetic and functional genomic studies is clustering. Cluster analysis traces its early development to the field of taxonomy which involves conceptually grouping species according to characteristics in a hierarchical fashion. In the 1960's taxonomists developed hierarchical clustering as a method for grouping species in an unbiased manner and participated in the founding of the Classification Society[63]. Hierarchical

clustering is compared to principal components analysis in the additional results section of this thesis. The application of ordination methods which reduce the dimensionality of data to make them effectively more manageable (i.e. principal components analysis) and hierarchical clustering for grouping data represents an important emergent feature of large data sets as these methods make observations possible which are not attainable with smaller data sets.

Data normalization is an important bioinformatic technique which is the major theme of Article One. Data normalization is the removal of undesired systematic variation from a dataset to facilitate statistical comparison. Variation in biological data can arise from many sources, not all of which are undesirable. Indeed the objective of all biological research is to draw conclusions about biological phenomena by measuring variation generated from biological sources despite confounding variation from unwanted sources. Sources of variation can be either systematic (affecting portions of the data in a predictable way) or stochastic. Systematic variation can sometimes be removed using correction factors estimated from the data itself without the necessity and expense of replicate sampling. The removal of systematic variation to facilitate data set comparisons is termed “normalization”. With the advent of microarray technology, one early area of intense research was in the application and optimization of normalization methods[11,12]. In expression microarrays (see also Introduction, Section One), systematic variation can arise from several sources. For example, differences in target DNA concentrations between spots during microarray manufacture, or uneven application of probe DNA in the experiment phase can cause within-array variation. Between-array variation can be caused by differences in exposure times of probes to the microarray, or differences in RNA extraction efficiency from the tissue. Article One proposes a between-array normalization based on the observation that gene expression levels follow a distinct mathematical distribution (see also Introduction Section Four). Algorithms for normalization of expression microarrays can be divided into three classes of methods: global, intensity-based, and location-based. The global methods were the first proposed and also the most primitive of the between-array normalization methods, which adjust all arrays to the same mean and standard deviation. However, these simple methods often led to unsatisfactory results. Later methods were therefore developed having normalization factors dependent on the intensity of the expression signal (quantile-quantile method, Loess curve fitting). In their most simple

implementation, correction factors used in both the global and intensity-based normalization methods are derived from the entire spectrum of measured gene expression levels. A variation on these methods derives the correction factor from a subset of data, either reference samples placed on the microarray for the purpose of normalization or so-called 'housekeeping genes'. Housekeeping genes are usually constitutive proteins, responsible for essential cell functions, which are assumed to be unaffected by experimental treatments and thus expressed at a consistent level between microarrays. The motivation for using housekeeping genes as the benchmark for normalization is that they are less influenced by extreme measurements and produce more consistent correction factors. This topic is further elaborated in Article One. The work presented in Article One was completed between 2000 and 2003, and reflects the state of knowledge and technology at that time. Many of the concepts developed for the normalization of expression microarrays are equally applicable to the genotyping microarrays that became first commercially available between 2003 (Affymetrix) and 2005 (Illumina).

Synopsis

The preceding five summaries of broad research areas provide a background for understanding the three articles that follow. Each of the three articles, in turn, integrates aspects of three research areas (Figure 1). Article One

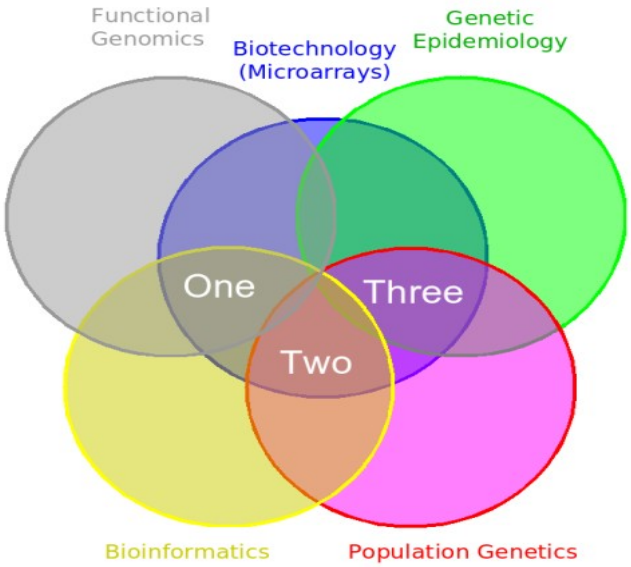

Figure 1. Integrated Areas of Research: A diagram putting the articles presented in this thesis (white text, One, Two, and Three) into the context of broader fields of biological research (colored circles).

integrates biotechnology, functional genomics, and bioinformatics by examining the nature of the distribution of gene expression levels measured in microarray experiments. The distributions are found to belong to a family of long-tailed distributions (power-law and log-normal distributions). This observation motivates the development of a normalization method which is evaluated against other available normalization methods. Article Two presents research in the overlapping fields of biotechnology, bioinformatics and population genetics. In this study, principal components analysis is applied to a large microarray generated genotype data set from the European human population to examine population genetic structure. Despite low levels of genetic differentiation between subpopulations, a strong continent-wide correlation between geographic and genetic distance is observed. Also, a widely used group of control sampled (the CEPH Caucasian individuals from Utah) are mapped onto the European genetic landscape. Article Three combines biotechnology with population genetics and genetic epidemiology by making observations on genetic similarity in the European population based on the same microarray genotype data used in the previous study. The structure of genetic relatedness between European individuals is examined and a genetic matching study design to increase the power of genetic association studies is proposed. A marker set designed to be useful in identifying the 'best' genetic-matching pairs is ascertained and evaluated.

Article 1

Can Zipf's law be adapted to normalize microarrays?

This article has been previously published as:

Lu, et. al. Can Zipf's law be adapted to normalize microarrays? BMC Bioinformatics. 2005 ;6:37.

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Can Zipf's law be adapted to normalize microarrays?

Tim Lu[1], Christine M Costello[1,3], Peter JP Croucher[1], Robert Häsler[1], Günther Deuschl[2] and Stefan Schreiber*[1]

Address: [1]Department of Medicine, Christian-Albrechts-University, Kiel, Germany, [2]Department of Neurology, University Hospital Schleswig Holstein, Kiel, Germany and [3]The Conway Institute for Biomolecular and Biomedical Research, University College Dublin, Ireland

Email: Tim Lu - t.lu@mucosa.de; Christine M Costello - christine.costello@ucd.ie; Peter JP Croucher - p.croucher@mucosa.de; Robert Häsler - r.haesler@mucosa.de; Günther Deuschl - g.deuschl@neurologie.uni-kiel.de; Stefan Schreiber* - s.schreiber@mucosa.de

* Corresponding author

Published: 23 February 2005

BMC Bioinformatics 2005, **6**:37 doi:10.1186/1471-2105-6-37

Received: 30 August 2004

Accepted: 23 February 2005

This article is available from: http://www.biomedcentral.com/1471-2105/6/37

Abstract

Background: Normalization is the process of removing non-biological sources of variation between array experiments. Recent investigations of data in gene expression databases for varying organisms and tissues have shown that the majority of expressed genes exhibit a power-law distribution with an exponent close to -1 (i.e. obey Zipf's law). Based on the observation that our single channel and two channel microarray data sets also followed a power-law distribution, we were motivated to develop a normalization method based on this law, and examine how it compares with existing published techniques. A computationally simple and intuitively appealing technique based on this observation is presented.

Results: Using pairwise comparisons using MA plots (log ratio vs. log intensity), we compared this novel method to previously published normalization techniques, namely global normalization to the mean, the quantile method, and a variation on the loess normalization method designed specifically for boutique microarrays. Results indicated that, for single channel microarrays, the quantile method was superior with regard to eliminating intensity-dependent effects (banana curves), but Zipf's law normalization does minimize this effect by rotating the data distribution such that the maximal number of data points lie on the zero of the log ratio axis. For two channel boutique microarrays, the Zipf's law normalizations performed as well as, or better than existing techniques.

Conclusion: Zipf's law normalization is a useful tool where the Quantile method cannot be applied, as is the case with microarrays containing functionally specific gene sets (boutique arrays).

Background

DNA microarrays have become a widely used biotechnology for assessing expression levels of tens of thousands of genes simultaneously in a single experiment [1,2]. Whether microarrays are being used for global tissue profiling or for differential expression studies, data normalization is an essential preliminary step before statistical analysis methods can be applied. The purpose of all normalization techniques is to transform the data to eliminate sources of variability stemming from experimental conditions, leaving only biologically relevant differences in gene expression for subsequent analysis. Normalization can be divided into two stages, intra-array normalization and inter-array normalization. Intra-array normalization deals with variability within a single array caused by factors such as differences in print-tip

characteristics, channel differences in two-dye systems, and spatial heterogeneity across the array surface [3-5] and should be carried out using accepted methods before inter-array normalization is applied. This paper assumes intra-array normalization has been performed and presents an inter-array normalization method for comparison of gene intensity levels between multiple microarrays to deal with variation caused by such factors as differences in RNA isolation efficiency, labeling efficiency, hybridization conditions, exposure times, and detection efficiencies.

It is now clear that simple inter-array normalization techniques, such as simple scaling to housekeeping genes or normalizing to a global mean, are not adequate for microarray data [6]. Housekeeping genes have been found to be more susceptible to modulation than previously thought [7]. Along with others [5], this paper underscores the potentially serious drawbacks of the global mean and other such methods. Recent literature has thus provided a plethora of more sophisticated normalization and analysis techniques as researchers struggle to cope with the task of microarray data analysis, some of which include maximum likelihood analysis [5], centralization [6], principal component analysis [8], analysis of variance [9] and Bayesian network analysis [10].

Analysis of publicly available large-scale SAGE gene expression data sets [11,12] and an intra-phyletic survey of genome wide Affymetrix microarray experiments [13] have indicated that the large majority of expressed genes exhibited power-law distributions, while some microarray expression data exhibit a more log-normal distribution [14]. Our normalization procedure was inspired by the observation that the intensities measured on our microarray system also followed a power law distribution and can therefore be described by a simple mathematical model. Zipf's law [15] is a power law function that states that the magnitude of an intensity measurement (y) is inversely proportional to the rank (r) of that data point in the data set,

$$y \propto r^c \quad (1)$$

where c is a coefficient close to -1. Our microarray data can be classified as a generalized form of Zipf's law because the coefficient (c) is not always close to -1 and, in fact, varies between individual microarrays, making simple linear normalization procedures, such as global normalization to the same mean, inappropriate. However, the normalization procedure proposed here demonstrates that by taking Zipf's law into account, it is possible to apply a simple intra-array normalization procedure such that all filters have the same coefficient c and proportionality.

We demonstrate the Zipf's law based normalization technique on microarray data sets representing both single channel and two channel technologies. In the single channel category, we produced two radio-labeled, nylon membrane based cDNA data sets, one commercial and one generated "in-house". Both systems contain a selection of genes chosen without regard to functional or pathway considerations, which make them especially appropriate for normalization using Zipf's law. These data sets were also normalized to a global mean (the mean of all microarrays) [16], and the quantile normalization method [17]. In addition we produced a two channel, fluorescently labeled, glass slide, oligo-based microarray data set generated 'in-house'. This microarray can be classified as a 'boutique' microarray because it consists of a selection of genes involved in apoptosis. This data set was normalized with a variant of the Zipf's law normalization method that uses a subset of the distribution as a proxy for normalizing the entire microarray. A comparison was then conducted against a variant of the loess normalization method that uses an *a priori* selection of 'housekeeping' genes as a proxy for normalization.

The finding that our microarray data distributions conform to a power law distribution agrees with predictions based on genome wide gene expression studies [11-13], however Hoyle, *et. al.* [14] observed that microarray distributions were log normally distributed with possible power law tails. To investigate this discrepancy, and to verify that our normalization technique could be useful in the normalization of data sets from other microarray systems, we also surveyed publicly available data sets from the NCBI Gene Expression Omnibus [18].

The two assumptions upon which the normalization method are based are the same as those used in other normalization methods [5,6], namely that in comparisons between similar tissues or cell lines under different experimental conditions i) most genes are not, or only moderately, regulated, and ii) approximately equal numbers of genes are up regulated as down regulated. Systems which conform to these two assumptions will be referred to as 'well-behaved' in this paper. While these assumptions probably hold for microarrays derived from a diverse sampling of genes, for example an EST library survey, they may not hold for microarrays containing genes specifically selected based on function or pathway (so called 'boutique' microarrays) as it is likely that most genes will be affected by the experimental treatments. One way to circumvent the restrictions resulting from these assumptions is to use a subset of data, or proxy, from the boutique array data set which fulfils the 'well-behaved' criteria. In developing a boutique microarray normalization technique, Wilson et. al. [4] have devised a method for selecting a subset of genes within a microarray data set

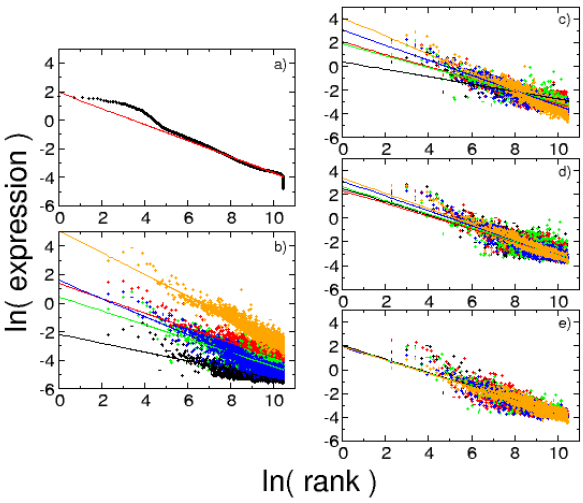

Figure 1

Unigene microarray log plots. Five human Unigene microarrays from the panel of thirty-one microarrays used in the sigmoidal colon experiments. Upper left to lower right: **a**. Log_e median gene intensity vs. log_e rank – conformity to Zipf's law is demonstrated by the linear regression line (in red) **b**. Five microarrays chosen to maximize pre-normalization variability, each plotted according to the gene ranks determined by their median gene intensity levels. **c**. The same five microarrays, normalized to a global mean, with regression lines. **d**. The same five microarrays, normalized with the quantile method, with regression lines. **e**. The same five microarrays, normalized taking Zipf's law into account, with regression lines. For plots b-d, a sub-sample of 10% of the data points are plotted for readability.

that have low variation between arrays and are well representative of the spectrum of intensities measured on the microarray. They term this *a priori* selected subset 'housekeeping' genes, however it should not be confused with the *a posteriori* set of genes typically envisioned when the term is used. Another possible proxy that could meet the 'well-behaved' criteria are control spots which are included on the microarray during it's manufacture. We tested our normalization method on data from a two channel boutique microarray experiment using two types of control spots as proxies for normalization (Positive and negative internal controls, and housekeeping genes). The Zipf's law normalization methods were then compared with the variant of the loess method developed by Wilson et. al. [4] using housekeeping genes.

Results

Verifying Zipf's Law

Before applying the described normalization method, the adherence of the reference curve (the median gene intensity data versus rank) to Zipf's law was verified. The most common method of verifying conformity to Zipf's law is a linear regression on the log_e-log_e transformed data set. Our regression showed a good fit, with a correlation coefficient of -0.98 and a slope of -0.56 for microarrays representing human colon (Figure 1a, Figure 6A, Table 1 set A), a correlation coefficient of -0.99 and a slope of -0.78 for rat brain microarrays (Figure 6B, Table 1 set B), and a correlation coefficient of -0.99 and a slope of -0.60 for the mouse apoptosis microarrays (Figure 6H, Table 1 set H). It should be noted that while the low ranking intensities may show a marked deviation from the regression line, this data typically accounts for a very small proportion of the total data and does not have a large affect on the regression curves.

Normalization results – single channel microarrays

A comparison of the Zipf's law normalization method to the simple method of setting all arrays to a global mean (the mean of all microarrays) and to the quantile method was conducted on the single channel microarray data sets. Five human Unigene microarrays from the panel of thirty-two microarrays used in the sigmoidal colon experiments were selected to represent the greatest variability in pre-normalized data observed in the experiment (Figure 1b). Normalization to a global mean (Figure 1c) yielded data sets that displayed a higher variability in the coefficient *c* of the Zipf's power function (formula 1) than that observed after normalization by the Zipf's law method (Figure 1e) or the quantile method (Figure 1d). The Zipf's method showed the lowest variation in the Zipf's exponent and had the lowest spread of the data around the ln(rank) vs. ln(intensity) line. Results of an identical log_e intensity versus log_e rank plot comparison in Clontech rat microarrays showed little difference between the quantile and Zipf's methods [see Additional file 1]. However it should be mentioned that this method of data plotting provides one view of the data which is especially favorable to the Zipf's law normalization method. Next we examine the results of the MA-plots, a technique that is especially favorable to the quantile normalization method.

In order to access the effectiveness of the normalization method, pairwise comparisons using MA-plots (sometimes called RI plots, or log ratio vs. log mean intensity plots) [19] were carried out on the raw data, and data normalized with the global mean method, quantile normalization and Zipf's law on both data set A & B (Figure 2 &3 respectively). With the raw data, the distribution of log-intensity ratios is not centered around zero which is as expected in an un-normalized data set. There is a noticeable intensity dependent effect, sometimes described as a 'banana' curve, which is characteristic of many microarray data sets. Normalization with the global mean method results in a shift of the center of the log-intensity ratio

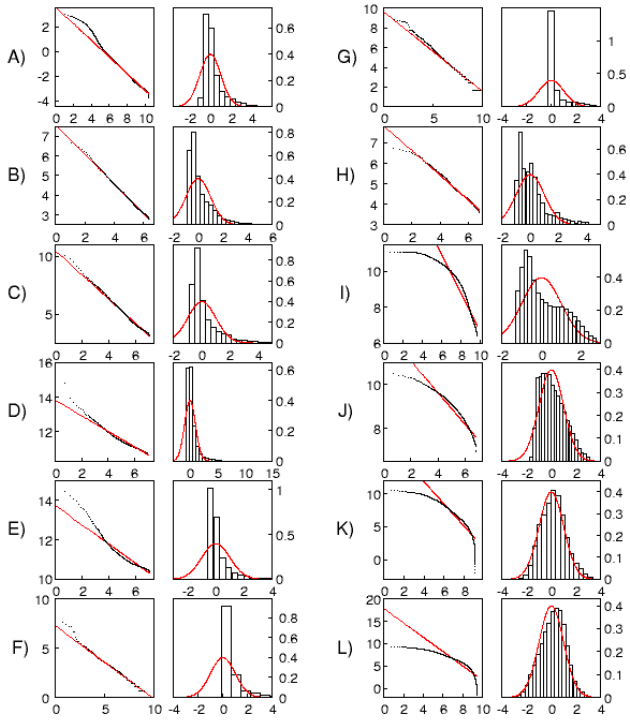

Figure 6

Data set comparison. Eleven microarray data sets (A-K) exhibiting varying degrees of conformation to power law and log normal distributions. On the left for each data set is a log mean intensity vs. log rank plot of the entire data set. Each array was sorted independently by intensity, and mean intensities for each rank over all arrays are plotted. A linear regression line is shown in red. Data sets with a linear distribution adhere well to a power law distribution. On the right for each data set is the distribution $(\ln(i) - \mu) / \sigma$ of the mean intensities used in the left hand plots, where i is the mean measured intensity for each rank and μ and σ are the mean and variance of i respectively. The standard normal curve N(0,1) is shown in red for comparison. Data sets that display a standard normal distribution adhere well to a log normal distribution.

distribution closer to zero, one important criterion for well normalized data, however, especially in the low log mean range, the bulk of the data points still deviate appreciably from zero. The intensity dependent effect is evident, with the low intensity end of the loess fit curving away from the zero axis. The intensity dependent effect is removed using the quantile method. The log intensity ratios of the data distributions normalized using Zipf's law are well centered around zero, but the intensity dependent effect is still apparent. In this case however, the bulk of the data lies very close to zero on the log-ratio scale. [see Additional file 2] This is due to the fact that Zipf's law normalization not only shifts the data distribution on the log ratio scale, but also rotates the whole distribution in log-ratio log-intensity space.

The Kolmogorov-Smirnov test is often used to determine whether data distributions differ significantly and provides a test statistic that measures the proportion of overlap between distributions which ranges from 0 (in the case of identical distributions) to 1 (for non-overlapping distributions) [20]. Mean Kolmogorov-Smirnov values (Table 2a, b) showed the expected trend, with the high values for raw, unnormalized data decreasing when global median normalization was applied, decreasing again after Zipf's law normalization, and reaching zero for both data sets under quantile normalization. It should be noted that the Kolmogorov-Smirnov test statistic will always be zero after quantile normalization because this method forces the data distributions of all microarrays to be identical.

Normalization Results – Two Channel Boutique Microarray

Plots of $\log_e$ intensity versus $\log_e$ rank fitted with linear regressions show that the Zipf's law normalization based on internal controls (Figure 4a) and on selected housekeeping genes (Figure 4c) have relatively similar coefficients *c* according to Zipf's power function (formula 1) as evidenced by the similarity in slopes of the regression lines. Loess normalization using selected housekeeping genes (Figure 4b) showed slightly more variation in *c* coefficients. The unnormalized raw data is also depicted (Figure 4d) along with two other normalization results, the loess method (Figure 4e) and the quantile method (Figure 4f). These are provided for reference only. Neither method can be validly applied to boutique arrays because both rely on the 'well-behaved' genes assumption.

It should be noted that much of the variation in *c* coefficients under the various normalization regimes is due to one channel (Cy3) on one microarray which had low median intensity and high variance due to low labelling efficiency (depicted in black in Figure 4). When normalized with the loess techniques (Figure 4c and 4f) the second channel (Cy5) on this array is adjusted to have a similar median intensity and variance, possibly skewing the results in favour of the Zipf's normalization techniques. To make the normalization method comparison unbiased, we eliminated this array from the analysis [see Additional file 3]. The Zipf's normalization based on internal controls (a) showed the lowest variation in *c* coefficients, the methods based on selected housekeeping genes (b, c) performed approximately equally well. Here again, raw (d), quantile normalized (e), and loess normalized (f) plots are provided for reference only.

Table 1: Data set comparison

Set	Microarray Platform	Number of Data Points	Number of Expts	R2	GEO platform	GEO experiment	Array type
A.	Human Unigene RZPD 1	34560	31	0.9877	GLP284	GSE1510	cDNA, membrane
B.	Clontech Atlas Rat cDNA Expression	588	39	0.9968	GPL158	GSE1509	cDNA, membrane
C.	Clontech Atlas Human 1.2 (I & II)	1176	10	0.9903	GPL127, GPL128	GSE751	cDNA, membrane
D.	Clontech Atlas Mouse 1.2	1159	12	0.9460	GPL144	GSE565	cDNA, membrane
E.	Clontech Atlas Human Cancer 1.2	1160	36	0.9109	GPL158	GSE796	cDNA, membrane
F.	NlaIII: Rattus norvegicus	76790	1	0.9982	GPL23	GSM1679	SAGE
G.	NlaIII: Homo sapiens	101677	1	0.9978	GPL4	GSM14771	SAGE
H.	Mouse Apoptosis	1024	5 × 2	0.994	--	--	cDNA, glass
I.	Caltech 16K cDNA mouse	908	58	0.8892	na	na	cDNA, glass
J.	Stanford Human Unigene	908	24	0.9081	na	na	cDNA, glass
K.	Affymetrix GeneChip Rat Genome	8799	24	0.8538	GPL85	GSE776	Oligo, glass
L.	Affymetrix GeneChip Human Genome	12625	24	0.7773	GPL91	GSE803	Oligo, glass

Eleven microarray data set comparison. Raw intensities, without background subtraction, were used. Controls and blanks were excluded. For Affymetrix chips (K and L), MM/PM ratios were used. For data set B two different Atlas arrays were analyzed together, when analyzed separately they gave similar results. For two channel array systems (I and J), each channel was treated as a separate array. For set I, only the cyanine-3 channel (spleen sample control) was used and for set J, both channels were used for analysis. Reference for data set J: Ross et. al. [31].

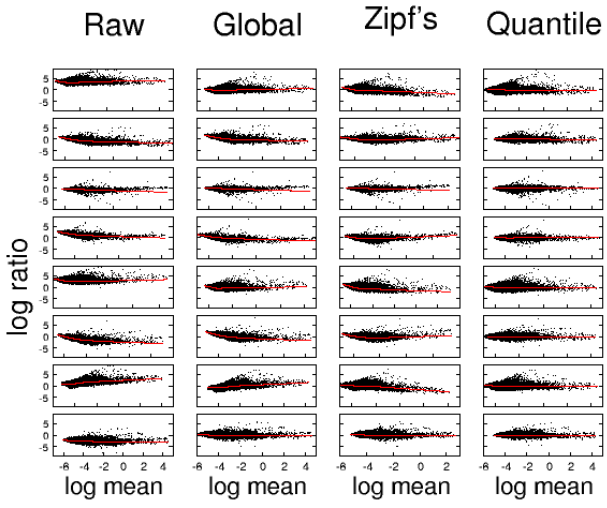

Figure 2
Unigene microarray MA plots. MA plots of Raw Unigene data compared to data normalized with the Global mean, Zipf's, and Quantile methods (columns). Each row of plots represents one pairwise comparison, only 8 of the possible 10 pairwise comparisons of the 5 microarrays used in figure 1are shown. Lowess curves are plotted in red.

We generated MA plots for each of the normalization methods we compared (Figure 5). Typically, MA plots are produced from data from each channel of a single microarray. In addition to these 'within-array' plots (the first three rows of graphs in Figure 5), we also examined 'between-array' plots to evaluate the potential of the normalization methods to allow us to perform across array comparisons. The Zipf's using internal controls was slightly more well centered around the zero log ratio axis than the methods using selected housekeeping genes, especially in between-array plots. The raw and loess normalized plots are provided for reference only.

Finally, to quantify the differences between distributions after normalization, pairwise Kolmogorov-Smirnov values were computed for both the complete boutique array data set (Table 2c) and after eliminating the array which contained a low median intensity and high variance due to low labelling efficiency (Table 2d). In addition to computing the Kolmogorov-Smirnov values for all possible between-array pairwise combinations, we also summarized just the within-array pairwise comparisons (in parenthesis in Table 2). Of the normalization methods which can be applied to boutique microarrays, the Zipf's method using internal controls produced the most similar data distributions when all possible between-array com-

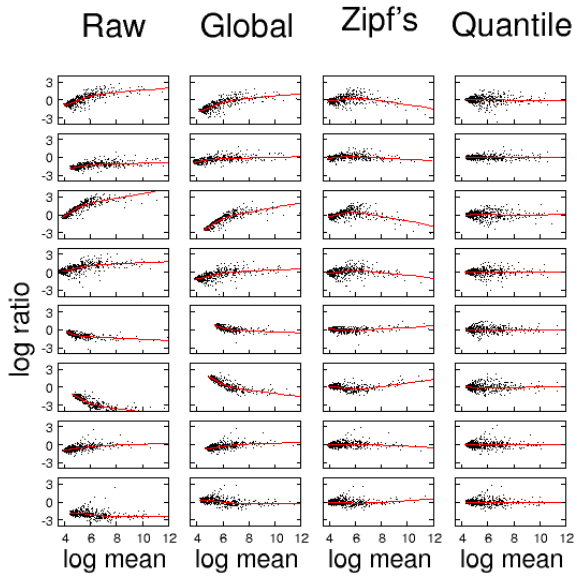

Figure 3
Clontech microarray MA plots. MA plots of Raw Clontech Rat data compared to data normalized with the Global mean, Zipf's, and Quantile methods (columns). Each row of plots represents one pairwise comparison, only 8 of the possible 10 pairwise comparisons of the 5 microarrays used in Additional file 1 are shown. Lowess curves are plotted in red.

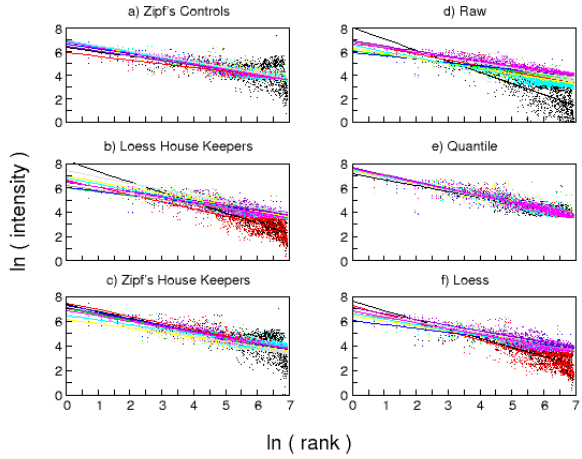

Figure 4
Boutique microarray log plots. Five mouse apoptosis boutique microarrays used in the mouse cell line experiments. Upper left to lower right: Log_e median gene intensity vs. log_e rank – **a**. Normalized according to Zipf's law, using internal positive and negative controls as proxies for the whole data set. **b**. Normalized with a loess curve fit using a selected set of housekeeping genes as proxies (see Methods). **c**. Normalized according to Zipf's law, using the same selected set of housekeeping genes as in b. as proxies **d**. The raw data. **e**. For comparison purposes only, normalized using the quantile method. **f**. For comparison purposes only, normalized using the standard loess method.

Table 2: Kolmogorov-Smirnov values

	Microarray Platform	Pairwise Combinations (within array)	Raw	Global Median	Zipfs	Quantile	Loess	Loess HK	Zipfs Control	Zipfs HK
a.	Clontech Atlas Rat cDNA Expression	465	0.539	0.484	0.119	0	na	na	na	na
b.	Human Unigene RZPD 1	703	0.662	0.225	0.060	0	na	na	na	na
c.	Mouse Apoptosis	45 (5)	0.548 (0.631)	**0.340 (0.318)**	**0.149 (0.167)**	**0 (0)**	**0.471 (0.042)**	0.487 (0.172)	0.182 (0.179)	0.303 (0.296)
d.	Mouse Apoptosis Subset	28 (4)	0.568 (0.667)	**0.303 (0.287)**	**0.111 (0.129)**	**0 (0)**	**0.317 (0.038)**	0.341 (0.190)	0.145 (0.128)	0.315 (0.291)

Three microarray data sets presented in this paper and seven normalization techniques were compared by computing the mean Kolmogorov-Smirnov values of all possible pairwise combinations of arrays within a data set. In the case of the two channel mouse apoptosis microarray, within-array pairwise comparisons were also computed and are shown in parenthesis (here n = the number of arrays, as each array has 2 channels). The symbol 'na' indicates that the normalization techniques which can only be carried out on two channel (loess) or boutique (loess HK, Zipfs Control, Zipfs HK) arrays were not performed on single channel arrays. Values in bold typeface were computed for reference purposes only – these normalization methods cannot be validly applied to boutique microarrays.

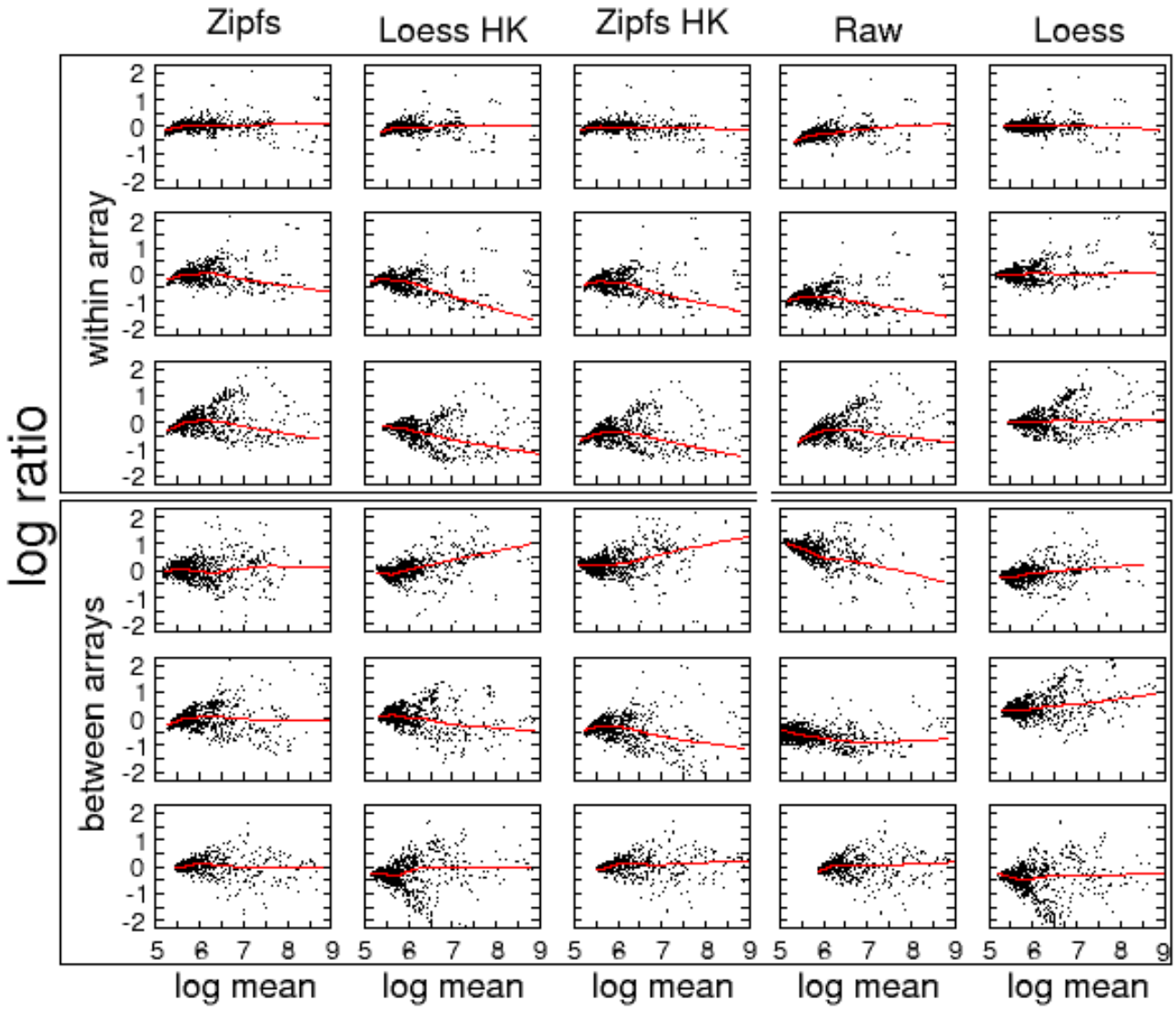

Figure 5
Boutique microarray MA plots. MA plots of the boutique data set comparing (in columns) Zipf's normalization using controls (Zipfs), Zipf's normalization using housekeepers (Zipfs HK), loess normalization using housekeepers (loess HK), raw data, and, for comparison purposes only, the standard loess normalization. Each row of plots represents one pairwise comparison, only 6 of the possible 45 pairwise comparisons of the 5 microarrays used in figure 4 are shown. The top three rows show within-array comparisons, and the bottom three rows show between-array comparisons. Lowess curves are plotted in red.

6, data sets A-E) and log normal distributions (Figure 6, data sets I-K). Of the six power law data sets, two (B and C) clearly followed Zipf's law distributions. The remaining four (data sets A, D, E, and H), while still power-law distributed, showing noticeable deviations from the distribution at the lower rank (higher intensity) portion of the distribution. Of the platforms that where recognizably log normal in distribution, two fluorescent dye labeled, oligo-based Affymetrics platforms (data sets K and L) followed the distribution most closely and two dye labeled, cDNA systems (data sets I and J) were perceptibly log normal. The two SAGE experiments (data sets F and G) which

were included for comparison purposes, exhibited Zipf's law distributions. Coefficients of determination (r^2) of the log mean intensity vs. log rank are a measure of conformation to a power-law distribution and ranged from 0.9968 to 0.7773 for microarray data sets, 0.9982 and 0.9978 for the SAGE experiments (Table 1).

Discussion

Zipf's law is based on observations made by linguist George Kingsley Zipf that the frequency of word occurrences in natural languages is proportional to the negative power of the rank order of the word. Beside the original findings in natural languages [15], Zipf's law has been found to apply to a plethora of natural phenomena, from the populations of cities to the impact factors of scientific journals as well as a variety of biological data, of which a review made available by Wentian Li [21] is an excellent online resource. It is important to point out, that being a phenomenological principle, Zipf's law does not imply that there is a universal underlying physical process at work. However, in much the same way that the Gaussian-Normal distribution occurs naturally in data and can be used to statistically test or otherwise manipulate the data, the fact that microarray data conforms to Zipf's law can be adapted for the purpose of microarray normalization.

Zipf's law is a power law function that states that the magnitude of an intensity measurement is inversely proportional to the rank of that data point in the data set, where c is a coefficient close to -1. Ranking is a method common in statistics, which has previously been used to analyze microarray data. Hoyle el. al. [14] used ranking as a method for evaluating microarray data and proposed the use of several statistics including χ^2 to quantify the agreement of the distribution to Benford's Law [22], and σ^2 as a quality control measure to detect such factors as low signal to background ratio, or mRNA probes extracted from mixed cell types. Ranking also figured prominently in the evaluation of a survey of inter-array normalization methods [23] where the statistics 'absolute rank deviation' and 'relative rank deviation' were used to select the method that produces the most 'well-normalized' data. The normalization procedure described in this paper is the first to combine these two ideas, namely that ranking can be used to judge the effectiveness of a normalization method, and that microarray data conforms to Zipf's law. We evolved these ideas into a novel and easily applicable normalization method and compared this method with existing methods to eliminate non-biological variation from microarray data sets.

In order to implement an appropriate data normalization technique, it is important to know the distribution of a given data set. Several publications have examined the data distributions that typically result from microarray experiments. In a survey of seventeen microarray data sets, sixteen of which were fluorescent dye labeled, Hoyle *et. al.* [14] reported that microarray data were found to have a log normal distributions with power law tails. More recent publications have reported that the abundance of expressed genes exhibit power-law distributions [11,13,24]. Results from our own data sets and a subsequent survey of publicly available data sets from both radioactively and fluorescently labeled platforms suggest that both types of distributions can be manifested in microarray data.

Comparisons between the Zipf's law and quantile normalization methods using MA plots showed that the quantile method effectively removes intensity dependant effects, sometimes referred to as 'banana' curves, from microarray data sets, while the Zipf's law method has no effect on the curved nature of the intensity dependent effect. This is not altogether unexpected as the quantile method was specifically designed to remove such effects. While the Zipf's method does not remove the curve from the intensity dependent effect, it does minimize negative consequences by rotating the data distribution such that the maximal number of data points lie on the zero of the log ratio axis. In this respect, the Zipf's law normalization technique can be considered inferior to the quantile method, however, it may still be a useful tool where the quantile method cannot be applied.

One such case, in which quantile normalization is inappropriate, is with so called 'boutique' microarrays where the genes spotted on the array represent a selected set of genes, for example from a specific pathway or those involved with a particular biological process or disease state. In such systems, most genes are expected to be differentially regulated when control and experimental samples are compared and the expected data distribution of control samples may be significantly different than that of experimental samples (in mean intensity for example). The quantile normalization method would effectively remove this difference by replacing the data distribution of each microarray with the mean distribution of all arrays. In contrast, the principle of normalization according to Zipf's law can also apply to arrays of this type if a group of control spots are included on the microarray. These control spots could be an external reference probe which hybridises to a concentration gradient of matching spots on the array, or internal positive (highly expressed genes) and negative (spotting buffer) control spots on the microarray, or an *a priori* selected set of housekeeping genes using a method such as that described by Wilson et. al. [4] or Schadt et. al. [25]. A linear model can be fitted to the control spots alone, and the normalization procedure can then be applied using the control spots as a proxy for the entire data distribution. The critical assumption in

using control spots in normalization is establishing their relationship to the experimental spots.

The results of our comparison between methods which are designed to normalize boutique microarray data show that Zipf's law normalization using internal control spots results in a relatively well normalized data set when compared to Zipf's law normalization using selected housekeeping genes and the modified loess method using selected housekeeping genes. In addition, the Zipf's law method produced data distributions which are more similar between arrays allowing for between-array comparisons which are advantageous in terms of both cost, because of the reduced number of microarrays that need to be run, and, statistical power, by allowing for greater numbers (n), experimental design permitting.

Conclusion

In summary, we examined the applicability of using Zipf's law as the basis for a novel normalization technique, which is applicable to both one channel microarray data and two channel microarrays. This method is shown to out-perform such methods as global normalization to the mean but would appear to be inferior to quantile normalization. The quantile method was superior to Zipf's law in removing intensity dependent effects commonly seen in microarray data. While the latter method cannot be applied to boutique arrays, we show that the Zipf's normalization method used with internal positive and negative controls or with selected housekeeping genes normalizes boutique arrays as well as currently existing methods. Additionally, data normalized with the Zipf's method using internal control spots seems more amenable to between-array gene intensity comparisons when compared to other methods.

Methods

Data acquisition

Data set A (Table 1) was generated using a global genome-wide cDNA clone set (Human UniGene clone set RZPD 1 Build 138, NCBI [26]), which consisted of ~33,792 cDNA clone inserts spotted in duplicate onto membranes [16]. These microarrays (n = 31) were hybridized with ^{33}P-labeled cDNA derived from total RNA extracted from biopsy material from the sigmoidal colon of normal (control, n = 11), and patients with Crohn's disease (condition A, n = 10) and ulcerative colitis (condition B, n = 10). To emphasize that our normalization technique can be used to normalize other array systems, the second array set used was a smaller, but widely used, commercially available microarray system. Data set B (Table 1) was generated by using Atlas Rat cDNA microarrays (Clontech, 588 genes) probed with rat brain tissue, from control (cerebellum n = 10, olive n = 10) and harmaline treated (cerebellum n = 10, olive n = 9) animals. A third microarray data set, data set H (Table 1) was included to demonstrate the normalization method on two channel fluorescent based (Cy3/Cy5) oligonucleotide systems. These custom produced boutique microarrays (n = 5) contained 1024 spots, and were used in a study to identify differences in apoptotic mechanisms in two different mouse cell lines. Microarrays were probed according to established protocols and exposed to imaging plates overnight (BAS-MS 2325) and scanned at a 50 μm resolution on a FLA-3000G phosphoimager (Raytest, Germany). Image gridding was carried out using VisualGrid® software [27], and intensity data was stored in a relational database and normalized and analyzed using database stored procedures and Perl scripts. All data was normalized from raw data, no background subtraction or other inter-array normalization was performed. Plots were generated using the Grace software package [28].

Normalization

Normalization was accomplished by transforming the data such that the coefficient c and proportionality of the Zipf's power function (formula 1) are identical for all microarrays. This is easily achieved using a regression model on the $\log_e$ intensity versus $\log_e$ rank transformed data, which has the general form,

$$\ln(y) = a + b\ln(r) + e \quad (2)$$

where y is the intensity, r is the rank, a is the regression constant (corresponding to proportionality in Zipf's power function), b is the regression coefficient (corresponding to the coefficient c in Zipf's power function), and e is an error coefficient, which is assumed to be normally distributed.

The first step in this three step procedure was to compute the median intensity of each gene over all microarrays to establish ranks, which were used as the 'reference' to which all microarrays were normalized. This was done by taking the median intensity (y_{med}) of each gene, over all microarrays on which it was measured, and sorting the resulting list of medians to obtain their median ranks (r_{med}). The regression model (2) is applied to the $\log_e$ median intensities and their ranks to estimate a_{med} and b_{med} using the least squares method,

$$\ln(y_{med}) = a_{med} + b_{med}\ln(r_{med}) \quad (3)$$

The ranking of genes by their median intensities effectively groups genes of similar overall expression level along the log rank axis. Under the assumptions that most genes are not differentially expressed, the reference curve generated from the median intensities should have an identical regression coefficient and constant to that of each individual microarray plotted using the ranks deter-

mined by the medians. For the genes which are differentially expressed, the median value represents a 'center' around which expression levels on each individual array may vary, and the neighbouring (by rank) genes, which do not (or only slightly) vary, act to stabilize the regression line and allow normalization to be performed.

In the second step of the normalization procedure, the regression model was applied individually to each microarray using the same ranking as the reference curve,

$$\ln(\gamma_k) = a_k + b_k \ln(r_{med}) \qquad (4)$$

This results in a set of coefficients a_k and b_k which are estimated individually for each array using the least squares method, where k is equal to the number of microarrays in one channel systems, and equal to 2 time the number of microarrays (one for each channel) in two channel systems. Data from two channel arrays were treated in the same way as one channel systems, i.e. each channel was treated independently.

In the third step, the difference between the expected gene intensity value on the *k*th array and that of the reference curve was applied as the normalization factor,

$$\gamma_k' = \exp\left(\ln(\gamma_k)\left(\ln(\gamma_{med})/\ln(\gamma_k)\right)\right) \qquad (5)$$

A scaling factor was applied to the raw data before normalization such that the values γ_k, $\hat{\gamma}_{med}$ and $\hat{\gamma}_k$ were always greater than one to avoid negative values after log transformation. After normalization, the same scaling factor was applied to the data to back transform to their original magnitude. For example, if the smallest raw value in the data set was 0.1, the unlogged raw data was multiplied by a scaling factor of 10 before normalization, and the unlogged normalized data was divided by the same scaling after normalization.

In the special case of our third microarray data set (see Methods: Data Acquisition) which was a boutique array, the same procedure as described above was applied with the following modifications. Each microarray contained 32 spots each of internal positive controls (GAPDH, glyceraldehyde-3-phosphate_dehydrogenase) and internal negative controls (spotting buffer). The medians of all gene intensities were computed (including internal positive and negative controls), and median ranks were assigned as described. However, only the medians of the 64 internal control spots were used to estimate a_{med} and b_{med}, and only the 64 internal control spots from each array were used to estimate a_k and b_k. In both cases, the ranks generated from the entire data set, were used. The normalization factor was then applied over the entire data set as described above.

An alternative to the used of internal control spots for the normalization of boutique microarrays was also explored. Wilson, et. al. [4] described a method wherein a set of 'housekeeping' genes is selected *a priori* from the data set by virtue of their low variance in intensity and such that the entire range of intensities observed on the microarrays is uniformly represented. We also applied the Zipf's law normalization technique to our boutique microarrays using the set of housekeeping genes selected using the method of Wilson, et. al.

In addition to the normalization method based on Zipf's law, all data sets were normalized to a global mean (the mean of logged intensities from all microarrays) and the quantile method. The quantile method is applied by ranking the genes in each array by intensity, taking the median intensity at each rank, and replacing each gene intensity with the median intensity corresponding to the same rank. All normalization methods were compared to each other and to the raw data distribution using box plots and MA plots (pairwise array comparisons of the log-intensity ratio (M) to the mean log-intensity (A)). The two channel boutique microarray data set allowed further normalization methods not possible on one channel array systems to be applied. We normalized this data set using the popular loess method [19], and a modified Loess method specifically designed for boutique arrays using selected housekeeping genes described by Wilson, et. al. [4].

Software

The Zipf's normalization procedure was initially implemented as an SQL stored procedure in a relational database. However, because this is not easily transferable to other systems, we provide two further implementations, a Perl script and an Excel macro [see Additional files 4, 5]. Implementations are available for download from our website [29] and as additional files accompanying this paper. Both the Perl script and Excel macro implement matrix algebra style computation, using either built-in functions or the Perl PDL module [30]. Normalization of two channel arrays with the loess method was performed using the marray package from R's Bioconductor [4]. Loess normalization using selected housekeeping genes and the selection of the housekeeping genes themselves was done with the tRMA package [19] which is publicly available for download on the internet. Sample data sets are also provided with this paper [see Additional files 6, 7, 8].

Normalization method comparison

To compare and evaluate the effectiveness of the various normalization methods applied in this paper, several well established methods were used along with some less common techniques. MA plots [19] are a convenient way to examine differences in fluorescent marker efficiency and

other dye effects in two channel microarray systems. In addition to the standard practice of generating within-array MA plots, we apply them additionally to one channel systems and between arrays in two channel systems to evaluate the extent to which a normalization procedure allows for multiple pairwise comparisons between microarrays. Plots of $\log_e$ intensity versus $\log_e$ rank fitted with linear regressions are a way to visually evaluate the normalization procedure according to the criteria of the Zipf's Law normalization. Specifically, all arrays have identical coefficients *c* and proportionality for the Zipf's power function when the slops and y-intercepts of the regression lines are identical. Finally, to quantify the similarity between microarray distributions after normalization, the mean Kolmogorov-Smirnov value was calculated over all possible pairwise combinations of microarrays within an experiment. In the case of two channel arrays, the mean of within-array Kolmogorov-Smirnov values was also computed (n = the number of arrays). It should be emphasized that even though the Kolmogorov-Smirnov values are technically a test statistic, no statistical test is performed. The values are here used only as a measure of similarity between microarray distributions.

Microarray platform comparison

The underlying premise of the Zipf's normalization method is that microarray data distributions follow a power law distribution such that the relationship between the log intensities and the log ranks is clearly linear. While this assumption holds true for the three data sets we present in this paper, to evaluate the general applicability of the method we also examined eight publicly available data sets (Table 1, data sets C-G, I, K-L) from the NCBI Gene Expression Omnibus [18], and one unpublished data set from an independently maintained website [31] (Table 1, data set J). The survey contains a variety of microarray system types (cDNA vs. Oligo based, radioactivity vs. dye labeled systems, academic vs. commercially produced) and two SAGE experiments for comparison. Two plots were generated for each data set to ascertain the conformity to the Zipf's power law distribution and the log normal distribution respectively. For each data set, a representative array was constructed by ranking the intensities within each array, and then mean over ranks were taken. To determine how well data sets follow the Zipf's power law distribution, log intensity vs. log rank plots were constructed and linear regressions were performed. Data distributions, which were very linear in form, closely follow the power law distribution. A second plot of the distribution of $(\log y - \mu) / \sigma$, where y is the mean intensity over ranks, and μ and σ^2 are the mean and variance, was made for each data set to visualize the conformity to log normal distribution.

List of abbreviations

EST – Expressed Sequence Tag

MA – log ratio (M) vs. mean log intensity (A)

NCBI – National Center for Biotechnology Information

RZPD – Deutsches Ressourcenzentrum für Genomforschung GmbH

SAGE – Serial analysis of gene expression

SQL – Structured Query Language

Authors' contributions

TL conducted the data analysis and implementation of algorithms, participated in the development of the normalization method and is principle author of this manuscript. CMC generated the Unigene and Clontech microarray data set, participated in the development of the normalization method and participated in manuscript preparation. PJPC conceived of and participated in the development of the normalization method. RH participated in the generation of microarray data sets and participated in the development of the normalization method. GD conceived of and coordinated neurology related aspects of this study. SS conceived of and coordinated gastrointestinal related aspects of this study.

Additional material

Additional File 1

Clontech microarray log plots *Five rat Clontech microarrays from the panel of thirty-nine microarrays probed with rat-brain tissue. Upper left to lower right:* ***a****.* Log_e *median gene intensity vs.* log_e *rank – conformity to Zipf's law is demonstrated by the linear regression line (in red)* ***b****. Five microarrays chosen to maximize pre-normalization variability, each plotted according to the gene ranks determined by their median gene intensity levels.* ***c****. The same five microarrays, normalized to a global median, with regression lines.* ***d****. The same five microarrays, normalized with the quantile method, with regression lines.* ***e****. The same five microarrays normalized taking Zipf's law into account, with regression lines. For plots b-d, a sub-sample of 50% of the data points are plotted for readability.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-6-37-S1.png]

Additional File 2

Mean of squared log ratios from MA plots in *Figure 2 In Figure 2, it is difficult to see that the distribution of the Zipf's normalized data is more closely centered around zero on the log ratio axis than the Globally normalized data. To quantify this, the mean of squared log ratios was computed for each MA plot. The positions of the values in this table correspond exactly to the positions of the plots in Figure 2. In 6 out of 8 cases, the mean of squared log ratio is smaller in the Zipf's normalized data than in the corresponding Globally normalized data.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-6-37-S2.doc]

Additional File 3

Boutique microarray log plots *Four mouse apoptosis boutique microarrays used in the mouse cell line experiments. This is the same data set as shown in Figure 4, with the array containing one channel with low expression intensities and high variability removed. Upper left to lower right: Log_e median gene intensity vs. log_e rank – **a**. Normalized according to Zipf's law, using internal positive and negative controls as proxies for the whole data set. **b**. Normalized with a loess curve fit using a selected set of housekeeping genes as proxies (see Methods). **c**. Normalized according to Zipf's law, using the same selected set of housekeeping genes as in b. as proxies **d**. The raw data. **e**. For comparison purposes only, normalized using the quantile method. **f**. For comparison purposes only, normalized using the standard loess method.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-6-37-S3.png]

Additional File 4

Requires: Microsoft Excel (Does not handle missing data values.)

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-6-37-S4.xls]

Additional File 5

Requires: Perl (which runs on many platforms), the PDL perl module (Handles missing data values if PDL is compiled correctly.)

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-6-37-S5.pl]

Additional File 6

Microarray type: Filter based cDNA from the RZPD Number of genes: 33,792 Number of microarrays: 31 Probed with: Total RNA from human sigmoidal colon. Within microarray normalization: None

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-6-37-S6.txt]

Additional File 7

Microarray type: Clonetech Atlas Rat cDNA 7738-1 Number of genes: 558 Number of microarrays: 33 Probed with: Total RND from rat cerebellum and olive. Within microarray normalization: None

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-6-37-S7.txt]

Additional File 8

Microarray type: custom made glass slide Number of genes: 1024 Number of microarrays: 5 Probed with: Total RND from mouse cell lines. Within microarray normalization: None

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-6-37-S8.dat]

Acknowledgements

The authors wish to thank Alexander Zein and Carl Manaster for critical input on data analysis techniques. We would like to express our appreciation to the clinicians and volunteers who provided tissue samples, and Hans Moises and Henrik Wilms for rat brain samples. We gratefully acknowledge the technical assistance of Brigitte Mauracher, and the invaluable assistance of the Max-Planck Institute for Molecular Genetics in Berlin, in particular Hans Lehrach, Holger Eickhoff and Elke Rohlfs. We also thank Sandra Freitag for advice on formulating the equations. This research was supported in part by a Training and Mobility of Researchers (TMR) grant, as well as grants from the German National Genome Research Program, the National Genome Research Network (NGFN) and the DFG (FOR423).

References

1. Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays.** *Nat Genet* 1999, **21(1 Suppl):**33-37.
2. Lander ES: **Array of hope.** *Nat Genet* 1999, **21(1 Suppl):**3-4.
3. Tsodikov A, Szabo A, Jones D: **Adjustments and measures of differential expression for microarray data.** *Bioinformatics* 2002, **18(2):**251-260.
4. Wilson DL, Buckley MJ, Helliwell CA, Wilson IW: **New normalization methods for cDNA microarray data.** *Bioinformatics* 2003, **19(11):**1325-1332.
5. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res* 2002, **30(4):**e15.
6. Zien A, Aigner T, Zimmer R, Lengauer T: **Centralization: a new method for the normalization of gene expression data.** *Bioinformatics* 2001, **17(Suppl 1):**S323-31.
7. Velculescu VE, Madden SL, Zhang L, Lash AE, Yu J, Rago C, Lal A, Wang CJ, Beaudry GA, Ciriello KM, Cook BP, Dufault MR, Ferguson AT, Gao Y, He TC, Hermeking H, Hiraldo SK, Hwang PM, Lopez MA, Luderer HF, Mathews B, Petroziello JM, Polyak K, Zawel L, Kinzler KW, *et al.*: **Analysis of human transcriptomes.** *Nat Genet* 1999, **23(4):**387-388.
8. Raychaudhuri S, Stuart JM, Altman RB: **Principal components analysis to summarize microarray experiments: application to sporulation time series.** *Pac Symp Biocomput* 2000:455-466.
9. Kerr MK, Martin M, Churchill GA: **Analysis of variance for gene expression microarray data.** *J Comput Biol* 2000, **7(6):**819-837.
10. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, **7(3-4):**601-620.
11. Furusawa C, Kaneko K: **Zipf's law in gene expression.** *Phys Rev Lett* 2003, **90(8):**088102. Epub 2003 Feb 26.
12. Ogasawara O, Kawamoto S, Okubo K: **Zipf's law and human transcriptomes: an explanation with an evolutionary model.** *C R Biol* 2003, **326(10-11):**1097-1101.
13. Ueda HR, Hayashi S, Matsuyama S, Yomo T, Hashimoto S, Kay SA, Hogenesch JB, Iino M: **Universality and flexibility in gene expression from bacteria to human.** *Proc Natl Acad Sci U S A* 2004, **101(11):**3765-9. Epub 2004 Mar 03.
14. Hoyle DC, Rattray M, Jupp R, Brass A: **Making sense of microarray data distributions.** *Bioinformatics* 2002, **18(4):**576-584.
15. Zipf GK: **The psycho-biology of language; an introduction to dynamic philology.** Boston, , Houghton Mifflin Company; 1935:ix , 2 *., [3]-336.

16. Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel H: **Normalization strategies for cDNA microarrays.** *Nucleic Acids Res* 2000, **28(10):**E47.
17. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19(2):**185-193.
18. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30(1):**207-210.
19. Dudoit S, Yang YH, Callow MJ, Speed TP: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** *Statistica Sinica* 2002, **12(1):**111-139.
20. Kendall MG, Stuart A, Ord JK: **Tests of fit based on the sample distribution function: Kolmogorov's Dn.** In *Kendall's advanced theory of statistics Volume 2*. Fifth edition. New York , Oxford University Press; 1987:1187-1188.
21. **Wentian Li's literature review of Zipf's Law** [http://www.nslij-genetics.org/wli/zipf/index.html]
22. Benford F: **The Law of Anomalous Numbers.** *Proc Am Philos Soc* 1936, **78:**551-572.
23. Kroll TC, Wolfl S: **Ranking: a closer look on globalisation methods for normalisation of gene expression arrays.** *Nucleic Acids Res* 2002, **30(11):**e50.
24. Kuznetsov VA, Knott GD, Bonner RF: **General statistics of stochastic process of gene expression in eukaryotic cells.** *Genetics* 2002, **161(3):**1321-1332.
25. Schadt EE, Li C, Ellis B, Wong WH: **Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.** *J Cell Biochem Suppl* 2001, **Suppl(37):**120-125.
26. **Website of the Deutsches Ressourcenzentrum für Genomforschung** [http://www.rzpd.de]
27. **Homepage of GPC Biotech, makers of VisualGrid®** [http://www.gpc-biotech.com]
28. **Grace plotting software** [http://plasma-gate.weizmann.ac.il/Grace]
29. **Original data sets and Zipf's normalization software** [http://www.mucosa.de/zipfs/zipfs_normalization.html]
30. **The Perl Data Language homepage** [http://pdl.perl.org]
31. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24(3):**227-235.

Article 2

Correlation between genetic and geographic structure in Europe.

This article has been previously published as:

Lao & Lu*, et. al. Correlation between genetic and geographic structure in Europe. Curr Biol. 2008 Aug 26;18(16):1241-8.

*shared primary authorship.

Current Biology *18*, 1241–1248, August 26, 2008 ©2008 Elsevier Ltd All rights reserved DOI 10.1016/j.cub.2008.07.049

Report

Correlation between Genetic and Geographic Structure in Europe

Oscar Lao,[1,22] Timothy T. Lu,[2,22] Michael Nothnagel,[2] Olaf Junge,[2] Sandra Freitag-Wolf,[2] Amke Caliebe,[2] Miroslava Balascakova,[3] Jaume Bertranpetit,[4] Laurence A. Bindoff,[5] David Comas,[4] Gunilla Holmlund,[6] Anastasia Kouvatsi,[7] Milan Macek,[3] Isabelle Mollet,[8] Walther Parson,[9] Jukka Palo,[10] Rafal Ploski,[11] Antti Sajantila,[10] Adriano Tagliabracci,[12] Ulrik Gether,[13] Thomas Werge,[14] Fernando Rivadeneira,[15,16] Albert Hofman,[16] André G. Uitterlinden,[15,16] Christian Gieger,[17,18] Heinz-Erich Wichmann,[17,18] Andreas Rüther,[19] Stefan Schreiber,[19] Christian Becker,[20] Peter Nürnberg,[20] Matthew R. Nelson,[21] Michael Krawczak,[2,23] and Manfred Kayser[1,23,*]

[1]Department of Forensic Molecular Biology
Erasmus University Medical Center Rotterdam
3000 CA Rotterdam
The Netherlands

[2]Institut für Medizinische Informatik und Statistik
Christian-Albrechts University Kiel
D-24105 Kiel
Germany

[3]Institute of Biology and Medical Genetics
University Hospital Motol and 2nd School of Medicine
Charles University Prague
CZ 150 06, Prague 5
Czech Republic

[4]Unitat de Biologia Evolutiva
Pompeu Fabra University
08003 Barcelona, Catalonia
Spain

[5]Department of Neurology
Haukeland University Hospital and Institute of Clinical Medicine
University of Bergen
5021 Bergen
Norway

[6]Department of Forensic Genetics and Forensic Toxicology
National Board of Forensic Medicine
SE 581 33 Linkoping
Sweden

[7]Department of Genetics, Development, and Molecular Biology
Aristotle University of Thessaloniki
GR-540 06 Thessaloniki
Greece

[8]Laboratoire d'Empreintes Génétiques
EFS-RA site de Lyon
69007 Lyon
France

[9]Institute of Legal Medicine
Medical University Innsbruck
A-6020 Innsbruck
Austria

[10]Department of Forensic Medicine
University of Helsinki
Helsinki FIN-00014
Finland

[11]Department of Medical Genetics
Medical University Warsaw
02-007 Warsaw
Poland

[12]Istituto di Medicina Legale
University of Ancona
I-60020 Ancona
Italy

[13]Molecular Neuropharmacology Group and Center for Pharmacogenomics Department of Neuroscience and Pharmacology
University of Copenhagen
2200 Copenhagen
Denmark

[14]Research Institute of Biological Psychiatry and Center for Pharmacogenomics
Mental Health Center Sct. Hans
Copenhagen University Hospital
DK-4000 Roskilde
Denmark

[15]Department of Internal Medicine, Genetics Laboratory
Erasmus University Medical Center Rotterdam
3000 CA Rotterdam
The Netherlands

[16]Department of Epidemiology
Erasmus University Medical Center Rotterdam
3000 CA Rotterdam
The Netherlands

[17]Institute of Epidemiology
Helmholtz Zentrum München - German Research Center for Environmental Health
D-85764 Neuherberg
Germany

[18]Institute of Medical Informatics, Biometry and Epidemiology
Ludwig-Maximilians University
D-81377 Munich
Germany

[19]Institut für Medizinische Molekularbiologie
Christian-Albrechts University Kiel
D-24105 Kiel
Germany

[20]Cologne Center for Genomics and Institut für Genetik
University of Cologne
D-50674 Cologne
Germany

[21]Genetics
GlaxoSmithKline
Research Triangle Park, North Carolina 27709

*Correspondence: m.kayser@erasmusmc.nl

[22]These authors contributed equally to this work
[23]These authors contributed equally to this work

Summary

Understanding the genetic structure of the European population is important, not only from a historical perspective, but also for the appropriate design and interpretation of genetic epidemiological studies. Previous population genetic analyses with autosomal markers in Europe either had a wide geographic but narrow genomic coverage [1, 2], or vice versa [3–6]. We therefore investigated Affymetrix GeneChip 500K genotype data from 2,514 individuals belonging to 23 different subpopulations, widely spread over Europe. Although we found only a low level of genetic differentiation between subpopulations, the existing differences were characterized by a strong continent-wide correlation between geographic and genetic distance. Furthermore, mean heterozygosity was larger, and mean linkage disequilibrium smaller, in southern as compared to northern Europe. Both parameters clearly showed a clinal distribution that provided evidence for a spatial continuity of genetic diversity in Europe. Our comprehensive genetic data are thus compatible with expectations based upon European population history, including the hypotheses of a south-north expansion and/or a larger effective population size in southern than in northern Europe. By including the widely used CEPH from Utah (CEU) samples into our analysis, we could show that these individuals represent northern and western Europeans reasonably well, thereby confirming their assumed regional ancestry.

Table 1. European Subpopulation Summary Statistics

Subpopulation	Code	Total No. Samples	Final No. Samples*	Sex Ratio (M:F)
Norway (Førde)	NO	52	52	1.74
Sweden (Uppsala)	SE	50	46	all male
Finland (Helsinki)	FI	47	47	0.74
Ireland	IE	37	35	4.29
UK (London)	UK	197	194	8.85
Denmark (Copenhagen)	DK	60	59	1.22
Netherlands (Rotterdam)	NL	292	280	all female
Germany I (Kiel)	DE1	500	494	1.08
Germany II (Augsburg)	DE2	500	489	1.02
Austria (Tyrol)	AT	50	50	all male
Switzerland (Lausanne)	CH	134	133	0.81
France (Lyon)	FR	50	50	2.13
Portugal	PT	16	16	0.78
Spain I	ES1	83	81	1.02
Spain II (Barcelona)	ES2	48	47	0.71
Italy I	IT1	107	106	1.38
Italy II (Marches)	IT2	50	49	all male
Former Yugoslavia	YU	58	55	1.90
Northern Greece	EL	51	51	1.43
Hungary	HU	17	17	0.54
Romania	RO	12	12	1.00
Poland (Warsaw)	PO	50	49	all male
Czech Republic (Prague)	CZ	53	45	0.96
Total		2,514	2,457	

Total number of samples, final number of samples after data cleaning, and the sex ratio (male:female) of the final sample data set for each subpopulation. * is after stringent quality control.

Results and Discussion

According to current theory, the autosomal gene pool of extant human populations in Europe lacks sharp discontinuities [1, 2], with the exception of known isolates such as the Finns [6, 7]. For classical genetic markers including, for example, erythrocyte antigens, changes in population genetic structure have been observed to follow a predominantly southeast-northwest gradient [1, 2], thereby apparently matching the Pleistocene settlement of Europe, the Neolithic expansion from the Fertile Crescent, and (at least in part) the postglacial resettlement of Europe during the Mesolithic. Such gradient was also observed with particular haplogroups derived from the nonrecombining part of the Y chromosome (NRY), but other NRY data revealed additional population structure in Europe that has been associated with various demographic events in prehistoric, historic, and modern times [8–10]. In contrast, the European mitochondrial DNA pool has been found to be rather homogeneous [11]. Here, we investigated the genetic structure of the European population by using 309,790 single-nucleotide polymorphisms (SNPs) in 2,457 individuals, ascertained at 23 sampling sites (henceforth referred to as "subpopulations") in 20 different European countries. The data emerged from the genotyping of 2,514 European samples with the GeneChip Human Mapping 500K Array, followed by stringent quality control (see Table 1 and Experimental Procedures for details) and represent the largest Europe-wide genetic study to date.

First, we quantified the amount of information that each SNP could potentially provide about an individual's subpopulation affiliation by using the ancestry informativeness index *In* (Figure S1 available online) [12]. The maximum *In* value (0.09) was observed for rs6730157 in the *RAB3GAP1* gene located about 68 kb away from the Lactase (*LCT*) gene. Furthermore, nine of the 20 (45%) most ancestry-informative SNPs, and 17 of the top 100 (Table S1), were from the *LCT* region and previously showed signatures of a selective sweep in CEU (Centre d'Etude du Polymorphism Humain from Utah) samples [13]. The average *In* across markers was 0.0064 (standard deviation: 0.0032), which represents only 0.93% of the maximum possible *In* of 0.69 in our study. (Note that this maximum would be attained if a SNP was fixed for one allele in 12 subpopulations and for the other allele in the remaining 11 subpopulations).

Second, we performed a principal-component analysis (PCA) in which the first two PCs were found to account for 31.6% and 17.3%, respectively, of the total variation, an amount similar to that reported in previous studies [1, 5]. In our study, the first two PCs revealed a SNP-based grouping of European subpopulations that was strongly reminiscent of the geographic map of Europe (Figure 1; Figure S2). The first PC aligned subpopulations according to latitude, with the two Italian subpopulations at one end and the Finnish subpopulation at the other. The second PC tended to separate subpopulations more according to longitude, with the Finnish subpopulation showing the largest values and the Irish and UK subpopulations showing the lowest values. The apparent geographic footing of the two PCs received additional support from an observed statistically significant positive correlation (Pearson $r^2 = 0.632$, two-tailed $p < 10^{-15}$) between the genetic distance (Euclidian distance between the median first two eigenvectors of the PCA) and the geographic (great-circle) distance between the analyzed subpopulations.

Third, we searched for genetic barriers [14] in our dataset by using the same genetic and geographic distance matrices. This analysis identified two statistically significant barriers for the 23 subpopulations. One barrier was observed between the Finnish and all other subpopulations (first PC considering FI against the rest: $r^2 = 0.074$, two-tailed $p < 10^{-15}$; second PC considering FI against the rest: $r^2 = 0.33$, two-tailed $p < 10^{-15}$) and the other one between the two Italian and all other subpopulations (first PC considering IT1 and IT2 against the

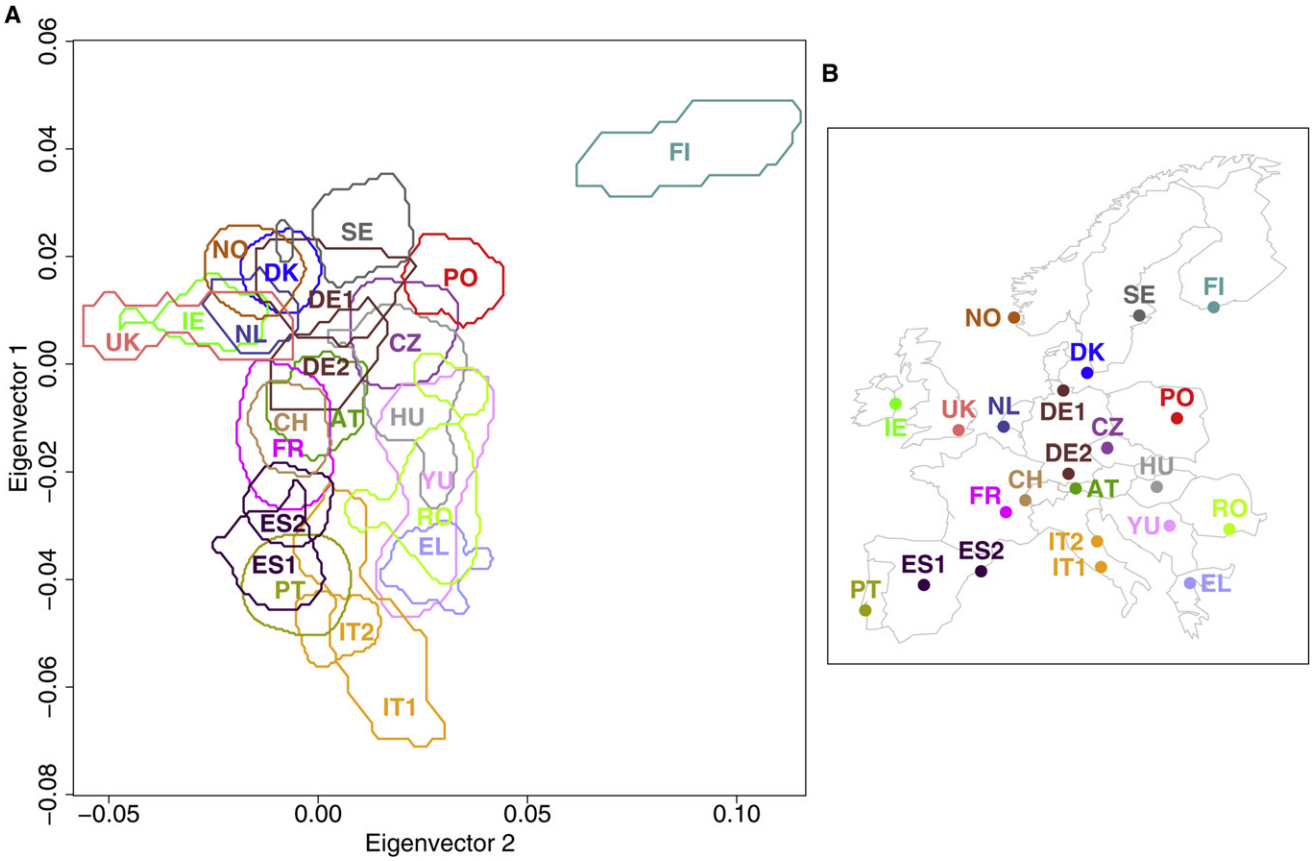

Figure 1. SNP-Based PCA of 2,457 European Individuals from 23 Subpopulations

(A) Kernel density plot of the first two dimensions of a SNP-based PCA using those 309,790 SNPs from the GeneChip Human Mapping 500K Array Set (Affymetrix) that passed quality control.

(B) Geographic distribution of the 23 subpopulations; capitals were used as the respective landmark if location information was either unspecific or lacking (see Table 1 for further sample details).

rest: $r^2 = 0.37$, two-tailed $p < 10^{-15}$; second PC considering IT1 and IT2 against the rest: $r^2 = 0.014$, two-tailed $p = 2.31 \times 10^{-9}$).

Fourth, we studied the geographic distribution of genetic diversity by computing mean heterozygosity and mean linkage disequilibrium (LD) based upon HR^2 [15] between markers at a distance < 10 kb for each subpopulation. Results from both analyses showed that the genetic diversity tended to be larger, and the LD smaller, in southern Europe as compared to northern Europe (Figure 2). Moreover, both analyses supported a genetic gradient of south-north orientation (r^2 adjusted for the number of data points between the mean observed heterozygosity and latitude: 0.76, $p = 3.80 \times 10^{-8}$; adjusted r^2 between HR^2 and latitude: 0.71, two-tailed $p = 4.33 \times 10^{-7}$) but not of west-east orientation (adjusted r^2 between heterozygosity and longitude: 0.03, two-tailed $p = 0.416$; adjusted r^2 between HR^2 and longitude: 0.099, two-tailed $p = 0.078$). Spatial autocorrelation analysis of both variables revealed statistically significant ($p < 0.05$) patterns compatible with a clinal distribution as indicated by the presence of positive and statistically significant autocorrelation values for small pair-wise distances and negative and statistically significant Moran's I values for large distances (see Figure 2). Bearing analysis [16] revealed for the heterozygosity measure the maximal angular correlations ($r = 0.69$) at 87° and the minimal ($r = -0.153$) at 165°, as well as for HR^2 the maximal at 55° ($r = 0.67$) and the minimal ($r = -0.167$) at 160°, thus also suggesting a south-to-north spatial distribution of both variable. These results are compatible with larger effective population sizes in the south than in the north of Europe and/or a population expansion from southern toward northern Europe. Hierarchical analysis of molecular variance (AMOVA) [17] revealed that clustering the individuals according to four geographic groups—north (NO, SE, FI), north-west/central (IE, UK, DK, NL, DE1, DE2, AT, CH, FR), east (HU, RO, PO, CZ), and south (PT, ES1, ES2, IT1, IT2, YU, EL)—explained an average of 0.17% (95% coefficient interval: 0.0% to 0.91%) of the total genetic variance, whereas individual subpopulation affiliation explained 0.25% (95% coefficient interval: 0.0% to 1.25%).

Overall, our study showed that the autosomal gene pool in Europe is comparatively homogeneous but at the same time revealed that the small genetic differentiation that is present between subpopulations is characterized by a significant correlation between genetic and geographic distance. Furthermore, the qualitative nature of these results is in close agreement with expectations based on human migration history in Europe. The major prehistoric waves of human migration in Europe followed south and southeastern to north and northwestern directions [1], including the first Paleolithic settlement of the continent by anatomically modern humans [18], most of the postglacial resettlement during the Mesolithic [19], and the farming-related population expansion during the Neolithic [18, 20]. Thus, both the level and the change in neutral autosomal

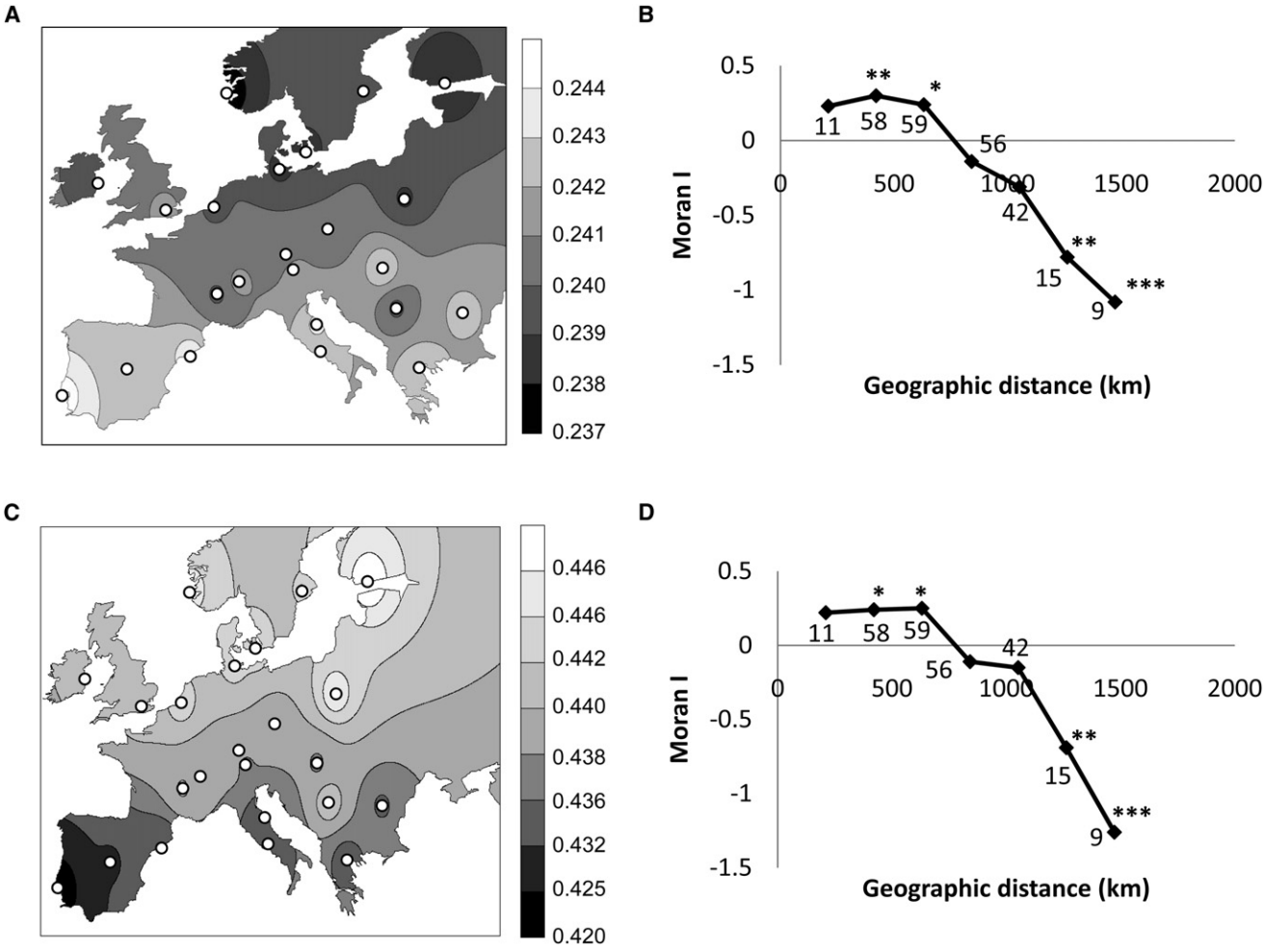

Figure 2. Geographic Distribution of Two Measures of Genetic Diversity across the European Population

(A and B) Isoline map (A) of Europe based on the mean observed heterozygosity in each of 23 European subpopulations with (B) corresponding spatial autocorrelogram.

(C and D) Isoline map (C) of Europe based on the mean observed linkage disequilibrium based on HR^2 in each of 23 European subpopulations with (D) corresponding spatial autocorrelogram. Both spatial autocorrelograms showed statistically significant departures from randomness ($p < 0.05$). For each distance class, the number of subpopulation pairs included and the statistical significance (*, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$) are provided.

variation in Europe can be expected to roughly follow southern-to-northern gradients as we observed, with the possible exception of population isolates as observed for the Finns. On the other hand, migration events in more recent (i.e., historic) times are presumed to have had a more homogenizing effect upon the previously established genetic landscape, as a result of their sporadic nature and haphazard geographic orientation [2]. This implies that genetic differences between extant European subpopulations can be expected to be small indeed. The genetic landscape described by the ~300,000 autosomal SNPs analyzed here closely resembles that previously obtained with 128 alleles from 49 classical markers (see Table 1.3.1 in [1]). This similarity is highlighted by a significant correlation ($r = 0.516$; two-tailed Mantel test $p = 0.0042$, performed with 10,000 Monte Carlo permutations) between the pair-wise F_{ST} values [21] computed for the 19 European subpopulations that overlapped between the two datasets (Danish, Dutch, Yugoslavian, Hungarian, Irish, Italian, Portuguese, Spanish, Swiss, English, German, Austrian, Finnish, French, Greek, Norwegian, Polish, Swedish, and Czechoslovakian). This notwithstanding, a stronger correlation between F_{ST} and great-circle geographic distances was observed for the subpopulations when the SNPs from our study were used ($r = 0.661$; two-tailed Mantel test $p = 0.00010$, performed with 10,000 Monte Carlo permutations) as compared to the classical markers ($r = 0.503$, two-tailed Mantel test $p = 0.00020$, performed with 10,000 Monte Carlo permutations).

Previous studies based on genome-wide SNP diversity reported differences between individuals of southern and northern/central European ancestry [3, 5, 6] and, to a lesser extent, between those of eastern and western European ancestry [3], which were not confirmed in our study. They mostly relied on the analysis of European Americans whose geographic assignment was determined from self-reported family records. Although genetic studies using European Americans can reveal important information about the genetic structure of the European ancestry of European Americans, caution must be exercised when drawing conclusions about the current genetic structure of Europe from European Americans because (1) European migrants may not have been representative of their country of origin, (2) the temporal difference introduced by sampling second- or third-generation descendants means

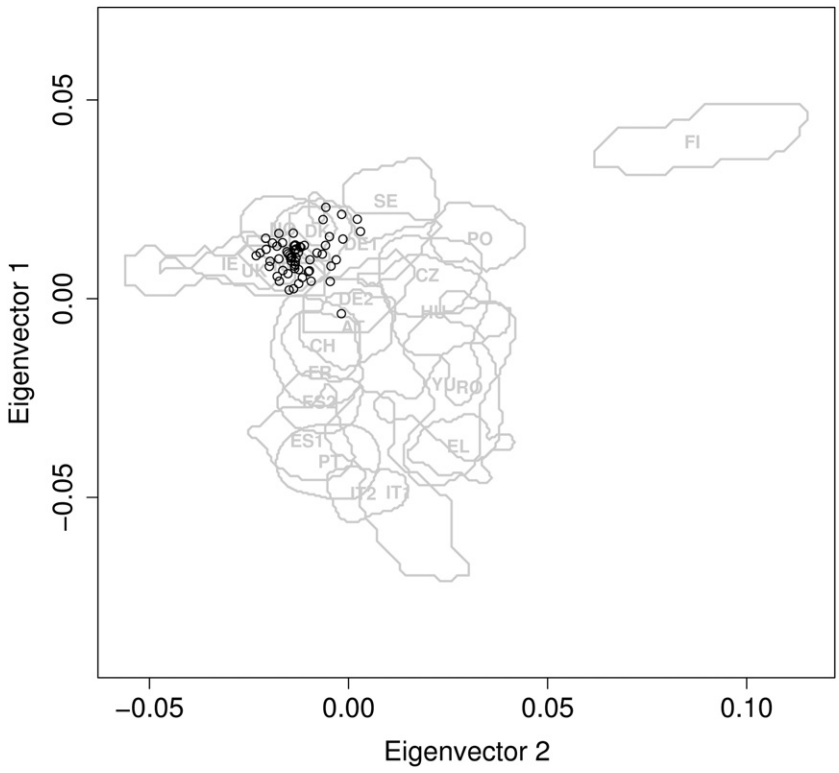

Figure 3. Position of CEPH-CEU Samples in a SNP-Based PCA Kernel-Density Plot of 23 European Subpopulations

CEU individuals (U.S. Americans of European descent from Utah) are plotted as open circles. For details, see Figure 1 and Table 1.

that allele-frequency estimates inevitably ignored recent population movements (i.e., WWII-related migrations), and (3) self-reported geographic origin is error prone [22]. Our study avoided these potential pitfalls by using large samples of individuals of genuinely European origin, as evidenced by the documentation of their respective place of birth or residence being in one of the named subpopulations, and with comprehensive continent-wide coverage.

It is of general interest to place the CEU samples, widely used in genetic epidemiological and population genetic studies as representing the European population, into the context of our findings. The CEPH-CEU panel comprises U.S. Americans who were collected in Utah in 1980 and who are assumed to have descended from migrants originating from northern and western parts of Europe [23]. The samples were also included in the International HapMap Project and formed the basis of selecting tagging SNPs used in current genome-wide association studies with Illumina SNP arrays. Whereas a previous study [3] confirmed the grouping of the CEPH-CEU samples with other northern and western European subpopulations, our study was capable of providing their most precise positioning on the European genetic map (Figure 3). It turned out that, while the CEPH-CEU panel was indeed largely representative of northwestern and central Europeans, parts of Scandinavia as well as southern and eastern Europe were not well represented by these samples (Figure 3). Estimated inflated false-positive rates for all subpopulations were largest in the Finns, followed by the two Italian subpopulations (see Table S2). This implies that researchers conducting genetic-association studies in at least these regions, using the CEPH-CEU samples as controls, may be at increased risk of false-positive associations. Our confirmation of the regional European origin of the CEPH-CEU samples also indicates that inferring the geographic origin of an unknown person from autosomal DNA markers, which is highly relevant in the forensic context, might now be feasible down to the level of European subregions, at least when a large number of genetic markers and a reference database, such as are applied here, are used.

Conclusions

Our comprehensive SNP genotype data from 23 European subpopulations, providing a dense coverage at both the geographic and genomic level and representing the largest Europe-wide genetic study to date, allowed us to describe the genetic structure of the European population with the highest resolution. Although the amount of differentiation within the European autosomal gene pool was found to be small, the existing genetic differences nevertheless correlated well with geographic distances. Furthermore, mean heterozygosity was larger, and mean linkage disequilibrium smaller, in southern than in northern European subpopulations, and both parameters exhibited a continuous clinal distribution across Europe. Overall, our results were compatible with expectations based on European population history, mainly the prehistoric population expansion from southern to northern Europe and/or a larger effective population size in the south as compared to the north of Europe. Our dataset also allowed placement of the widely used CEPH-CEU samples onto the European genetic landscape, essentially confirming their genetic ancestry in northern and western Europe.

Experimental Procedures

Samples and Genotyping

The GeneChip Human Mapping 500K Array Set (Affymetrix) was used to genotype 500,568 SNPs in 2,514 individuals from 23 different sampling sites (henceforth termed "subpopulations") located in one of 20 different European countries. Genotyping according to the instructions provided by the manufacturer was carried out at one of seven specialized centers: the Cologne Center for Genomics at the University of Cologne (Germany) for DE1, NO, SE, FI, AT, FR, ES2, IT2, EL, PO, and CZ; the Helmholtz Zentrum München - German Research Center for Environmental Health for DE2; the genetics laboratory of the Department of Internal Medicine, Erasmus MC (Netherlands) for NL; and the RH Microarray Centre Rigshospitalet, Copenhagen University Hospital (Denmark) for DK (see Table 1 for abbreviation explanations). Samples from the GlaxoSmithKline-sponsored POPRES project (IE, UK, CH, PT, ES1, IT1, YU, HU, and RO) were genotyped at Expression Analysis (Durham, NC, USA) and at Gene Logic (Gaithersburg, MD, USA) (see Table 1 for abbreviation explanations). Some samples belonged to existing control population studies, with detailed descriptions available elsewhere: KORA [24] for DE2, PopGen [25] for DE1, the Rotterdam Study [26–28] for NL, and POPRES (drawn from the LOLIPOP and CoLaus studies) for IE, UK, CH, PT, ES1, IT1, YU, HU, and RO [29–31]. Samples were drawn randomly from these pools or, in the case of POPRES, were ascertained on the basis of sample-size requirements. European migrants from non-European regions were not included in the initial analysis. For 11 of the subpopulations (NO, SE, FI, AT, FR, ES2, IT2, EL, PO, CZ, and DK), samples were

obtained from healthy unrelated volunteers: Norwegian samples (NO) from blood donors of the Førde region, Swedish samples (SE) from the Uppsala region [32], Finnish samples (FI) from the Helsinki area with parents and grandparents originating from various regions in Finland, Austrian samples (AT) from the Tyrol region with parents originating from Tyrol, French samples (FR) from blood donors of Lyon with parents originating from the Rhône Alpes area, Spanish samples (ES2) from Catalonia of blood donors from rural areas who speak Catalan as their mother tongue and who had regional Catalan ancestry for at least two generations [33], Italian samples (IT2) from blood donors of the upland of the Marches region [34], Greek samples (EL) from the north of the country [35], Polish samples (PO) from the Warsaw region of central Poland [36], Czech samples (CZ) from the central Bohemian region in and around Prague, and Danish samples (DK) from the Danish Blood Donor Corps in the Copenhagen area. In addition, GeneChip Human Mapping 500K Array data from CEPH-CEU samples were retrieved from the Affymetrix website (http://www.affymetrix.com).

Quality Assessment and Control Procedure

Array-based SNP genotypes were subjected to stringent quality control: First, each individual was required to have a genotype call rate $\geq$ 93%, with the dynamic model (DM) algorithm with a confidence score of 0.26, and a per-individual call rate $\geq$ 95% for all individuals genotyped by the same facility, with the Bayesian robust linear model with Mahalanobis distance classifier (BRLMM) algorithm with a confidence score of 0.5. The call rate was defined here as the proportion of unambiguous genotypes among either all SNPs (per-individual call rate) or all individuals (per-marker call rate), respectively. Markers that were monomorphic (1.4% of the total), that were located on the X chromosome (2.1%), or that had a per-marker call rate $\leq$ 90% in at least one genotyping facility (5.7%) were excluded, as were those showing a significant ($p \leq 0.05$) deviation from Hardy-Weinberg equilibrium (HWE) in at least one subpopulation (31.3%). HWE was tested by means of a χ^2 test, or by Fisher's exact test when the observed or expected number of a given genotype was less than 5. This method was preferred over others that have been shown to be more powerful [37] because the computational requirements of these methods increase exponentially with sample size and were thus too resource intensive for our study. The average proportion of heterozygous genotypes at X chromosomal markers was estimated per individual in order to detect false gender assignments. Male subjects can be expected to show X chromosomal heterozygosity proportions $\leq$ 1%, reflecting the overall genotyping error rate, and female subjects should show proportions near the average heterozygosity (26%) of the analyzed X chromosomal SNPs. Average identity-by-state (IBS) distances were calculated for a given set of markers as the average genetic dissimilarity between pairs of individuals. Analysis of IBS values within subpopulations allowed us to detect two types of outliers: (1) cognate relatives, i.e., individuals that were genetically more similar than expected to another member of the same subpopulation, and (2) "aliens," i.e., individuals that were far less genetically similar than expected to the rest of the subpopulation. Formally, cognate relatives were defined as pairs of individuals having a pair-wise IBS value larger than the so-called "Tukey outlier criterion" when compared with the rest of pairs of individuals of the same subpopulation, i.e., the median IBS plus three times the interquartile range (IQR) in that subpopulation. In this case, the partner with the lower call rate was excluded. Aliens were defined as individuals with at least 60% of their pair-wise IBS values below the median minus three times the IQR. These two criteria led to the exclusion of 56 individuals from further analysis (Table 1). One individual identified as female had an average proportion of heterozygous X chromosomal markers of only 0.6% and was thus excluded from further analysis. In total, quality control left 2,457 individuals (97.6%) and 309,790 markers (62.4%) for inclusion in subsequent analysis. AMOVA [17] was performed to ascertain the magnitude of variation attributable to the respective genotyping center or subpopulation. The mean amount of genetic variance explained among genotyping centers was 0.095% (95% confidence interval: 0% to 0.71%), whereas subpopulation affiliation explained 0.63% of the variance (95% confidence interval: 0% to 2.86%). As expected, the largest amount of genetic variation was explained by differences between individuals (99.72%; 95% confidence interval: 98.61% to 100.00%). Data are available on request from the authors according to the regulations of the participating studies and sample cohorts.

Statistical Data Analyses

The ancestry-informativeness index *In* was estimated for each marker as described eslewhere [12]. Principal-component analysis was performed with the *Eigensoft* program with the default settings [38]. Population-wise kernel densities were computed from the first two PCs with the *adehabitat* R package [39] and subjected to least-squares crossvalidation [40] that used 80% of individuals per subpopulation for training. Pearson correlation coefficients were computed for the genetic distance between the subpopulations (represented by the respective median over all individuals in that subpopulation of the first two eigenvectors) and the great-circle geographic distance. The statistical significance of these correlation coefficients was assessed by means of a Mantel test [41]. Barrier analysis was performed on the basis of the Monmonier's algorithm [14]. Locus-wise AMOVA [17] was conducted after clustering the European subpopulations by genotyping center as well as by the use of four geographic groups. Negative percentages of explained variation were settled to 0. Both mean heterozygosity and mean linkage disequilibrium computed by means of HR^2 [15] were computed with a subsample of ten individuals per population in order to adjust for possible influence of sample size [42]. Spatial autocorrelation and Bearing analyses were performed with the software PASSAGE 1.1 [43]. Isoline maps were performed with the Golden Surfer 8 software [44], with the inverse-distance method used for interpolation points. Isoline levels were defined to include the value of at least one of the 23 populations with intervals of 0.001 in the case of heterozygosity and 0.002 in the case of HR^2. For evaluation of the extent to which the CEPH-CEU samples are representative of the subpopulations used in the present study, marker-wise tests of association (Fisher's exact test) were performed each time with the CEPH-CEU samples as "controls" and a given subpopulation as "cases." The false-positive rate was defined as the percentage of markers yielding a p value < 0.05. If the CEPH-CEU samples were representative of a subpopulation, the false-positive rate would be around 0.05, whereas higher false-positive rates indicate that the CEPH-CEU samples may not be representative of the respective subpopulation.

Supplemental Data

Supplemental Data include two tables and two figures and can be found with this article online at http://www.current-biology.com/cgi/content/full/18/16/1241/DC1/.

Acknowledgments

All volunteers are gratefully acknowledged for sample donation. We thank the following colleagues for their help and support: J. Kooner and J. Chambers of the LOLIPOP study and D. Waterworth, V. Mooser, G. Waeber, and P. Vollenweider of the CoLaus study for providing access to their collections via the GlaxoSmithKline-sponsored Population Reference Sample (POPRES) project; K. King for preparing the POPRES data; M. Simoons, E. Sijbrands, A. van Belkum, J. Laven, J. Lindemans, E. Knipers, and B. Stricker for their financial contribution to the generation of the Rotterdam Study dataset; P. Arp, M. Jhamai, W. van IJken, and R. van Schaik for generating the Rotterdam Study dataset; T. Meitinger, P. Lichtner, G. Eckstein, and all other members of the Helmholtz Zentrum München genotyping staff for generating the KORA Study dataset; H. von Eller-Eberstein for management of the PopGen project; F.C. Nielsen, R. Borup, C. Schjerling, H. Ullum, E. Haastrup, and numerous colleagues at the Copenhagen University Hospital Blood Bank for assistance in making the Danish data available; and S. Brauer for DNA sample management. We would additionally like to thank Affymetrix for making the GeneChip Human Mapping 500K Array genotypes of the CEPH-CEU trios publicly available and the Centre d'Etude du Polymorphisme Humain (CEPH) for the original sample collection. We are grateful to three anonymous reviewers for their comments, which stimulated us to improve the manuscript. This work was supported by the Netherlands Forensic Institute to M.Ka.; Affymetrix to M.Ka. and M.Kr.; the German National Genome Research Network and the German Federal Ministry of Education and Research to H.-E.W., S.S., M.Kr., and P.N. (01GR0416 to P.N.); the Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, and the Munich Center of Health Sciences as part of LMUinnovativ to H.-E.W.; the Netherlands Organization for Scientific Research (NWO 175.010.2005.011) to A.G.U.; the European Commission to A.G.U. (GEFOS; 201865) and A.S. (LD Europe; QLG2-CT-2001-00916); the Czech Ministry of Health (VZFNM 00064203 and IGA NS/9488-3) to M.M.; Helse-Vest, Regional Health Authority Norway to L.A.B., the Swedish National Board of Forensic Medicine (RMVFoU 99:22, 02:20) to G.H.; and the Academy of Finland to A.S. (80578, OMLL) and J.P. (109265 and 111713). None of the funding organizations had any influence on the design, conduct, or conclusions of the study.

Received: May 2, 2008
Revised: July 9, 2008
Accepted: July 10, 2008
Published online: August 7, 2008

References

1. Cavalli-Sforza, L.L., Menozzi, P., and Piazza, A. (1994). The History and Geography of Human Genes (Princeton, NJ: Princeton University Press).
2. Sokal, R.R., Harding, R.M., and Oden, N.L. (1989). Spatial patterns of human gene frequencies in Europe. Am. J. Phys. Anthropol. *80*, 267–294.
3. Bauchet, M., McEvoy, B., Pearson, L.N., Quillen, E.E., Sarkisian, T., Hovhannesyan, K., Deka, R., Bradley, D.G., and Shriver, M.D. (2007). Measuring European population stratification with microarray genotype data. Am. J. Hum. Genet. *80*, 948–956.
4. Price, A.L., Butler, J., Patterson, N., Capelli, C., Pascali, V.L., Scarnicci, F., Ruiz-Linares, A., Groop, L., Saetta, A.A., Korkolopoulou, P., et al. (2008). Discerning the ancestry of European Americans in genetic association studies. PLoS Genet *4*, e236.
5. Tian, C., Plenge, R.M., Ransom, M., Lee, A., Villoslada, P., Selmi, C., Klareskog, L., Pulver, A.E., Qi, L., Gregersen, P.K., et al. (2008). Analysis and application of European genetic substructure using 300 K SNP information. PLoS Genet *4*, e4.
6. Seldin, M.F., Shigeta, R., Villoslada, P., Selmi, C., Tuomilehto, J., Silva, G., Belmont, J.W., Klareskog, L., and Gregersen, P.K. (2006). European population substructure: Clustering of northern and southern populations. PLoS Genet *2*, e143.
7. Sajantila, A., Salem, A.H., Savolainen, P., Bauer, K., Gierig, C., and Paabo, S. (1996). Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. Proc. Natl. Acad. Sci. USA *93*, 12035–12039.
8. Roewer, L., Croucher, P.J., Willuweit, S., Lu, T.T., Kayser, M., Lessig, R., de Knijff, P., Jobling, M.A., Tyler-Smith, C., and Krawczak, M. (2005). Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. Hum. Genet. *116*, 279–291.
9. Rosser, Z.H., Zerjal, T., Hurles, M.E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., Barbujani, G., et al. (2000). Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am. J. Hum. Genet. *67*, 1526–1543.
10. Kayser, M., Lao, O., Anslinger, K., Augustin, C., Bargel, G., Edelmann, J., Elias, S., Heinrich, M., Henke, J., Henke, L., et al. (2005). Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. Hum. Genet. *117*, 428–443.
11. Simoni, L., Calafell, F., Pettener, D., Bertranpetit, J., and Barbujani, G.V. (2000). Geographic patterns of mtDNA diversity in Europe. Am. J. Hum. Genet. *66*, 262–278.
12. Rosenberg, N.A., Li, L.M., Ward, R., and Pritchard, J.K. (2003). Informativeness of genetic markers for inference of ancestry. Am. J. Hum. Genet. *73*, 1402–1422.
13. Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol. *4*, e72.
14. Manni, F.C., Guérard, E., and Heyer, G.E. (2004). Geographic patterns of (genetic, morphologic, linguistic) variation: How barriers can be detected by "Monmonier's algorithm.". Hum. Biol. *76*, 173–190.
15. Sabatti, C., and Risch, N. (2002). Homozygosity and linkage disequilibrium. Genetics *160*, 1707–1719.
16. Falsetti, A.B., and Sokal, R.R. (1993). Genetic structure of human populations in the British Isles. Ann. Hum. Biol. *20*, 215–229.
17. Excoffier, L., Smouse, P.E., and Quattro, J.M.V. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics *131*, 479–491.
18. Belle, E.M., Landry, P.A., and Barbujani, G. (2006). Origins and evolution of the Europeans' genome: Evidence from multiple microsatellite loci. Proc Biol Sci *273*, 1595–1602.
19. Torroni, A., Bandelt, H.J., Macaulay, V., Richards, M., Cruciani, F., Rengo, C., Martinez-Cabrera, V., Villems, R., Kivisild, T., Metspalu, E., et al. (2001). A signal, from human mtDNA, of postglacial recolonization in Europe. Am. J. Hum. Genet. *69*, 844–852.
20. Chikhi, L., Nichols, R.A., Barbujani, G., and Beaumont, M.A.V. (2002). Y genetic data support the Neolithic demic diffusion model. Proc. Natl. Acad. Sci. USA *99*, 11008–11013.
21. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution Int. J. Org. Evolution *38*, 1358–1370.
22. Burnett, M.S., Strain, K.J., Lesnick, T.G., de Andrade, M., Rocca, W.A., and Maraganore, D.M. (2006). Reliability of self-reported ancestry among siblings: Implications for genetic association studies. Am. J. Epidemiol. *163*, 486–492.
23. Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.M., and White, R. (1990). Centre d'etude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome. Genomics *6*, 575–577.
24. Lowel, H., Doring, A., Schneider, A., Heier, M., Thorand, B., Meisinger, C., and Group, M.K.S. (2005). The MONICA Augsburg surveys–basis for prospective cohort studies. Gesundheitswesen *67* (*Suppl 1*), S13–S18.
25. Krawczak, M., Nikolaus, S., von Eberstein, H., Croucher, P.J., El Mokhtari, N.E., and Schreiber, S. (2006). PopGen: Population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. Community Genet. *9*, 55–61.
26. Hofman, A., Breteler, M.M., van Duijn, C.M., Krestin, G.P., Pols, H.A., Stricker, B.H., Tiemeier, H., Uitterlinden, A.G., Vingerling, J.R., and Witteman, J.C. (2007). The Rotterdam Study: Objectives and design update. Eur. J. Epidemiol. *22*, 819–829.
27. Hofman, A., Grobbee, D.E., de Jong, P.T., and van den Ouweland, F.A. (1991). Determinants of disease and disability in the elderly: The Rotterdam Elderly Study. Eur. J. Epidemiol. *7*, 403–422.
28. Kayser, M., Liu, F., Janssens, A.C., Rivadeneira, F., Lao, O., van Duijn, K., Vermeulen, M., Arp, P., Jhamai, M.M., van Ijcken, W.F., et al. (2008). Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. Am. J. Hum. Genet. *82*, 411–423.
29. Kooner, J.S., Chambers, J.C., Aguilar-Salinas, C.A., Hinds, D.A., Hyde, C.L., Warnes, G.R., Gomez Perez, F.J., Frazer, K.A., Elliott, P., Scott, J., et al. (2008). Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides. Nat. Genet. *40*, 149–151.
30. Nelson, M.R., Bacanu, S.A., Mosteller, M., Li, L., Bowman, C.E., Roses, A.D., Lai, E.H., and Ehm, M.G. (2008). Genome-wide approaches to identify pharmacogenetic contributions to adverse drug reactions. Pharmacogenomics J., in press. Published online February 26, 2008. 10.1038/tpj.2008.4.
31. Sandhu, M.S., Waterworth, D.M., Debenham, S.L., Wheeler, E., Papadakis, K., Zhao, J.H., Song, K., Yuan, X., Johnson, T., Ashford, S., et al. (2008). LDL-cholesterol concentrations: A genome-wide association study. Lancet *371*, 483–491.
32. Karlsson, A.O., Wallerstrom, T., Gotherstrom, A., and Holmlund, G. (2006). Y-chromosome diversity in Sweden - a long-time perspective. Eur. J. Hum. Genet. *14*, 963–970.
33. Plaza, S., Calafell, F., Helal, A., Bouzerna, N., Lefranc, G., Bertranpetit, J., and Comas, D. (2003). Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. Ann. Hum. Genet. *67*, 312–328.
34. Onofri, V., Alessandrini, F., Turchi, C., Fraternale, B., Buscemi, L., Pesaresi, M., and Tagliabracci, A. (2007). Y-chromosome genetic structure in sub-Apennine populations of Central Italy by SNP and STR analysis. Int. J. Legal Med. *121*, 234–237.
35. Kondopoulou, H., Loftus, R., Kouvatsi, A., and Triantaphyllidis, C. (1999). Genetic studies in 5 Greek population samples using 12 highly polymorphic DNA loci. Hum. Biol. *71*, 27–42.
36. Ploski, R., Wozniak, M., Pawlowski, R., Monies, D.M., Branicki, W., Kupiec, T., Kloosterman, A., Dobosz, T., Bosch, E., Nowak, M., et al. (2002). Homogeneity and distinctiveness of Polish paternal lineages revealed by Y chromosome microsatellite haplotype analysis. Hum. Genet. *110*, 592–600.
37. Schaid, D.J., Batzler, A.J., Jenkins, G.D., and Hildebrandt, M.A. (2006). Exact tests of Hardy-Weinberg equilibrium and homogeneity of disequilibrium across strata. Am. J. Hum. Genet. *79*, 1071–1080.
38. Patterson, N., Price, A.L., and Reich, D. (2006). Population structure and eigenanalysis. PLoS Genet *2*, e190.
39. Calenge, C. (2006). The package "adehabitat" for the R software: A tool for the analysis of space and habitat use by animals. Ecol. Modell. *197*, 516–519.

40. Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis (Boca Raton, Florida: Chapman & Hall / CRC Press).
41. Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. Cancer Res. *27*, 209–220.
42. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature *451*, 998–1003.
43. Rosenberg, M.S. (2001). PASSAGE: Pattern Analysis, Spatial Statistics, and Geographic Exegesis. 1.1 Edition, A.S.U. Department of Biology, ed. (Tempe, AZ.).
44. Golden Software. (2007). Surfer Version 8.08.3267. Colorado, USA.

Article 3

An evaluation of genetic matched pair study design applied to genome-wide SNP genotyping data from European populations.

This article has been previously published as:

Lu & Lao*, et. al.. An evaluation of the genetic-matched pair study design using genome-wide SNP data from the European population. Eur J Hum Genet. 2009 Jan 21;[online advanced publication] doi:10.1038/ejhg.2008.266

* shared primary authorship.

European Journal of Human Genetics (2009), 1 – 9

www.nature.com/ejhg

ARTICLE

An evaluation of the genetic-matched pair study design using genome-wide SNP data from the European population

Timothy Tehva Lu[1,23], Oscar Lao[2,23], Michael Nothnagel[1], Olaf Junge[1], Sandra Freitag-Wolf[1], Amke Caliebe[1], Miroslava Balascakova[3], Jaume Bertranpetit[4], Laurence Albert Bindoff[5], David Comas[4], Gunilla Holmlund[6], Anastasia Kouvatsi[7], Milan Macek[3], Isabelle Mollet[8], Finn Nielsen[9], Walther Parson[10], Jukka Palo[11], Rafal Ploski[12], Antti Sajantila[11], Adriano Tagliabracci[13], Ulrik Gether[14], Thomas Werge[15], Fernando Rivadeneira[16,17], Albert Hofman[17], André Gerardus Uitterlinden[16,17], Christian Gieger[18], Heinz-Erich Wichmann[18,19], Andreas Ruether[20], Stefan Schreiber[20], Christian Becker[21], Peter Nürnberg[21], Matthew Roberts Nelson[22], Manfred Kayser[2,23] and Michael Krawczak*[,1,23]

[1]*Institut für Medizinische Informatik und Statistik, Christian-Albrechts-Universität Kiel, Kiel, Germany;* [2]*Department of Forensic Molecular Biology, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands;* [3]*Department of Biology and Medical Genetics, University Hospital Motol and 2nd School of Medicine, Charles University Prague, Prague, Czech Republic;* [4]*Institute of Evolutionary Biology (UPF-CSIC), CEXS-UPF-PRBB, Universitat Pompeu Fabra, Barcelona, Spain;* [5]*Department of Neurology, Haukeland University Hospital and Department of Clinical Medicine, University of Bergen, Bergen, Norway;* [6]*Department of Forensic Genetics and Forensic Toxicology, National Board of Forensic Medicine, Linköping, Sweden;* [7]*Department of Genetics, Development and Molecular Biology, Aristotle University of Thessaloniki, Thessaloniki, Greece;* [8]*Laboratoire d'Empreintes Génétiques, EFS-RA site de Lyon, Lyon, France;* [9]*Department of Clinical Biochemistry and Center for Pharmacogenomics, University of Copenhagen, Copenhagen, Denmark;* [10]*Institute of Legal Medicine, Medical University Innsbruck, Innsbruck, Austria;* [11]*Department of Forensic Medicine, University of Helsinki, Helsinki, Finland;* [12]*Department of Medical Genetics, Warsaw Medical University, Warsaw, Poland;* [13]*Istituto di Medicina Legale, Università di Ancona, Ancona, Italy;* [14]*Molecular Neuropharmacology Group and Center for Pharmacogenomics, Department of Neuroscience and Pharmacology, University of Copenhagen, Copenhagen, Denmark;* [15]*Research Institute of Biological Psychiatry, Mental Health Center Sct. Hans, Copenhagen University Hospital, and Center for Pharmacogenomics, University of Copenhagen, Copenhagen, Denmark;* [16]*Department of Internal Medicine, Genetics Laboratory, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands;* [17]*Department of Epidemiology and Biostatistics, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands;* [18]*Institute of Epidemiology, Helmholtz Zentrum München – German Research Center for Environmental Health (GmbH), Neuherberg, Germany;* [19]*Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany;* [20]*Institut für Medizinische Molekularbiologie, Christian-Albrechts-Universität Kiel, Kiel, Germany;* [21]*Cologne Center for Genomics and Institut für Genetik, Universität zu Köln, Köln, Germany;* [22]*Genetics, GlaxoSmithKline, Research Triangle Park, NC, USA*

Genetic matching potentially provides a means to alleviate the effects of incomplete Mendelian randomization in population-based gene–disease association studies. We therefore evaluated the genetic-matched pair study design on the basis of genome-wide SNP data (309 790 markers; Affymetrix GeneChip Human Mapping 500K Array) from 2457 individuals, sampled at 23 different recruitment sites across

***Correspondence: Professor Dr M Krawczak, Institut für Medizinische Informatik und Statistik, Christian-Albrechts-Universität Haus 31, Arnold-Heller-Straße 3, Kiel 24105, Germany. Tel: +49 431 597 3200; Fax: +49 431 597 3193; E-mail: krawczak@medinfo.uni-kiel.de**
[23]These authors contributed equally to this work.
Received 26 September 2008; revised 3 December 2008; accepted 10 December 2008

Europe. Using pair-wise identity-by-state (IBS) as a matching criterion, we tried to derive a subset of markers that would allow identification of the best overall matching (BOM) partner for a given individual, based on the IBS status for the subset alone. However, our results suggest that, by following this approach, the prediction accuracy is only notably improved by the first 20 markers selected, and increases proportionally to the marker number thereafter. Furthermore, in a considerable proportion of cases (76.0%), the BOM of a given individual, based on the complete marker set, came from a different recruitment site than the individual itself. A second marker set, specifically selected for ancestry sensitivity using singular value decomposition, performed even more poorly and was no more capable of predicting the BOM than randomly chosen subsets. This leads us to conclude that, at least in Europe, the utility of the genetic-matched pair study design depends critically on the availability of comprehensive genotype information for both cases and controls.
European Journal of Human Genetics advance online publication, 21 January 2009; doi:10.1038/ejhg.2008.266

Keywords: population structure; matching; association; ancestry; microarray

Introduction

In both classical epidemiology and clinical research, potential confounders are usually controlled for by one of two different means, matching or randomization. In genetic studies, however, including the large number of genome-wide association (GWA) studies that have recently been published,[1–3] only so-called 'Mendelian' randomization has been employed to control for genetic confounders, whereas matching by genotype has not played an important role.[4] Nevertheless, there has always been some awareness among genetic epidemiologists that Mendelian randomization may fail, thereby leading to false positive reports of disease genes or to biased effect size estimates.[5] One possible cause of such failure may be systematic differences in terms of the rate at which individuals with a particular phenotype or genotype are sampled from genetically distinct populations. Therefore, two statistical methods to retrospectively rectify genetic imbalances in case-control studies were developed in the late 1990s, both of which rely upon genotyping loci that are unrelated to the genetic variants under study (ie unlinked and not in linkage disequilibrium). The 'genomic control' approach[6] uses marker genotypes to correct the employed test statistic, whereas 'structured association'[7] infers the number of populations represented in a sample, and then assigns each individual to one of these populations with a certain probability.

With the possibility to effectively genotype large numbers of single nucleotide polymorphisms (SNPs) in large numbers of individuals, using microarray technology,[8] the effects of imperfect Mendelian randomization can, in principle, also be alleviated by genetic matching. If individuals from different samples such as cases and controls were as closely matched as possible in terms of their identity-by-state (IBS) status at a large number of SNPs, it may be surmised that most systematic population genetic differences would be eliminated between the ensuing sub samples. However, genetic matching would have to be based on markers from outside the genomic region under study to avoid over-matching. This implies that, in practise, repeated matching may be necessary if multiple or even GWA assessments are due. In any case, genetic matching could of course be accomplished efficiently with the use of genome-wide microarray data, but such a costly strategy may not be necessary if a set of 'best genetic match' (BGM) markers could be established in advance that are capable of capturing the major population genetic characteristics of relevant extant populations. Once a set of BGM markers has been found, it can be used in two ways: either to retrospectively confirm whether two samples of interest were genetically well-matched or to select members of matched samples prospectively, before any additional genotyping.

Recruitment of phenotypically well-characterized control samples is one of the major bottlenecks of genetic epidemiological and pharmacogenetic research. The use of common controls across different association studies has proven to be an efficient solution to this problem, pioneered at a local level by the Wellcome Trust Case Control Consortium (WTCCC),[3] and since adopted, for example, by the US-American Genetic Association Information Network (GAIN)[1] and the German National Genome Research Network ('Nationales Genomforschungsnetz', NGFN).[9] However, the number and geographical distribution of control samples required for the common controls approach to be feasible at a broader geographical level are currently unknown.

In the present study, we investigated three issues related to the genetic-matched pair study design, using genome-wide SNP data from across Europe: (1) the prospects of identifying a small subset of SNPs that accurately predict the 'best' genome-wide matching partner of a given

individual, (2) the distribution of 'best' genetic-matching partners between the European subpopulations and (3) the inter-individual variability in terms of the uniqueness of the 'best' genetic-matching partner. To this end, we analyzed the genotypes of 309 790 markers obtained from the GeneChip Human Mapping 500K Array Set in 2457 individuals, ascertained at one of 23 recruitment sites. The European population is important in this context, not only because of the historical interest in these people and their descendants in the Americas, Australia and elsewhere, but also because they are a major focus of both genetic epidemiological and pharmacogenetic research.[1,3]

Material and methods

Samples, genotyping and quality control

The GeneChip Human Mapping 500K Array (Affymetrix) was used to genotype 500 568 SNPs in 2514 individuals from 23 different sampling sites (henceforth, termed 'subpopulations'), distributed over 20 different European countries. Subpopulation sizes ranged from 12 to 500 individuals (Table 1). Sex ratios differed markedly between subpopulations, with some comprising only females or males, respectively. Genotyping was carried out at six different facilities. For further details, see Lao *et al.*[10]

Array-based SNP genotypes were subjected to stringent quality control as described earlier.[10] Briefly, markers, which had a genotype call rate $\geq$93%, were monomorphic, located on the X chromosome or had a per marker call rate $\leq$90% in at least one genotyping facility were excluded, as were those showing a significant ($P<0.05$) deviation from Hardy–Weinberg equilibrium (HWE) in at least one subpopulation. Individuals deemed genetic outliers to their subpopulation of origin, based on low average IBS to the remaining individuals, were omitted from the respective subpopulation. In total, quality control left 2457 individuals (97.6%) and 309 790 markers (62.4%) for inclusion in subsequent analyses. The set of quality controlled markers will henceforth be referred to as marker set C. Ascertainment of a marker set for genetic matching was carried out with internal validation, using 2/3 of the members of each subpopulation (ie, 1638 randomly chosen individuals) as the training set, and using the remainder (819 individuals) as the validation set (Table 1).

All data were stored as either flat files or in a customized database with an interface to the R statistical software. All data analysis, except for the IBS estimation, was done in R version 2.4.1[11] using customized scripts. IBS calculations and selection of marker sets were carried out using custom C++ programs. All software is available from the authors on request.

Best genetic match marker set

For the ascertainment of a marker subset M of C that would allow us to identify 'best' genetic-matching partners, we will use a set-specific criterion, Δ(M) that is related to the IBS between given individuals and their matching partners, as selected on the basis of M (see below). In this context, we will use the term 'best overall match' (BOM) to denote that individual or group of individuals who maximize the average pair-wise IBS with the individual of interest for the complete marker set C. Ideally, we would want to ascertain a subset of markers that consistently lead to the selection of matching partners with an IBS with the reference individual that is close to the IBS between the reference individual and its BOM.

More formally, if the genotype (g), of a given SNP is encoded by the dose of one of its two alleles (ie, as 0, 1 or 2), then the IBS between any two individuals x and y equals $1-|g(x)-g(y)|/2$ for that SNP. Here, $g(x)$ and $g(y)$ denote the genotypes of x and y, respectively. For a marker set M, let $i_M(x,y)$ be the average IBS, taken over all markers in M, and let $i_M(x)$ denote the maximum $i_M(x,y)$, taken over all individuals y other than x. Finally, if M$\subseteq$N are two nested marker sets, let $i_{M,N}(x)$ be the average $i_N(x,y)$ taken over all y for which $i_M(x,y)=i_M(x)$. For a marker set M$\subseteq$C, Δ(M) is defined as the average difference $|i_C(x)-i_{M,C}(x)|$, taken over all individuals x and weighted by the inverse of the size of the subpopulation to which x belongs.

We used forward selection from marker set C to ascertain marker sets that successively minimized the Δ criterion. The ensuing marker sets will be referred to as the best genetic match (BGM) marker sets. Upper and lower baselines for Δ were computed as follows. The upper baseline was obtained from randomly chosen marker sets of varying size (10–100 in steps of 10), with 1000 sets sampled for each set size value. The lower baseline was obtained from marker sets that theoretically should have captured most of the genetic variation present in the individuals under study, ie sets for which any additional marker would have been in strong linkage disequilibrium with the markers already included. Each chromosome was thus divided into bins of 20 kb, based on the mean swept radius of 500 kb estimated for the European population.[12,13] The swept radius is the distance at which the average association between two markers, measured by r^2, is reduced to approximately one-third (more precisely, e^{-1}) of its initial value. A bin size of 20 kb therefore ensures an average r^2 of $e^{-10/500}=0.98$ between markers in the bin. Markers were then randomly selected from bins, one at a time, and Δ calculated for the resulting marker set. The described selection process was repeated 1000 times and the mean Δ value taken as the lower baseline, ie the expectation of Δ at r^2-based saturation.

Ancestry-sensitive marker set

To compare the BGM set, which focuses on inter-individual genetic variation with a marker set that was ascertained with the aim to highlight inter-population variation, we generated an ancestry-sensitive marker (ASM) set using the

Table 1 European subpopulation summary statistics

Subpopulation	*Code*	*No. samples*	*Final no. samples*	*No. training*
Norway (Førde)	NO	52	52 (0.63)	35
Sweden (Uppsala)	SE	50	46 (1.00)	31
Finland (Helsinki)	FI	47	47 (0.43)	31
Ireland	IE	37	35 (0.80)	23
UK (London)	UK	197	194 (0.90)	129
Denmark (Copenhagen)	DK	60	59 (0.56)	39
Netherlands (Rotterdam)	NL	292	280 (0.00)	187
Germany I (Kiel)	DE1	500	494 (0.52)	329
Germany II (Augsburg)	DE2	500	489 (0.51)	326
Austria (Tyrol)	AT	50	50 (1.00)	33
Switzerland (Lausanne)	CH	134	133 (0.44)	89
France (Lyon)	FR	50	50 (0.68)	33
Portugal	PT	16	16 (0.44)	11
Spain I	ES1	83	81 (0.51)	54
Spain II (Barcelona)	ES2	48	47 (0.43)	31
Italy I	IT1	107	106 (0.58)	71
Italy II (Marche)	IT2	50	49 (1.00)	33
Former Yugoslavia	YU	58	55 (0.65)	37
Northern Greece	EL	51	51 (0.59)	34
Hungary	HU	17	17 (0.35)	11
Romania	RO	12	12 (0.50)	8
Poland (Warsaw)	PO	50	49 (1.00)	33
Czech Republic (Prague)	CZ	53	45 (0.51)	30
Total		2514	2457	1638

Subpopulation, site of sample origin, with more specific location details given in parentheses; No. samples, total number of samples genotyped; Final no. samples, number of samples that passed stringent quality control, with proportion of males in parenthesis (for details, see text); No. training, size of the training set used for marker selection.

singular value decomposition (SVD) method with redundant marker reduction described by Paschou *et al*.[14,15] Global allele frequencies were used to interpolate missing data as suggested by the authors. Some 228 individuals were eliminated from the training set during PCA analysis with Eigensoft2[16] using the standard criterion of having an ancestry coefficient >6 standard deviations in at least one of the eigenvector axes. SVD was carried out with SVDLIBC (version 1.34, http://tedlab.mit.edu/~dr/SVDLIBC), a C library based on the SVDPACK library.[17] Rank-revealing QR matrix decomposition was carried out in Octave version 2.0.17[18] to reduce the redundancy of the first 5000 markers, ordered by the first SVD eigenvector. This resulted in a set of the same size (ie 100 markers) as the BGM set.

Distribution of best genetic match pairs

A count matrix was generated that contains, for each pair of subpopulations, the number of times an individual in the first subpopulation had their BOM in the second population. Cell counts were tested for a deviation from the null hypothesis that BOMs were drawn randomly from subpopulations using a two-tailed exact test as implemented in the R routine *binom.test*. A plot of directed graphs representing the relationships between individuals and their BOMs was generated using Graphviz.[19]

False positive rates

Thresholds for the false positive rates of population-based gene–disease associations in Europe were determined from contrived case-control experiments, using PLINK version 1.03[20] on all markers in set C (Fisher's exact test on allele frequencies). These mock studies were carried out for all pair-wise combinations of subpopulations, each time labeling one subpopulation as 'cases' and the other as 'controls'. The percentage of markers with *P*-values <0.05 was reported. As the variance of the *P*-value is inversely related to sample size, false positive rates were not estimated for subpopulations with sample sizes <20 (PT, HU and RO; see Table 1 for subpopulation abbreviations).

Results

Best genetic match and ancestry sensitive marker sets

Two subsets of markers (BGM and ASM) were ascertained from the complete marker set using either IBS-based forward selection or SVD with redundant marker reduction, respectively. As the decrease in Δ as a function of marker set size levelled off very rapidly (see Figure 1), BGM marker selection was terminated at 100 SNPs (Supplementary Table 1). For the sake of comparability, the ASM set was chosen so as to contain the same number of markers as the BGM set (Supplementary Table 2). Interestingly, the top 5000 markers of the provisional ASM set included various SNPs annotated to genes known to stratify the European

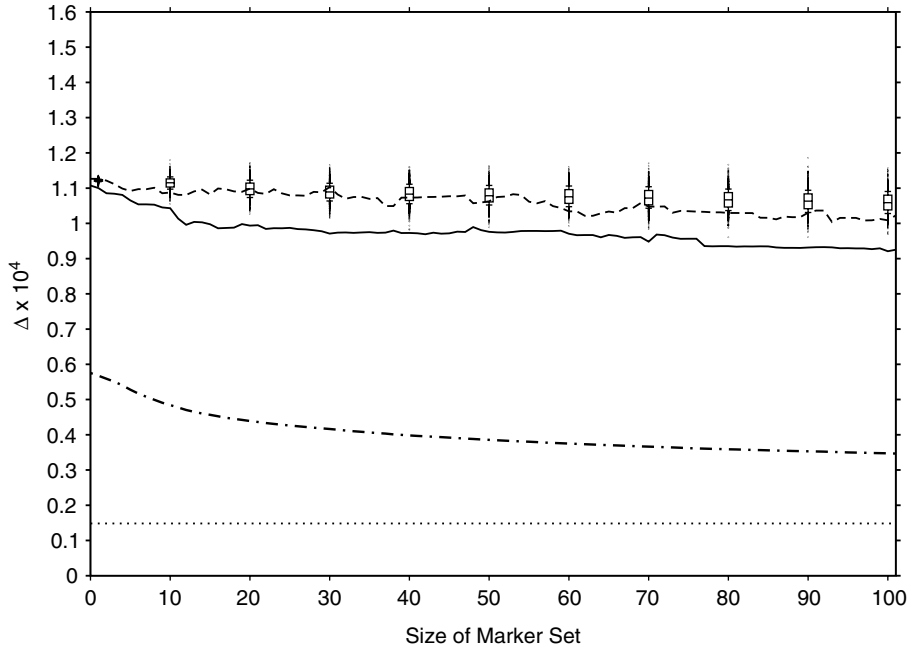

Figure 1 IBS-based forward selection of best genetic match (BGM) marker sets. The upper baseline for Δ is illustrated by box-whisker plots, each generated from 1000 random selections of a marker set of given size. The lower baseline for Δ (dotted line) is provided by a marker set for which any additional markers could be expected to be in strong linkage disequilibrium ($r^2 > 0.98$) with at least one marker already included in that set (for details, see text). Selection of the BGM marker sets is depicted by a solid line; the performance of ASM sets of various sizes is illustrated by a dashed line. All Δ values were calculated from the validation set of individuals. The training set Δ values obtained for the BGM marker sets are included for reference (dash-dotted line).

gene pool as a result of recent positive selection acting differently in different geographic regions, including *HERC2*[21] (ranked 7), *OCA2*[22] (ranked 33), *LCT*[23] (ranked 262) and *TYRP1*[24] (ranked 1138).

A graphical representation of the forward selection process leading to the BGM set is provided in Figure 1. In the validation set, the Δ criterion decreased by ~10% until it levelled off at ~20 markers, and decreased only marginally thereafter. Although forward selection on the training set showed a promising reduction in Δ value, the validation Δ for the 100 top markers comprising the BMG set was still at 9.3×10^{-5}, which is 14.3% lower than the upper (random) baseline but exceeds the lower baseline of 1.5×10^{-5} by a factor of six. This implies that the genome-wide similarity of two European individuals is hard to predict with sufficient accuracy on the basis of a small, specifically selected marker set, and that the little benefit that can be gained in this respect already arises from 100 markers or even fewer. By comparison, the capacity of the ASM set for BOM prediction was found to be indistinguishable from the upper (random) baseline, ie, it performed no better than randomly drawn marker sets.

Distribution of best overall matches (BOMs)

A significant amount of genetic similarity between the European subpopulations is revealed by an assessment of the subpopulation of origin of BOMs (Table 2). In a considerable proportion of cases (1868/2457 or 76.0%), the BOM of a given individual belonged to a different subpopulation than the individual itself. That this was particularly so when individuals or BOMs came from subpopulations with large sample sizes (DE1, DE2 and NL) was presumably due to the wider range of genetic diversity captured by these samples, but may also reflect their concurrent geographic location in central Europe. On the other hand, for some relatively isolated subpopulations (FI and IT2) the source of the BOM was mostly the subpopulation itself, reflecting their separation also seen in genetic barrier analysis and, in the case of the Finns, principle component analysis.[10] Closer inspection at the individual level revealed that some individuals were disproportionately more often selected as BOMs than others (Figure 2). Thus, of the 2457 individuals examined, 1860 (75.7%) were never deemed a BOM at all. This is significantly higher than the expected number (1553.3, 63.2%) if BOMs were drawn at random ($\chi^2 = 165.1$, 1 df, $P < 0.001$). At the same time, 120 individuals were chosen as BOMs at least five times, which is a significant excess over expectation (9.0, 0.36%, $\chi^2 = 1401.9$, 1 df, $P < 0.001$). The subpopulation of origin of the 10 most frequently ascertained BOMs was generally among those central Europeans who also had the largest sample size (DE1 five,

Table 2 Count matrix of BOM (best overall match) affiliation

	NO	*SE*	*FI*	*IE*	*UK*	*DK*	*NL*	*DE1*	*DE2*	*AT*	*CH*	*FR*	*PT*	*ES1*	*ES2*	*IT1*	*IT2*	*YU*	*EL*	*HU*	*RO*	*PO*	*CZ*	*Total*
NO	**8**	2	0	0	0	0	3	**25**	7	0	0	0	0	0	0	0	3	0	0	0	0	4	0	52
SE	6	1	0	0	2	2	4	22	5	0	2	0	0	0	0	0	2	0	0	0	0	0	0	46
FI	1	0	**39**	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	47
IE	1	0	0	4	**12**	3	3	3	3	0	0	2	0	0	0	0	4	0	0	0	0	0	0	35
UK	2	0	0	8	27	**23**	40	62	15	0	1	3	0	0	0	0	10	1	0	0	0	2	0	194
DK	1	0	0	0	0	**10**	13	23	7	0	0	0	0	0	0	0	2	0	0	0	0	3	0	59
NL	4	1	0	1	14	**45**	**94**	79	16	0	3	1	0	0	0	0	8	0	1	0	0	11	2	280
DE1	19	1	0	4	21	**74**	60	**230**	54	0	4	1	0	0	0	0	10	0	0	0	0	16	0	494
DE2	9	0	0	5	24	**68**	83	179	90	3	10	1	0	0	0	0	8	0	0	0	0	9	0	489
AT	0	0	0	0	3	**8**	7	15	13	0	0	0	0	0	0	0	3	0	0	0	0	1	0	50
CH	2	1	0	1	18	**15**	26	36	20	2	8	1	0	0	0	0	1	0	0	0	0	2	0	133
FR	0	0	0	1	3	4	11	16	8	0	1	2	0	0	0	0	1	1	0	0	0	2	0	50
PT	1	0	0	0	0	2	3	5	2	0	1	1	0	1	0	0	0	0	0	0	0	0	0	16
ES1	1	1	0	1	9	6	16	21	8	1	5	8	0	0	0	1	1	0	0	0	0	2	0	81
ES2	1	1	0	0	4	3	8	6	6	0	3	3	0	0	**9**	2	0	0	1	0	0	0	0	47
IT1	1	2	0	1	6	3	19	28	15	2	10	5	0	0	2	5	0	2	2	0	0	1	2	106
IT2	2	0	0	0	0	1	3	9	1	0	0	0	0	0	0	0	**33**	0	0	0	0	0	0	49
YU	3	1	0	0	4	0	8	20	7	0	1	0	0	0	0	0	1	**10**	0	0	0	0	0	55
EL	2	2	0	0	2	1	7	17	8	3	1	0	0	0	0	1	0	0	**7**	0	0	0	0	51
HU	1	0	0	0	1	2	4	8	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	17
RO	0	0	0	0	1	2	1	2	2	0	2	1	0	0	0	1	0	0	0	0	0	0	0	12
PO	3	0	0	0	1	13	2	16	5	0	0	0	0	0	0	0	0	0	0	0	0	**8**	1	49
CZ	0	0	0	2	2	3	11	15	3	1	1	1	0	1	0	0	1	0	0	0	0	0	4	45
Total	68	13	39	28	154	288	426	841	295	12	53	31	0	2	11	10	88	14	11	0	0	64	9	

Row, subpopulation of origin of reference individual; Column, subpopulation of origin of BOM of reference individual.
Underlined values are significantly higher than random expectation (*P*-value ≤ 0.05), bold values are statistically significant after Bonferroni correction (FWER ≤ 0.05).

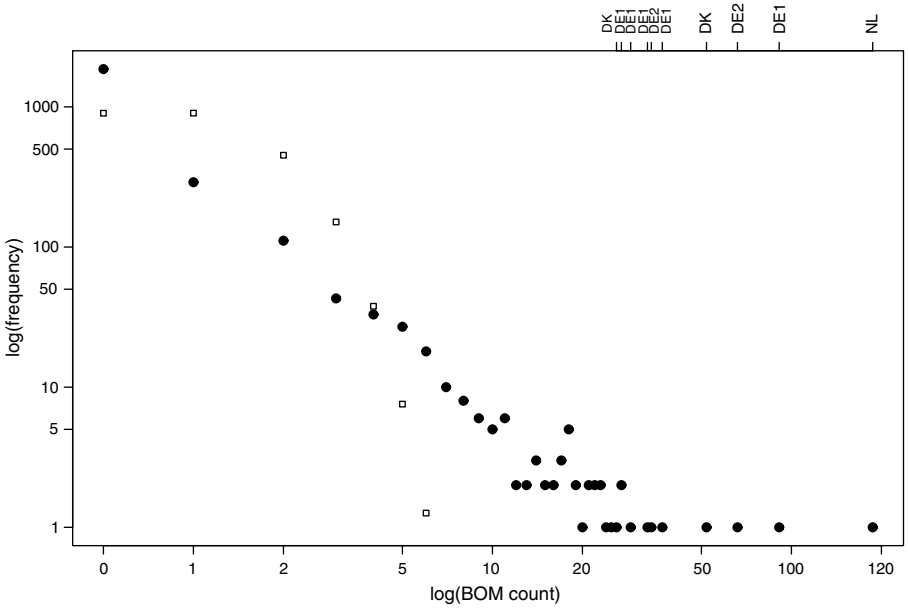

Figure 2 Distribution of the number of times an individual was deemed a BOM. The observed distribution is marked by circles. Also included is a Poisson distribution with the same mean as the sample mean (marked by squares), which approximately corresponds to the theoretical expectation if best overall matching (BOM) were selected at random. The codes of the subpopulation of origin of the 10 most frequently selected BOMs are given at the upper right edge of the plot.

DE2 two and NL one), with the notable exception of DK (59 individuals, yet holding two of the top 10 positions; Figure 2). Interestingly, barring of the 10 most frequently chosen BOMs left the number of times the BOM was found outside the subpopulation of origin of the individual of interest virtually unchanged (1862/2457 or 75.8%,

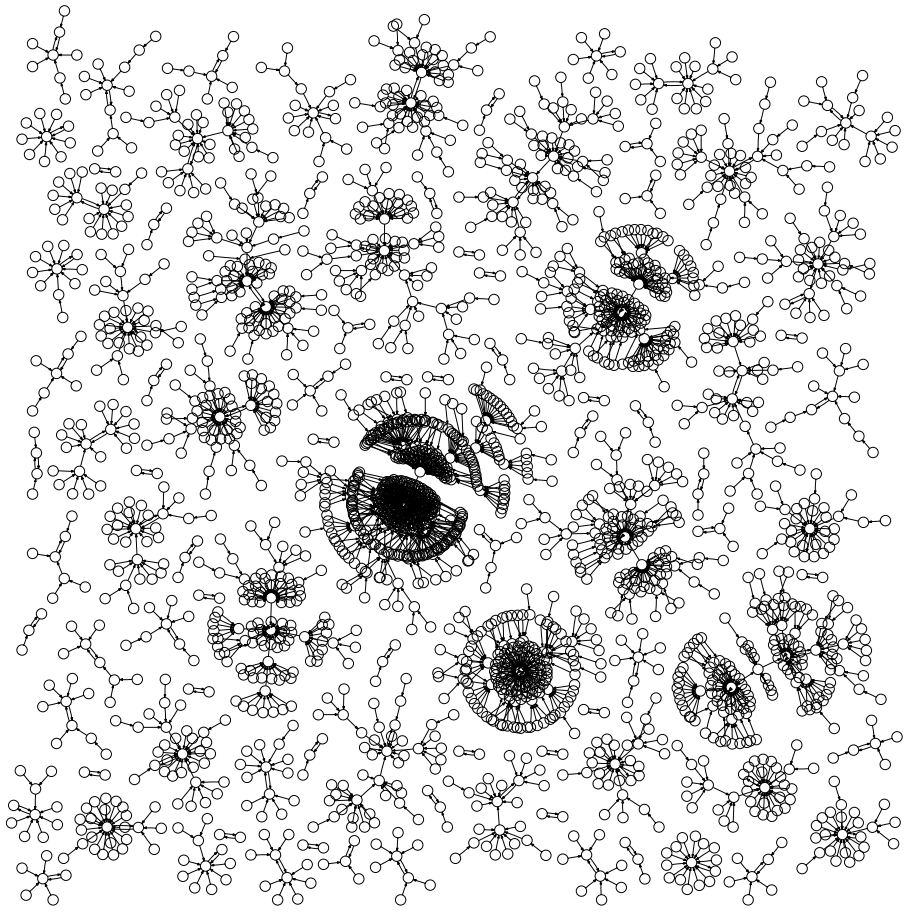

Figure 3 Directed graph illustrating the best overall matching (BOM) relationships between individuals. Circles represent individuals (2457 total) and arrows point towards the respective BOM. The most frequently selected BOM (centre of the plot) was selected for 187 individuals.

Supplementary Table 4). A graphical representation of the BOM relationships between individuals is provided in a directed graph illustrating the complexity of networks of matches (Figure 3).

False positive rates

Although it is admittedly unlikely that a researcher would actually carry out a population-based gene–disease association study in which cases and controls were sampled from different countries, without adjusting for population origin in one way or another, measurement of the false positive rates expected from such undertaking is of general interest as a gauge of the magnitude of stratification pertaining in the European population. Mock false positive rates for pairs of subpopulations (Supplementary Table 3) ranged from 0.039 (CZ and PO) to 0.208 (DE1 and IT1), with a median of 0.070. Subpopulations sampled from the same political country often had false positive rates indicative of little or no population stratification, although this was not always the case (DE1–DE2: 0.089). Many neighboring countries also had false positive rates close to those expected under the null hypothesis, indicating the absence of major population differences as well (eg UK-IE: 0.042, NL-DK: 0.051, EL-YU: 0.047, CH-AT: 0.039, FR-DE2: 0.051).

Discussion

This is the first study to evaluate the genetic (ie, IBS-) matched pair study design with genome-wide SNP data of a large number of European individuals from across the continent. The high number of best genetic-matching partners found in different subpopulations corroborates earlier reports of a considerable amount of genetic similarity between the European subpopulations,[4,10,14,25–27] particularly those in close geographic proximity. The surprising inter-individual variability observed in terms of the number of times a person was chosen as the best genetic-matching partner of others does not necessarily imply that the relationship between genetic and geographic distance in a

given sample hinges on a small number of people. Thus, when the most frequently chosen matching partners were barred in our analysis, the proportion of best matches found outside the subpopulation of origin of the respective index person remained virtually unchanged.

We observed that the best genetic-matching partner for a genome-wide marker set such as the Affymetrix GeneChip Human Mapping 500K Array cannot be predicted from a small, specifically selected subset of markers alone, but that the information required to make such predictions is distributed evenly across all markers. This leads us to conclude that, at least in Europe, the utility of the genetic-matched pair study design depends critically on the availability of comprehensive genotype information for both cases and controls. In practise, this would mean that shared controls should ideally be genotyped for all relevant genome-wide marker sets, thereby allowing the chromosome-specific choice of best matching partners for given case individuals on the basis of the remainder of the genome.

A distinction must obviously be made between ASM, collections of which have been described in recent papers,[14,25–28] and the BGM marker set that we attempted to generate. As the genetic within-subpopulation variation in Europe is much greater than the between-subpopulation variation, it is not unlikely for any two individuals from different subpopulations to be genetically more similar to each other than any two individuals from the same subpopulation. In this sense, an ASM marker set consists of markers that differentiate subpopulations, whereas a BGM marker set should contain variants that highlight genetic similarity at the individual level. Although the two concepts are complimentary, the marker sets fit to each task need not be the same, and the existence of one set does not necessitate the existence of the other. Obviously, markers that arose on early branches of the corresponding, region-specific coalescence tree of the extant Europeans would provide good ASM, but they cannot at the same time identify nearest neighbors at the tips of the tree. Such identification requires a much higher resolution of the tree topology, and therefore many more markers. Consequently, no adequately sized BGM set could be constructed in our study and the ASM set selected with established methodology was no more capable of identifying the best genetic-matching partner of an individual than a randomly chosen marker set.

Recently, two independent applications of genetic matching have been reported in the context of GWA studies,[4,29] both of which relied on information derived from PCA of genotypes to match individuals. In the first study, using US-American type 1 diabetes patients and German controls, Luca *et al*[4] carried out 'full' matching wherein matches consist of clusters of individuals that contain at least one case and one control. Matching was based upon a distance measure with the top eigenvectors as coordinates, weighted by the eigenvalues to exaggerate differences in dimensions of greater importance. In the second study, Heath *et al*[29] undertook a PCA on a large pan-European group of individuals and proposed a method to predict the population affiliation of a sample of unknown origin from the eigenvector matrix of its genotypes. As both methods are likely to reduce spurious genetic differences between cases and controls in disease association studies, basing their matching criteria on eigenvectors from PCA is strongly reminiscent of selecting ASM. However, as we have shown above, matching with ASM is less efficient than best overall genetic matching particularly in Europe, where the within-subpopulation genetic variation is known to be much greater than the between-subpopulation variation. Indeed, the conclusion by Luca *et al*[4] that some individuals remain 'unmatchable' by their approach is not surprising bearing in mind that ASM can only capture a miniscule proportion of the actual inter-individual genetic differences in a given population.

The false positive rates derived in our study from mock genetic case-control experiments represent an upper limit to the likely consequences of sharing samples in continent-wide scientific collaborations. In this respect, the rate estimates also rationalize collaborative genetic epidemiological and pharmacogenetic research in Europe; from the data we have compiled, it seems as if research projects combining cases from neighboring subpopulations and matching them against common control samples, such as those provided by the WTCCC,[3] GAIN[1] and NGFN,[9] may indeed be valid.

In conclusion, we found that the pattern of pair-wise genetic matching in the European population was more complex than anticipated. Best genetic matches occurred frequently across the continent in our study, and disproportionately often involved a small group of individuals. Ascertainment of a subset of markers that accurately predicts best overall genetic matches turned out to be infeasible.

Acknowledgements

All sample donors are gratefully acknowledged for their participation. We thank the following colleagues for their help and support: J Kooner and J Chambers of the LOLIPOP study and D Waterworth, V Mooser, G Waeber and P Vollenweider of the CoLaus study for providing access to their collections through the GlaxoSmithKline-sponsored Population Reference Sample (POPRES) project; K King for preparing the POPRES data; M Simoons, E Sijbrands, A van Belkum, J Laven, J Lindemans, E Knipers and B Stricker for their financial contribution to the Rotterdam study; P Arp, M Jhamai, W van IJken and R van Schaik for generating the Rotterdam study dataset; T Meitinger, P Lichtner, G Eckstein and all genotyping staff at the Helmholtz Zentrum München for generating the KORA study dataset; H von Eller-Eberstein for providing access to the PopGen data; R Borup, C Schjerling, H Ullum, E Haastrup and numerous colleagues at the Copenhagen University Hospital Blood Bank for making the Danish data available; and S Brauer for DNA sample management. We also wish to thank Affymetrix for making

the GeneChip Human Mapping 500K Array genotypes of the CEPH-CEU trios publicly available, and the Centre d'Etude du Polymorphisme Humain (CEPH) for the original sample collection. This work was supported by the Netherlands Forensic Institute (M Ka), Affymetrix (M Ka and M Kr), the German National Genome Research Network and the German Federal Ministry of Education and Research (H-EW, SS, M Kr and PN); the Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg and the Munich Center of Health Sciences as part of LMUinnovativ (H-EW), the Netherlands Organization for Scientific Research (AGU: NWO 175.010.2005.011), the European Commission (AGU: GEFOS; 201865, AS: LD Europe; QLG2-CT-2001-00916); the Czech Ministry of Health (MM: VZFNM 00064203 and IGA NS/9488-3), Helse-Vest, Regional Health Authority Norway (LAB), the Swedish National Board of Forensic Medicine (GH: RMVFoU 99:22, 02:20) and the Academy of Finland (AS: 80578, OMLL, JP: 109265 and 111713). None of the funding organization had any influence on the design, conduct or conclusions of the study.

References

1 GAIN Collaborative Research GroupManolio TA, Rodriguez LL, Brooks L *et al*: New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet* 32007; **9**: 1045–1051.
2 Hirschhorn JN: Genetic approaches to studying common diseases and complex traits. *Pediatr Res* 2005; **57**: 74R–77R.
3 The Wellcome Trust Case Control Consortium: Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. *Nature* 2007; **447**: 661–678.
4 Luca D, Ringquist S, Klei L *et al*: On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *Am J Hum Genet* 2008; **82**: 453–463.
5 Davey Smith G, Ebrahim S: What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? *BMJ* 2005; **330**: 1076–1079.
6 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
7 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945–959.
8 Wang WY, Barratt BJ, Clayton DG, Todd JA: Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 2005; **6**: 109–118.
9 Wichmann HE, Gieger C, Illig T, MONICA/KORA_Study_Group: KORA-gen – resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* 2005; **67**: 26–30.
10 Lao O, Lu TT, Nothnagel M *et al*: Correlation between genetic and geographic structure in Europe. *Curr Biol* 2008; **18**: 1241–1248.
11 R Development Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, 2008.
12 Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok PY, Collins A: The optimal measure of allelic association. *Proc Natl Acad Sci USA* 2001; **98**: 5217–5221.
13 Wollstein A, Herrmann A, Wittig M *et al*: Efficacy assessment of SNP sets for genome-wide disease association studies. *Nucleic Acids Res* 2007; **35**: e113.
14 Paschou P, Drineas P, Lewis J *et al*: Tracing sub-structure in the European American population with PCA-informative markers. *PLoS Genet* 2008; **4**: e1000114+.
15 Paschou P, Ziv E, Burchard EG *et al*: PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet* 2007; **3**: 1672–1686.
16 Patterson N, Price AL, Reich D: Population structure and eigenanalysis. *PLoS Genet* 2006; **2**: e190.
17 Berry MW: Large scale singular value computations. *Int J Supercomput Appl* 1992; **6**: 13–49.
18 Eaton JW: *GNU Octave Manual*. Network Theory Unlimited: Bristol, 2002.
19 Gansner ER, North SC: An open graph visualization system and its applications to software engineering. *Softw Pract Exp* 2000; **30**: 1203–1233.
20 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
21 Kayser M, Liu F, Janssens AC *et al*: Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. *Am J Hum Genet* 2008; **82**: 411–423.
22 Duffy DL, Montgomery GW, Chen W *et al*: A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am J Hum Genet* 2007; **80**: 241–252.
23 Bersaglieri T, Sabeti PC, Patterson N *et al*: Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 2004; **74**: 1111–1120.
24 Voight BF, Kudaravalli S, Wen X, Pritchard JK: A map of recent positive selection in the human genome. *PLoS Biol* 2006; **4**: e72.
25 Bauchet M, McEvoy B, Pearson LN *et al*: Measuring European population stratification with microarray genotype data. *Am J Hum Genet* 2007; **80**: 948–956.
26 Price AL, Butler J, Patterson N *et al*: Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet* 2008; **4**: e236.
27 Seldin MF, Shigeta R, Villoslada P *et al*: European population substructure: clustering of northern and southern populations. *PLoS Genet* 2006; **2**: e143.
28 Tian C, Hinds DA, Shigeta R *et al*: A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. *Am J Hum Genet* 2007; **80**: 1014–1023.
29 Heath SC, Gut IG, Brennan P *et al*: Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Hum Genet* 2008; **16**: 1413–1429.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)

Additional Results

Some findings produced during the course of my research were never published for various reasons. This may be because the results are not “earth shattering” enough, or because they did not fit well into the overall theme set forth in the articles, or perhaps the results were incomplete upon submission of the article, or they constituted negative results which are so lamentably under-represented in the scientific literature. Here, I take the opportunity to present some interesting results produced in the course of my work which have not been previously published. The data set used in all of the following analyses is that presented in Articles One and Two of this thesis. The hierarchical clustering analysis and the multidimensional scaling were performed as part of the initial exploratory analysis of the data set. The sampling strategy experiments were carried out with regard to observations which conflicted with previously published results. Selection experiments were conducted with the goal of discovering new genetic loci under selection. Observations about the distributions of the ancestry sensitive marker set (ASM) and the best genetic match (BGM) marker set were made to better understand why these ascertained marker sets proved to be less useful than anticipated. All data analyses was performed at the Institut für Medizinische Informatik und Statistik on the Universitätsklinikum Schleswig-Holstein Campus, Kiel, Germany between 2007 and 2009.

Hierarchical Clustering of European Population Data

Various methods falling under the general heading of data clustering have been successfully performed on genetic data sets from large populations such as the one described in Articles Two and Three of this thesis. Hierarchical clustering has been used to effectively differentiate populations on a global or intra-continental scale[64]. Studies have found that as few as 50 randomly selected SNP markers can be used to cluster individuals on an inter-continental scale[65,66]. Model-based clustering methods including K-means clustering[67] have been used to identify population structure at global[68,69] as well as regional levels. Based on the success of hierarchical structuring reported in these and other studies, we decided that this method should be applied to the genome-wide European data

set.

An identity by state matrix containing all individual-individual genetic distances was computed using all quality controlled markers from the genome-wide scan. Several different hierarchical clustering algorithms were applied to the matrix data using the *hclust* routine in the R statistical analysis software[70]. Of the algorithms tested, two produced reasonable clusters: the average linkage clustering algorithm (which minimizes the average distance of objects between clusters) and Ward's method (which minimizes the variance of objects within each cluster). A cutoff criterion for defining cluster groups was applied to clusters generated by each algorithm using the *cluster.stats* routine in the R statistical analysis software. To evaluate the robustness of the clusters, a bootstrapping method, programed in R, was developed and the co-occurrence of individuals within clusters was evaluated at each re-sampling.

Unfortunately, the clusters generated in these experiments and tested by bootstrapping were unstable. While hierarchical clustering was quite powerful when applied to the large datasets in other studies, it was inadequate for resolving genetic structure in the comprehensive and diverse European data set used in this thesis (Figure 2). One hypothesis explaining this result is that the clustering methods applied are most effective under two conditions: 1) when the data being clustered consists of multiple relatively distinct groups, as is the case with genotype data collected on a global scale from discrete populations, or 2) when the structure underlying the clusters consists of a small number of somewhat differentiated groups, as is the case in the more regional studies. The European population data used in this thesis consists of many subpopulations with highly overlapping population structure and the clustering methods were therefore unable to successfully differentiate the subpopulations under these conditions. Another hypothesis as to the failure of the clustering method is presented in the next section.

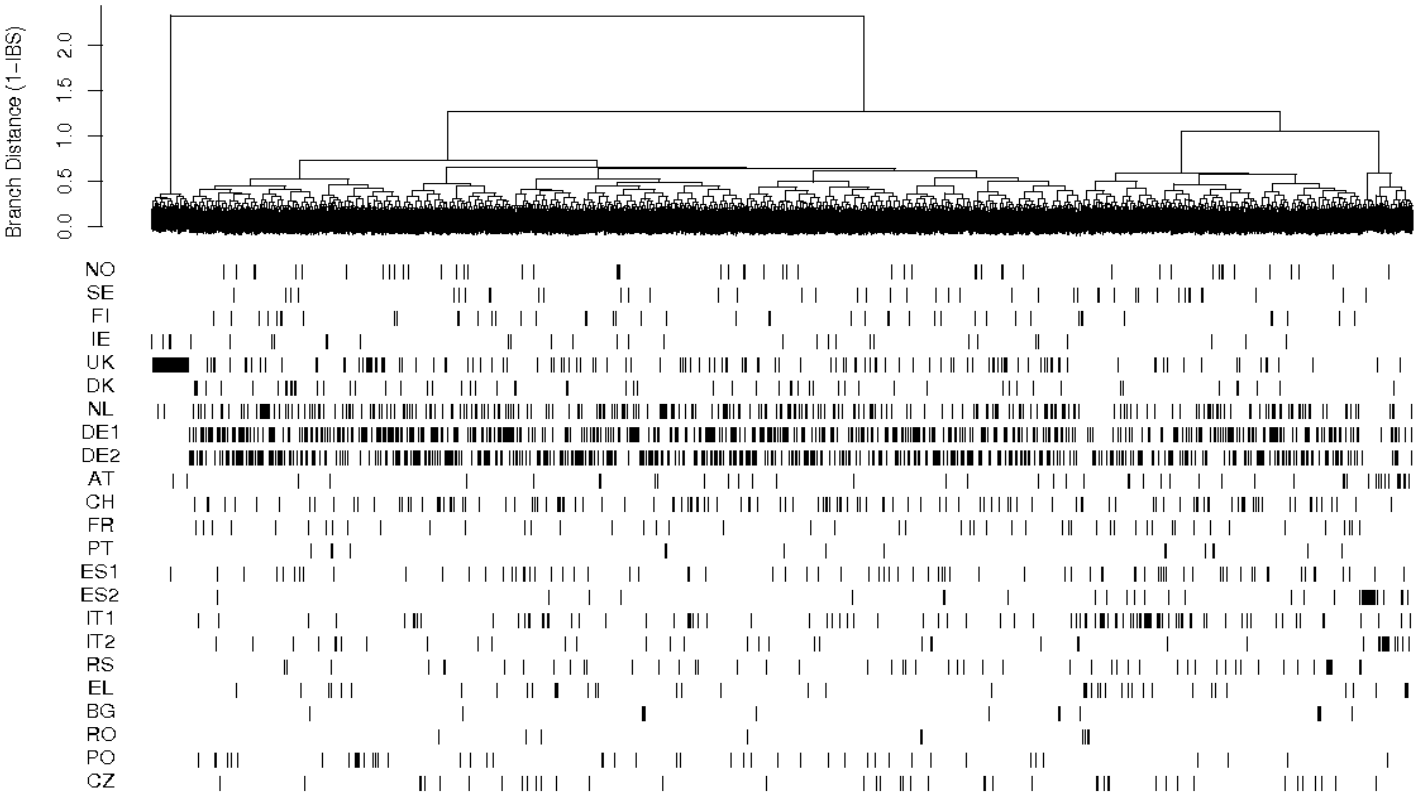

Figure 2. An example of hierarchical clustering of the human European population based on identity by state distances using the average linkage clustering algorithm. The positions of individuals from each population are indicated by vertical lines beneath the cluster tree. Country codes are identical to those used in Articles Two and Three. A custom script developed specifically for this project was used to generate this plot.

Multidimensional Scaling of European Population Data

Multidimensional Scaling (MDS) is an ordination analysis method which, like Principal Components Analysis, has also found widespread use in population genetic studies. MDS is typically employed in the analysis of population-based distance matrices (i.e. pairwise F_{st} or Φ_{st}) to quantify genetic structure. I have used MDS in published studies to generate multi-dimensional plots based on Y-chromosome marker data depicting population differentiation in the German[71] and Norwegian[72] populations. Other studies have used distances between individuals rather than populations for MDS. These studies use identity-by-state (IBS) matrices from autosomal marker genotypes to examine population structure in northern Europe[73] and Japanese[74] populations, and to test for population outliers in genome-wide association studies on a British[75] population and a population of US Americans of European descent[76]. The application of this analytical method in the aforementioned studies inspired the decision to undertake MDS analysis on the IBS matrix computed for the European population featured in Articles Two and Three. This analysis was performed using the MASS package from the R statistical software suite[70]. Unfortunately, MDS failed to achieve a satisfactory degree of separation of individuals from this data set (Figure 3).

In light of the similar failing of the hierarchical clustering algorithms on the

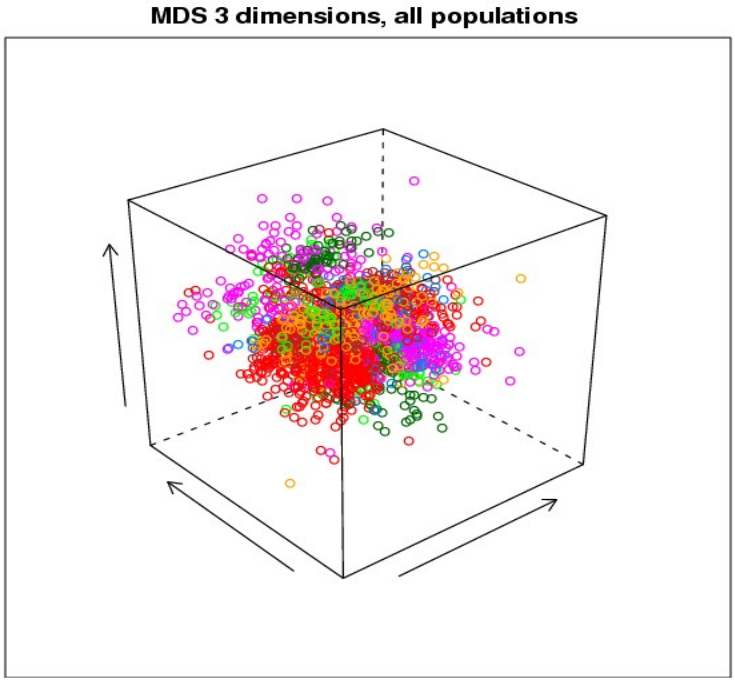

Figure 3. Multidimensional Scaling Analysis of the European population. A three dimensional plot.

same IBS matrix, it may be hypothesized that the failure of both analysis methods is due to the loss of information occuring when an IBS matrix is generated from individual genotype data. Technically, MDS analysis is a special case of principal components analysis[77], so with the proper data transformation, similar results should be achievable with both methods. The success of the PCA analysis on the European data set was attributable in large part a method published by Nick Patterson and David Reich[20]. Their method did not involve first transforming individual genotypes into an IBS matrix. They instead applied MDS directly to the genotype matrix after the genotype values have been appropriately centered and normalized. The intuitive leap that led to this improved method was a deep insight on their part and has been subsequently adapted for use in MDS analysis[78,79].

Effects of Sampling Strategy on Observed Genetic Structure

When Article Two was submitted for publication in May of 2008, it was one of many studies appearing around the time that analyzed a genome wide survey of genetic markers on samples of Europeans to determine population structure.

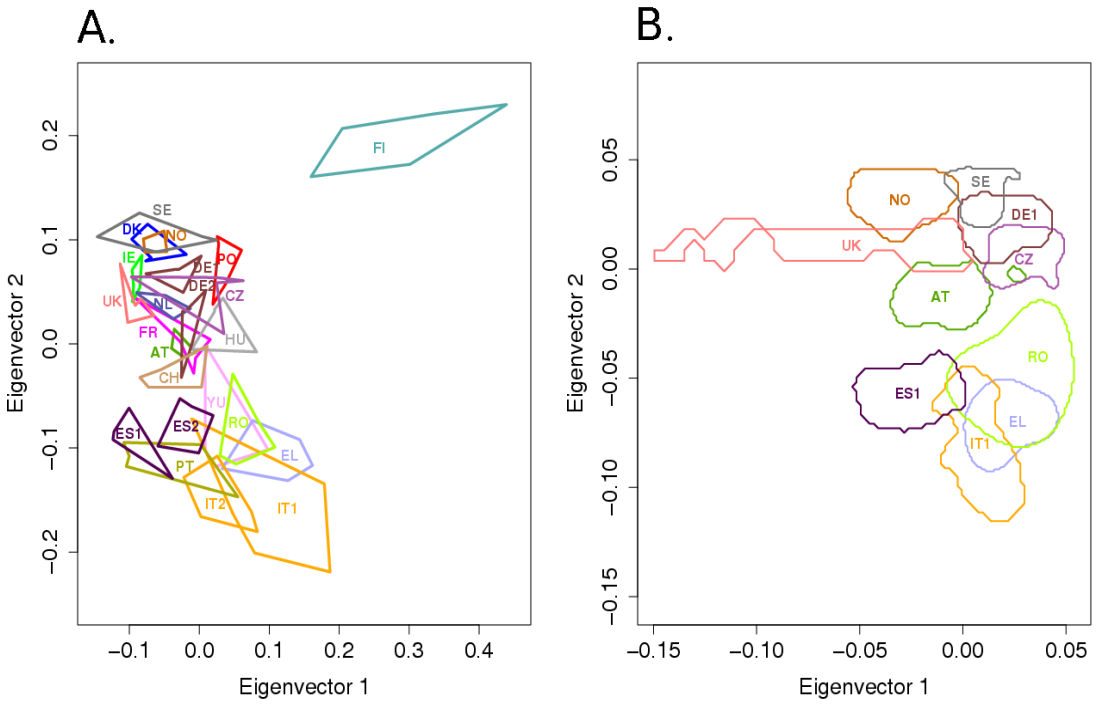

Figure 4. Subsampling results on the human European population. A. 5 samples from each population. B. Six northern vs. four southern populations.

One of the precursors to these studies was published in 2006 by Michael F. Seldin *et. al.*[21] Along with the typical genetic clines observed previously in other studies, they reported “evidence for major difference in population structure of northern and southern Europe”. The study presented in Article Two notes a continuous gradient from south to north, with no apparent clustering. To try to resolve this contradiction, a sub-sampling experiment was designed and executed to determine if the result observed by the Seldin *et. al.* could be an artifact of sampling bias. Two causes for the observed clusters seemed likely 1) that the sample sizes were not large enough to give evidence of a smooth gradient or 2) that the population sampling distribution did not cover the geographic region densely enough. To determine if such sampling effects influenced the results, two types of subsamples were taken from our population. In the first experiment, a limited number of samples (n=5) from each population was taken. In the second, a selection, using the entire subpopulations, of southern (ES1, IT1, EL, RO) and northern (UK, NO, SE, DE1, CZ, AT) subpopulations was made. The sample selection criteria were designed to reflect those used by Seldin *et. al.* PCA analysis was done on both re-sampled data sets. PCA analysis was performed with the *Eigensoft* program using default settings[20] and plotted using the R statistical analysis software[70] with kernel densities drawn by the *adehabitat* package[80] with an 80% boundary. These experiments revealed a cluster pattern in both circumstances indicating that sample size (Figure 4.A.) and sampled population distribution (Figure 4.B.) do influence the observation of genetic structure. Ideally, the robustness of the limited sampling portion of the experiment would be tested by rigorous re-sampling. This was not done in a formal manner as it was decided not to pursue this line of argument against the Seldin *et. al.* paper in the published manuscript.

Evidence of Selection in OCA2/HERC2 and LCT/MCM6 Genes

Previous studies on the European population have shown evidence of selection at two distinct genetic loci: one locus controlling eye color (OCA2/HERC2), and one controlling persistence of the lactase gene (LCT). Using the data set presented in Articles Two and Three, an exploratory analysis was undertaken to search for new, previously undiscovered genes under selection in the European population. Initially, the entire European sample population was

divided into two groups, representing northern and southern Europe. PLINK[81] was used to perform association tests over the entire genome using one group as mock-cases and the other group as mock-controls. This procedure was repeated using an east-west division as well as northeast-southwest and northwest-southeast divisions. In the case of the north-south population division, exploratory analysis showed evidence of possible selection pressure in genetic regions on chromosomes 2 and 15 (Figure 5). These were verified to be the OCA2 (Figure 6) and LCT (Figure 7) loci which had already been well reported in the literature. All plots were generated using the *UCSC Genome Browser* web interface. The fact that validation of previously observed results is an under appreciated part of the scientific process, will not hinder those results from being reported here.

The locus on chromosome 15 (15q12-13) associated with the brown/blue eye color phenotype was first described in the Danish population in 1996[82]. Subsequent association studies replicated this finding in Americans of European descent[83], Australians of European descent[84,85], and in the Dutch population[86]. These studies were able to quantify the amount of variation attributed to this trait locus (QTL) to approximately 75%. The two genes of interest located in this region are oculocutaneous albinism II (OCA2) and hect domain & RLD 2 (HERC2). The function of these two genes has not been precisely elucidated. They are both involved in processes related to melanine, one of the predominant pigments in determining skin, hair, and eye color. OCA2 is though to be involved in the trafficking of tyrosinase or its substrates[87] which are important in melanosome maturation. A mutation in HERC2, which is located just upstream adjacent to OCA2, has been implicated as an inhibitory regulator for OCA2[87]. Strong association between eye color and the HERC2 locus has been demonstrated in the Dutch population[88] and in Australians of European descent[89]. Selective advantages hypothesized for the OCA2 related mutations which lead not only to blue eye color, but to a paler skin phenotype are: 1) enhanced vitamin D synthesis in the lower sunlight conditions found at high latitudes, and 2) sexual selection, implying that prospective mates find this phenotype to be more attractive.

Genetic variation in the lactase gene, which is responsible for digestion of milk sugar (disaccharide lactose, or simply lactose), has been much more thoroughly studied than the OCA2/HERC2 genes due to the interest of evolutionary biologists, anthropologists, and molecular geneticists,. Lactase is a membrane bound intestinal enzyme found in micovilli projecting into the lumen of

the small intestine. Lactase hydrolyzes lactose to monosaccharide galactose and glucose which are then absorbed. The first scientific evidence that levels of lactase are highly variable in adult humans came in the 1960's despite the fact that individual differences in ability to digest milk have long been observed[90]. In most mammals, the ability to digest lactose subsides upon maturation, however some humans retain this ability into adulthood. This trait displays dominant autosomal transmission[91]. Early studies speculated that lactase persistence is common only in certain populations with long traditions of pastoralism[92]. The distribution of lactase persistence in Europe, Asia and Africa has since been well documented[90] and manifests itself as a gradient in Europe and India and as a patchy distribution throughout Africa and in the Middle East. The presence of lactase persistence in the European population has been shown to be almost completely correlated with a single haplotype which is most likely to have been the result of a single mutational event. An associated locus, MCM6, contains two regulatory regions for LCT. However, the African and Middle Eastern populations show evidence of lactase persistence due to multiple mutation events. The hypothesized selective advantage for the LCT mutant in northern Europe are 1) an additional nutritional advantage conferred by milk, and 2) the consumption of vitamin D counteracting an increased risk of developing rickets and osteomalacia which can occur more frequently at higher latitudes due to low exposure to sunlight. These selection hypothesis remain untested.

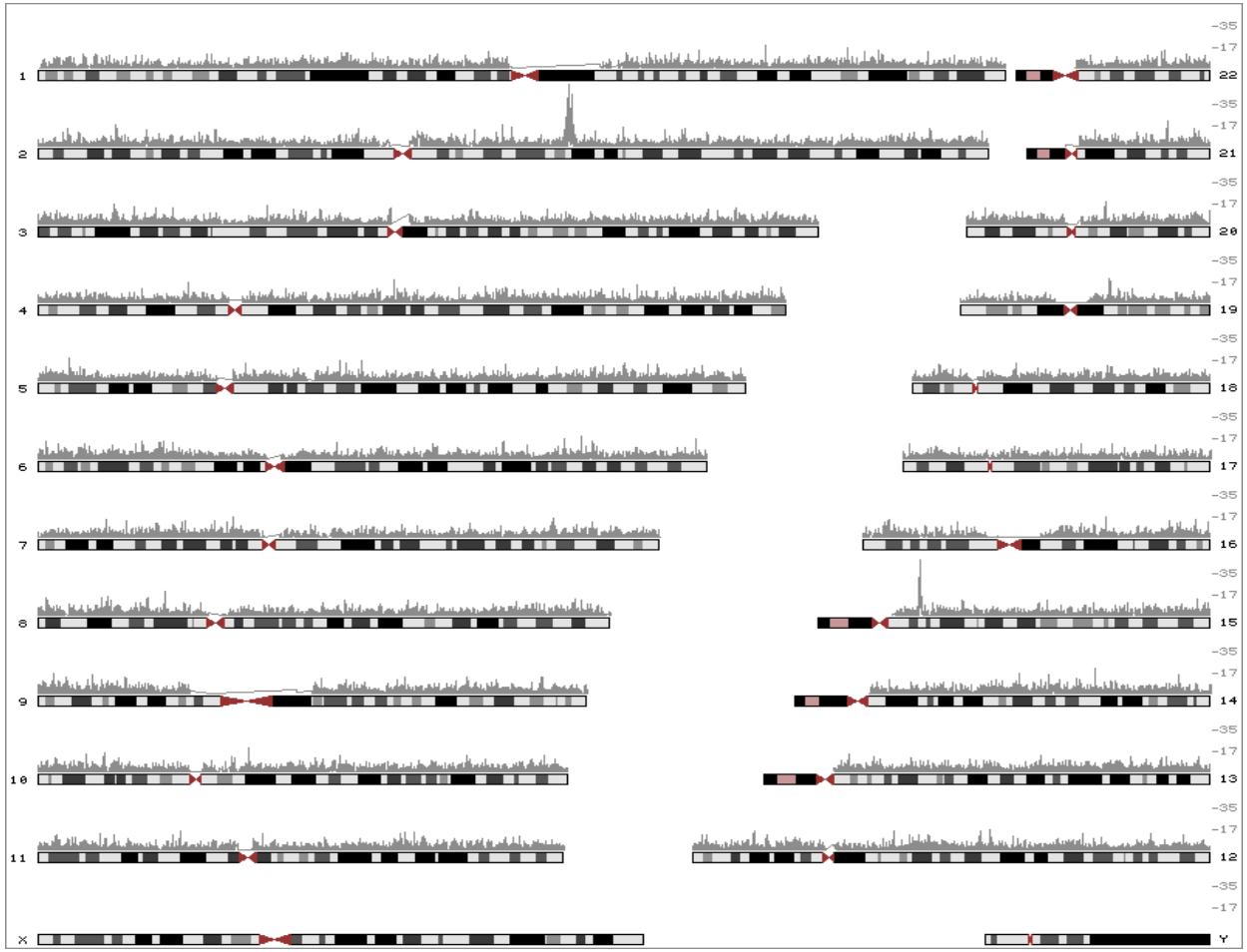

Figure 5. Genome wide association scan for genetic differences between northern and southern Europeans. Note the peaks on *chromosome* 2 and 15.

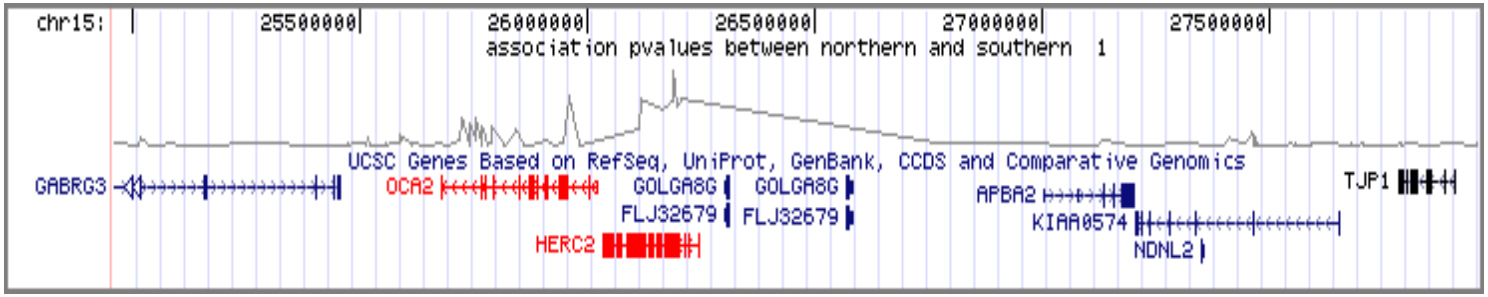

Figure 6. Genome scan result: Magnification of chromosome 15 region showing the association peak relationship to the OCA2 and HERC2 genes (red).

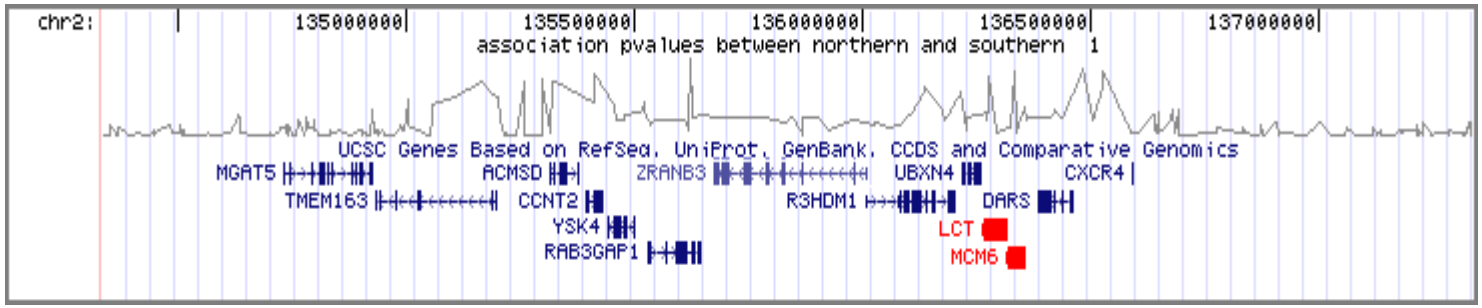

Figure 7. Genome scan result: Magnified region of chromosome 2 showing the relationship of the association peaks to the LCT and MCM6 genes (red).

Distribution of the PCA-Correlation Coefficient in an Ancestry Sensitive Marker Set

One of the as-yet unpublished results of the research on population genetic structure of the European population is the ascertainment of an Ancestry Sensitive Marker (ASM) set. An ASM set is defined as a subset of markers that captures much of the genetic structure represented in the complete set of markers surveyed. If the ASM set is small in size and has an information content comparable to that of the complete marker set, it can be utilized and result in savings in genotyping expense. One of the methods used to ascertain an ASM set for the European population was adopted from the work of Paschou *et. al.*[93] who described a procedure for computing a PCA-correlation coefficient (PCAcc) for each marker reflecting its correspondence to the top principal components of a PCA. By selecting an ASM set consisting of the markers with the highest PCAcc for our data set, it was hoped that a useful set of ASM markers could be ascertained. This was done by sorting all the markers according to their PCAcc, larger coefficients indicating better correlation with the top principal components and by extension, higher information content for population differentiation. Though this method seemed promising, experiments showed that marker sets small enough to be used as ASM sets could not represent the genetic structure of the complete marker set with sufficient resolution. Sometime after the publication of Article Two, an analysis was undertaken to estimate the individual contributions of each marker to the information contained of the PCA top principal components. It was obvious that the distribution of PCAcc was a long-tail distribution (Figure 8.A., see the Discussion and Conclusions Section for more about long-tail distributions) so visual inspection was done to determine whether the distribution displayed power-law or log-normal properties. All plots were generated using the R statistical analysis software[39] using the built-in *plot*, *hist* and *dnorm* functions. A data set with a power-law distribution can be recognized as having a linear appearance with a negative slope on a log-log plot (Figure 8.B., also compare with Article One, Figure 6). To visualize the distributions correspondence to a log-normal distribution, a histogram of the normalized absolute deviations $(\log(i) - \mu)/\sigma$ was compared to the standard normal curve N(0,1) (Figure 8.C., also compare to Article One, Figure 6). A data set with a log-normal distribution has a histogram similar to the standard normal curve.

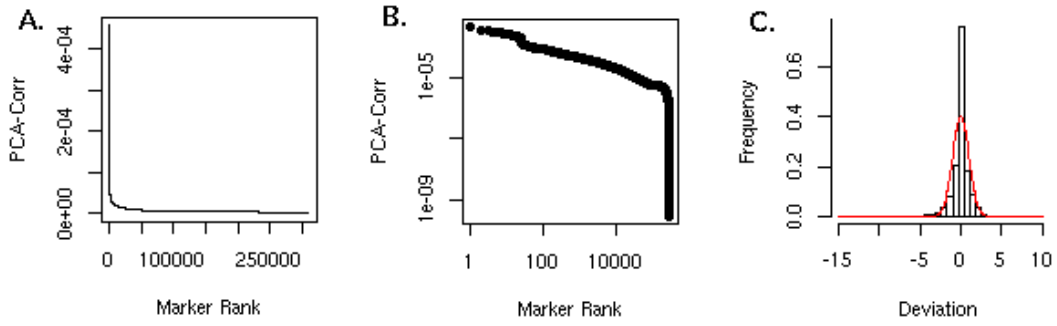

Figure 8. Plots depicting properties of the ASM marker sets PCA-correlation coefficient distribution. **A.** Correlation vs. marker rank showing the long-tail **B.** Log(correlation) vs. log(marker rank) showing non-linearity, therefore poor correspondence to a power-law distribution **C.** Histogram of the normalized absolute deviations showing good correspondence with a log-normal distribution.

The most striking feature of the log-log plot (Figure 8.B.) is the rapid drop in magnitude of the PCAcc in the region of 150,000 markers (half of the data set). The distribution of markers before the bend shows an additional kink at marker 25. Although the section before the bend appears to be linear on the plot, closer inspection reveals a curved distribution deviating from the regression line (data not shown). Inspection of the histogram of absolute deviations (Figure 8.C.) shows that the distribution has marked log-normal properties.

The observation that this distribution is a long-tail distribution can be used to understand why the ASM marker set selection was unsuccessful. The distribution suggests that as each marker is added to the ASM set, it contributes incrementally less additional information to the resolution of the genetic structure. If a small number of markers provides insufficient resolution, then very large numbers of markers will need to be added to make appreciable improvements. The further significance of this power-law distribution is presented in the Discussion and Conclusions.

Distribution of Information in the Best Genetic Match Marker Set

One of the primary results presented in Article Three (the examination of genetic similarity in the European population) is that no small set of markers can predict the best genetic match with the degree of accuracy approaching the complete set of 300k markers. To come to this conclusion it was necessary to

incrementally measure the predictive accuracy of the smaller set of markers, referred to as the Best Genetic Match (BGM) set, as each marker was added. In Figure 1 of Article Three, Δ represents the inverse of the accuracy, or the inaccuracy of the BGM set in predicting the best genetic match. As more markers are added to the set, the inaccuracy (Δ) of the prediction of the set decreases. It is more convenient to think of the inverse of Δ as being a measure of the amount of information present in the BGM marker set. As more markers are added to the marker set, the information present in the marker set grows, as does its predictive accuracy. Sometime after the journal publication of Article Three, we completed an analysis of the distribution of the information contribution of each genetic marker in the BGM marker set. Because the forward analysis to generate BGM marker set was very computationally intensive, only 130 markers were ascertained for the set before the selection process was terminated. All plots were generated using the R statistical analysis software[39] using the built-in *plot*, *hist* and *dnorm* functions.

An initial inspection of the distribution of information content shows that the curve is likely to be a long-tail curve (Figure 9.A.). Further analysis shows that the

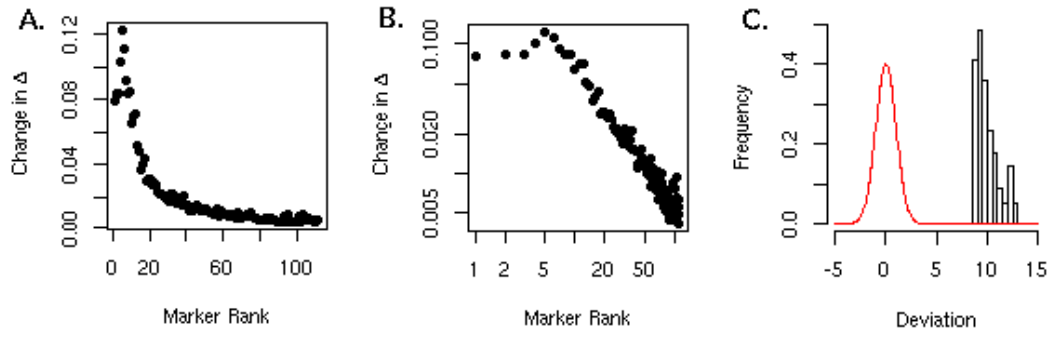

Figure 9. Plots depicting properties of the BGM marker sets "Change in Delta" distribution. **A.** Change in delta vs. marker rank showing the long-tail (extends to 300k markers). **B.** Log(change in delta) vs. log(marker rank) showing linearity, therefore good correspondence to a power-law distribution **C.** Histogram of the normalized absolute deviations showing poor correspondence with a log-normal

distribution closely resembles a power-law distribution (Figure 9.B.) (power regression, $y = 0.714 \cdot x^{-1.054}$, $r^2=0.9613052$) and not a log-normal distribution (Figure 9.C.). The first four markers have a relatively constant information contribution, which would have allowed for the efficient selection of an accurately

predictive marker set had it continued through the rest of the distribution. The observed long-tail properties of the distribution indicate that the increase in predictive performance of the BGM marker set gets smaller with each marker that is added to the set, in other words the rate of increase in accuracy of the marker set decreases as it gets larger, and explains the failure in finding an accurate BGM marker set. The further significance of this distribution is presented in the Discussion and Conclusions section.

Long-tails, Microarrays, and Marker Sets

Gaining a better understanding of natural processes is the basic motivation behind all natural scientific work. It is often the case that scientific progress occurs in a haphazard manner with discoveries and advances made at an unpredictable pace and in unexpected directions. Due to the nature of this discovery processes, it is at times difficult to take notice of certain similarities inherent in the phenomena being observed. Sometimes, however, an underlying structure which highlights a common feature in unrelated processes can be brought forth to provide an explanation for the observed phenomena. The properties of data set distributions are an example of an underlying structure which can emphasize such similarities.

The three articles presented in this thesis explore three different questions in related disciplines of biology: functional genomics, population genetics and genetic epidemiology. Each of the three examines a different quantitative aspect of genetic information, one the level of expression of a gene, another the level of informativeness of a genetic marker in discriminating between individuals, and yet another the level in ability of a genetic marker to predict the best genetic match. As with all quantitative measures, each of these measurements, when taken in aggregate, creates a distribution. The distributions of quantitative measures often follow mathematically well- described models. Perhaps the best known of these is the normal or Gaussian distribution which describes the distribution of many natural phenomena. The distributions describing pattern of data we observed in all three of our studies are from a class called *long-tail*, *fat-tail*, or *heavy-tail*

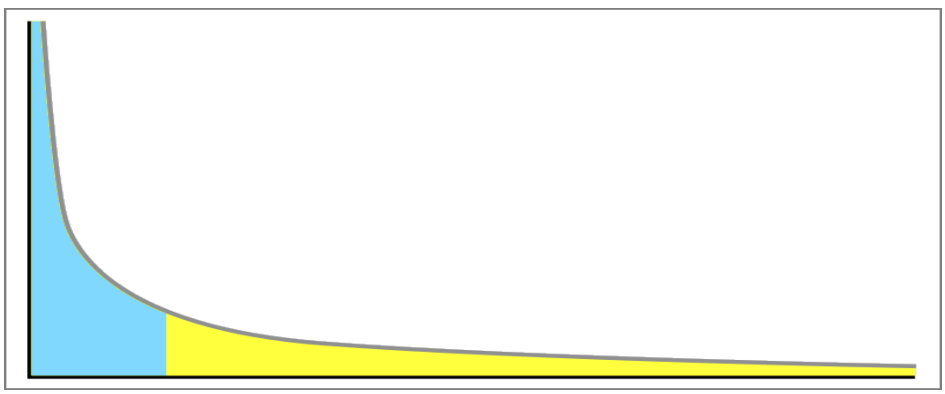

Figure 10. An example of a long-tail distribution, featuring the long tail (yellow) and a smaller number of large values (blue).

distributions (Figure 10), examples of which include the power law and the log normal distributions. As was described in Article One, long-tail distributions feature an initial region with a small number of large values and a long tail with larger number of small values. Because the bulk of the data is in the tail, the tail exerts a large influence and is a characteristic feature of this type of distribution. Before further discussing the implications of long-tail distributions on the data presented here, it will be helpful to review the relevant results and relate them to the five fields of study introduced at the beginning of this thesis.

Microarray technology is a recurring theme in this thesis, a kind of thread that ties together all of the research presented here. This work began with an examination of a fundamental property of biological organisms, the distribution of gene expression levels as measured by expression screening microarrays. In all types of arrays examined, the distribution of expression levels was shown to be long-tail distributions exhibiting varying degrees of conformation to either the power-law or log-normal distributions. Insights on the nature of these distributions were used to develop a normalization technique which conform to a version of the power-law distribution for discrete variables called Zipf's law, also known as the zeta distribution. The technique was compared to other normalization techniques in common use and was found to be especially appropriate for a specific type of microarray used in functional genomic studies called 'boutique' expression arrays which rely on standardized controls or housekeeping genes for normalization.

Making use of the same basic biotechnology, another common application of microarrays was examined, that of genotyping arrays. Here the focus switched to the field of population genetics by way of an extensive set of human genotype data made available through a large number of collaboration partners throughout Europe. The results of this study produced an impressive genetic to geographic correspondence with PCA analysis at a never before observed level of detail, further clarifying our understanding of European population structure. We generated a genetic landscape mirroring the geographic landscape of the European population using approximately three hundred thousand genetic markers. One contribution of the study was an improvement to genetic map visualization techniques through the application of kernel densities to population plots. The search for an ancestry sensitive marker (ASM) set gave evidence that a small set of markers cannot be used to accurately reconstruct the genetic structure present in the complete data set. This implies that the genetic

information that reflects population differentiation is spread across the entire genome and each marker makes only a small contribution to the overall structural information, exceptions being markers under selection (except for markers under selection such as OCA2 and HERC2).

Continuing with the European genetic data set, a different investigative thread was taken up which incorporated the field of genetic epidemiology. This time, instead of examining the markers informative for genetic differences, those associated with genetic similarities between individuals in the European population were investigated. This line of research resulted in several interesting discoveries about the genetic relationships between individuals on the continental level. First, considering that the data set consisted of groups of individuals sampled from many populations throughout Europe, it might be naïvely assumed that individuals from the same population are most likely to be genetically similar to each other. However, it was found that the most genetically similar matches were more likely (76%) to be made with an individual from outside the population. Another interesting observation about the distribution of the number of best overall genetic matches was that certain rare individuals were matched to many individuals. These individuals had an unusually “average” genetic makeup that allowed them to be selected as the best match for many others. A study design based on the idea of genetic matching was proposed in Article Three. The design has superior power to traditional population-based association (case/control) studies because it reduces much of the stochastic “genetic noise” by pre-matching for similarity before tests for statistical association are performed. Having a small set of markers with which to predict the “Best Genetic Match” was deemed very useful in this case. The task of searching for such a set was computationally intensive and utilized a massively parallel algorithm developed and run on a computer cluster, however, the analysis was unable to find a small set of markers to accurately predict BGM. This led to the conclusion that genetic information neccessary to identify the best genetic match is spread throughout the genome, with each marker making a small incremental contribution. This is similar to the situation observed with the ASM set of the population genetics study.

Returning to the topic of data distributions, expression microarrays were known to exhibit a long-tail distribution (Article One, Figure 6) before this series of studies was undertaken. However, the observation that the utility of genetic markers in differentiating European individuals (Additional Results, Figure 7) and

the predictive ability of markers in selecting the best genetic match (Additional Results, Figure 8) also follow long-tail distributions is a new finding. In fact, the scientific questions leading to the discovery of these distributions were, in both cases, premised on the hope that the distributions did not have long-tails. Both studies investigated a subset of markers containing the most informative portions of a larger set. However, upon acquisition of an appropriately-sized set of markers it was determined that the acquired set lacked the power to be useful for accurate prediction. At this point, a decision had to be made whether or not to continue expanding the marker set in order to improve predictive accuracy. An understanding of the nature of long-tail distributions can help to guide this decision. Because it is known that the information necessary to make the marker set useful is in the long-tail, but including the long-tail makes the marker set too large to be practical, we can conclude that no marker set of appropriate size can be ascertained. The suggestion to use another method, stratagem, or algorithm to obtain an alternative marker set can be similarly countered. Therefore, in general, to adequately characterize genetic properties of relationships between individuals in the human population and gain a deeper understanding of genetic differences and similarities it is necessary to take into account the many small contributions of numerous genetic loci spread throughout the genome. It is noteworthy that this conclusion was reached investigating the underlying nature of the mathematical distribution of genetic information contained in these markers.

Outlook

In light of the attention focused on genome-wide genetic studies in the past few years and the contributions made by the research presented here to both population genetics and genetic epidemiology, it is fruitful here to explore opportunities for future research in these fields.

In the field of population genetics, the four major forces affecting allele frequencies are mutation, selection, migration, and genetic drift. An interesting application for the genome-wide data set used in this thesis would be to try to measure these parameters in the European population. Somatic mutation rate measurement is the most prohibitive of the four, as it involves long-term studies that repeatedly genotype participating individuals, which is not possible with the currently collected data. A cursory examination for effects of selection (Additional

Results) uncovered two loci under selective pressure. Migration rates could be estimated by examining allele dispersion. This could be accomplished on a per marker basis by superimposing allele occurrence on the coordinates generated in the PCA analysis and producing genetic maps for each marker upon which spatial autocorrelation analysis could be performed. These would represent a sort of genetic “snapshot” of the dynamic processes of allelic migration in time. Hundreds of thousands of genetic snapshots could then be conglomerated, generating a map of genetic migration. The last parameter which could be examined is genetic drift, which is influenced by the estimated effective population size. Effective population size is a population genetic principle that is used to describe an ideal hypothetical population that responds to the effects of genetic drift in the same way as the actual population, though the actual population may differ from the effective population in many respects. In human populations, two important influences on the effective population size are population bottlenecks and subsequent population expansions which are believed to have occurred several times in the past[94]. Other factors influencing effective population size on a less episodic, more temporally continuous scale are unequal contributions of gametes to the next generation, inbreeding, and overlapping generations. An accurate estimation of effective population size would provide insights into the European population's history, as well as its spread and development. It would also allow predictions of the expected rate of allele frequency changes in the future. Because the accuracy of estimated effective population size depends on the number of pairs of unlinked biallelic markers used in the computation, the unprecedented size of the European genome-wide data set (tens of millions of marker pairs), provides a unique opportunity to accurately estimate this parameter. Previously, studies have been performed using only the much smaller HapMap sample of individuals (circa 60 individuals from Europe) on some 20 million marker pairs. The completion of large scale sampling and genome-wide genotyping of other populations would allow further estimates of population genetic parameters and provide additional interesting points of comparison.

All studies published to date on the European population have focused predominantly on the western half of the continent and less so on the eastern half. Thus, both sample size and the number of sites sampled have been biased in favor of the west. It would therefore be of interest to augment the research done thus far with increased sampling from under-represented regions of Europe to

provide better estimations of population differentiation. Along similar lines, and if money were no object, genome-wide genotyping with complete population coverage could be conducted in other human populations and used for comparison to the European population. Comparing an area of similar geographic scale and population density to that of the European study would be desirable. Also, it would be prudent to choose an area with an indigenous population that has not experienced massive migrational influx, which would eliminate the Americans and Australia from the prospective study locations. Good candidates might be a region of sub-Saharan Western Africa with a population density comparable to that of Europe, or perhaps some subregion of Asia, perhaps in China. Additional studies in the two regions would shed light on the operation of many population genetic mechanisms because these those regions represent extremes when compared with the European population. The population histories of the two regions are very different, as African populations are more fragmented and China, with its large and widespread distribution of Han Chinese (a single ethnic group) is more genetically homogeneous. One might expect that an extensive survey of autosomal markers would reveal a much more diverse and structured population in Africa and a less structured one in China when compared to population structures determined in the European study. A comprehensive genetic study in the two regions would be far reaching in its ability to provide a better understanding of the four driving factors of genetic change: migration, selection, mutation, and genetic drift. Current projects focus on either microsatellites[95] or sex-linked[96] markers and have been modest in both the numbers of markers and the extent of sampling, making a large scale autosomal marker survey of African and Chinese populations a matter of scientific interest, even if not financially feasible.

In the field of genetic epidemiology, we proposed the genetic matched-pair study to provide more power in the analysis of population-based association data. After the publication of our proposed study design, Guan *et. al.*[97] published an application and evaluation of the same study design. Their simulation studies showed that, as predicted, our study design does indeed reduce false-positive rates and retains high power to detect disease-associated markers as compared with traditional association study analysis methods. They also demonstrated that the study design adequately controls for population stratification, quote: “Effectively it treats every sample as a single population and compares it to the

most similar counterpart." This property is especially useful in populations where admixture is an important consideration, for example the North American population. Based on these results, this study design could be immediately adopted by the research community for use with existing data from genome-wide association studies that have appeared in the literature over the past few years. No new data would need to be collected, as the cases could be matched to the pools of existing controls for which all the necessary genetic data has already been collected, or from existing pools of controls maintained by the Welcome Trust[38], NIH[39], and PopGen[40], among others. Given the inclination to high rates of false positives of the previously applied association study methods, there are two possible outcomes when comparing the already discovered association leads with those generated by the genetic matching study design. Either the lead will be verified giving weight to its status as an actual lead, or it will not be validated and should probably be discarded as a false positive, or at least considered with considerable prejudice. Also of interest would be the new leads generated under the genetic match study design which were not apparent in the original analysis, however these would still need to be verified before being considered as bona fide association leads.

In summary, exciting advances in the integrated fields of genetic epidemiology, population genetics, functional genomics, bioinformatics and biotechnology are currently being made, breaking new ground and exploring new territory with discoveries and technological advancements to bring us a better understanding of the processes underlying human genetic phenomena.

References

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995;270(5235):467-470.

2. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nat. Genet*. 1999;21(1 Suppl):33-37.

3. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*. 1998;280(5366):1077-1082.

4. Diamandis EP. Sequencing with Microarray Technology--A Powerful New Tool For Molecular Diagnostics. *Clin Chem*. 2000;46(10):1523-1525.

5. Wang H, Hubbell E, Hu J, et al. Gene structure-based splice variant deconvolution using a microarry platform. *Bioinformatics*. 2003;19(1 Suppl):i315-322.

6. Mockler TC, Chan S, Sundaresan A, et al. Applications of DNA tiling arrays for whole-genome analysis. *Genomics*. 2005;85(1):1-15.

7. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921.

8. CRICK FH. On protein synthesis. *Symp. Soc. Exp. Biol*. 1958;12:138-163.

9. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet*. 2003;33 Suppl:228-237.

10. Syvanen A. Toward genome-wide SNP genotyping. *Nat Genet*. 2005;37(Suppl):S5-10.

11. Parmigiani G, Garett ES, Irizarry RA, Zeger SL. *The Analysis of Gene Expression Data*. 1st ed. Springer; 2003:504.

12. Simon RM, Korn EL, McShane LM, et al. *Design and Analysis of DNA Microarray Investigations*. 1st ed. Springer; 2004:199.

13. Hartl DL. *A Primer of Population Genetics.* 1st ed. Sinauer; 1988.

14. Menozzi P, Piazza A, Cavalli-Sforza L. Synthetic maps of human gene frequencies in Europeans. *Science*. 1978;201(4358):786-792.

15. Sokal RR, Menozzi, P. Spatial autocorrelation of HLA frequencies in Europe support demic diffusion of early farmers. *American Naturalist* . 1982;119:1-17.

16. Novembre J, Johnson T, Bryc K, et al. Genes mirror geography within Europe.

Nature. 2008;456(7218):98-101.

17. Sokal RR. The continental population structure of Europe. *Annual Review of Anthropology*. 1991;20(1):119-140.

18. Lell JT, Wallace DC. The Peopling of Europe from the Maternal and Paternal Perspectives. *Am J Hum Genet.* . 2000;67(6):1376–1381.

19. Seielstad MT, Minch E, Cavalli-Sforza LL. Genetic evidence for a higher female migration rate in humans. *Nat Genet*. 1998;20(3):278-280.

20. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet*. 2006;2(12):e190.

21. Seldin MF, Shigeta R, Villoslada P, et al. European population substructure: clustering of northern and southern populations. *PLoS Genet*. 2006;2(9):e143.

22. Bauchet M, McEvoy B, Pearson LN, et al. Measuring European population stratification with microarray genotype data. *Am J Hum Genet*. 2007;80(5):948-56.

23. Tian C, Hinds DA, Shigeta R, et al. A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. *Am J Hum Genet*. 2007;80(6):1014-23.

24. Henig RM. *The Monk in the Garden: The Lost and Found Genius of Gregor Mendel, the Father of Genetics*. illustrated edition. Mariner Books; 2001:304.

25. Motulsky AG. Genetics of complex diseases . *J Zhejiang Univ Sci B.* . 2006;7(2):167–168.

26. Keavney B. Genetic epidemiological studies of coronary heart disease. *Int. J. Epidemiol.* 2002;31(4):730-736.

27. Farrer LA, Cupples LA. Estimating the probability for major gene Alzheimer disease. *Am J Hum Genet.* . 1994;54(2):374–383.

28. Pérez-Tur J. Parkinson's disease genetics: a complex disease comes to the clinic. *Lancet Neurol*. 2006;5(11):896-897.

29. Ziegler A, Koenig IR. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications*. Wiley VCH; 2006:361.

30. Nicholas FW. Simple segregation analysis: a review of its history and terminology. *J Hered*. 1982;73(6):444-450.

31. Ott J. *Analysis of Human Genetic Linkage*. 3rd ed. The Johns Hopkins University Press; 1999:416.

32. Li W. Three lectures on case-control genetic association analysis. *Brief. Bioinformatics*. 2008;9(1):1-13.

33. Thomas DC, Witte JS. Point: Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations? *Cancer Epidemiol Biomarkers*

Prev. 2002;11(6):505-512.

34. Wacholder S, Rothman N, Caporaso N. 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*. 2002;11(6):513-520.

35. Gordon D. *Advances in Family-Based Association Analysis: 15 Years of Practical Experience with the Original Transmission Disequilibrium Test*. S Karger AG; 2008:76.

36. Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999;55(4):997-1004.

37. Pritchard JK, Donnelly P. Case-Control Studies of Association in Structured or Admixed Populations. *Theoretical Population Biology*. 2001;60(3):227-237.

38. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661-78.

39. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat Genet*. 2007;39(9):1045-1051.

40. Krawczak M, Nikolaus S, von Eberstein H, et al. PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet*. 2006;9(1):55-61.

41. Wichmann H, Gieger C, Illig T. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*. 2005;67 Suppl 1:S26-30.

42. Qu H, Grant SF, Bradfield JP, et al. Association of RASGRP1 with type 1 diabetes is revealed by combined follow-up of two genome-wide studies. *J. Med. Genet*. 2009. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19465406 [Accessed May 26, 2009].

43. Orozco G, Eyre S, Hinks A, et al. Association of CD40 with rheumatoid arthritis confirmed in a large UK case-control study. *Ann. Rheum. Dis*. 2009. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19435719 [Accessed May 26, 2009].

44. Wu C, Shete S, Chen W, et al. Detection of disease-associated deletions in case-control studies using SNP genotypes with application to rheumatoid arthritis. *Hum. Genet*. 2009. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19415332 [Accessed May 26, 2009].

45. Franke A, Balschun T, Karlsen TH, et al. Sequence variants in IL10, ARPC2 and multiple other loci contribute to ulcerative colitis susceptibility. *Nat. Genet*. 2008;40(11):1319-1323.

46. Hofmann S, Franke A, Fischer A, et al. Genome-wide association study identifies ANXA11 as a new susceptibility locus for sarcoidosis. *Nat. Genet*. 2008;40(9):1103-1106.

47. Buch S, Schafmayer C, Völzke H, et al. A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease. *Nat. Genet*. 2007;39(8):995-999.

48. Gao X, Martin ER, Liu Y, et al. Genome-wide linkage screen in familial Parkinson disease identifies loci on chromosomes 3 and 18. *Am. J. Hum. Genet*. 2009;84(4):499-504.

49. Schaefer AS, Richter GM, Groessner-Schreiber B, et al. Identification of a shared genetic susceptibility locus for coronary heart disease and periodontitis. *PLoS Genet*. 2009;5(2):e1000378.

50. Samani NJ, Erdmann J, Hall AS, et al. Genomewide association analysis of coronary artery disease. *N. Engl. J. Med*. 2007;357(5):443-453.

51. Baum AE, Akula N, Cabanero M, et al. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. *Mol Psychiatry.* . 2008;13(2):197–207.

52. Sharma P, Fatibene J, Ferraro F, et al. A Genome-Wide Search For Susceptibility Loci to Human Essential Hypertension. *Hypertension*. 2000;35(6):1291-1296.

53. Pearson TA, Manolio TA. How to Interpret a Genome-wide Association Study. *JAMA*. 2008;299(11):1335-1344.

54. Nature Insight on Functional Genomics. *Nature*. 2000;405(6788):819-65.

55. India P. Milestones in gene expression. *Nature Cell Biology*. 2005;7:1149.

56. Nacher J, Akutsu T. Sensitivity of the power-law exponent in gene expression distribution to mRNA decay rate. *Physics Letters A*. 2006;360(1):174-178.

57. Hoyle DC, Rattray M, Jupp R, Brass A. *Making sense of microarray data distributions*. Oxford Univ Press; 2002:576-584.

58. Kuznetsov VA. Distribution associated with stochastic processes of gene expression in a single eukaryotic cell. *EURASIP Journal on applied signal processing*. 2001;2001(4):285-296.

59. Costello CM, Mah N, Häsler R, et al. Dissection of the inflammatory bowel disease transcriptome using genome-wide cDNA microarrays. *PLoS Med*. 2005;2(8):e199.

60. Mah N, Thelin A, Lu T, et al. A comparison of oligonucleotide and cDNA-based microarray systems. *Physiol Genomics*. 2004;16(3):361-70.

61. Schulze HA, Häsler R, Mah N, et al. From model cell line to in vivo gene expression: disease-related intestinal gene expression in IBD. *Genes Immun*. 2008;9(3):240-8.

62. Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical magazine*. 1901;2(6):559-572.

63. Lorr M, Lyerly SB. *Conference on Cluster Analysis of Multivariate Data, New Orleans, LA., December 9, 10 and 11.*. Catholic Univ of America Washington DC; 1967.

64. Gao X, Starmer J. Human population structure detection via multilocus genotype clustering. *BMC Genet*. 2007;8:34.

65. Allocco DJ, Song Q, Gibbons GH, Ramoni MF, Kohane IS. Geography and genography: prediction of continental origin using randomly selected single nucleotide polymorphisms. *BMC Genomics*. 2007;8:68.

66. Turakulov R, Easteal S. Number of SNPS loci needed to detect population structure. *Hum Hered*. 2003;55(1):37-45.

67. Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*. 2000;155(2):945-959.

68. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet*. 2004;36(5):512-517.

69. Rosenberg NA, Pritchard JK, Weber JL, et al. Genetic Structure of Human Populations. *Science*. 2002;298(5602):2381-2385.

70. Team RC. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2007. Available at: http://www.R-project.org.

71. Krawczak M, Lu, Timothy T, Willuweit, Sascha, Roewer, Lutz. Genetic Diversity in the German Population. In: *Encyclopedia of Life Sciences (ELS)*. Chichester: John Wiley & Sons, Ltd; 2008.

72. Dupuy BM, Stenersen M, Lu TT, Olaisen B. Geographical heterogeneity of Y-chromosomal lineages in Norway. *Forensic Sci Int*. 2006;164(1):10-9.

73. Salmela E, Lappalainen T, Fransson I, et al. Genome-Wide Analysis of Single Nucleotide Polymorphisms Uncovers Population Structure in Northern Europe. *PLoS ONE*. 2008;3(10):e3519.

74. Yamaguchi-Kabata Y, Nakazono K, Takahashi A, et al. Japanese Population Structure, Based on SNP Genotypes from 7003 Individuals Compared to Other Ethnic Groups: Effects on Population-Based Association Studies. *Am J Hum Genet.* . 2008;83(4):445–456.

75. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661-678.

76. Nalls MA, Simon-Sanchez J, Gibbs JR, et al. Measures of Autozygosity in Decline: Globalization, Urbanization, and Its Implications for Medical Genetics. *PLoS Genet*. 2009;5(3):e1000415.

77. Jackson JE. *A user's guide to principal components*. 1st ed. Wiley-

Interscience; 1991:569.

78. Li Q, Yu K. Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet. Epidemiol*. 2008;32(3):215-226.

79. Miclaus K, Wolfinger R, Czika W. SNP selection and multidimensional scaling to quantify population structure. *Genetic Epidemiology*. 2009; (Online Publication):DOI:10.1002/gepi.20401.

80. Calenge C. The package "adehabitat" for the R software: A tool for the analysis of space and habitat use by animals. *Ecological Modelling*. 2006;197(3-4):516-519.

81. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. 2007;81(3):559-575.

82. Eiberg H, Mohr J. Assignment of genes coding for brown eye colour (BEY2) and brown hair colour (HCL3) on chromosome 15q. *Eur. J. Hum. Genet*. 1996;4(4):237-241.

83. Frudakis T, Terravainen T, Thomas M. Multilocus OCA2 genotypes specify human iris colors. *Hum. Genet*. 2007;122(3-4):311-326.

84. Zhu G, Evans DM, Duffy DL, et al. A genome scan for eye color in 502 twin families: most variation is due to a QTL on chromosome 15q. *Twin Res*. 2004;7(2):197-210.

85. Duffy DL, Montgomery GW, Chen W, et al. A three-single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation. *Am. J. Hum. Genet*. 2007;80(2):241-252.

86. Posthuma D, Visscher PM, Willemsen G, et al. Replicated linkage for eye color on 15q using comparative ratings of sibling pairs. *Behav. Genet*. 2006;36(1):12-17.

87. Eiberg H, Troelsen J, Nielsen M, et al. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Human Genetics*. 2008;123(2):177-187.

88. Kayser M, Liu F, Janssens ACJW, et al. Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. *Am. J. Hum. Genet*. 2008;82(2):411-423.

89. Sturm RA, Duffy DL, Zhao ZZ, et al. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *Am. J. Hum. Genet*. 2008;82(2):424-431.

90. Ingram CJE, Mulcare CA, Itan Y, Thomas MG, Swallow DM. Lactose digestion and the evolutionary genetics of lactase persistence. *Hum. Genet*. 2009;124(6):579-591.

91. Ferguson A, Maxwell JD. Genetic aetiology of lactose intolerance. *Lancet*. 1967;2(7508):188-190.

92. McCracken RD. Lactase deficiency: an example of dietary evolution. *Current Anthropology*. 1971;12(4/5):479.

93. Paschou P. PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations. *PLoS Genetics*. 2007;3(9):e160.

94. Tenesa A, Navarro P, Hayes BJ, et al. Recent human effective population size estimated from linkage disequilibrium. *Genome research*. 2007;17(4):520.

95. Rosenberg NA, Mahajan S, Ramachandran S, et al. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet*. 2005;1(6):e70.

96. Behar DM, Rosset S, Blue-Smith J, et al. The Genographic Project Public Participation Mitochondrial DNA Database. *PLoS Genet*. 2007;3(6):e104.

97. Guan W, Liang L, Boehnke M, Abecasis GR. Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genetic Epidemiology*. 2009; (Online Publication):DOI:10.1002/gepi.20403.

Acknowledgments

I thank my advisors: Michael Krawczak for supporting me in innumerably ways over the many years during which this research was undertaken, Stefan Screiber for inviting me to Germany and, in giving me a start here, made this work possible, and Manuela Ditmar for taking me on as a candidate and helpfully guiding me through the promotion process.

I thank Oscar Lao and Michael Kayser for their support as research partners.

Thanks to Robert Häsler for many interesting discussions over coffee.
To my proofreaders Martin Kerick, Robert Häsler, Sandra Freitag-Wolf, Paul Detwiler, and Gerda Rädle for their sharp eyes and quick but sure advice.

I dedicate this thesis to Mom and Dad for all they have done for me.

Finally, thanks to all my co-authors for their contributions (listed alphabetically): Miroslava Balascakova, Christian Becker, Jaume Bertranpetit, Laurence A. Bindoff, Amke Caliebe, David Comas, Christine M. Costello, Peter J.P. Croucher, Günther Deuschl, Sandra Freitag-Wolf, Ulrik Gether, Christian Gieger, Robert Haesler, Albert Hofman, Gunilla Holmlund, Olaf Junge, Manfred Kayser, Anastasia Kouvatsi, Michael Krawczak, Oscar Lao, Milan Macek, Isabelle Mollet, Matthew R. Nelson, Michael Nothnage, Peter Nürnberg, Walther Parson, Jukka Palo, Rafal Ploski, Fernando Rivadeneira, Andreas Rüther, Antti Sajantila, Stefan Schreiber, Adriano Tagliabracci, André G. Uitterlinden,Thomas Werge, Heinz-Erich Wichmann.

Curriculum vitae

Timothy Te Hua Lu

Olshausenstrasse 16
24118 Kiel
Germany

Work Phone: 0431-597-1617
Home Phone: 0431-560-1246
Email: t.lu@mucosa.de

Citizenship: United States of America

Education

Masters Degree: Biology - Fall 1990 - Spring 1993
San Diego State University - San Diego CA, USA
Masters Thesis: Population Genetic Structure of the Brown Alga Cystosiera osmundacea

Bachelors Degree: Biology - Fall 1986 - Spring 1990
University of California at Irvine - Irvine CA, USA

Work History

Job title: Scientific Staff/PhD Candidate July 2003 - present
Place of Employment: Institute für Med. Informatik und Statistik(IMIS) Kiel, Germany.
The IMIS provides statistical support for the University Clinic, Kiel and works closely with the Institute for Clinical Molecular Biology (IKMB) as a co-member of the Center for Molecular Biosciences.
Work Experience:

- Research in the fields of genetic epidemiology and population genetics.
- Statistical analysis of genetic data.
- Parallel software implementation for analysis of genome-wide SNP marker data.

Job Title: Scientific Staff/Computer Programmer July 2000 - July 2003
Place of Employment: Institute für Klinische Molekularbiologie Kiel, Germany.
The IKMB focuses on researching the genetic causes of mainly inflammatory diseases. It is the expression screening center and one of the genotyping centers of the German National Genome Research Network (NGFN).
Work Experience:

- Normalization and statistical analysis of expression microarray data
- Programming and maintenance of genetic analysis software, LIMS, custom laboratory software and database data.

Job Title: Computer Programmer March 1999 - July 2000

Place of Employment: MP3.Com San Diego, CA, USA.

MP3.com was an online music provider for independent artists to distribute and promote their music, and enabled consumers to search, sample and download music.

Work Experience:

- Statistical analysis of web site traffic.
- Release and maintenance of the music charts and associated web pages.

Job Title: Technical Support Supervisor January 1998 - March 1999

Place of Employment: Axys Pharmaceuticals La Jolla, CA, USA.

The same as below. This company was a merger between Sequana Therapeutics and Arris Pharmaceuticals.

Work Experience:

- Quality control testing of proprietary LIMS system.
- User support of proprietary and commercial bioinformatics software (Macintosh).
- Relational database management with data maintenance and programming duties.
- Supervised two other technicians.

Job Title: Technical Support Fall 1996 - January 1998

Place of Employment: Sequana Therapeutics La Jolla, CA, USA.

A company focused on high-throughput genotyping (microsatellite) of large cohorts and positional cloning in an effort to identify disease causing genes in polygenic diseases.

Work Experience:

- Same as above duties as Technical Support, without the supervisory duties

Job Title: Laboratory Technician Summer 1994 - Fall 1996

Place of Employment: Genset La Jolla, CA, USA.

The oligonucleotide synthesis facility for the American region and a subsidiary of Genset SA, Paris, France.

Work Experience:

- Oligonucleotide synthesis and quality control.
- Testing and establishment of a proprietary oligonucleotide synthesis platform.
- Head technician of the synthesis group.

Job Title: Laboratory Technician Fall 1993 - March 1994

Place of Employment: San Diego State University San Diego, CA, USA.

This marine ecology research lab was dedicated to the study of population genetics and ecology in algae and marine plants.

Work Experience:

- Population genetic and ecological studies of marine plants.
- Starch gel electrophoresis of allozymes for genetic analysis of Zostera marina.
- SCUBA diver performing field work and sample collection.

Publications

	Thesis Chapters																		
Area of Research	1	2	3	Articles published under secondary authorship															
Biotechnology	•	•	•	•	•	•	•	•	•										
Population Genetics		•	•							•	•	•	•						
Genetic Epidemiology			•	•	•	•								•	•	•	•		
Functional Genomics	•						•	•	•										
Bioinformatics	•	•															•	•	
Publication reference number	14	2	1	3	4	8	7	12	15	5	12	13	19	6	16	17	11	18	

1. **Lu TT**, Lao O, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Nielsen F, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann H, Ruether A, Schreiber S, Becker C, Nurnberg P, Nelson MR, Kayser M, Krawczak M. An evaluation of the genetic-matched pair study design using genome-wide SNP data from the European population. Eur J Hum Genet. 2009 Jan 21; doi:10.1038/ejhg.2008.266

2. Lao O, **Lu TT**, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, Balascakova M, Bertranpetit J, Bindoff LA, Comas D, Holmlund G, Kouvatsi A, Macek M, Mollet I, Parson W, Palo J, Ploski R, Sajantila A, Tagliabracci A, Gether U, Werge T, Rivadeneira F, Hofman A, Uitterlinden AG, Gieger C, Wichmann H, Rüther A, Schreiber S, Becker C, Nürnberg P, Nelson MR, Krawczak M, Kayser M. Correlation between genetic and geographic structure in Europe. Curr Biol. 2008 Aug 26;18(16):1241-8.

3. Franke A, Fischer A, Nothnagel M, Becker C, Grabe N, Till A, **Lu T**, Müller-Quernheim J, Wittig M, Hermann A, Balschun T, Hofmann S, Niemiec R, Schulz S, Hampe J, Nikolaus S, Nürnberg P, Krawczak M, Schürmann M, Rosenstiel P, Nebel A, Schreiber S. Genome-Wide Association Analysis in Sarcoidosis and Crohn's Disease Unravels a Common Susceptibility Locus on 10p12.2 [Internet]. Gastroenterology. 2008 Jul 18;[cited 2008 Oct 8] Available from: http://www.ncbi.nlm.nih.gov/pubmed/18723019

4. Franke A, Balschun T, Karlsen TH, Hedderich J, May S, **Lu T**, Schuldt D, Nikolaus S, Rosenstiel P, Krawczak M, Schreiber S. Replication of signals from recent studies of Crohn's disease identifies previously unknown disease loci for ulcerative colitis. Nat Genet. 2008 Jun ;40(6):713-5.

5. Krawczak M, **Lu, Timothy T**, Willuweit, Sascha, Roewer, Lutz. Genetic Diversity in the German Population. In: Encyclopedia of Life Sciences (ELS). Chichester: John Wiley & Sons, Ltd; 2008.

6. Nothnagel M, **Lu TT**, Krawczak M. Hypotheses in genome-wide association scans. Eur J Hum Genet. 2008 Oct ;16(10):1174-5.

7. Schulze HA, Häsler R, Mah N, **Lu T**, Nikolaus S, Costello CM, Schreiber S. From model cell line to in vivo gene expression: disease-related intestinal gene expression in IBD. Genes Immun. 2008 Apr ;9(3):240-8.

8. Buch S, Schafmayer C, Völzke H, Becker C, Franke A, von Eller-Eberstein H, Kluck C, Bässmann I, Brosch M, Lammert F, Miquel JF, Nervi F, Wittig M, Rosskopf D, Timm B, Höll C, Seeger M, ElSharawy A, **Lu T**, Egberts J, Fändrich F, Fölsch UR, Krawczak M, Schreiber S, Nürnberg P, Tepel J, Hampe J. A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for

human gallstone disease. Nat Genet. 2007 Aug ;39(8):995-9.

9. Dupuy BM, Stenersen M, **Lu TT**, Olaisen B. Geographical heterogeneity of Y-chromosomal lineages in Norway. Forensic Sci Int. 2006 Dec 1;164(1):10-9.

10. Dresske B, Haendschke F, Lenz P, Ungefroren H, Jenisch S, Exner B, El Mokhtari NE, **Lu T**, Zavazava N, Faendrich F. WOFIE stimulates regulatory T cells: a 2-year follow-up of renal transplant recipients. Transplantation. 2006 Jun 15;81(11):1549-57.

11. Franke A, Wollstein A, Teuber M, Wittig M, **Lu T**, Hoffmann K, Nürnberg P, Krawczak M, Schreiber S, Hampe J. GENOMIZER: an integrated analysis system for genome-wide association data. Hum Mutat. 2006 Jun ;27(6):583-8.

12. Costello CM, Mah N, Häsler R, Rosenstiel P, Waetzig GH, Hahn A, **Lu T**, Gurbuz Y, Nikolaus S, Albrecht M, Hampe J, Lucius R, Klöppel G, Eickhoff H, Lehrach H, Lengauer T, Schreiber S. Dissection of the inflammatory bowel disease transcriptome using genome-wide cDNA microarrays. PLoS Med. 2005 Aug ;2(8):e199.

13. Roewer L, Croucher PJP, Willuweit S, **Lu TT**, Kayser M, Lessig R, de Knijff P, Jobling MA, Tyler-Smith C, Krawczak M. Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. Hum Genet. 2005 Mar ;116(4):279-91.

14. **Lu T**, Costello CM, Croucher PJP, Häsler R, Deuschl G, Schreiber S. Can Zipf's law be adapted to normalize microarrays? BMC Bioinformatics. 2005 ;6:37.

15. Mah N, Thelin A, **Lu T**, Nikolaus S, Kühbacher T, Gurbuz Y, Eickhoff H, Klöppel G, Lehrach H, Mellgård B, Costello CM, Schreiber S. A comparison of oligonucleotide and cDNA-based microarray systems. Physiol Genomics. 2004 Feb 13;16(3):361-70.

16. Stenzel A, **Lu T**, Koch WA, Hampe J, Guenther SM, De La Vega FM, Krawczak M, Schreiber S. Patterns of linkage disequilibrium in the MHC region on human chromosome 6p. Hum Genet. 2004 Mar ;114(4):377-85.

17. Croucher PJP, Mascheretti S, Hampe J, Huse K, Frenzel H, Stoll M, **Lu T**, Nikolaus S, Yang S, Krawczak M, Kim WH, Schreiber S. Haplotype structure and association to Crohn's disease of CARD15 mutations in two ethnically divergent populations. Eur J Hum Genet. 2003 Jan ;11(1):6-16.

18. Hampe J, Wollstein A, **Lu T**, Frevel HJ, Will M, Manaster C, Schreiber S. An integrated system for high throughput TaqMan based SNP genotyping. Bioinformatics. 2001 Jul ;17(7):654-5.

19. **Lu TT**, Williams SL. Genetic diversity and genetic structure in the brown alga Halidrys dioica (Fucales: Cystoseiraceae) in Southern California. Marine Biology. 1994 Dec 1;121(2):363-371.

Erklärung

I hereby do declare that, aside from the guidance provided by my thesis advisors, this dissertation is completely and in its entirety, my own work. All work presented here has been carried out in accordance with good scientific practices. The three articles presented in this dissertation have been previously published. My contributions to work presented in published journal articles upon which I appear as primary author or primary co-author is summarized below. I have not submitted this thesis, either in the past or simultaneously, to any other doctoral program and I have never enrolled in any doctoral program other than the present one.

Kiel,.....

(Timothy Te Hua Lu)

Article1:

Lu, et. al. Can Zipf's law be adapted to normalize microarrays? BMC Bioinformatics. 2005 ;6:37.

I conducted all data analysis, wrote all original computer software, was primary developer of normalization method, and was principle author of the manuscript.

Article 2:

Lao & Lu, et. al. Correlation between genetic and geographic structure in Europe. Curr Biol. 2008 Aug 26;18(16):1241-8.

I conducted IBS quality control analysis, principle component analysis, CEPH-CAU sample analysis, conceived and implemented kernel density mapping of populations, and was principle co-author of the manuscript.

Article 3:

Lu & Lao, et. al.. An evaluation of the genetic-matched pair study design using genome-wide SNP data from the European population. Eur J Hum Genet. 2009 Jan 21;[online advanced publication] doi:10.1038/ejhg.2008.266

I conducted all data analysis, wrote all original computer software, and was principle co-author of the manuscript.