Patterns of Linkage Disequilibrium in the Human Genome

By

Shau Neen Liu-Cordero

B.A. Genetics
University of California Berkeley, 1993

SUBMITTED TO THE DEPARTMENT OF BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

MAY 2002

Signature of Author: _____
Department of Biology
May 24 2002

Certified by: _____
Eric S. Lander
Professor of Biology
Thesis Supervisor

Accepted by: _____
Terry Orr-Weaver
Professor of Biology
Co-Chair Graduate Committee

In Memory of my Father

I miss you, and I'm sorry you were not able

to witness the completion of my Ph.D. Thesis.

Acknowledgements

Etchell, you are amazing in so many ways. I have learned so much from you. I simply could not make it through life without you.

Eric, your intensity and enthusiasm towards everything you set your mind to have always been an inspiration to me. You have shown a great deal of patience and understanding, and you have significantly shaped me as a scientist and a person.

Mark Daly, you have been extensively involved in every project that I have ever done in this lab; I am grateful for all your help and all that you have taught me. David Altshuler, your advice and support have helped to keep me focused and moving; thank you for your ideas and leadership. Kristin Ardlie, you were the most pleasant and generous person that I ever worked with; thank you for all the "no worries". Terry Orr-Weaver and David Page, my thesis committee members of many years, thank you for pressing me to stay on track, shaping my thesis, and defining the limits of what I could finish. Also, thanks to all the great colleagues that I have worked with over the years who are too numerous to name.

I'd like to thank Jim, Chris, Laura, Karen, Mike, Vic, Nika, and Darin for being my friends and providing unconditional support throughout the years. Todd and Christine, you almost kept me sane in an insane work environment. Thanks to the FTJ crew for allowing me to live the high life.

Table of Contents

Patterns of Linkage Disequilibrium in the Human Genome

By

Shau Neen Liu-Cordero

Submitted to the Department of Biology
on May 24, 2002 in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in
Biology at the Massachusetts Institute of Technology

ABSTRACT

Although enormous progress has occurred in the field of human genetics, the cloning of complex trait mutations remains a challenging and unresolved process. This continuing difficulty is responsible for an ever-increasing awareness of the phenomenon of linkage disequilibrium (LD). The principle behind LD is relatively simple. Over the lifetime of a population, the genetic markers that are adjacent to an ancestral mutation will recombine less often than more distant markers. Therefore, the ancestral alleles of the markers closest to the mutation should be most frequent in a collection of disease chromosomes. The allelic association should decrease as the distance from the ancestral disease mutation increases. This thesis is a collection of ideas and experiments aimed at dissecting the behavior of LD in the human genome. Specific studies examine LD in a variety of populations including isolated founder populations, as well as globally diverse population samples. A large number of regions throughout the genome are investigated using both pairwise comparisons of markers, as well as multimarker haplotypes. The X chromosome is more closely scrutinized because of its unique population history, as well as the advantages afforded to haplotyping due to hemizygosity of the X chromosome in males. Major conclusions include the observation that LD between pairs of markers is highly variable even at extremely close distances and multimarker haplotypes better serve to resolve the underlying haplotype structure of the genome. The genome appears to be structured as blocks of limited haplotype diversity that do not exhibit much internal recombination but which are separated by segments that show little or no LD. The lack of LD between haplotype blocks appears to be due to clustering of recombination events into specific hotspots. The size of the blocks and haplotype diversity varies slightly by population. In addition, the identity of the haplotypes varies between populations. The existence of 3-4 major haplotypes for specific regions in a diverse human population sample is a surprising finding that was originally believed to have only existed in very special isolated and young populations.

Thesis Supervisor: Eric Lander

Title: Professor of Biology

CHAPTER 1

Introduction

Linkage Disequilibrium: A Guided Tour

Overview

In the past decade, immense strides have been taken to advance the study of human genetic diseases. Major developments in the mapping of genes and cloning of disease causing mutations have been catalyzed largely by the creation of whole genome resources. The process of systematically discovering the mutations causing inherited disease based solely on their position in the genome was once seen as an insurmountable task. Successively better installments of polymorphism maps, gene maps, and genome sequence have been instrumental in implementing this strategy. This genetic mapping strategy is essential for most human traits due to the lack of any prior information about the biological pathways involved. The vast majority of successful attempts to map and clone genetic disease mutations in humans have come in the category of single gene disorders -- completely penetrant with a clearly defined phenotype and characterized by a simple Mendelian inheritance pattern. Most of the genes and mutations for these disorders were identified by traditional linkage analysis. In linkage analysis, regions of the genome that cosegregate with the disease phenotype are identified. This is accomplished by following the inheritance of polymorphic markers in a collection of families affected with the disease. All affected members should share the region surrounding the disease mutation, and the boundaries of this region can be delimited by the observation of crossover events between the surrounding marker alleles and the disease allele. However the number of these observed crossovers is normally fairly low. Consequently, the interval surrounding the disease locus remains large enough to make cloning the causative mutation a daunting process. The advent of denser genetic marker maps, gene maps, and, ultimately, genome sequence has increased the feasibility of this task by many orders of magnitude. In addition, the interval can be refined by examining a large number of affected families or by obtaining very large multigenerational families. However, for most traits, the feasible resolution in genetic linkage mapping still remains limited (Boehnke, 1994; Kruglyak, 1999).

The realization of the limits of resolution in linkage mapping precipitated disease gene mappers to search for other approaches to refine the interval surrounding their disease mutation of interest. Mapping by the method of association became more popular

due to the successes of this method in the HLA region of the genome. Instead of following cosegregation in families, association studies detect differences in the frequency of genetic variants between unrelated affected individuals and unaffected controls. In this population sample, one looks for an association between a phenotype and a specific marker allele. If the allele is found in excess in the affected individuals, then that allele is "associated with" the disease phenotype. A positive association result can be interpreted in one of two ways. The allele that is in association with the disease phenotype may in fact be the causative allele of the disease, in which case, one would ideally expect that the association between the phenotype and the disease-causing variant should be complete. Alternatively, the associated allele may not cause the disease but may simply be a neutral polymorphism that happens to reside nearby the disease causing allele in the genome. This phenomenon is known as linkage disequilibrium. In this case, throughout the history of the disease population, recombination has not had enough time to occur between the two loci. An increasing amount of recombination between the disease locus and the neutral polymorphism would gradually reduce the level of the association until they appeared to be unlinked and completely non-associated. In the most basic application of association studies, case-control studies, there are some tricky aspects of examining allele frequencies in unrelated individuals that may cause biases in the results if the control group is not well matched demographically to the affected individuals. This ethnic admixture bias can cause false positive results in associations (Lander and Schork, 1994) and will be described in greater detail in a subsequent section of this introduction.

Linkage disequilibrium will arise in a population when most individuals affected with the disease in question carry relatively few ancestral mutations at a disease-causing locus. In addition, a specific allele of a marker must have been present on one of those ancestral chromosomes and must reside near enough to the disease-causing locus that the correlation has not yet been eroded by recombination during the history of the population. In the most favorable scenarios, most affected individuals in the population will share the same mutant allele at the disease locus. Therefore, the whole population could be regarded essentially as one enormous family. The degree of linkage disequilibrium detected between the disease locus and nearby markers can then be used to reduce the

genetic interval surrounding the disease locus. In contrast to the limited number of crossover events available in linkage mapping in families, the population approach would employ the much larger numbers of crossover events that have occurred in all the generations since the founding of the population or the first appearance of the mutation in the population. This approach would therefore theoretically provide a much finer resolution for mapping. This concept forms the basis for linkage disequilibrium (hereafter referred to as LD) mapping. **(Figure 1)**

In actuality, LD mapping is only as simple as described above in a young, isolated population. LD does not display a simple relationship with recombination, physical distance, or a variety of other factors. Therefore, it's not surprising that the first successful applications of LD mapping in human population emerged from mapping simple monogenic disorders such as cystic fibrosis, myotonic dystrophy, and diastrophic dysplasia (Hastbacka et al., 1992; Jorde et al., 1994; Jorde et al., 1993; Kerem et al., 1989). LD mapping in these and other monogenic diseases was instrumental in the cloning of the disease mutations. Although human geneticists have been victorious in the arena of rare, simple monogenic disorders, the diseases that affect larger proportions of the population usually follow a more complex pattern of inheritance. Elucidating the genetic basis of common afflictions such as diabetes, asthma, hypertension, and psychiatric disorders is a much more complicated task.

The Challenge of Mapping Human Complex Traits

Efforts of human geneticists have become more and more focused on these more complex and common traits. Linkage studies have not been nearly as successful for these diseases due to a large number of confounding factors (Lander and Schork, 1994). Genetic heterogeneity is one of the most devastating complexities for linkage mapping. Mutations in any one of a number of different genes may be sufficient to cause a disease phenotype. Consequently, many of the affected families in a mapping experiment may have different loci in the genome segregating with the disease. Allelic heterogeneity will affect association studies more. In these cases, there are many different disease-causing alleles at the same locus in the genome, so that any particular one of those alleles will not be associated with the disease in all affected individuals. Rare diseases in young, isolated

founder populations have the highest chance of having reduced heterogeneity of both types. Incomplete penetrance can also be a major problem. Individuals who inherit a disease allele at a certain locus may not display the disease phenotype. The converse situation can be seen in phenocopy, where the disease is manifested in spite of the fact that no risk allele has been inherited. This situation may arise due to environmental factors that are difficult to control in human disease studies because they are predominantly not controlled in human populations. Other complexities include polygenic inheritance, in which combinations of multiple genes is necessary to acquire the disease phenotype, and "epigenetic" methods of inheritance such as mitochondrial inheritance or imprinting. Essentially, all these factors preclude a perfect correlation between genotype and phenotype. The most difficult of the common diseases display combinations of these complexities, and the inferred genomic regions are large and do not necessarily include any mutation with 100% confidence.

Due to the difficulties inherent in mapping these complex traits, the question arises: could the power of LD mapping be used to overcome some of the complexities and discover factors involved in the etiology of common human diseases? Attention became focused on creating resources and designing experiments to search for associations in complex diseases. Bearing in mind the two possible interpretations of a positive association ((1) the allele is the disease causing mutation itself or (2) the allele is in LD with the disease allele), one of two general strategies can be employed in the design of association studies: direct or indirect (Collins et al., 1997) **(Figure 2a).** Both strategies depend on the validity of a hypothesis that supposes that common genetic variants underlie susceptibility to common diseases. This common variant-common disease (CDCV) hypothesis has been only been validated by a small number of examples. The CDCV hypothesis is not proven to be generally valid and is still an active topic of debate. The direct strategy requires a comprehensive collection of common variation in human genes. This collection of variants would be assumed to contain the causative mutation, and frequencies of these variants would be compared in patients and controls in search of an excess (or deficiency in the case of a protective allele) in patients (Cargill et al., 1999; Halushka et al., 1999). The advantage of this method is that the experiments could be hypothesis driven, choosing variants that are most likely to be involved in the

disease. Additionally, a detected association to a causative allele would be much stronger than an association to a nearby marker, so the sample sizes for these studies would not have to be as large. Unfortunately, it is difficult to know where a causative mutation would be located in a gene. The coding regions are often considered the most likely locations, but that may not necessarily be true in a disease with complex inheritance. Furthermore, mutations may be impossible to recognize from only primary sequence data. Important variants may be buried in introns or upstream control regions as well. The indirect strategy for association studies relies on detecting an association between the disease phenotype and markers that are in LD with the disease mutation. The indirect approach avoids the need for a comprehensive collection of supposed susceptibility alleles. However, the strength of the association decreases as the level of LD decreases. This approach would therefore require an extremely dense map of single nucleotide polymorphisms (SNPs) across the genome. SNPs are the most preferable types of markers for these types of studies, as opposed to simple sequence length polymorphism or other kinds of repeats. These markers are extremely prevalent, with the genome expected to contain as many as ten million common SNPs. SNPs have much lower mutation rates and, therefore, are more appropriate for studies in older populations. The use of SNPs is also advantageous because they will tend to have similar properties to disease mutations since they are exactly the types of polymorphisms expected to be causative mutations.

Currently, at least 2 million SNPs have been discovered in the human genome at an average density of one SNP every 1.9 kb in available genome sequence (Sachidanandam et al., 2001). Genome-wide screening for LD in a common disease population with this dense SNP map resource is a tantalizing possibility. However, even the earliest studies (reviewed in the following sections) indicated that the levels of LD were quite variable between different regions of the genome. Consequently, even with the available resources, more information about the structure of LD in the human genome would be necessary in order to determine if these experiments are even feasible. It is still unclear how many -- and which -- of the millions of SNPs should be genotyped and which populations should be studied. It is likely that the answer would depend on the characteristics of the specific disease of interest. However, we have only begun to

explore these uncertainties or to even understand what other uncertainties exist. The goal of this thesis is to present and answer many of the questions concerning the patterns and behavior of LD in the human genome, understand the forces that created those patterns, and then apply that knowledge to determine the best methods to elucidate the factors underlying human disease.

The remainder of this introduction will review in detail the development of the ideas and questions surrounding the use of LD in disease mapping and human demographic history studies. The concurrent technological advances and creation of human genome resources will also be described since they were intimately connected to the types of experiments that could be performed at various points in time. In this introduction, a chronological approach is provided to describe the major advances in LD mapping. Furthermore, this history of LD studies is presented to focus more clearly on the gradual shaping of ideas surrounding the use of LD in mapping of human traits, which is embodied in and paralleled by the chapters in this thesis.

Properties of Linkage Disequilibrium

Factors that Affect Linkage Disequilibrium

As described above, LD is the nonrandom association of alleles at adjacent loci. LD arises in a population when a new disease mutation, or any kind of mutation, occurs on an existing chromosome that is already carrying a specific allele at a nearby locus. This mutation event will give rise to a particular combination of the two alleles found at the adjacent loci. A haploid combination of alleles for a pair of markers, or a number of markers from a whole region or chromosome segment, is often simply referred to as "a chromosome". This notation is simply applied to represent the particular configuration of alleles on a single molecule of DNA across a genomic region, and not necessarily across an entire chromosome. In the simple case of a pair of markers, the configuration of alleles at the two adjacent loci will decay gradually over time eventually giving rise to all four possible haplotypes (**Figure 2b**). If there hasn't been enough time for this particular configuration of alleles to completely decay, then the two alleles will appear on the same

"chromosome" more often than would be expected if they were unlinked and segregating independently. These two alleles are therefore in linkage disequilibrium with each other.

The most straightforward explanation for the decay of LD is that meiotic crossover events occur between the SNPs over time. The extent and patterns of LD would be much simpler to resolve if crossovers were the only factor affecting the relationship of LD to physical distance. However, there are many determinants of the levels of LD across the genome. Recurrent mutations, in which the same nucleotide is mutated multiple times, at either locus, can also reduce the level of LD. For most SNPs though, the mutation rate is low enough that it most likely does not play a major role the process of LD decay. Recurrent mutation would play a much greater role in breaking down LD between microsatellite repeat polymorphisms. Some SNPs may have much higher mutation rates, for instance, those at CpG dinucleotide sites. Even extremely tightly-linked markers, with an absence of historical recombination events, would demonstrate a minimal degree of LD if one or both of those markers had a high frequency of mutation events. Also, in the case of very closely spaced SNPs, the process of gene conversion can decrease the magnitude of LD. In a gene conversion event, a stretch of one copy of a chromosome is used to convert the sequence of the other copy during meiosis. If this tract of sequence contains a polymorphism, then there can be an allele change. The distance over which gene conversion acts in humans is not known, but it should contribute to the apparent rate of recombination and the breakdown of LD primarily when the distance between markers is not significantly greater than the length of a gene conversion tract. Gene conversion may be an important cause of highly variable LD seen at short distances (Ardlie et al., 2002; Frisse et al., 2001). At larger distances, variable LD would be consistent with extremely variable recombination rates. Across the human genome, recombination rates can vary by more than an order of magnitude (Payseur and Nachman, 2000). Recombinational hotspots can occur, where recombination events are limited to relatively small regions with long interspersed segments virtually devoid of recombinational activity.

Demographic histories can vary widely for different populations and for each region of the genome. Much of the variability in extent of LD has its origin in these diverse population histories. These forces can be highly stochastic and difficult to trace

over time. In the study of evolutionary and demographic processes, there are a great number of components to consider, each having a significant degree of uncertainty. Some of the many features that are known to affect population structures, and therefore affect LD levels, are inbreeding and non-random mating, migration, selection, admixture population bottlenecks and genetic drift. Genetic drift is the process by which gene and haplotype frequencies can fluctuate over the generations due to random sampling of gametes. If a population is small and not growing at an appreciable rate, chromosomal haplotypes can be lost by drift and, therefore, LD would be expected to increase. By the same token, a population which is growing very rapidly would tend to lose fewer haplotypes. As a result, LD would be expected to decrease over time (Terwilliger et al., 1998). Selection acting on mutations having advantageous or deleterious effects can act to increase LD surrounding those mutations by decreasing the haplotype diversity in those regions. Decreased haplotype diversity simply means that there will be fewer combinations of alleles available to recombine with each other, and therefore associations between those alleles will be greater. Selection for a mutation causing an advantageous trait in a population can increase the frequency of the haplotype on which that mutation resides, thus increasing the LD in the region. In this "hitchhiking" scenario, the extent of LD across the region depends on the rate in which the haplotype increases in frequency at the expense of other haplotypes. Chromosomes with haplotypes that contain deleterious mutations, when eliminated, also decrease haplotype diversity and thus increase LD. These demographic processes have been well documented over the years by population geneticists, but the exact manner in which they affect LD or how the LD generated by these processes can be used to map disease genes has not been well characterized.

Questions Surrounding the Patterns of Linkage Disequilibrium in the Human Genome

With so many factors affecting the levels of LD, a huge range of questions emerges, which can be broken into a number of distinct categories. 1) Identifying the patterns of LD across the genome; 2) Applying this knowledge of LD patterns to map human complex traits; 3) Designing experiments to delimit the extent and patterns of LD; 4) Determining how these patterns of LD were created in the genome; 5) Examining these patterns of LD to better understand the structure and history of the human

population. Although these aspects are all interdependent, a common problem with LD studies is that they are often designed to try to answer all of these questions at once. Experiments aimed at elucidating LD structure attempted to define the limits of detectable LD in physical and genetic distance, determine the variability across the genome and across populations, and understand the effect of allele frequencies.

Pairs of markers certainly cannot distinguish between all existing chromosomal haplotypes in a population. A combination of alleles for a pair of markers defines a haplotype. However, longer haplotypes can be constructed from three or more markers, and a particular two-marker haplotype can exist on multiple longer haplotypes **(Figure 3)**. The more markers that are examined, the more the underlying haplotype structure can be resolved. The number of markers it would take to resolve all existing haplotypes is likely to be different in each region of the genome and to vary across populations. Therefore, an issue emerges concerning the difference between pairwise LD comparisons and longer multimarker haplotypes. Does LD vary more when examined in pairs of markers? How many markers are necessary to genotype before the full haplotype diversity is ascertained? In order to recognize how to use this information to perform LD mapping in complex disease, one would have to determine the number and density of markers necessary. In addition, it would be important to choose the allele frequencies of these markers that would be most applicable to the disease of interest. The issue would arise concerning whether or not it would be better to search across the genome in a random fashion or to just focus on genes, i.e. the indirect or direct approach to LD mapping. What would be the appropriate resource to create? SNP discovery in one population does not provide the best resource for studies in different populations. Experiments designed to use LD to map disease genes, in theory, should be much different than experiments designed to determine the structure of LD in the region of interest. Traditionally, both types of experiments were one and the same and sought to use the maximum number of SNPs that were currently available. Currently, designing these experiments has become a significant issue because of the overwhelming number of SNPs now available. It would be an infeasible task to type all of the SNPs in a sample of any size, let alone a comprehensive search across populations.

An area of study less directly applicable to gene mapping, but of equal interest, addresses how various forces act to create the patterns of LD. Recombination is the primary force acting on LD, but how does it act in concert with processes that shape population structures? This set of issues will be especially related to those that concern the history of the human population. What groups of SNPs and haplotypes are common throughout the human population and which are unique? Is there greater nucleotide and haplotype diversity in some populations and how does this impact LD mapping in diseases? It is apparent that the task at hand is quite complicated and all of the questions are interrelated. Many diverse experiments are required. However, The study of LD had its origins long before human geneticists sought to take advantage of LD for disease gene mapping. Many of these questions were first addressed in a much simpler organism, *Drosophila Melanogaster.*

Origins of Linkage Disequilibrium Concepts and Early Experiments

Lessons from Drosophila

Most of the early experiments regarding how LD behaves in a population were not performed in humans but came from the field of *Drosophila* genetics. Accordingly, the emphasis was not on gaining insight for disease mapping, but on determining what forces are responsible for shaping LD. Recent studies continue this trend and still provide some of the most detailed explanations for the patterns of LD. Well-characterized lab strains of *Drosophila melanogaster* have been widely studied because they provide an excellent genetic model system. However, naturally occurring populations are more preferable for studying the characteristics of LD, as it would be impossible to replicate all of the forces that contribute to LD in an inbred lab strain. Single chromosomes extracted from natural populations by selective breeding can be maintained in inbred lines, and haplotypes can be determined in these strains. In this manner, LD studies have been performed in different wild *Drosophlia* populations. Comparative studies of LD levels among populations or among genes have shown a large degree of variation in different genomic regions (Zapata and Alvarez, 1993). In many cases, he determinative factors have been shown to be variability in recombination rate,

including crossovers and gene conversion events (Langley et al., 2000; Schaeffer and Miller, 1993). In *Drosophila*, this regional variation in recombination rate has been correlated with variability in polymorphism rates (Begun and Aquadro, 1994; Begun and Aquadro, 1992). A recent study suggests that positive selection may be responsible for decreased nucleotide diversity due to fixation of haplotypes that hitchhiked with an advantageous mutation (Parsch et al., 2001), leading to decreased recombination and increased LD.

One *Drosophila* chromosome which has a unique evolutionary history is the fourth chromosome. The fourth chromosome has been thought for many years to be completely non-recombining. A recent study has shown a low level rate of recombination on the chromosome (Wang et al., 2002). This study showed that, although the polymorphism rate on the chromosome was extremely low, an irregular pattern of polymorphic regions was distributed between long regions absent of polymorphism. This example underscores the possible correlation between mutation and recombination. This raises questions about whether or not regions in humans with low recombination rates will necessarily be correlated with segments of high linkage disequilibrium and low variation. For example, the human X chromosome has much lower polymorphism and recombination rates than the autosomes, establishing a very interesting issue about the variability of LD on chromosomes having very different evolutionary histories. LD might be expected to extend further on the X chromosome due to selection, different population history, and smaller effective population size. Many lessons can be learned from *Drosophila* in the design of experiments in human population. *Drosophila* provided the first examples of how different forces, such as recombination rates and selection, act on LD to create highly variable patterns in different regions. The following sections will describe how the same picture is gradually coming into focus in humans.

Linkage Disequilibruim and Early Disease Mapping Experiments in Human Populations

The pioneering experiments in Drosophila framed the pertinent issues to be addressed in humans. However, the driving force behind LD studies in humans was the mapping of disease genes. As a result, much of the subtlety behind the forces that shape

LD were not initially considered because the resources in humans were too scarce to perform the ideal experiments. In addition, there are inherent limitations of humans as a genetic model system which precluded human geneticists from implementing the same sort of precise experiments as in *Drosophila*. However, there exists some knowledge about the history of the human population and, in the case of some special populations, very precise genealogical records. The ability to gain phenotypic information from medical records and the existence of a vast natural study population to draw from allows entire human populations to become elegant model genetic systems with their own unique properties.

Studies of the major histocompatibility complex brought a great deal of attention to the use of LD as a strategy for detecting associations to human diseases. Polymorphisms in the human leukocyte antigen (HLA) system were found to be associated with a wide variety of autoimmune and immunologically-mediated diseases as well as infectious diseases. Early studies used serological protein variation before more modern genotyping methods were invented, but studies using other types of polymorphisms have confirmed these early studies. Associations were particularly strong and easily detectable in the HLA genomic locus due to the extensive LD in the region. Linkage disequilibrium exists over most of the HLA region, a distance of at least several megabases. The early successes with HLA sparked a trend in human disease mapping studies. However, the results from other regions were often not nearly as strong or significant as in HLA. The large extent of LD in the HLA region allowed for a less dense coverage of polymorphisms to screen the region. Furthermore, HLA provided the most variation in any genomic region at the time to the scientific community. These advantages were not readily available for other genomic loci. Although crucial for the implication of HLA in many diseases, the extremely high levels of LD in the region would also work against the discovery of the specific variants that caused the diseases. Only much later would more fine-scale LD mapping assist in the cloning of human disease genes.

At the end of the 1980's and beginning of the 1990's, studies in mapping simple monogenic disorders yielded some the first successful examples of LD mapping used clone disease mutations in human populations. Positional cloning was still in its infancy.

The problem of reducing the mapped interval delimited by linkage studies was the main focus. At this time, genetic or physical maps available in humans were extremely limited. Researchers had to discover many polymorphisms on their own, use them to resolve the smallest interval surrounding their disease mutation, and then clone the region and create their own physical maps. Most successful attempts to clone disease mutations had the advantage of significant chromosomal abnormalities. Cystic fibrosis was the earliest example of positional cloning without the assistance of such large aberrations (Collins, 1995; Kerem et al., 1989). LD mapping played an important role in the cloning of the cystic fibrosis (CF) mutation. Restriction fragment length polymorphisms (RFLPs) were the markers of choice at this point in time. CF was primarily studied in a broad and genetically heterogeneous Caucasian disease sample. However, this disadvantage was somewhat alleviated by the fact that the major causative mutation was a three base pair deletion and was present in about 70% of CF patients. Consequently, the haplotype diversity in the region surrounding this mutation was somewhat limited, making it easier to apply LD mapping. If there were a larger number of lower frequency mutations existing on different haplotype backgrounds, it would have been a greater challenge to detect associations. It would have been extremely optimistic to think that more than a few diseases would have this mutational allelic spectrum, especially for more complex diseases. The advantage of reduced genetic heterogeneity became more apparent and the search for ways to reduce this heterogeneity began. The idea emerged that it may be advantageous to use special populations with reduced genetic diversity resulting from specific demographic histories. These isolated founder populations proved to be extremely valuable in mapping and cloning of many human diseases.

Isolated Founder Populations

The early part of the 1990's witnessed the first examples of the elegant use of whole populations as model genetic systems. In a number of populations, such as Finland, Iceland, and Sardinia, a large proportion of the present day population is descended from a limited set of founder individuals. This restricted genetic diversity owes its existence to the occurrence of a bottleneck sometime during the history of the population. A bottleneck is simply a period of time during which the size of a population

is reduced **(Figure 4)**. The size of this reduction can vary depending on the specific population history or event causing the reduction. The greater the reduction in size, the fewer the number of founders that give rise to the population, leading to a more homogenous resultant population. The decreased heterogeneity will be even more pronounced if the bottleneck is followed by a rapid population expansion, because each of the chromosomes that make it through the bottleneck will greatly increase in number. Therefore, it will be a much rarer event for different chromosomes to come together in a randomly mating population. In addition to the number of founders and the amount and rate of growth subsequent to the bottleneck, the length of the period during which the population size is reduced also will have an effect on genetic diversity. The smaller the population, the greater the chance there is to lose haplotypes by genetic drift **(Figure 4)**. This chance will also increase with increasing time of the population bottleneck (Kruglyak, 1999). Another advantage of some of these founder populations is the lack of any appreciable contribution of external genetic variation due to migration. Most populations are admixed with varying levels of contribution from other sources that may or may not be traceable. Therefore, controls that are taken from isolated founder populations are more likely to closely match their counterpart cases taken from the same genetically homogeneous, non-admixed population.

In many ways, studying isolated founder populations is even more advantageous over other genetic model systems. A population was formed, and it is possible to look 100, 1000, or even 5000 generations later. One doesn't have to wait for recombination events to occur as in a genetic cross in model organisms. All of the recombination events throughout the population history may be available for fine structure mapping. Although it is not possible to observe the results of each meiosis over time, a great deal of information can be acquired by looking at the sum total of all meioses that occurred over the history of the population. This recombinational history is the reason that isolated founder populations are so valuable for LD mapping, especially in the case of a rare variant. In the case of a rare variant, the allele may have entered the population in only one founder. This means that, at the time of the bottleneck, there would be only one haplotype surrounding the disease variant and that haplotype would stretch across the entire chromosome. Recombination would break down the haplotype over time, but the

extent of LD would be far greater in this situation than if there were many copies of the disease allele that entered the population, or if there had not been a bottleneck at all. It can be even more valuable if extensive genealogical and medical records are available. One may be able to trace almost completely the ancestry of every individual; such is the case for the Saguenay Lac St. Jean region in Quebec and the North American Hutterites (Heyer and Tremblay, 1995; Ober et al., 1998). Isolated founder populations were regarded at the time as a powerful tool for the problems of human disease mapping due to their beneficial attributes: decreased genetic heterogeneity due to bottlenecks, subsequent genetic drift, and limited migration into the population; increased LD due to limited copies of disease alleles entering the population; and often excellent knowledge of population history, genealogies and medical histories.

The Example of Finland

In the earliest efforts to use LD mapping for rare monogenic disorders in isolated founder populations, Finland was the principal subject. Finland appeared to provide an ideal population for gene mapping. The majority of the present day population of Finland, approximately 5 million inhabitants, is thought to have descended from a relatively small set of founders that migrated to the southwestern region of Finland about 2000 years ago, or approximately 100 generations ago. The size of the population remained relatively small, and there were some internal migrations north about 500 years ago. About 250 years ago there was a large population expansion, and a large portion of the population moved into towns and cities in the last 50 years. Therefore, in addition to the overall population being an excellent resource for LD mapping, there may also be some more recently created population substructure that may provide even further homogeneity and longer tracts of LD (Kere, 2001). The Kainuu region in the eastern central part of Finland has a current population of about 95,000. The region was founded by about 2000 individuals only around 25 generations ago. Many of those individuals were likely to have been related, and therefore the effective number of founding chromosomes in this portion of Finland may have been as low as 100. LD would be expected to be greater in Kainuu than Finland as a whole and more genetically homogeneous as well. Accordingly, long ancestral haplotypes many megabases in length

have been discovered in this subregion of Finland (Hoglund et al., 1995). Finland has substantial church parish records allowing ancestries to be traced as far back as twelve generations. The founder effect and genetic drift are remarkably evident in the substantial differences in disease frequencies between Finland and the remainder of Europe. These disease frequencies may be higher or lower and are indicative of random drift processes.

Many of these rare Finnish diseases were mapped, including diastrophic dysplasia, progressive myoclonus epilepsy, cartilage-hair hypoplasia, congential chloride diarrhea, and Batten disease (Hastbacka et al., 1992; Hoglund et al., 1995; Lehesjoki et al., 1993; Mitchison et al., 1995). In the most famous example of LD mapping in an isolated founder population, diastrophic dysplasia in Finland, 95% of affected individuals carried the same haplotype, which was present at only 3% on unaffected Finnish chromosomes. This demonstrated the striking degree of homogeneity that can exist in the Finnish population. In addition, the linkage mapping had only localized the mutation to approximately a megabase sized genomic region. By looking at many markers within this region, the likely region in which the mutation resided was reduced to around 60 kb by examining the many historical recombination events in the region. The mutation itself was subsequently cloned, and the utility of populations such as Finland was quickly applied to many different disease mapping studies.

Although the critical region for diastrophic dysplasia could be refined to about 60 kb, detectable LD extended as far as a megabase. However, this was an extremely rare haplotype, and LD would not necessarily be expected to extend as far in a collection of chromosomes from the general Finnish population. Such a sample of individuals is examined in **Chapter 2**. In this study, evidence is provided for extensive LD in a broad sample of the Finnish population. LD was detectable across 600 kb, and extremely limited haplotype diversity was discovered in the region. In one subregion, which spanned more than 50 kb, just two different haplotypes comprised of 5 markers account for the majority of haplotype diversity. In addition, there were few rare haplotypes in the subregion, again highlighting the extremely limited effective number of founders in the Finnish population. Another study of anonymous genome regions also found detectable

LD between microsatellite markers that were more than a megabase apart (Peterson et al., 1995).

During this same time period, reasonably dense microsatellite marker maps of the human genome were being created. These maps were constructed by using a resource of many large families created by the Centre d'Etude Polymorphisme Humaine (CEPH), which is still very widely used today. Although large numbers of these markers had to be run on gels, which was a time consuming and laborious process, methods for increasing throughput in these experiments were beginning to appear. Microtiter plates and multichannel pipettors made it easier to complete large-scale experiments. The notion that increasing the numbers of markers and samples provided more complete information became widely accepted (Kruglyak and Lander, 1995).

In some linkage disequilibrium experiments, regions implicated in disease were studied in CEPH individuals instead of isolated founder populations in order get a view of LD in a general population (Jorde et al., 1994; Jorde et al., 1993; Watkins et al., 1994). What was found was a correlation between LD and physical distance, with wide variability in the levels of LD between and within different regions. This observation of variability foreshadowed what would be discovered later in much larger data sets.

Family-based controls

The successful associations and LD mapping studies in isolated founder populations sought to do more than to take advantage of the genetic homogeneity for increased matching of cases and controls. The concept of family-based controls was introduced. Trios consisting of a father, mother and child, with the child affected with a disease, were collected and used in association studies. By typing markers in this trio, one can determine the phase of the alleles of the markers using the family structure, and haplotypes could be constructed. In addition, the untransmitted chromosomes of the father and mother, presumably not containing any disease mutation, could be used as unaffected controls perfectly matched for ethnic ancestry (**Figure 5**). At the time, in the early and mid 1990's a large number of false associations were accumulating in the literature. Many studies failed to account for the possibility of an ethnic admixture problem. In a mixed population, any allele that is more common in one ethnic group will

yield a positive association with any trait that also happens to be higher in that ethnic group, even though the allele has no causative effect on the trait. The problem is that the effect of this can be subtle, and the variable ancestry can reside within single individuals as well. Therefore, the population can be a homogeneous population with mixed ancestry and the problem could possibly still arise depending on the markers that were used. The use of Finland and other genetically homogeneous isolated founder populations as well as the use of untransmitted chromosomes as perfectly matched controls provided the cleanest and most robust association studies at the time.

The Consequences of Different Population Histories

The level of LD found in different populations can be quite variable. There are isolates that were founded 100 or more generations ago that have started from a variable, but generally relatively few, number of founders and undergone large expansions. Examples of populations with these characteristics include Finland, Sardinia, Iceland, and Japan (Wright et al., 1999). Some extremely recent isolates exist as well, such as Costa Rica, regions of Quebec, Newfoundland, and the Hutterites, which were all founded less than 50 generations ago. In the case of the Hutterites, only 10-12 generations ago, an extremely small number of founders settled in the western US and Canada. At least one population exists, the Scandinavian Saami, which is a small population that has remained stable in size over a long period. The effects of genetic drift would be expected to be and are in fact extreme in this population (Laan and Paabo, 1997; Laan and Paabo, 1998; Terwilliger et al., 1998). The LD created by this drift may be quite useful in mapping studies. The whole range of histories of genetic isolates exists in the many populations of the world, and it is important to distinguish the utility of each of these populations for LD mapping. The more recent the isolate, the farther you would expect to see LD extend, and ,conversely, the older isolates would have a lesser extent of LD. Some of these populations may be young but founded by a larger number of individuals. In such cases, one would expect to see greater genetic heterogeneity but longer tracts of LD as well.

In Finland, greater homogeneity exists, but the extent of LD is less. This presents a dilemma in the practice of LD mapping. In young populations with higher levels of LD, it would be easier to initially detect an association if you were scanning the genome,

and fewer markers would be required. However, since the population is younger and there has been less time for historical recombination events, the limit of resolution in LD mapping would be much higher. In an older population, it would be possible perform much more fine structure mapping due to the lesser extent of LD, but initially locating the disease would be much more difficult. The ideal population would thus be somewhere in the middle, young enough to provide considerable, but not too much, LD. The history of Finland comes closest to providing this situation. However, a more preferable approach would be to take a two-tiered approach, using a younger population for the initial mapping and an older population for the fine scale mapping. The problem with this approach is that it is nearly impossible to find two populations fitting this description that would have enough cases of a disease. Furthermore, that disease would have to have the same etiology in both populations. This highlights another disadvantage of using isolated founder populations. Since the frequencies of some diseases are so drastically altered in some populations, it is questionable as to whether or not the findings would be applicable to other populations. Of course, in the world of human disease studies, any information about the mechanism of a disease is a significant achievment. An additional disadvantage of using founder populations is that, in many cases, the sample size of many diseases coming from some founder populations is simply not great enough to perform a mapping experiment.

The early experiments in LD mapping focused, by necessity, on rare and/or monogenic diseases in special populations using fairly mutable microsatellite repeat markers. Over time the focus shifted to nucleotide variation and more complex diseases in more general populations. This shift yielded a great deal of information but also present some difficult challenges.

Increasing Focus on Linkage Disequilibrium Studies

Focus on Human Nucleotide Variation

In the mid to late 1990's, the focus shifted to more general Caucasian and global populations. Studies done in isolated founder populations were considered the most crisp and well-controlled disease studies that existed, but most large patient populations had

been collected over a long period of time in more broad, mixed populations. In many isolated populations it was difficult to collect enough samples to study a range of common diseases. As the focus shifted more toward complex diseases, the ability to perform well-controlled LD studies in more general populations became necessary. In addition, many of the bottlenecks that gave rise to founder populations may not have been small enough to significantly reduce the heterogeneity of common disease mutations. The mutation exists on many chromosomal copies, so more copies of the mutation enter the population. Since the mutation would be introduced by multiple founders, the recombinational history of the variant extends back to its origin in the general population from which the founder population is derived (Kruglyak, 1999; Risch and Merikangas, 1996). Isolated founder populations, although extremely useful for rare monogenic disorders, may not have quite the same advantages of increased LD surrounding common disease variants. One recent study of common mutations on the X chromosome discovered a pattern of LD that was extremely uneven. Regions of extensive LD were found adjacent to regions with minimal LD, reminiscent of the data from *Drosophila* chromosome 4 (Taillon-Miller et al., 2000). Taillon-Miller et al. discovered that the levels of LD for these regions were virtually indistinguishable between the general Finnish and the CEPH populations, although it is not clear that this finding also applies to common disease chromosomes. Other results from isolated founder populations concerning long range LD around common alleles are contradictory as well (Arnason et al., 2000; Eaves et al., 2000; Mohlke et al., 2001). Nonetheless, isolated founder populations still retained a very important advantage of increased genetic homogeneity, but this factor remained largely ignored as attention was mainly diverted to more general population samples for further LD studies.

In the investigation of complex traits in a more general Caucasian or globally diverse sample, it is necessary to stress the importance of using markers that are less mutable. Microsatellite repeat markers may have been sufficient for younger populations, but the high degree of recurrent mutation at these marker loci makes them less useful for examining older populations. Presumably, a population that is more mixed in ethnic ancestry was founded earlier and has been subject to much more migration. This is one of the factors that catalyzed efforts in the discovery and use of single

nucleotide polymorphisms (SNPs). SNPs are more mutationally stable, much more prevalent throughout the genome, and easier to type in large numbers. Methods were being created to genotype SNPs without having to run gels. SNPs were beginning to be arrayed on chips, slides, typed by standard sequencing, and run through denaturing HPLC (dHPLC) columns. These same techniques were also being used to screen for new markers on a large-scale level.

Conclusions about LD came from experiments primarily focused on the collection of DNA sequence variation in the human genome. It became apparent that increasing numbers of polymorphisms were necessary to carry out LD experiments and association studies properly and to more completely understand the structure of the human genome. This was especially true in the case of mapping complex diseases in broad population samples. Unfortunately, there was still very little reliable information on the extent of LD and SNP resources were still sparse. LD studies depended on having enough variation to detect the patterns, and there was no indication that there was a dense enough set of markers to accomplish the task. Obtaining full genome sequence became a more feasible endeavor and efforts were focused on collecting the full repertoire of sequence variation in a population for particular genes. During this time period, most efforts turned to exhaustive searches for all nucleotide variation in and around specific genes that were selected as candidate genes for a particular disease. In this way, LD was examined and the disease was studied in the same experiment. Initiatives for creating extensive community based resources included a first generation SNP genetic map of the genome (Wang et al., 1998) and significant efforts to characterize and catalog variation in large numbers of genes in a search for putative causative mutations (Cargill et al., 1999; Halushka et al., 1999).

Experimental design for searching for associations and examining LD was fairly straightforward in genes: simply screen all available sequence, coding and noncoding, and look for associations of markers to disease and at LD between markers. However, it was difficult to design experiments solely to examine the extent and patterns of LD in a presumably neutral region in a non-disease sample. There was no real guiding information on how to design these kinds of experiments. Details were elusive, such as the density of markers to look at or even an order of magnitude on the range of physical

distance to examine. In addition, most known polymorphism was focused only in coding regions of genes, making it impossible to perform these experiments even if one knew how to design them. **Appendix I** of this thesis describes an attempt to gather polymorphisms and assess LD in a 12 kb region on the X chromosome. At the time there was no evidence to show that LD would be detectable further than 5-10 kb in a general Caucasian population. Results from simulations, given well-accepted assumptions about human demographic history, estimated that useful levels of LD might not extend past 5 kb (Kruglyak, 1999). Complete sequence data from the 12 kb region uncovered an extreme lack of sequence variation, not yielding even enough polymorphism to study LD patterns. As described earlier, the X chromosome has much lower polymorphism and recombination rates than the autosomes, and this feature may be due to selection, different population history, and smaller effective population size, there being only three X chromosomes to every four autosomes in the population. This study provided an initial clue about the extremely low polymorphism rates on the X chromosome in the Caucasian population. This lower polymorphism rate could be indicative of a greater extent of LD on the X chromosome as well. This issue is explored in **Chapter 5**. The results of this sequencing experiment on the X chromosome also highlighted the need for looking at multiple regions on many different chromosomes in order to obtain a more complete picture of both LD and polymorphism rates across the genome.

Gene Sequencing Studies

Most measures of LD during this time came from a number of single gene sequencing studies. Although these studies used disparate and not-well-characterized populations, the variation in LD between and within loci could not be ignored. The major gene studies of the time, lipoprotein lipase (LPL), ACE, β-globin, ApoE, and DMD are summarized in a review by Preworski et al. (Przeworski et al., 2000). In some cases, LD was detectable across many kb, and in others LD was negligible even at extremely close distances of hundreds of base pairs. Pairwise comparisons of LD levels were inconsistent within the same gene regions depending on the specific pair of markers that was examined and their allele frequencies.

A problem with drawing general conclusions from studies done in different groups was the inconsistencies in experimental design. Data was accumulated in various regions in different populations with different sample sizes, using different numbers of markers with different allele frequencies and calculated with different measures of LD. Still, these studies provided the best initial look at LD in genes, the results of which painted a hazy picture of how LD could be used to discover disease mutations.

It became clear that variability from region to region was going to be the rule rather than the exception. The disturbing finding of a low level of intragenic LD in the LPL gene provided an impetus to determine if this was a more general phenomenon. A survey of many regions across the genome was needed which used the same types of markers in the same population and possessed a consistent measure of significant LD. The problem was that there were not enough SNPs available across the genome to perform a large-scale survey of many regions. One resource that did exist was a first generation SNP genetic map. In the process of discovering SNPs, many small segments of sequence were obtained from STSs, which occasionally contained more than one SNP. This fortuitous resource provided many examples of SNP pairs scattered throughout the genome. **Chapter 3** describes an experiment that sought to utilize this resource to obtain a more general view of LD for different regions of the genome. Instead of discovering and examining many polymorphisms within a single gene or restricted genomic region, we sought to broaden the investigation to many loci each containing only two or a small number of SNPs. These SNP pairs were separated by very short distances, an average of 124 bp, and would not address the upper limits of LD. However, almost complete LD should be observed for markers spaced at 124 bp due to a lack of recombination over such a short interval, given the current accepted human population history model. The experiment was designed to assess variability across the genome at what would be expected to be complete LD. Nonetheless, the results revealed that a significant fraction of pairs showed incomplete LD. Since an extreme alteration of estimates of effective human population size would be necessary to conform to this data, we concluded instead that the discrepancy was due to gene conversion which increased the apparent rate of recombination between the closely spaced loci. This experiment thus defined the lower extent in distance of useful LD. In addition, the importance of giving consideration to

allele frequencies in calculating LD was illustrated. **Appendix II** shows that LD statistics can provide a biased estimate due to inadequate sampling. These results highlighted the need for an experiment that assayed many distance categories for a large collection of regions.

In order to identify the magnitude and extent of LD in the genome, the desirable data to acquire seemed to be a complete "LD curve" which would describe the average LD from many regions for a complete range of physical distances. From this curve, a threshold could perhaps be determined for detecting LD between variants and thus would describe the density of markers necessary to cover the genome at an adequate level to perform genome-wide association studies. One study attempted to construct such a curve for 19 genes distributed throughout the genome. Reich et al. identified common SNPs (common being defined in this case as rare allele frequency >0.35) by sequencing 32 individuals from a globally diverse population (Reich et al., 2001). Common polymorphisms are more likely to be present in many populations, although the frequencies of these polymorphisms are likely to vary substantially between populations. 2 kb regions were then resequenced at 5, 10, 20, 40, 80, and 160 kb from each of the core common SNPs in a CEPH Utah population and SNPs were identified. A subset of these newly discovered SNPs were genotyped in European and Nigerian populations. Pairwise comparisons of LD exhibited a considerable variation within and between gene regions, but for useful levels of LD the average value extended out to 60kb, an estimate that was much larger than previously thought based on other gene studies and simulations (Kruglyak, 1999). This result appeared to set a rather optimistic limit on the number of markers that would be needed to cover the genome at approximately 50,000. LD declined much more rapidly in the Nigerian sample, prompting the authors to postulate a narrow population bottleneck in the founding part of the European population.

Although that hypothesis is the subject of debate, this study as a whole advanced the field by providing a large amount of data that was needed to understand the average extent and distribution of pairwise LD across genomic regions. However, this study simultaneously underscored several problems with the current mode of thought on the concept of LD mapping. First, although this study provided a measure of the average physical distance over which LD is detectable, the extreme variation in LD does not

allow such a measure to be useful to mapping disease mutations in any specific region of the genome. LD in one region may be either much more significant or nonexistent due to the other forces described earlier such as gene conversion, uneven distribution of mutation and recombination, including hotspots, and complicated population history processes, which cause a breakdown in the relationship between LD and physical distance. Abecasis et al. estimated that only about 45% of variation in LD was due to physical distance, the remainder might be accounted for by the factors listed above (Abecasis et al., 2001). Therefore, perhaps the average extent of LD is not the measure that is really the most desirable. Second, as there are obvious population differences in the levels of LD, results from LD survey experiments will not necessarily be generalized to any study population, and LD would have to be assessed before mapping for each region individually in each population. Finally, LD is not consistent within regions for pairwise comparisons even in extremely well-controlled experiments. Different pairs of markers behave differently based on the individual characteristics of a particular SNP, indicating that SNP pairs are not the unit of variation that history has acted upon. Longer haplotypes give a much better description of the genomic structure in a region. As noted before, a pairwise combination of alleles can exist on multiple haplotypes (**Figure 3**). Therefore, these pairwise comparisons actually consider multiple historical events as the same event, which creates an inherent uncertainty in the results. The upshot from this whole period of LD experiments is that LD is extremely variable within and between regions and populations. Furthermore, pairwise marker comparisons are not the most ideal, but haplotypes are a much more important measure of the underlying genomic structure. The number and length of distinct multimarker haplotypes and how that differs from region to region would be the matter that dominated the most recent period of experiments concerning patterns of LD and their application to mapping of disease genes.

Quantifying Linkage Disequilibrium

Up to this point in this introduction, general terms such as the "levels" or "extent" of LD have been used. What are the measures that are used to quantitate these levels? LD is most commonly described for alleles of pairs of markers, and there are numerous measures that have been used over time. Most measures of disequilibrium calculate the

difference between the observed frequency of a two-locus haplotype and the expected frequency under the hypothesis of random segregation of the alleles. For two alleles at each of two loci, A and B, (A, a and B, b respectively), the frequency of each of the alleles at these two loci can be represented as P_A, P_B, P_a, and P_b, and the frequency of the haplotype which contains alleles A and B is P_{AB}. One of the earliest and simplest measures, D, is defined by $D=P_{AB}-P_A-P_B$. There are many variations to this measure of LD, mostly to avoid the dependence of D on allele frequency (Devlin and Risch, 1995). The most commonly used variation of D is D', in which D is divided by the maximum value that D can obtain with the existing allele frequencies at the two loci. If only three of the possible 4 haplotypes are present in the population, for example AB, Ab, and aB, then LD is said to be complete and D'=1. Different values of D'<1 have been used to indicate an acceptable magnitude of LD in various studies, providing difficulty in comparing most studies. There are many other measures, such as the commonly used r^2, which are reviewed elsewhere (Ardlie et al., 2002; Weiss and Clark, 2002). D' is the measure that is predominantly described throughout this thesis.

Focus on Haplotype Diversity and Recombination Patterns

In the last one to two years, there has been an explosion in interest in LD studies. Much of the previous attention to the patterns of LD was held in the realm of population geneticists, whose major focus was to characterize the demography of the human population in order to decipher migration patterns and bottlenecks. The continuing difficulty in cloning complex disease mutations has amplified the importance of LD mapping as an indispensable tool. Since resources have been somewhat slim in the past, much of the major work in understanding LD came from larger labs that had significant sequencing capabilities for SNP discovery. However, in the last couple of years, creation of a vast SNP resource, sequencing of the human genome, and mapping millions of SNPs onto the sequence have sparked a surge in interest and has allowed much more in-depth studies of LD to occur. An increased awareness in the relevant issues and problems of LD mapping has also contributed to a change in the kinds of studies that are done. Advances in genotyping technologies provided the ability to look at large numbers of SNPs in many individuals. Variability is still the key word in results from LD studies.

Credible reports of LD range from markers hundreds of kilobases apart to extremely weak LD at relatively short distances (Ardlie et al., 2002; Weiss and Clark, 2002). Study populations, while predominantly European derived, still vary considerably between studies. There are still debates over the most useful LD statistic or over what level constitutes useful or sufficient LD for disease mapping. However, a number of studies have indicated a significant paradigm shift in the conceptualization of LD mapping. While analysis has historically concentrated on individual markers, the most recent focus has been on the underlying haplotype structure for different regions throughout the genome.

As described before, individual polymorphisms each have unique attributes which cause them to be prone to inconsistent behavior, owing to distinct population histories. Acquiring haplotype data is a more laborious process. One must collect genotypes from a denser set of markers in a specific region, and these genotypes must be examined in families or trios **(Figure 5)** in order to assign phase to the specific alleles of the markers. The availability of a denser SNP map, and, in some cases, an extreme quantity of sequencing for SNP discovery, allowed for these experiments to be performed. The issues that emerged concerned the number and frequency of haplotypes, or haplotype diversity, present in a population in a specific region and the length over which these haplotypes extend. From a wave of publications, the pattern that began to emerge was that even in a general population sample, a limited set of common haplotypes could be discerned (Daly et al., 2001; Goldstein, 2001; Jeffreys et al., 2001; Johnson et al., 2001; Rioux et al., 2001). There may be a large number of more rare haplotypes, but these comprise a smaller percentage of the overall haplotype diversity. Therefore, the hazy and complicated picture that existed when only pairs of markers were investigated became simplified as a more ordered chromosomal organization began to be resolved. In addition to less intricate haplotype diversity, the patterns of LD appeared to be parsed into discrete blocks of haplotypes characterized by a high level of LD within the block and a breakdown of LD between blocks **(Figure 6)**.

Daly et al. revealed a genomic structure in a 500-kb expanse of chromosome 5q31 that consisted of 11 segments of high LD and extremely limited haplotype diversity interspersed with apparent clusters of recombination events. Each block only has 2 to 4

common haplotypes that account for greater than 90 percent of chromosomes. These haplotypes do not appear to be derived from each other by recombination. Exceedingly restricted intra-block recombination suggests that the haplotype block might be a meaningful unit of genomic structure. Greater haplotype diversity exists for regions which span the intervals containing putative hotspots of recombination (**Figure 6**). The size of the blocks in terms of physical distance and numbers of markers is variable. Block size spans between 3 and 92 kilobases and each block contains 5 or more common SNPs. This same pattern is also seen in 216 kb of the HLA region class II region (Jeffreys et al., 2001), and another study reveals a block pattern that appears to exist across all of chromosome 21 (Patil et al., 2001). Although LD is very strong across blocks in these studies, some degree of decay of LD can be seen within the blocks and some level of LD can be observed between blocks.

A New Model of Human Haplotype Structure

In the process of examining genomic structure in a more detailed way, at the level of dense haplotypes, the major questions surrounding LD mapping have evolved. The initial objective to define a numerical correlation between LD and physical distance and the determination of the number of markers necessary to cover the genome has been replaced. The new goal consists of an effort to define the borders of these haplotype blocks and to assess the overall haplotype diversity. If LD is complete across blocks of the genome, then only one or a few markers from each block may be chosen as surrogates or representatives of the entire block, thus reducing the complexity of LD mapping experiments. The challenge remains as to which markers to choose and how to define what constitutes a discrete block. If in fact there is no breakdown of LD across a block then the same problem would occur as seen in LD mapping in an extremely young founder population. The mapping of a disease mutation could not be resolved closer than the boundaries of a block. This could present a major challenge depending on the size of the block. In addition, for a mutation occurring on a common haplotype, it would be extremely difficult to discern between a set of apparently equivalent variants which one is the actual mutation.

A survey of many regions of the genome was necessary to determine if this punctate pattern of LD was a general phenomenon and to further characterize the properties of these blocks. **Chapter 4** of this thesis describes a study of the haplotype structure in 54 distinct regions of the genome. Thousands of SNPs were genotyped on a large sample of chromosomes from several populations. The block structure was evident in most regions and in all populations. Haplotype diversity appeared to be limited to an average of four to six common haplotypes which comprised the majority of diversity in the sample. The average size of a haplotype block was determined to be approximately 22 kb in Caucasian and Asian samples. Not surprisingly, the average block size, approximately 11 kb, was considerably smaller in African and African-American populations. This haplotype structure does seem to be a widespread phenomenon in the genome, verifying the results of Daly et al. and other studies. The implications of this study for future experiments are discussed in more detail in the Perspectives chapter, **Chapter 6**.

Linkage disequilibrium is an interesting phenomenon on a less practical level as well. LD mapping will become a pivotal component of mapping complex disease mutations. Although, how can more be learned about the forces that shaped the haplotype structure of the genome? **Chapter 4** and other studies provide valuable insight, but only describe a pattern of *historical* recombination events. It would be preferable to get a more complete picture of recombination by gaining direct evidence of recombination hotspots. The experiments in **Chapter 5** are designed to look at the shaping of a chromosomal haplotype by crossover events. A haplotype can be labeled with a unique event and thus frozen in time. The pattern of recombination can be observed thereafter by examining only the chromosomes that were affected by the unique event. Polymorphic Alu insertions provide a stable and unique event in history. By typing SNPs in a region surrounding the Alu insertion, it can be determined if specific crossover events occur in local hotspots or are distributed more evenly across the chromosome. In many ways, this is similar to a pulse-chase experiment that began many generations ago. The haplotype diversity and LD patterns can then be compared between the insertion chromosomes and the non-insertion chromosomes to observe how specific recombination patterns compare with the full historical recombination patterns.

Understanding the way recombination specifically shapes haplotype patterns is a small but important step towards understanding how LD is shaped in human populations.

Conclusion

A tremendous amount has been learned about linkage disequilibrium in a relatively short period of time. More importantly, a framework has been initially constructed to understand what kind of experiments should be performed and what were the relevant issues. The developing field has increasingly approached a clear understanding of the behavior and patterns of linkage disequilibrium as well as the forces that shape those patterns. The chapters of this thesis closely parallel the theoretical, experimental, and technological progress of the field over time. *Drosophila* geneticists were focused on the determinants of linkage disequilibrium in nature. In human studies, the application of linkage disequilibrium to mapping disease genes was the ultimate goal. A testament to the progress achieved in the field is that human genetics have returned to the origins of LD studies, rekindling an interest in the fundamental forces responsible for shaping linkage disequilibrium. It is the understanding of these fundamental forces that will ultimately lead to a complete evaluation of the structure of the genome and human populations.

References

Abecasis, G. R., Noguchi, E., Heinzmann, A., Traherne, J. A., Bhattacharyya, S., Leaves, N. I., Anderson, G. G., Zhang, Y., Lench, N. J., Carey, A., Cardon, L. R., Moffatt, M. F., and Cookson, W. O. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet *68*, 191-197.

Ardlie, K. G., Kruglyak, L., and Seielstad, M. (2002). Patterns of Linkage Disquilibrium in the Human Genome. Nature Reviews Genetics *In press*.

Arnason, E., Sigurgislason, H., and Benedikz, E. (2000). Genetic homogeneity of Icelanders: fact or fiction? Nat Genet *25*, 373-4.

Begun, D. J., and Aquadro, C. F. (1994). Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of drosophila: selection and geographic differentiation. Genetics *136*, 155-71.

Begun, D. J., and Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature *356*, 519-20.

Boehnke, M. (1994). Limits of resolution of genetic linkage studies: implications for the positional cloning of human disease genes. Am J Hum Genet *55*, 379-90.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalayanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q., and Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet *22*, 231-8.

Collins, F. S. (1995). Positional cloning moves from perditional to traditional. Nat Genet *9*, 347-50.

Collins, F. S., Guyer, M. S., and Charkravarti, A. (1997). Variations on a theme: cataloging human DNA sequence variation. Science *278*, 1580-1.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. Nat Genet *29*, 229-232.

Devlin, B., and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics *29*, 311-22.

Eaves, I. A., Merriman, T. R., Barber, R. A., Nutland, S., Tuomilehto-Wolf, E., Tuomilehto, J., Cucca, F., and Todd, J. A. (2000). The genetically isolated populations of Finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes [see comments]. Nat Genet *25*, 320-3.

Frisse, L., Hudson, R. R., Bartoszewicz, A., Wall, J. D., Donfack, J., and Di Rienzo, A. (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet *69*, 831-43.

Goldstein, D. B. (2001). Islands of linkage disequilibrium. Nat Genet *29*, 109-11.

Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., and Chakravarti, A. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet *22*, 239-47.

Hastbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., and Lander, E. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet *2*, 204-11.

Hastbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., and Lander, E. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland [published erratum appears in Nat Genet 1992 Dec;2(4):343]. Nat Genet *2*, 204-11.

Heyer, E., and Tremblay, M. (1995). Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. Am J Hum Genet *56*, 970-8.

Hoglund, P., Sistonen, P., Norio, R., Holmberg, C., Dimberg, A., Gustavson, K. H., de la Chapelle, A., and Kere, J. (1995). Fine mapping of the congenital chloride diarrhea gene by linkage disequilibrium. Am J Hum Genet *57*, 95-102.

Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet *29*, 217-22.

Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C., Clayton, D. G., and Todd, J. A. (2001). Haplotype tagging for the identification of common disease genes. Nat Genet *29*, 233-7.

Jorde, L. B., Watkins, W. S., Carlson, M., Groden, J., Albertsen, H., Thliveris, A., and Leppert, M. (1994). Linkage disequilibrium predicts physical distance in the adenomatous polyposis coli region. Am J Hum Genet *54*, 884-98.

Jorde, L. B., Watkins, W. S., Viskochil, D., O'Connell, P., and Ward, K. (1993). Linkage disequilibrium in the neurofibromatosis 1 (NF1) region: implications for gene mapping. Am J Hum Genet *53*, 1038-50.

Kere, J. (2001). Human population genetics: lessons from Finland. Annu Rev Genomics Hum Genet *2*, 103-28.

Kerem, B., Rommens, J. M., Buchanan, J. A., Markiewicz, D., Cox, T. K., Chakravarti, A., Buchwald, M., and Tsui, L. C. (1989). Identification of the cystic fibrosis gene: genetic analysis. Science *245*, 1073-80.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet *22*, 139-44.

Kruglyak, L., and Lander, E. S. (1995). High-resolution genetic mapping of complex traits. Am J Hum Genet *56*, 1212-23.

Laan, M., and Paabo, S. (1997). Demographic history and linkage disequilibrium in human populations. Nat Genet *17*, 435-8.

Laan, M., and Paabo, S. (1998). Mapping genes by drift-generated linkage disequilibrium. Am J Hum Genet *63*, 654-6.

Lander, E. S., and Schork, N. J. (1994). Genetic dissection of complex traits. Science *265*, 2037-48.

Langley, C. H., Lazzaro, B. P., Phillips, W., Heikkinen, E., and Braverman, J. M. (2000). Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the Drosophila melanogaster X chromosome. Genetics *156*, 1837-52.

Lehesjoki, A. E., Koskiniemi, M., Norio, R., Tirrito, S., Sistonen, P., Lander, E., and de la Chapelle, A. (1993). Localization of the EPM1 gene for progressive myoclonus epilepsy on chromosome 21: linkage disequilibrium allows high resolution mapping. Hum Mol Genet *2*, 1229-34.

Mitchison, H. M., O'Rawe, A. M., Taschner, P. E., Sandkuijl, L. A., Santavuori, P., de Vos, N., Breuning, M. H., Mole, S. E., Gardiner, R. M., and Jarvela, I. E. (1995). Batten disease gene, CLN3: linkage disequilibrium mapping in the Finnish population, and analysis of European haplotypes. Am J Hum Genet *56*, 654-62.

Mohlke, K. L., Lange, E. M., Valle, T. T., Ghosh, S., Magnuson, V. L., Silander, K., Watanabe, R. M., Chines, P. S., Bergman, R. N., Tuomilehto, J., Collins, F. S., and Boehnke, M. (2001). Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in Finns. Genome Res *11*, 1221-6.

Ober, C., Cox, N. J., Abney, M., Di Rienzo, A., Lander, E. S., Changyaleket, B., Gidley, H., Kurtz, B., Lee, J., Nance, M., Pettersson, A., Prescott, J., Richardson, A., Schlenker, E., Summerhill, E., Willadsen, S., and Parry, R. (1998). Genome-wide search for asthma susceptibility loci in a founder population. The Collaborative Study on the Genetics of Asthma. Hum Mol Genet *7*, 1393-8.

Parsch, J., Meiklejohn, C. D., and Hartl, D. L. (2001). Patterns of DNA sequence variation suggest the recent action of positive selection in the janus-ocnus region of Drosophila simulans. Genetics *159*, 647-57.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P., and Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science *294*, 1719-23.

Payseur, B. A., and Nachman, M. W. (2000). Microsatellite variation and recombination rate in the human genome. Genetics *156*, 1285-98.

Peterson, A. C., Di Rienzo, A., Lehesjoki, A. E., de la Chapelle, A., Slatkin, M., and Freimer, N. B. (1995). The distribution of linkage disequilibrium over anonymous genome regions. Hum Mol Genet *4*, 887-94.

Przeworski, M., Hudson, R. R., and Di Rienzo, A. (2000). Adjusting the focus on human variation. Trends Genet *16*, 296-302.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199-204.

Rioux, J. D., Daly, M. J., Silverberg, M. S., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S., Kulbokas, E. J., O'Leary, S., Winchester, E., Dewar, K., Green, T., Stone, V., Chow, C., Cohen, A., Langelier, D., Lapointe, G., Gaudet, D., Faith, J., Branco, N., Bull, S. B., McLeod, R. S., Griffiths, A. M., Bitton, A., Greenberg, G. R., Lander, E. S., Siminovitch, K. A., and Hudson, T. J. (2001). Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet *29*, 223-8.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516-7.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., and Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928-33.

Schaeffer, S. W., and Miller, E. L. (1993). Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of Drosophila pseudoobscura. Genetics *135*, 541-52.

Taillon-Miller, P., Bauer-Sardina, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. P., and Kwok, P. Y. (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28 [see comments]. Nat Genet *25*, 324-8.

Terwilliger, J. D., Zollner, S., Laan, M., and Paabo, S. (1998). Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. Hum Hered *48*, 138-54.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lander, E. S., and et al. (1998). Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome. Science *280*, 1077-82.

Wang, W., Thornton, K., Berry, A., and Long, M. (2002). Nucleotide variation along the Drosophila melanogaster fourth chromosome. Science *295*, 134-7.

Watkins, W. S., Zenger, R., O'Brien, E., Nyman, D., Eriksson, A. W., Renlund, M., and Jorde, L. B. (1994). Linkage disequilibrium patterns vary with chromosomal location: a case study from the von Willebrand factor region. Am J Hum Genet *55*, 348-55.

Weiss, K. M., and Clark, A. G. (2002). Linkage disequilibrium and the mapping of complex human traits. Trends Genet *18*, 19-24.

Wright, A. F., Carothers, A. D., and Pirastu, M. (1999). Population choice in mapping genes for complex diseases. Nat Genet *23*, 397-404.

Zapata, C., and Alvarez, G. (1993). On the detection of nonrandom associations between DNA polymorphisms in natural populations of Drosophila. Mol Biol Evol *10*, 823-41.

Figure 1 Decay of linkage disequilibrium around a mutation

LD mapping is based on a relatively simple concept. A mutation responsible for a trait arises on a specific chromosomal haplotype or "chromosome" in an ancestral population (the existence of a mutation on a single chromosome can also be due to the introduction of a single copy of the chromosome into a population bottleneck, as described in **Figure 4**). In this figure, the ancestral chromosome with the mutation is colored gray. All other chromosomes that existed at the time of the mutation are represented by one green chromosome. Crossover events occur between all the chromosomes over time in the population. Present-day individuals displaying the trait carry chromosomes that contain the mutation. Each of these chromosomes has a region surrounding the mutation that is descended from the original mutated "gray chromosome". The haplotype structure outside of this region is derived from any of the other chromosomes that existed in the population. All of these descendant chromosomes when examined together should define a region surrounding the mutation which consists of the overlap of the gray regions. This is the region which has the highest level of LD because the least number of crossover events have occurred in this region. The level of LD should decrease with increasing distance away from the mutation because a smaller proportion of mutated chromosomes will contain segments of the original gray haplotype farther away from the mutation. On chromosomes with the mutation in affected individuals, more crossover events will have occurred the farther away from the mutation one looks. A curve can be constructed that illustrates this decline of LD with distance. The curve will be different for various traits, regions of the genome, and populations, but the basic concept is the same. the mutation should lie nearby the region of highest LD.

Figure 1

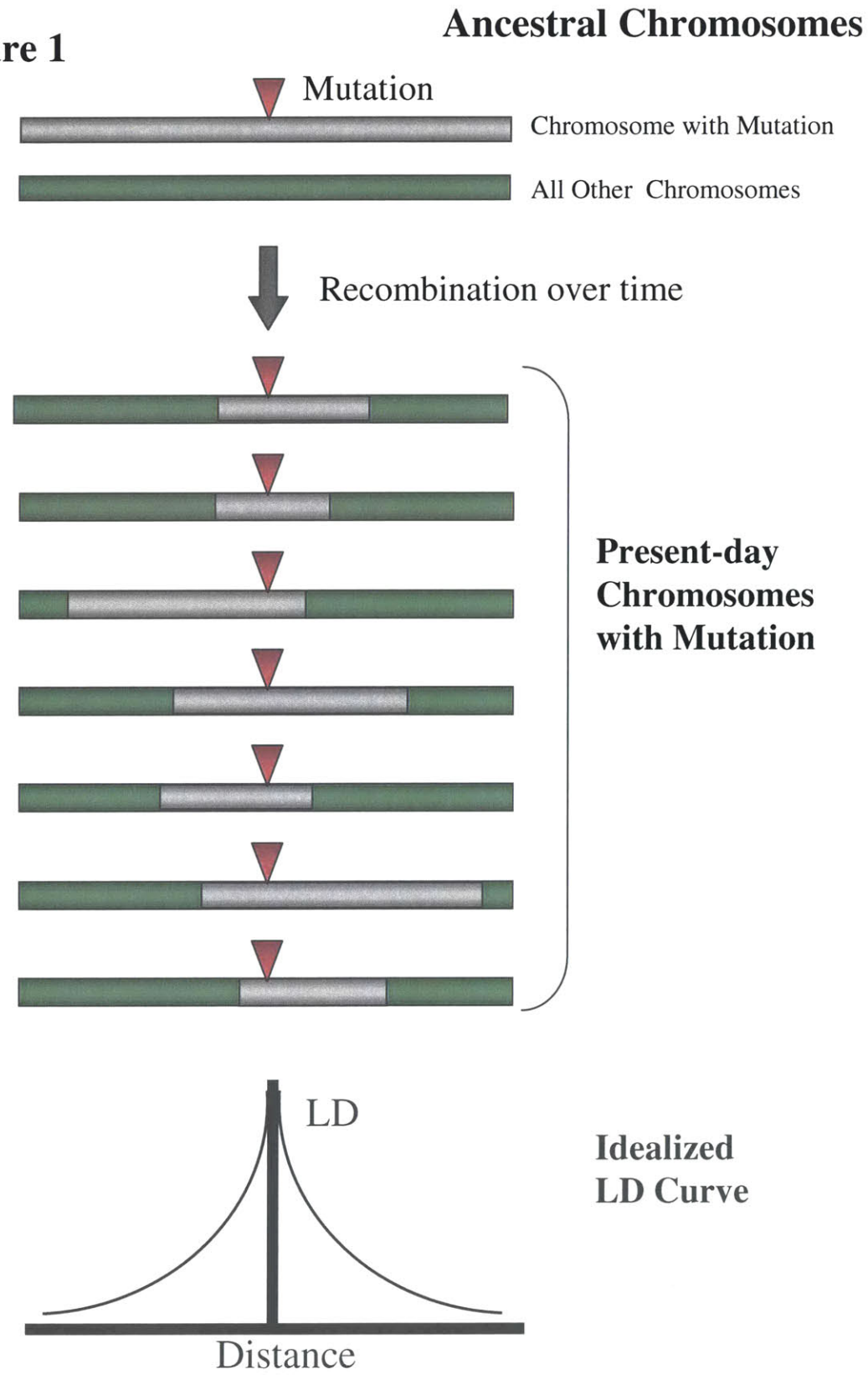

Ancestral Chromosomes
Mutation
Chromosome with Mutation
All Other Chromosomes
Recombination over time
Present-day Chromosomes with Mutation
LD
Distance
Idealized LD Curve

Figure 2 Linkage disequilibrium allows indirect associations between mutations and nearby markers

A) LD can be detected between closely-linked markers and a mutation

A disease or trait mutation is illustrated as a red triangle and is found within a gene represented by the green box. A direct association can be detected between the causative red triangle mutation and a phenotype. If this mutation has not yet been discovered, then an indirect association can be observed with nearby markers. These closely-linked markers may be used as surrogates for this mutation if crossover events have not disrupted the original combination of alleles of these markers and the mutation. Markers that are farther away will have a higher probability of crossover events occurring between them and the mutation and therefore LD is less likely to be detected with these more distant markers.

B) Linkage disequilibrium between pairs of markers

The decay of LD in the simple case of a pair of markers is illustrated by considering just the disease locus and one adjacent marker locus. The disease mutation (D) occurred on the chromosome containing allele 1 of the marker locus. At this point in time, all four possible haplotypes do not exist in the population. The mutation only occurs on the same chromosomal haplotype containing allele 1 of marker 1, allowing the observation of an association between the disease and allele 1. The orange, yellow, and blue chromosomes represent the three existing haplotypes in the population. If only three of the four haplotypes exist then a recombination event between the loci can create the fourth haplotype. The recombinant fourth haplotype is represented as the half yellow-half blue chromosome as it is derived by a recombination event between the disease locus and marker 1 that places allele 2 on the same chromosome as D. These recombination events also create a half yellow-half blue chromosome that consists of the non-disease allele (d) and allele 1 that is not illustrated because it already exists in the population as the orange haplotype. As the recombinant fourth chromosome increases in frequency in the population, LD between the markers will eventually decline and the association between alleles at the two markers will be disrupted.

Figure 2

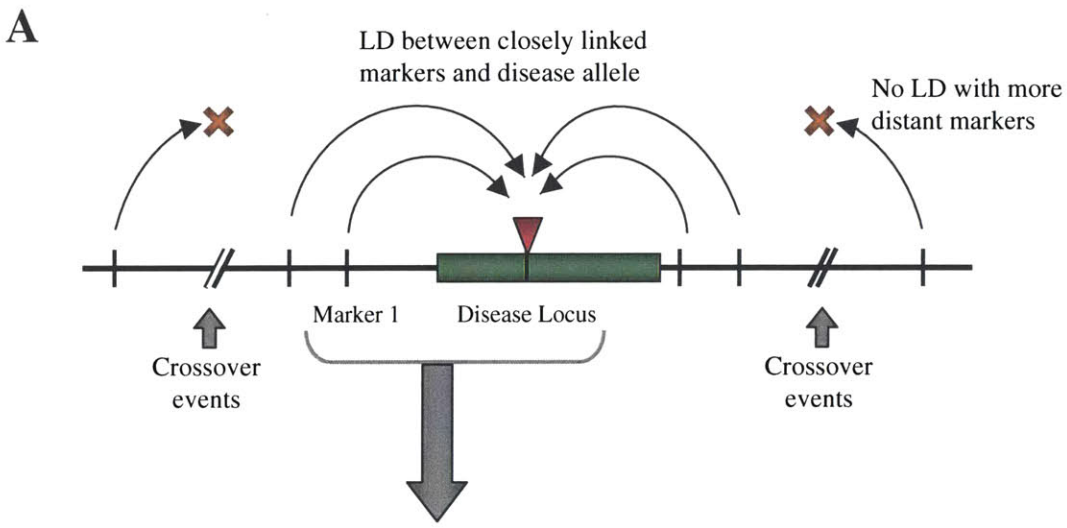

A
LD between closely linked markers and disease allele
No LD with more distant markers
Marker 1
Disease Locus
Crossover events
Crossover events
B
A Pair of Loci
4 possible haplotypes:
Marker 1
Disease Locus
1
d
3 haplotypes in existing population
1
D
2
d
Disease associated with allele 1
Recombination creates 4th haplotype over time
2
D

Figure 3

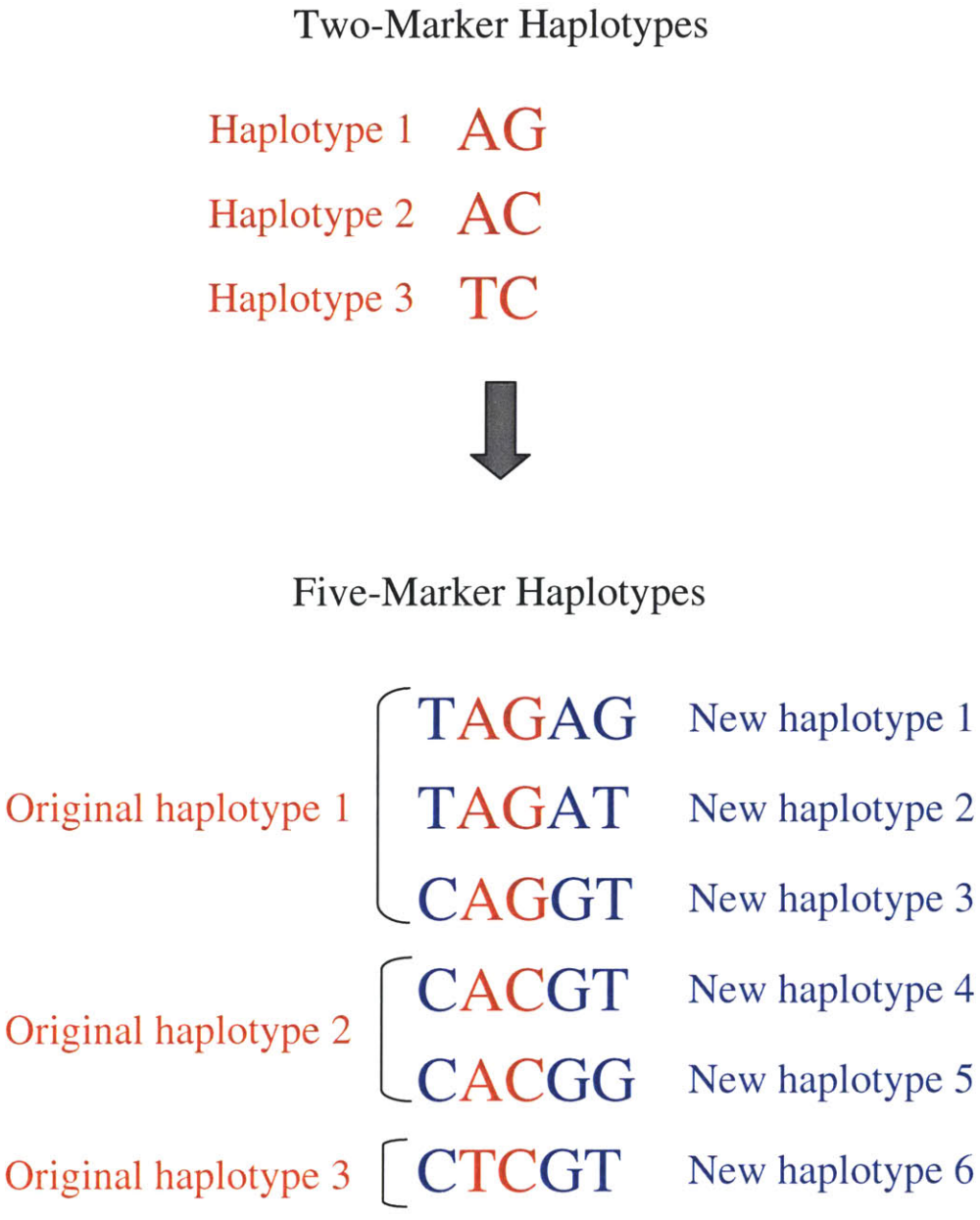

Two-Marker Haplotypes
Haplotype 1 AG
Haplotype 2 AC
Haplotype 3 TC
Five-Marker Haplotypes
Original haplotype 1
TAGAG New haplotype 1
TAGAT New haplotype 2
CAGGT New haplotype 3
Original haplotype 2
CACGT New haplotype 4
CACGG New haplotype 5
Original haplotype 3
CTCGT New haplotype 6

Figure 4 Genetic homogeneity in isolated founder populations created by bottlenecks

A small number of chromosomes are taken from an original source population to create a new and separate population. The original population is illustrated as a diverse population by the many different colors of the existing chromosomes in that population. In this example only four of these colors, representing four different haplotypes, enter the bottleneck. Genetic drift eliminates one of these four haplotypes represented by the chromosome with the red X. The longer the period of time that the bottleneck exists, the greater the chance there is of losing chromosomes by this process. In addition, the other three chromosomes recombine with each other over time. A recombination event is illustrated between the blue and red chromosomes. A population expansion can take place with these chromosomes as the only source genetic material for all of the descendants. The resultant population will therefore be much more genetically homogeneous than the original source population. Consequently, all chromosomes that exist in the present-day population are illustrated as some combination of the yellow, red and blue chromosomes. The amount of time over which the recombination events can occur will affect the amount of LD discernable in the present-day population. In reality, related individuals may enter a bottleneck together, thus carrying multiple copies of the same chromosome. It is thereby useful to describe the *effective* number of founders in a population based solely on the overall allelic spectrum instead of the actual number of individuals founding a population.

Figure 4

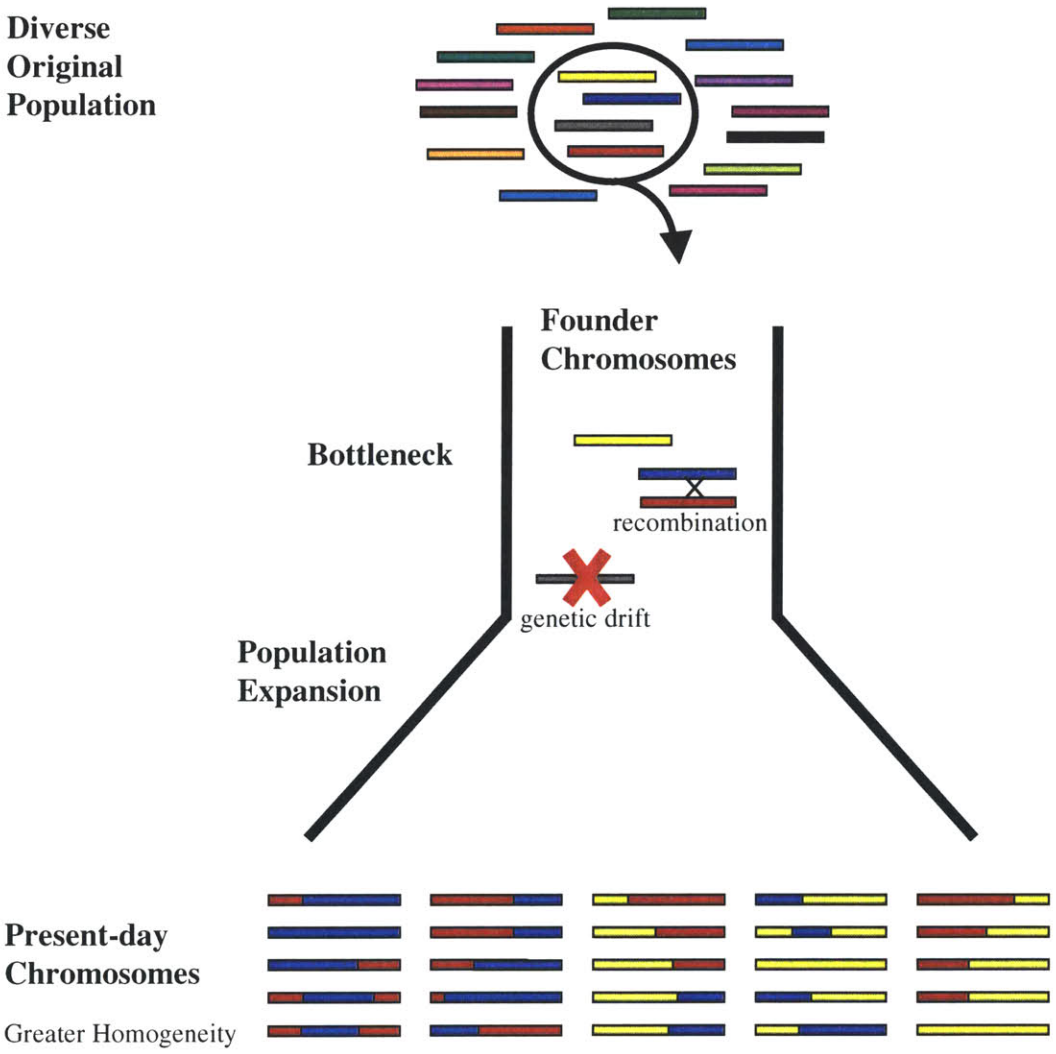

Diverse
Original
Population
Founder
Chromosomes
Bottleneck
recombination
genetic drift
Population
Expansion
Present-day
Chromosomes
Greater Homogeneity

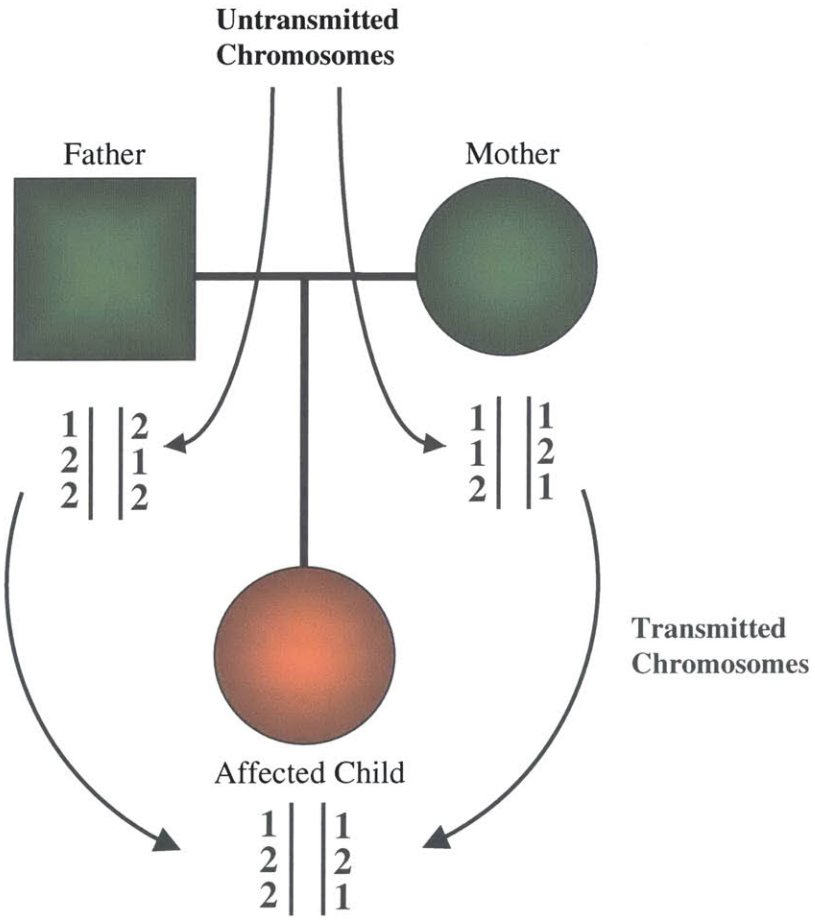
Figure 5 The use of father, mother, and child trios in haplotyping and association studies

In the trio described in this example, the mother and father (colored green) transmit a three marker haplotype to their affected child (colored orange). Each marker represents a SNP with the two alleles called 1 and 2. The mother transmits a 121 (alleles are listed for markers 1, 2, and 3, respectively) haplotype and the father transmits a 122 haplotype. The two chromosomes that are untransmitted are the 212 and 112 haplotypes. The mother and father will transmit a chromosome containing the disease-causing mutation more frequently than a chromosome not containing the mutation to an affected child. Therefore, the frequency of transmission of the different haplotypes can be examined. If a certain haplotype is transmitted significantly more frequently to an affected child, then that haplotype is associated with the disease and the mutation may lie nearby to the markers residing on that haplotype.

Another use of trios is to assign phase to marker combinations. In this example, haplotypes can be determined for the mother, father, and child for alleles of all three markers. This can only achieved by examining the genotypes of all three individuals at each marker and will not always be possible. In the case where the mother, father and child are all heterozygous for a marker, the phase will not be determinable. The child in this example is only heterozygous for marker 3. Since the father is homozygous for allele 2 of marker 3, the child's allele 2 must have come from the father and allele 1 must have come from the mother. We already know that allele 1 of marker 1 and allele 2 or marker 2 have to be on the same haplotype in both the mother and father because both parents transmitted both of those alleles to the child. Therefore, in the father, allele 2 of marker 3 must be on the same chromosome as allele 1 of marker 1 and allele 2 of marker 2. Likewise, in the mother, allele 1 on marker 3 must be with allele 1 of marker 1 and allele 2 of marker 2. This information reveals that the untransmitted haplotype must be 212 in the father and 112 in the mother.

Figure 5

Trios



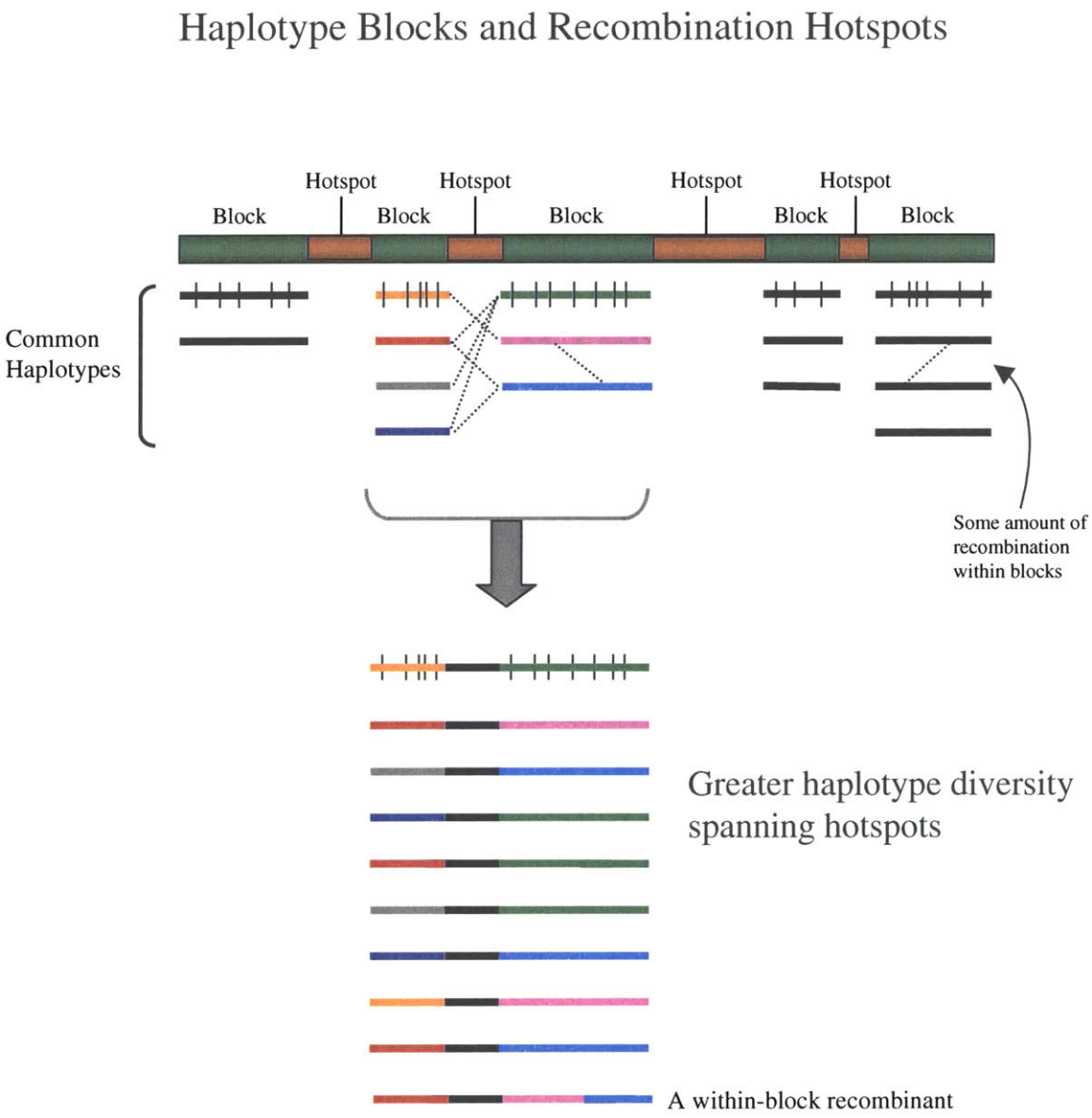
Example:

2 → marker 1 allele 2
1 → marker 2 allele 1
2 → marker 3 allele 2

Figure 6 Haplotype blocks, recombinational hotspots, and haplotype diversity

An idealized example of block structure is shown in this diagram. Blocks of markers that exhibit complete LD across all markers within the block are represented as green segments. Intervals which demonstrate possible clustered crossover events are labeled as hotspots, and are represented by brown segments. Both the blocks and the hotspots can be variable in physical distance. Variable numbers of SNPs are shown for each block, with no correlation between the numbers of SNPs and block size, illustrating the random distribution of SNPs in the genome. Within each block 2 or more common haplotypes are shown. A larger number of rare haplotypes can exist within each block. An example of haplotype diversity across blocks is illustrated for Block 2, Hotspot 2, and Block 3. Block 2 has the yellow, red, gray and blue haplotypes. Block 3 consists of the green, pink, and blue haplotypes. Haplotypes that span Hotspot 2 are represented by combinations of these colors corresponding to the specific crossover events that are shown to occur within the hotspot. The hotspot segment is represented by a black segment to illustrate that crossover events can occur anywhere within that segment. The crossover events are represented by dotted lines except for the haplotypes that are lined up directly across from each other across hotspot 2 (crossover events that give rise to those haplotypes are not shown). A crossover event within a block is shown between the pink and blue haplotypes. This type of event is less common, but does occur since all blocks do not always exhibit perfect LD across the whole block. A block definition or threshold needs to be delineated to describe the underlying haplotype structure in the human genome.

Figure 6


Haplotype Blocks and Recombination Hotspots
Hotspot
Hotspot
Hotspot
Hotspot
Block
Block
Block
Block
Block
Common
Haplotypes
Some amount of
recombination
within blocks
Greater haplotype diversity
spanning hotspots
A within-block recombinant

CHAPTER 2

Genetic analysis of the TCRB region in the Finnish population: Examination of association with multiple sclerosis and linkage disequilibrium

Shau Neen Liu-Cordero, Mark Daly, Mary Pat-Reeve-Daly, Tomi Pastinen, Patrick Charmley, Lee Rowen, Leroy Hood, Leena Peltonen and Eric S. Lander

Contributions: I was responsible for the experimental design and for genotyping of all markers except for the HLA typing, which was performed by Tomi Pastinen in the Leena Peltonen's group in Helsinki. The DNA samples were collected and prepared in Finland by Leena Peltonen's group as well. TCRB markers were primarily characterized by Pat Charmley in Lee Hood's group at UW Seattle, but I discovered and characterized several of the markers. The data processing and analysis and statistical analysis were performed by Mark Daly, Mary Pat Reeve-Daly and myself at the Whitehead Institute.

Preface

The experiments and analysis contained in this chapter were performed about 7 years ago. The results were never published, although a fairly complete manuscript existed. Reporting a body of work this old presents challenges. It would be extremely difficult to provide an introduction and discussion that justifies these experiments from the viewpoint of the current scientific atmosphere. I have chosen to present the work as it was in 1995 with the background and significance and discussion pertaining to the state of the field at the time when the experiments were conceived and completed. This study was initially designed as an association study to explore the role of the T-cell receptor beta locus in multiple sclerosis, but after the subsequent negative result, the data became quite useful for examining linkage disequilibrium in the Finnish population. The description of the rationale for these experiments is kept intact, as the study would be designed differently given what is currently known about TCRB and the etiology of multiple sclerosis as well as the advances in the tools and technology that have occurred in the intervening years. Although the technology and resources available at the time are now outdated, many of the ideas which led to the design of this study are now becoming widely recognized, accepted and applied. The inclusion of this body of work in this thesis is an appropriate measure to illustrate the development of ideas surrounding linkage disequilibrium. In order to address the historical significance of these results regarding all the studies that have taken place since that time, I have included a brief secondary discussion which reviews the contents of the chapter from a modern perspective using current knowledge.

Introduction

Multiple Sclerosis (MS) is a chronic inflammatory autoimmune disease characterized by T-cell mediated myelin destruction causing lesions in the white matter of the central nervous system. These lesions can result in impaired nerve conduction leading to a variety of debilitating symptoms such as motor paralysis, impaired vision, ataxia, and bowel or bladder dysfunction. MS is a disease with a complex etiology. The age of onset is extremely variable, most frequently ranging from 20 to 40 years. The prevalence is also notably variable, ranging from .01% to .2%, being highest in temperate regions and Caucasians. A genetic contribution to MS susceptibility is suggested by the high concordance rate in monozygotic twins (26%) compared with dizygotic twins (2.3%) and nontwin siblings (1.9%) (Bernard and Kerlero de Rosbo, 1992). However, the inheritance pattern is not consistent with the segregation of a single gene and is most likely multifactorial, heterogeneous, and incompletely penetrant. Genes coding for components of the immune system are the obvious candidate genes in MS. Typically, the first such genes tested in autoimmune diseases are those in the HLA region. The HLA region is one of the most polymorphic and well characterized regions of the genome, and many studies of the HLA region has provided the best insights into the etiology of autoimmune diseases. Many autoimmune diseases show varying degrees of increased relative risk with particular HLA alleles. An HLA association with MS has been well established in a variety of populations (Ebers and Sadovnick, 1994; Haegert and Marrosu, 1994; Oksenberg et al., 1993). However, the question still remains, are there other regions like HLA which are less well characterized yet may be just as important in the development of MS or autoimmune diseases in general?

Variation in the T-cell receptor has also long been suggested to be involved in a variety of autoimmune diseases including MS. However the evidence has been much less convincing than HLA (Hillert and Olerup, 1992; Oksenberg and Steinman, 1989; Ransohoff, 1992; Robinson and Kindt, 1992). One of the most common studies has been to examine the repertoire of T-cell receptors in T-cells which have been isolated from affected tissues instead of mapping inherited polymorphism at the chromosomal DNA level (Usuku et al., 1992). Although the major histocompatiblity complex has been established to be a genetic component of susceptibility to multiple sclerosis, the case for the T-cell receptor is less clear. For both the TCRA and TCRB

complexes there is a morass of conflicting association studies. A review of association studies carried out for TCRB and multiple sclerosis is given in **Table 1**. There are no less than fifteen studies claiming either positive or negative results for linkage or association using TCRB markers. There are at least five studies that represent each side of this issue (References provided in Table 1).

These conflicts are not surprising given the problems inherent in association studies. One major problem, one which affects any genetic analysis, is the lack of genetic polymorphism. Until recently there has not been much polymorphism discovered in the TCRB complex. Genetic markers have consisted almost exclusively of widely spaced and not very informative RFLPs and some other single base pair polymorphisms. Since associations are population based, case-control studies, sufficient resolution requires distinguishing most of the chromosomes in the population, not just the few chromosomes that may be segregating in a large pedigree. Therefore, more polymorphic and informative microsatellite markers can be especially preferable for association studies. In addition, since the same allele of a particular marker can be present on a number of different chromosomes in a population, it is also necessary to use multiple closely-spaced markers and construct haplotypes in order to gain the required resolution. None of the association studies carried out for TCRB and MS have used microsatellite markers, and the majority of studies were done with two or three RFLPs. When these multiple RFLPs were used, haplotypes were assigned mainly in individuals who were homozygous for one or two markers. The DNA samples that were obtained for some of the studies were selected from populations far too heterogeneous and ancient to make any claims about the whole complex, which is more than half a megabase in size, based on a single marker.

A better strategy in the mapping of complex traits is to use populations with reduced heterogeneity and increased levels of linkage disequilibrium. Isolated founder populations may possess these attributes. Finland is one of the best described isolated founder populations. Genetic studies in Finland benefit from a reduced genetic heterogeneity due to a relatively recent bottleneck, as well as a large population to draw from with a well described history **(Figure 1)**. The majority of the present day population of Finland, approximately 5 million inhabitants, is thought to have descended by a relatively small set of founders that migrated in to the southwestern region of Finland about 2000 years ago, or approximately 100 generations. The

size of the population remained relatively small, and there were some internal migrations north about 500 years ago (de la Chapelle, 1993). About 250 years ago there was a large population expansion, and a large portion of the population moved into towns and cities in the last 50 years. Therefore, in addition to the overall population being an excellent resource for LD mapping, there may also be some more recently created population substructure that could be utilized to provide even further homogeneity and longer tracts of LD. Finland has substantial church parish records allowing ancestries to be traced as far back as twelve generations. The founder effect and genetic drift are remarkably evident in the substantial differences in disease frequencies between Finland and the remainder of Europe. These disease frequencies may be higher or lower and are indicative of random drift processes.

A recent example of successful LD mapping for a rare, monogenic trait in Finland is diastrophic dysplasia (DTD) (Hastbacka et al., 1992). In Finland, 95% of individuals affected with DTD carry the same haplotype, which was present at only 3% on unaffected Finnish chromosomes. This demonstrated the striking degree of homogeneity that can exist in the Finnish population. In addition, the linkage mapping had only localized the mutation to approximately a megabase of DNA. By looking at many markers within this region, the likely region in which the mutation resided was reduced to around 60 kb by examining the many historical recombination events in the region. Although the critical region for diastrophic dysplasia could be refined to about 60 kb, detectable LD extended as far as a megabase. However, this was an extremely rare haplotype, and LD would not necessarily be expected to extend as far in a collection of chromosomes from the general Finnish population. The utility of the Finnish population has yet to be realized for complex traits. In addition, the haplotype diversity and extent of linkage disequilibrium has yet to be described for Finland as a whole as opposed to a limited set of chromosomes from a small number of rare disorders.

Population admixture poses the largest potential problem to association studies. Traditionally, association studies have consisted of a comparison of unrelated affected and unaffected individuals. However, if the unaffected control group is not matched well in ethnic ancestry, a spurious positive association may arise due to population admixture. In a mixed population, any allele that is very common in a specific ethnic subgroup in that population can show a spurious positive association with a specific trait that also occurs at a higher frequency in

that particular ethnic group. The best way to control for admixture in association studies is to use the genotypes of the parents of affected individuals to construct haplotypes of both the transmitted and untransmitted chromosomes. The untransmitted chromosomes should provide an internal unaffected control which is perfectly matched for ethnic ancestry. Neither the positive nor negative results from studies done without dealing with the problems described above lend a great deal of evidence for the involvement of TCRB in multiple sclerosis. The low significance levels in the positive studies and the lack of sufficient resolution in all studies outlined in **Table 1** point to the need for a more rigorous study using the better resources that are now available.

In this study, we have sought to overcome many of the problems inherent to association studies and provide one of the most rigorous association studies done to date. This study also represents the most thorough examination of the TCRB complex and multiple sclerosis. The entire 685-kilobase genomic sequence of the TCRB complex is now available (Rowen et al., 1996) and we have a dense set of microsatellite markers that cover the region (Charmley et al., 1995). We have also collected a large sample of MS patients from Finland. As has been described above, the Finnish population is one of the best populations for genetic studies, being a relatively young and isolated population. The relatively young age of the Finnish population allow the detection of linkage disequilibrium at relatively large distances, making it much more likely that we would detect an association with an anonymous microsatellite marker located at some distance from the actual mutation. In order to increase our ability to detect an association, we have also constructed haplotypes with multiple closely spaced markers in order to unambiguously resolve unique chromosomes. By using more informative markers, internal controls, and a young and homogeneous population, our results conclusively show that TCRB is not involved in multiple sclerosis in Finland and suggest that TCRB may not be as generally involved in autoimmune disease as the HLA region.

Subjects, Materials, and Methods

Population Samples

The population sample consists of 81 unaffected parents and affected child trios from Finland. Twenty-one of the trios are derived from multiplex MS families with 2-6 affected cases

per pedigree. These families are identical to those previously used in Finnish genetic mapping studies (Kuokkanen et al., 1996; Tienari et al., 1992; Tienari et al., 1993). These 21 pedigrees were selected on the basis of multiple affected members according to Poser's diagnostic criteria and included both nuclear families and extended pedigrees. The 21 trios taken from these families were chosen to be unrelated to each other so as to count as independent samples for the association study. Fourteen of these families originated from the province of Vaasa in Western Finland where the prevalence of MS is twice as high as elsewhere in Finland and where the familial occurrence of MS is increased up to 30% (Kinnunen et al., 1983; Wikstrom, 1975). The remainder of the trios consists of singleton families whose origins are restricted to any one location in Finland but are distributed throughout the whole country. These remaining trios included both unaffected parents and one child with MS who satisfied Poser's criteria for clinically definite or laboratory supported definite MS.

TCRB Sequence and Markers

Simple sequence repeats were discovered and characterized as described in Charmley et al., 1995. Markers were tested for polymorphism in a small panel of Finnish chromosome. Twelve of the microsatellite repeats discovered in the TCRB sequence were ultimately appropriate for use in our Finnish MS sample. Allele frequencies were also determined in a panel of CEPH chromosomes for allele frequency comparison in a more general European population.

PCR and Genotyping

Genomic DNA (20 ng) prepared from peripheral blood lymphocytes was amplified in 10 µl PCR reactions using 2 µM labeled primers. A single primer in each reaction was end-labeled with either γ-^{32}dATP or fluorescent dye (Research Genetics). PCR reactions were assembled using Packard MultiPROBE robotic pipetting stations, overlaid with 40 µl mineral oil, and run 16 microtiter-format plates at a time (1,536 reactions per run), on Intelligent Automation Systems TC1600 thermocyclers. Thirty cycles of PCR were performed (94°C for 30 s, 57°C for 30 s, and 72°C for 60 s) and 0.5-1.0 µl of each reaction were loaded onto denaturing 6% polyacrylamide gels and electrophoresed for up to 4 h in 1X TBE. Gels were run using ABI 377 or ABI373A

Sequencers (Perkin Elmer) for fluorescent detection; or on standard polyacrylamide gel apparatuses, and exposed to autoradiographic film for radioactive detection. Radioactive gels were manually scored in duplicate and fluorescent gels were processed using the ABI software GENESCAN/GENOTYPER (Perkin Elmer).

Results

TCRB markers, genotyping and haplotype construction

Nearly 700 kb of contiguous genomic sequence from the TCRB complex is available (genbank accession no. L36092) (Rowen et al., 1996). Possessing the complete genomic sequence of the region provides a number of significant advantages. First, the entire sequence can be searched for simple sequence repeats, which can then be characterized for their suitability as genetic markers. We have previously reported the characterization of a large number of markers from the TCRB region (Charmley et al., 1995). Second, with the complete sequence, the exact physical distances between the markers can also be determined. Using this information, we could choose a set of relatively evenly spaced markers at a density which we could be confident would allow us to detect an association. As an initial screen, we used a set of 12 markers at an average spacing of around 50 kb (**Figure 2**). The markers consisted of di-, tri-, and tetranucleotide repeats with heterozygosities ranging from .26 to .86.

We studied a sample of 81 unrelated individuals affected with MS along with their parents. Twenty-one of these trios are derived from multiplex families originating in the high-risk Vaasa province in Western Finland, and the remaining trios originate from throughout Finland where the risk of MS is similar to that of Europe as a whole. The majority of the affected individuals follow a relapsing/remitting progression of the disease, but several also have experienced a secondary progressive phase. Singleton families, which consist of unaffected parent-affected child trios, represent the sample composition of choice for an association study. Instead of examining the distribution of alleles of a single marker in affected and unaffected individuals, we obtained the genotypes of the parents to create haplotypes consisting of multiple markers. Each trio yields 4 unique chromosomes for any one marker or haplotype. The 81

parents and child sets were genotyped by a radioactive PCR assay to yield 324 haplotyped chromosomes, one half of which represents the affected, or transmitted, gametes, and the other half consists of the untransmitted parental gametes.

Lack of Association of Multiple Sclerosis with TCRB markers and haplotypes

We employed the transmission disequilibrium test (TDT) to search for any excess or deficiency in the segregation of certain alleles or haplotypes to the affected individuals as compared to the untransmitted chromosomes. The TDT is a test for LD and linkage between a marker and a disease susceptibility (Ewens and Spielman, 1995; Spielman et al., 1993). The TDT is performed by examining the transmission of marker alleles from parents heterozygous for the marker to affected offspring. In this study only one offspring is analyzed, but the test can be applied to multiple offspring as well. In fact, the TDT possesses the advantage of not requiring data on multiple affected family members or unaffected siblings. There must exist, however, an association due to LD in order to detect linkage between a marker and disease locus. As described before, the TDT is also a valid test of association and linkage even in the presence of population admixture and subdivision. The frequencies of alleles of each single marker in affected and unaffected chromosomes are shown in **Table 2** and the TDT test is applied. When we looked at alleles of single markers, there was only one allele with borderline significant transmission disequilibrium in affected individuals. However, this was not reproduced in any of the haplotype data described below.

Haplotypes consisting of multiple markers provide a more accurate basis for describing unique chromosomes in a population. A single marker will contain alleles that are present on a number of different chromosomes. As more markers are added to a haplotype the underlying chromosomal structure will be resolved more clearly. The ultimate data is the complete genomic sequence of a region, but it is not clear how many markers are needed to define all of the haplotypes in the region and provide the best surrogate for the complete sequence. A standard test of association was used for multiple-marker haplotypes. The untransmitted chromosomes are treated as control chromosomes in a case-control study (with the additional advantage of better matching of ethnic ancestry). The same result was obtained for comparing the frequencies

of haplotypes of multiple adjacent loci in affected and unaffected chromosomes. Haplotypes of different lengths were examined — for example, haplotypes of two, three or four adjacent markers. Since up to twelve markers were examined, windows of haplotypes of each length had to be considered. For example, for three-marker haplotypes, there are ten triplets as a sliding window moves along a set of twelve markers. The frequencies for all haplotypes for each triplet were examined in transmitted and untransmitted chromosomes. The data for the most frequent triplet haplotypes are shown in **Table 3**. For the longer haplotypes, e.g. haplotypes of 8 to 12 markers in length, there are fewer possible adjacent sets of consecutive markers. However, there were significantly more unique haplotypes in our data set for each of the longer haplotypes, and therefore, the sample size for any particular haplotype was not very large. For all haplotype lengths and windows, the frequencies of each haplotype in transmitted and untransmitted chromosomes revealed no significant associations (data not shown). In fact the frequencies in transmitted and untransmitted chromosomes were strikingly similar. This data along with the TDT results suggest that TCRB does not play a significant role in susceptibility to MS in Finland.

HLA association confirmed in Finnish MS sample

We also sought to confirm the previously established HLA associations in our patient sample. We obtained DRB1, DQA1, and DQB1 genotypes for the patient sample. We see highly significant transmission disequilibrium of the commonly seen MS associated haplotype (Wansen et al., 1997). These HLA data provide a positive result, which lends support to the negative result observed for TCRB.

Using molecular genotyping data can be used to subdivide the phenotype as well. It is a frequent practice in autoimmune diseases to stratify the sample by HLA types. A common problem in complex traits mapping is the definition of the phenotype. It is difficult to know at what level a certain phenotype can be further subcategorized to create a more homogeneous sample. Although this is a preferable situation, it may have a negative effect on obtaining a significant result by decreasing the sample size. In our sample, the majority of MS cases exhibit the relapsing/remitting form of the disease. If subdividing the phenotype in this way corresponds

to a genetic subdivision as well, then it should increase the likelihood of detecting an existing association in our sample. In order to maximize our chances of observing any association between MS and the TCRB complex or an interaction between HLA and TCRB, we split the TCRB genotypes by the major HLA subtypes. This HLA stratification had no effect significant effect on the transmission disequilibrium result without the subdivision **(data not shown)**.

Linkage Disequilibrium within the TCRB region

One concern is that the linkage disequilibrium would not extend far enough to detect a significant association between the disease mutation and any of the markers. Indeed this is one reason why a genome scan for linkage disequilibrium is not currently feasible even in the Finnish population. One would need to genotype individuals with thousands of markers in a population around 100 generations in age. In order to address this concern, we examined marker-to-marker linkage disequilibrium.

We detected highly significant linkage disequilibrium between pairwise combinations of alleles of different markers across entire region **(Table 4)**. Based on the transmission disequilibrium results above, this region is presumably not involved in MS in Finland. Therefore, we could combine the data for all transmitted and untransmitted chromosomes and treat it as a set of random unaffected chromosomes from the Finnish population. The p-values for linkage disequilibrium underestimates the actual significance level, for the number of degrees of freedom was set at the total number of combinations of alleles in pairwise comparisons of markers even though not all possible allele combinations existed in our data set. This result confirms that our collection of markers is sufficiently dense to study the entire TCRB region in Finland. It is not clear that this will be true for other populations.

In addition to pairwise combinations of alleles, we could see linkage disequilibrium with longer haplotypes. As in the association study, all windows of 3 to 12 markers were examined for common haplotypes and LD. There is no reliable statistic for calculating LD in haplotypes of three or more markers. However it is possible to calculate the effective number of distinct chromosomes in the population. This number is simply the inverse of the kinship coefficient which is given by the sum of the square of the allele (or haplotype) frequencies. When we

examined one region with the five most closely spaced markers, markers N through J, we found only 43 haplotypes of which only 3 major haplotypes account for over 50% of the chromosomes (**Table 5**). These five markers span a distance of over 50 kilobases. More than half of these haplotypes, 23, are present in only 1 copy and 7 are present in two copies. Only 12 remaining haplotypes are present in more than 2 individuals. Of these 12 haplotypes, several can even be derived by one recombination or mutation event from another common haplotype. Considering all 43 haplotypes, their frequencies translate into approximate number of 5.2 effective founders for the Finnish population for this genomic region. For a broad collection of chromosomes this is a significantly smaller number than would be expected. When the entire region is examined a much larger number of haplotypes is evident due, in part, to recombination, but significant linkage disequilibrium is still evident (**data not shown**). Therefore, not only is LD extensive enough in the population to detect an association with the existing density of markers, but this result also illustrates that the general Finnish population displays a high level of homogeneity, at least for this region of the genome.

Discussion

The involvement of TCRB in MS has long been the subject of debate. In this study we provided the most conclusive evidence on this subject. We have definitively shown that TCRB is not involved in MS in Finland. The frequency of MS in Finland is similar to that of the rest of Europe, and the same predisposing HLA haplotype can be found in Finland as in numerous other populations. So there is no reason to believe that the Finnish population differs significantly from other populations. However, the particularly high incidence of MS in the Vaasa region of Western Finland could be a cause for concern. We analyzed the genotypes for families from the Vaasa region and the rest of the sample separately from each other. This subdivision did not reveal a significant association in either set of data (**data not shown**). Therefore, a lack of association in this special region of Finland doesn't seem to have diluted the overall sample. It is also not likely that there are enough MS predisposing TCRB haplotypes in Finland to mask an association. However, locus heterogeneity would preclude our ability to detect an association if TCRB was only involved in a small percentage of MS cases in Finland, but we have shown that

it is certainly not a major locus and there is no compelling data from other studies to state otherwise.

There are a large number of studies claiming a positive association of multiple sclerosis with various markers within the TCRB complex. There are numerous drawbacks to the design of these studies, as well as the studies which claim to have seen no association, which are outlined in **Table 1**. All of these studies used RFLP markers or other single base-pair polymorphisms located in genes. Microsatellites are known to be more mutable than single nucleotide polymorphisms, mutation rates being on the order of 10^{-3} to 10^{-4} per generation. However, ancestral founder chromosomes in the Finnish population will be at most 100 generations old. Therefore, the mutation rate of the microsatellite markers will not significantly affect our analysis unless the chromosomes entered the population multiple times. Although RFLPs are less mutable than microsatellites, they are also less informative. A small number of these markers do not provide enough information to be able to resolve all the chromosomes in their data sets. Even with multiple RFLPs, if the parents of affected individuals are not utilized then the true haplotypes can not be discerned, and in essence only single marker tests are being performed. Seboun et al. was the only study to use the untransmitted alleles of the parents as controls, distinguishing these "family-normal" chromosomes from controls coming from unrelated healthy individuals (Seboun et al., 1989). However, in this study haplotypes were not examined even though the parents' genotypes were available. This same group later conducted a study using three RFLPs and the proper internal controls in which they failed to confirm the positive association. In all of the other positive associations, haplotypes were only constructed in the least informative situations, when there were single, double or triple homozygotes for two or three adjacent loci. When the parents of affected individuals are not available, the only way to unambiguously assign phase is to look at the homozygotes. Beall et al., 1993, had the advantage of utilizing six RFLPs across the variable region of TCRB, but only constructed haplotypes in single or double homozygotes for a maximum of two adjacent loci (Beall et al., 1993).

Another major drawback in the design of these studies is the choice of both the affected population and, more importantly, the control population. In populations that are too heterogeneous and too ancient, even a large number of markers covering the TCRB complex may not be enough to detect an association because recombination has occurred for a long enough

time to degrade any detectable linkage disequilibrium between a disease mutation and the markers. In older populations, it is not valid to make claims about the entire TCRB complex based on two or three RFLPs. With the use of Finland in this study we sought to diminish the problems of genetic heterogeneity and increase our chances of detecting an association due to the high level of LD present in this population. In addition, the MS sample that has been collected for this study primarily consists of patients with a relapsing/remitting form of the disease. This narrowing of the phenotype in our sample is also an attempt to diminish the effects of heterogeneity. In 12 out of 16 studies described in Table 1 the patient samples are either comprised of a mix of relapsing/remitting and chronic progressive forms of the disease or the phenotype of the patients is not described at all. Possessing an understanding of the phenotype is an extremely important aspect of mapping disease genes, especially one as complex as MS.

The choice of an improper control population can lead to spurious associations due to population admixture. Beall et al., 1989, used an unrelated Caucasian control group to compare against a Caucasian patient sample (Beall et al., 1989), but there was no information provided about the origin of those individuals. In a study that claimed to localize an MS mutation to a 175 kb region of TCRB, Beall et al., 1993, utilized 35% unrelated Caucasian CEPH parents, 45% unrelated blood donors at the NIH, and 20% unrelated blood donors in the Seattle area as control populations for a Caucasian affected sample (Beall et al., 1993). Martinez-Naves et al. used Caucasian-Spaniards for both the MS sample and the control population, but there can still exist population substratification in a group of individuals with a common ethnic ancestry (Martinez-Naves et al., 1993). In studies which either claim or deny association, the use of inappropriate control populations that may not be representative of the affected population is a serious concern for the validity of the results. For various research groups using different study populations, or even the same research group studying different samples, conflicting results are to be expected if the results are not properly controlled for ethnic admixture.

In addition to using perfectly matched internal controls, we have demonstrated that we can detect linkage disequilibrium over a distance that spans multiple markers, thus providing a control for our ability to detect an association. We would have likely detected linkage disequilibrium between the disease locus and one of the markers. This is due to the fact that one can detect linkage disequilibrium at a much greater distance on a disease chromosome as

opposed to a collection of normal chromosomes. A disease chromosome is descended from a small number of ancestral chromosomes whereas a collection of normal chromosomes represent a much larger number of meiotic events. Therefore, a set of disease chromosomes represents a much smaller portion of the total chromosomes in a population and should have a greater level of kinship between them. If there were many predisposing alleles of MS at the TCRB locus, it would hinder our ability to identify an association. However, the strong HLA association suggests that MS in Finland is not too heterogeneous, at least at one major locus, to preclude detection of associations. More to the point, the effective number of founders for multiple clusters of markers within TCRB is extremely low. The five-marker cluster described in Table 4 displays a strikingly high kinship coefficient. This is further evidence of the level of genetic homogeneity in our study population. Since we can detect both significant association to HLA and significant linkage disequilibrium between the TCRB markers, the negative result at TCRB is not likely to reflect a problem with our study population or an artifact of the experimental design.

The ability to see LD over a region as large as TCRB is a very interesting observation in its own right. It was unclear from studies of rare Finnish monogenic disorders whether or not long-range detection of LD would be possible in more common diseases or even in a collection of chromosomes from the general Finnish population. This initial data set suggests that a small number of relatively long haplotypes may exist around common disease mutations that don't have a prohibitively large number of disease causing alleles.

Experimental autoimmune encephalomyelitis (EAE) in mice and rats is generally considered to be a model for multiple sclerosis. In a recent study, Sundvall et al. carried out a genome scan for linkage in EAE mice and found 2 major non-MHC susceptibility loci (Sundvall et al., 1995). No evidence for involvement of TCR in EAE was found, confirming a previous EAE study. Although EAE is only a rodent model for the human disease, we believe that this result also lends support to our claim that TCRB is not involved in multiple sclerosis.

In summary, we hope to set new standards for conducting association studies. This includes the study of an appropriate population, the use of proper internal controls, the use of extensive genome sequence to discover a dense set of markers, and applying these markers to the construction of multimarker haplotypes. The data we have presented for multiple sclerosis and

TCRB illustrate the high significance and conclusive results that can be obtained by employing these strategies. As more of the genome sequence becomes available, more markers will be discovered, and more candidate regions can be analyzed in this manner for a large variety of diseases and other traits.

Discussion with Modern Perspective

Although these experiments were performed much before the rest of the work in this thesis, the ideas and study design contained in this chapter still represents a paradigm of rigor in the practice of association studies. Although the resources and standards of scale have changed dramatically over time, the goals described in this study still remain today.

Use of full sequence information to discover markers has been the goal of both individual efforts and large-scale public and private resource building in the intervening period since this study was performed. The markers that are used are now largely SNPs, because microsatellites are seen as too mutable to be useful for more broad populations and common diseases. Current thought is focused on the fact that mutations that were common in a population would enter a bottleneck in many copies unless the bottleneck was extremely small. It is argued that Finland's bottleneck may not have had a large effect on LD around common variants in the population. Interestingly that would mean that the results for LD and homogeneity that we observed in this study would not just be applicable to Finland, but to the larger European and Caucasian populations as well. While this further validates the results, it also raises interesting points about demographic history and the underlying haplotype structure in the human genome. Since many of the common chromosomes we were studying may have entered the population in multiple copies, the relationship between these haplotypes may go back much further in time than initial founding of the Finnish population. The highly mutable microsatellite markers would then have a greater number of generations over which to mutate. However, since we detect highly significant LD and an extreme degree of homogeneity for the TCRB region, especially among certain clusters of markers, it appears as if this positive result could be conservative because of the mutability of the markers.

Due to the prevailing thought on common variants entering bottlenecks, the utility of Finland as an advantageous population for the mapping of complex traits has been questioned. However, this doubt has been greatly exaggerated. Finland still has many advantages as a homogeneous population. First, the number of founder chromosomes carrying a disease allele is still likely to be small. Second, over a long period of time it has remained largely isolated to migrations. Third, there is a great deal of population substructure such that some bottlenecks within Finland are much more recent and may have been sufficiently small to restrict the

repertoire of common disease haplotype. One of these subpopulations, Kainuu, is discussed in the introduction.

The realization that the extreme LD and homogeneity observed in this study may apply to the more general Caucasian population is quite interesting. **Table 4** shows a plot of pairwise LD that exhibits a similar block-like structure to regions studied in **Chapter 5**. In addition, the distance over which at least one of these blocks extends, 50 kb, is similar to the distances observed in many regions of the genome. The haplotype diversity of three major haplotypes and many minor haplotypes shown in **Table 5** is also similar to that seen in a general Caucasian population. I believe that we were seeing the first evidence of haplotype blocks in this study, but we were not yet aware of the generality of the observation. The diversity in this region in Finland is probably less than in a general population, but we would not have predicted at the time of this study that the results would necessarily reflect a broader Caucasian population.

Also apparent from this study was the idea that studying single candidate genes was not the best way to perform association studies. Although TCRB was an extremely good candidate gene, it was a daunting prospect to proceed one region at a time until a positive result was obtained. The problem was that not enough polymorphism was available to perform truly comprehensive studies. Large initiatives to collect more polymorphisms were therefore undertaken.

The goals of this study as they were described are still the goals of modern association studies. The high standard described above for the design of the experiments still represents the best practices available. The use of many markers in creating haplotypes to define chromosomes in associations and LD studies has only recently become a widely practiced procedure. Although it is often difficult to collect a sufficient number of parents of affected individuals, the use of untransmitted chromosomes to control for admixture is still the best way to ensure the validity of results in association studies.

References

Beall, S. S., Biddison, W. E., McFarlin, D. E., McFarland, H. F., and Hood, L. E. (1993). Susceptibility for multiple sclerosis is determined, in part, by inheritance of a 175-kb region of the TcR V beta chain locus and HLA class II genes. J Neuroimmunol *45*, 53-60.

Beall, S. S., Concannon, P., Charmley, P., McFarland, H. F., Gatti, R. A., Hood, L. E., McFarlin, D. E., and Biddison, W. E. (1989). The germline repertoire of T cell receptor beta-chain genes in patients with chronic progressive multiple sclerosis. J Neuroimmunol *21*, 59-66.

Bernard, C. C., and Kerlero de Rosbo, N. (1992). Multiple sclerosis: an autoimmune disease of multifactorial etiology. Curr Opin Immunol *4*, 760-5.

Biddison, W. E., Beall, S. S., Concannon, P., Charmley, P., Gatti, R. A., Hood, L. E., McFarland, H. F., and McFarlin, D. E. (1989). The germline repertoire of T-cell receptor beta-chain genes in patients with multiple sclerosis. Res Immunol *140*, 212-5; discussion 245-8.

Briant, L., Avoustin, P., Clayton, J., McDermott, M., Clanet, M., and Cambon-Thomsen, A. (1993). Multiple sclerosis susceptibility: population and twin study of polymorphisms in the T-cell receptor beta and gamma genes region. French Group on Multiple Sclerosis. Autoimmunity *15*, 67-73.

Charmley, P., Beall, S. S., Concannon, P., Hood, L., and Gatti, R. A. (1991). Further localization of a multiple sclerosis susceptibility gene on chromosome 7q using a new T cell receptor beta-chain DNA polymorphism. J Neuroimmunol *32*, 231-40.

Charmley, P., Concannon, P., Hood, L., and Rowen, L. (1995). Frequency and polymorphism of simple sequence repeats in a contiguous 685-kb DNA sequence containing the human T-cell receptor beta-chain gene complex. Genomics *29*, 760-5.

Ciulla, T. A., Robinson, M. A., Seboun, E., Doolittle, T. H., Hayashi, T., Kindt, T. J., and Hauser, S. L. (1988). Molecular genotypes of the T-cell receptor beta chain in families with multiple sclerosis. Ann N Y Acad Sci *540*, 271-6.

de la Chapelle, A. (1993). Disease gene mapping in isolated human populations: the example of Finland. J Med Genet *30*, 857-65.

Ebers, G. C., and Sadovnick, A. D. (1994). The role of genetic factors in multiple sclerosis susceptibility. J Neuroimmunol *54*, 1-17.

Ewens, W. J., and Spielman, R. S. (1995). The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet *57*, 455-64.

Fugger, L., Sandberg-Wollheim, M., Morling, N., Ryder, L. P., and Svejgaard, A. (1990). The germline repertoire of T-cell receptor beta chain genes in patients with relapsing/remitting multiple sclerosis or optic neuritis. Immunogenetics *31*, 278-80.

Haegert, D. G., and Marrosu, M. G. (1994). Genetic susceptibility to multiple sclerosis. Ann Neurol *36 Suppl 2*, S204-10.

Hansen, T., Ronningen, K. S., Ploski, R., Kimura, A., and Thorsby, E. (1992). Coding region polymorphisms of human T-cell receptor V beta 6.9 and V beta 21.4. Scand J Immunol *36*, 285-90.

Hastbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., and Lander, E. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland [published erratum appears in Nat Genet 1992 Dec;2(4):343]. Nat Genet *2*, 204-11.

Hillert, J., Leng, C., and Olerup, O. (1991). No association with germline T cell receptor beta-chain gene alleles or haplotypes in Swedish patients with multiple sclerosis. J Neuroimmunol *32*, 141-7.

Hillert, J., and Olerup, O. (1992). Germ-line polymorphism of TCR genes and disease susceptibility--fact or hypothesis? Immunol Today *13*, 47-9.

Kinnunen, E., Wikstrom, J., Porras, J., and Palo, J. (1983). The epidemiology of multiple sclerosis in Finland: increase of prevalence and stability of foci in high-risk areas. Acta Neurol Scand *67*, 255-62.

Kuokkanen, S., Sundvall, M., Terwilliger, J. D., Tienari, P. J., Wikstrom, J., Holmdahl, R., Pettersson, U., and Peltonen, L. (1996). A putative vulnerability locus to multiple sclerosis maps to 5p14-p12 in a region syntenic to the murine locus Eae2. Nat Genet *13*, 477-80.

Lynch, S. G., Rose, J. W., Petajan, J. H., Stauffer, D., Kamerath, C., and Leppert, M. (1991). Discordance of T-cell receptor beta-chain genes in familial multiple sclerosis. Ann Neurol *30*, 402-10.

Martinez-Naves, E., Victoria-Gutierrez, M., Uria, D. F., and Lopez-Larrea, C. (1993). The germline repertoire of T cell receptor beta-chain genes in multiple sclerosis patients from Spain. J Neuroimmunol *47*, 9-13.

Oksenberg, J. R., Begovich, A. B., Erlich, H. A., and Steinman, L. (1993). Genetic factors in multiple sclerosis. Jama *270*, 2362-9.

Oksenberg, J. R., Gaiser, C. N., Cavalli-Sforza, L. L., and Steinman, L. (1988). Polymorphic markers of human T-cell receptor alpha and beta genes. Family studies and comparison of frequencies in healthy individuals and patients with multiple sclerosis and myasthenia gravis. Hum Immunol *22*, 111-21.

Oksenberg, J. R., and Steinman, L. (1989). The role of the MHC and T-cell receptor in susceptibility to multiple sclerosis. Curr Opin Immunol *2*, 619-21.

Ransohoff, R. M. (1992). T-cell receptor germline genes and multiple sclerosis susceptibility: an unfinished tale. Neurology *42*, 714-8.

Robinson, M. A., and Kindt, T. J. (1992). Linkage between T cell receptor genes and susceptibility to multiple sclerosis: a complex issue. Reg Immunol *4*, 274-83.

Rowen, L., Koop, B. F., and Hood, L. (1996). The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. Science *272*, 1755-62.

Seboun, E., Robinson, M. A., Doolittle, T. H., Ciulla, T. A., Kindt, T. J., and Hauser, S. L. (1989). A susceptibility locus for multiple sclerosis is linked to the T cell receptor beta chain complex. Cell *57*, 1095-100.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet *52*, 506-16.

Sundvall, M., Jirholt, J., Yang, H. T., Jansson, L., Engstrom, A., Pettersson, U., and Holmdahl, R. (1995). Identification of murine loci associated with susceptibility to chronic experimental autoimmune encephalomyelitis. Nat Genet *10*, 313-7.

Tienari, P. J., Salonen, O., Wikstrom, J., Valanne, L., and Palo, J. (1992). Familial multiple sclerosis: MRI findings in clinically affected and unaffected siblings. J Neurol Neurosurg Psychiatry *55*, 883-6.

Tienari, P. J., Wikstrom, J., Koskimies, S., Partanen, J., Palo, J., and Peltonen, L. (1993). Reappraisal of HLA in multiple sclerosis: close linkage in multiplex families. Eur J Hum Genet *1*, 257-68.

Usuku, K., Joshi, N., and Hauser, S. L. (1992). T-cell receptors: germline polymorphism and patterns of usage in demyelinating diseases. Crit Rev Immunol *11*, 381-93.

Vandevyver, C., Buyse, I., Philippaerts, L., Ghabanbasani, Z., Medaer, R., Carton, H., Cassiman, J. J., and Raus, J. (1994). HLA and T-cell receptor polymorphisms in Belgian multiple sclerosis patients: no evidence for disease association with the T-cell receptor. J Neuroimmunol *52*, 25-32.

Wansen, K., Pastinen, T., Kuokkanen, S., Wikstrom, J., Palo, J., Peltonen, L., and Tienari, P. J. (1997). Immune system genes in multiple sclerosis: genetic association and linkage analyses on TCR beta, IGH, IFN-gamma and IL-1ra/IL-1 beta loci. J Neuroimmunol *79*, 29-36.

Wei, S., Charmley, P., Birchfield, R. I., and Concannon, P. (1995). Human T-cell receptor V beta gene polymorphism and multiple sclerosis. Am J Hum Genet *56*, 963-9.

Wikstrom, J. (1975). Studies on the clustering of multiple sclerosis in Finland II: microepidemiology in one high-risk county with special reference to familial cases. Acta Neurol Scand *51*, 173-83.

Table 1 Summary of conflicting TCRB and MS association studies

Comprehensive synopsis of existing studies TCRB studies. Positive associations on linkage results are shown in a), and negative results are listed in b). The specific markers that were used are included as well as haplotyping status. If parents were not used then multiple marker haplotypes could only be obtained for single or double homozygotes. Details of the MS sample and control samples are also given. A relapsing/remitting course is abbreviated by R/R, and a chronic progressive course is abbreviated as CP. For several studies the MS sample was subdivided by a major HLA subtypes, and this stratified sample was further examined for TCRB association.

Figure 1 Finland as a model genetic system

Details of the historical settlement patterns of Finland are outlined.

Figure 2 Microsatellite repeat markers in the TCRB region

The markers used in this studied are located throughout the more than 600 kilobases (kb) of the TCR beta region. The repeated sequence is listed above the markers indicated in red, as are the heterozygosities of these markers in a CEPH population sample.

Table 2 Transmission disequilibrium for TCRB markers

Transmission disequilibrium is listed for each of the alleles of the 12 markers. There are an average of 8.5 alleles for each microsatellite marker. The numbers of transmitted and untransmitted chromosomes are given, as are the chi-squared values for each allele and the overall chi-squared for each marker.

Table 3 Three-marker haplotypes in transmitted and untransmitted chromosomes

Haplotypes for each window of three markers were constructed for affected and unaffected chromosomes. The marker names are listed in the first column and the alleles of these markers for each common haplotype. In this example common is defined as greater than or equal to three occurrences of the haplotype.

Table 4 Linkage disequilibrium for pairs of markers

The log of the p-value is given for each pairwise comparison of markers. The highest negative numbers are therefore the most significant. There is a major block of markers that are in very significant LD between markers D and G.

Table 5 Low haplotype diversity reveals a small number of effective founders in the Finnish population

Haplotypes for the seventh window of 5 markers are listed. The two major haplotypes are 44647, and 79434, which together comprise more than 50% of the total chromosomes. The kinship coefficient (CK) is provided for each haplotype, the total of all haplotypes, the 2 most frequent haplotypes and the 12 most frequent haplotypes. The effective number of founders is listed for the all haplotypes combined, 5.23, the 2 most frequent haplotypes, 5.59, and the 12 most frequent haplotypes, 5.27.

Table 1a) Summary of conflicting TCRB and MS association studies : Positive Results

AUTHORS	Claim Assoc	Claim Linkage	Markers	Haplotyped?	Type of MS	MS sample	Controls	HLA stratified?
Seboun 1989	+	+	1 constant region marker and 1 variable region marker	No - even though they had the parents' genotypes	R/R	17 familes with one affected child and 34 multiplex families	Untransmitted Chromosomes	NO
Martinez-Naves et al. 1993	+	N/A	Vb8 RFLP; Vb11 RFLP; Cb RFLP	Only for double or triple homozygotes: maximum of three adjacent loci	both CP and R/R	97 unrelated Caucasian Spaniards from Northern Spain - 11 with CP MS and 86 with R/R MS	93 unrelated Caucasian Spaniards	NO
Charmley et al. 1991 (Hood)	+	N/A	Markers from Beall et al. 1989 + an additional marker: Vb15 RFLP	Only for single or double homozygotes: maximum of two adjacent loci - controls were fully haplotyped	CP	Same sample as Beall et al. 1989	unrelated CEPH individuals	See Beall et al. 1989
Beall et al. 1989/ Biddison et al. 1989 (Hood)	+	N/A	Vb8 RFLP; Vb11 RFLP; Cb RFLP	Only for double or triple homozygotes: maximum of three adjacent loci	CP	40 Caucasians	100 Caucasians; an additional 43 DR2+ Caucasians	84% of MS patients DR2+
Beall et al. 1993 (Hood)	+	N/A	6 RFLP's across 600 kb of variable region	Only for single or double homozygotes: maximum of two adjacent loci	95% CP	83 Caucasian	197 Caucasians: 35% unrelated CEPH; 45% unrelated blood donors (NIH); 20% blood donors from Seattle	51 DR2+/ 32 DR2-, normal individuals not HLA-typed

Table 1b) Summary of conflicting TCRB and MS associations : Negative results

AUTHORS	Claim Assoc	Claim Linkage	Markers	Haplotyped?	Type of MS	MS sample	Controls	HLA stratified?
Fugger et al. 1990	-	N/A	Same as Beall et al. 1989	Only for double or triple homozygotes: maximum of three adjacent loci	R/R and ON	37 /R/R patients; 20 monosymptomatic optic neuritis patients. All patients from University Hospital, Lund, Sweden	99 unselected, unrelated healthy Danes	78% of MS patients DR2+; 60% of ON patients DR2+; normal individuals not HLA-typed
Hillert et al. 1991	-	N/A	Vb8 RFLP; Vb11 RFLP; Cb RFLP	Only for double or triple homozygotes: maximum of three adjacent loci	both CP and R/R	100 unrelated Swedish patients: 23 CP and 77 R/R - some of R/R had a secondary chronic progressive phase	100 unrelated healthy persons - origin not given	By DR2; DR4, DQw8(for CP); DRw17, DQw2(for R/R); normal individuals also HLA-typed
Hansen 1992	-	N/A	2 Vb coding region polymorphisms	NO	not reported	77 Norwegian patients? see Spurkland, A., et al. 1991	200 Norwegian healthy individuals?	NO
Vandevyver 1994	-	N/A	Vb8 RFLP; Vb11 RFLP; Cb RFLP	Only for single, double or triple homozygotes: maximum of three adjacent loci	CP	71 unrelated Belgian patients	67 randomly selected healthy individuals of Euro-Caucasian descent	DR2+/ DR2-
Oksenberg et al. 1988	-	N/A	2 RFLP's not mapped	NO	not reported	28 unrelated Caucasian patients from Neurology Cinics at Stanford	70 normal unrelated Caucasian individuals	NO
Briant 1993	borderline for Vb and Vb/Cb haplotype; - for Cb alone	N/A	1 Cb RFLP and 1 Vb RFLP	Only for single or double homozygotes: maximum of two adjacent loci	not reported	48 pairs of monozygotic and dizygotic twins - at least one of each pair affected; 63 additional unrelated patients	Same control group as for Beall et al. 1989	NO
Ciulla et al. 1988 (Seboun)	-	N/A	3 RFLP's - one constant region and 2 variable region	Legitimate haplotypes of 3? see text	not reported	27 patients	Untransmitted chromosomes. And 10 ethnically diverse control families?	DR2+/ DR2-
Peltonen Lab 1995	-	N/A	1 CA-repeat and 3 minisequencing markers	NO	both CP and R/R	22 multiplex families from Vaasa	Random Finnish Individuals	NO
Wei, Charmley et al. 1995	-	N/A	14 TCRBV-gene polymorphisms	NO	both CP and R/R	48 patients seen at the Virginia Mason Clinic: 33 R/R; 9 CP; 6 unknown	Healthy unrelated employees from the medical center	DR2+/ DR2- for both MS patients and controls
Lynch et al. 1991	N/A	-	2 RFLP's from J and C region	YES	not reported	96 individuals from 14 multiplex familes - 5 members abnormal MRI	N/A	NO

Figure 1

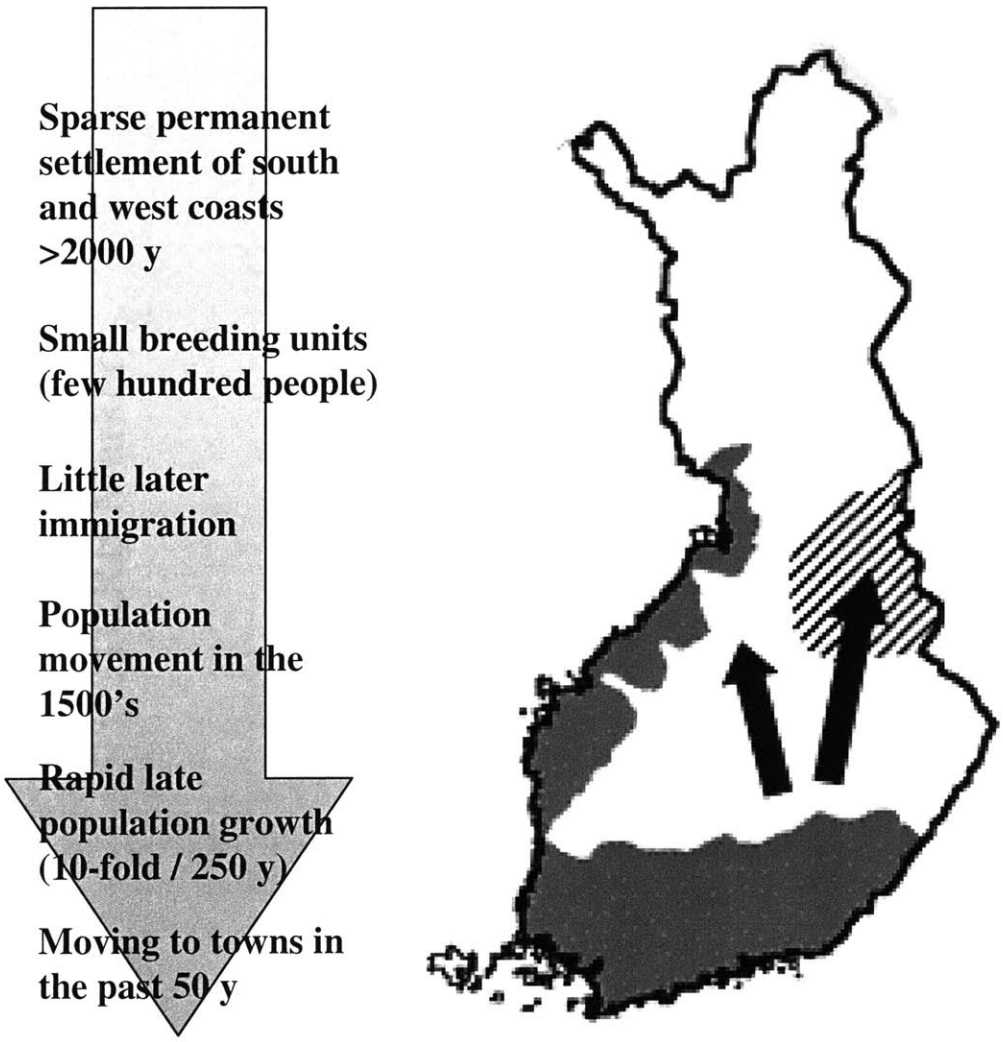

Sparse permanent settlement of south and west coasts >2000 y
Small breeding units (few hundred people)
Little later immigration
Population movement in the 1500's
Rapid late population growth (10-fold / 250 y)
Moving to towns in the past 50 y

Figure 2

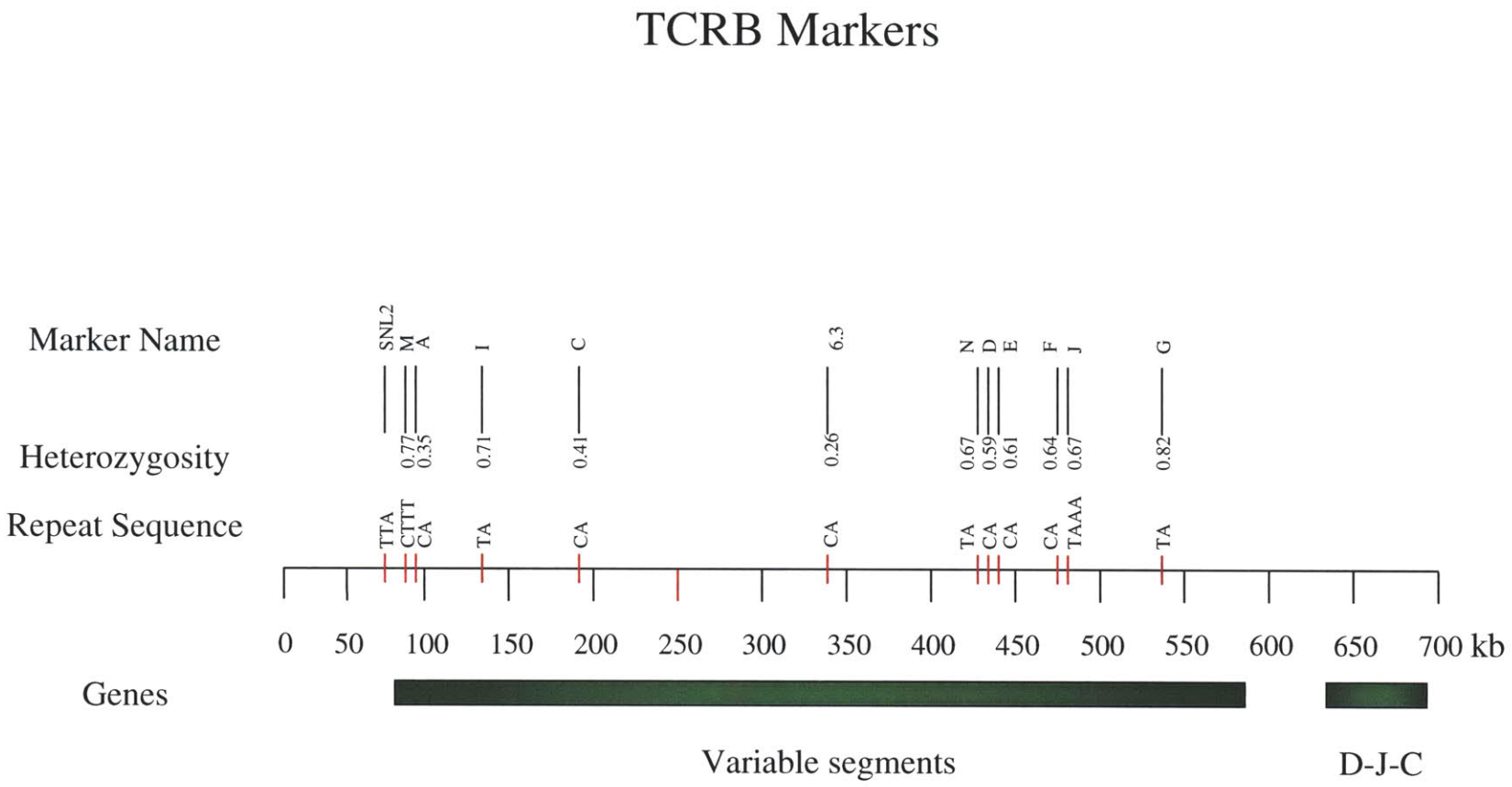

TCRB Markers
Marker Name
Heterozygosity
Repeat Sequence
SNL2
M
A
I
C
6.3
N
D
E
F
J
G
0.77
0.35
0.71
0.41
0.26
0.67
0.59
0.61
0.64
0.67
0.82
TTA
CTTT
CA
TA
CA
CA
TA
CA
CA
CA
TAAA
TA
0
50
100
150
200
250
300
350
400
450
500
550
600
650
700 kb
Genes
Variable segments
D-J-C

Table 2) Transmission disequilibrium for TCRB markers

Marker SNL2

Total X^2 = 8.09 (6df) P= 0.23

Allele	Trans	Untrans	Chi-sq
2	1	2	0.333
3	1	0	1.000
4	5	2	1.286
5	8	9	0.059
6	32	46	2.513
7	51	38	1.899
8	0	1	1.000

Marker M

Total X^2 = 6.95 (12df) P= 0.71

Allele	Trans	Untran	Chi-sq
1	7	5	0.333
2	0	1	1.000
5	4	4	0.000
8	4	1	1.800
12	19	22	0.220
15	9	12	0.429
16	2	0	2.000
18	16	14	0.133
19	16	12	0.571
20	4	6	0.400
22	31	33	0.063
23	0	2	2.000
26	2	2	0.000

Marker A

Total X^2 = 14.41 (6df) P= 0.025

Allele	Trans	Untran	Chi-sq
2	3	4	0.143
3	2	3	0.200
4	31	23	1.185
5	8	24	8.000
6	9	3	3.000
7	4	1	1.800
8	7	6	0.077

Marker I

Total X^2 = 6.33 (9df) P= 0.71

Allele	Trans	Untran	Chi-sq
3	0	1	1.000
5	2	1	0.333
7	4	3	0.143
9	2	3	0.200
11	4	10	2.571
13	22	17	0.641
15	3	6	1.000
17	2	1	0.333
23	26	26	0.000
25	42	39	0.111

Marker C

Total X^2 = 10.13 (10df) P= 0.43

Allele	Trans	Untran	Chi-sq
0	0	1	1.000
1	1	1	0.000
3	6	9	0.600
5	21	21	0.000
7	7	2	2.778
8	2	8	3.600
9	9	6	0.600
10	35	36	0.014
11	23	20	0.209
12	0	1	1.000
13	2	1	0.333

Marker v.6.3

Total X^2 = 10.22 (4df) P= 0.04

Allele	Trans	Untran	Chi-sq
1	14	4	5.556
2	5	5	0.000
5	8	16	2.667
6	0	1	1.000
7	0	1	1.000
7	0	1	1.000

Marker N

Total X^2 = 7.05 (6df) P= 0.32

Allele	Trans	Untran	Chi-sq
2	1	4	1.800
3	13	10	0.391
4	33	46	2.139
5	1	0	1.000
6	11	8	0.474
7	37	28	1.246
8	1	1	0.000

Marker D

Total X^2 = 7.79 (5df) P= 0.17

Allele	Trans	Untran	Chi-sq
2	2	1	0.333
3	1	0	1.000
4	25	38	2.683
6	12	12	0.000
7	0	1	1.000
9	32	20	2.769

Marker E

Total X^2 = 6.13 (5df) P= 0.19

Allele	Trans	Untran	Chi-sq
3	6	4	0.400
4	34	25	1.373
5	10	6	1.000
6	26	41	3.358
7	1	1	0.000

Marker F

Total X^2 = 3.48 (3df) P= 0.32

Allele	Trans	Untran	Chi-sq
2	2	3	0.200
3	34	24	1.724
4	28	38	1.525
5	13	12	0.040

Marker J

Total X^2 = 8.29 (8df) P= 0.41

Allele	Trans	Untran	Chi-sq
1	0	1	1.000
2	2	4	0.667
3	9	8	0.059
4	23	15	1.684
5	11	8	0.474
6	4	11	3.267
7	31	34	0.138
9	1	0	1.000
10	3	3	0.000

Marker G

Total X^2 = 18.95 (18df) P= 0.39

Allele	Trans	Untran	Chi-sq
5	0	1	1.000
7	2	5	1.286
8	1	0	1.000
9	4	2	0.667
10	3	5	0.500
11	4	7	0.818
12	4	1	1.800
13	11	12	0.043
14	8	16	2.667
15	19	17	0.111
16	21	25	0.348
17	4	3	0.143
18	1	4	1.800
19	2	3	0.200
20	8	6	0.286
21	3	0	3.000
22	13	13	0.000
25	1	0	1.000
26	32	21	2.283

Table 3)

Three-marker haplotypes in transmitted and untransmitted chromosomes

	Three Marker Haplotype	Affected chromosomes	Unaffected chromosomes	Difference	X^2
TRIPLE 1 (SNL2-M-A)	6 22 4	28	25	3	0.2
	7 12 4	20	15	5	0.7
	7 18 4	15	11	4	0.6
	7 19 4	15	5	10	5
TRIPLE 2 (M-A-I)	12 4 13	10	10	0	0
	12 4 23	5	5	0	0
	15 4 25	4	7	-3	0.8
	18 4 25	12	13	-1	0
	19 4 25	12	5	7	2.9
	22 4 23	10	11	-1	0
	22 4 25	21	18	3	0.2
TRIPLE 3 (A-I-C)	4 13 11	4	5	-1	0.1
	4 13 5	13	11	2	0.2
	4 23 10	5	5	0	0
	4 23 3	7	6	1	0.1
	4 23 5	7	5	2	0.3
	4 25 10	41	42	-1	0
TRIPLE 4 (I-C-6.3)	13 11 5	7	5	2	0.3
	13 5 5	11	12	-1	0
	23 10 5	4	5	-1	0.1
	23 5 5	6	5	1	0.1
	23 9 5	5	4	1	0.1
	25 10 5	53	47	6	0.4
	25 11 5	6	4	2	0.4
	7 11 5	4	3	1	0.1
TRIPLE 5 (C-6.3-N)	10 5 3	5	5	0	0
	10 5 4	37	32	5	0.4
	10 5 7	14	12	2	0.2
	11 5 4	8	11	-3	0.5
	11 5 7	12	6	6	2
	5 5 4	12	18	-6	1.2
	9 5 7	7	5	2	0.3
TRIPLE 6 (6.3-N-D)	5 3 4	11	6	5	1.5
	5 4 4	64	74	-10	0.7
	5 6 4	5	8	-3	0.7
	5 7 6	11	11	0	0
	5 7 9	26	20	6	0.8
TRIPLE 7 (N-D-E)	3 4 6	11	6	5	1.5
	4 4 6	64	76	-12	1
	6 4 6	5	8	-3	0.7
	6 9 4	7	2	5	2.8
	7 6 3	3	3	0	0
	7 6 4	4	3	1	0.1
	7 6 5	5	6	-1	0.1
	7 9 4	26	25	1	0
TRIPLE 8 (D-E-F)	4 6 4	78	84	-6	0.2
	6 4 5	4	4	0	0
	6 5 5	5	6	-1	0.1
	9 4 3	30	24	6	0.7
TRIPLE 9 (E-F-J)	4 3 10	3	3	0	0
	4 3 3	6	5	1	0.1
	4 3 4	17	13	4	0.5
	4 4 7	5	4	1	0.1
	5 5 5	4	4	0	0
	6 4 6	4	8	-4	1.3
	6 4 7	74	82	-8	0.4
TRIPLE 10 (F-J-G)	3 4 26	17	8	9	3.2
	4 7 13	11	12	-1	0
	4 7 14	9	11	-2	0.2
	4 7 15	19	16	3	0.3
	4 7 16	19	22	-3	0.2

Table 4) Linkage disequilibrium for pairs of markers

Linkage Disequilibrium Log of p (χ^2)												
	SNL2	M	A	I	C	v6.3	N	D	E	F	J	G
SNL2	***	-25	-10	-1	-16	-4	-1	-1	-6	-2	-1	-2
M		***	-14	-15	-10	-3	-2	-5	-1	-5	-14	-5
A			***	-4	-5	-1	-2	-1	-1	-1	-1	-5
I				***	-36	-1	-20	-1	-1	-3	-3	-6
C					***	-13	-12	-5	-7	-5	-13	-5
v6.3						***	-1	-4	-4	-2	-12	-4
N							***	-81	-35	-39	-34	-33
D								***	-58	-101	-49	-35
E									***	-48	-36	-35
F										***	-71	-39
J											***	-39
G												***

Table 5) Low haplotype diversity reveals a small number of effective founders in the Finnish population

QUINT 7 haplotype	Number of chromosomes	All haplotypes		2 haplotypes		12 haplotypes	
		p^2	$(1/\Sigma p^2)$	p^2	$(1/\Sigma p^2)$	p^2	$(1/\Sigma p^2)$
24632	1	2.74E-05					
24647	1	2.74E-05					
29434	1	2.74E-05					
32657	1	2.74E-05					
34634	1	2.74E-05					
34647	11	3.32E-03				3.32E-03	3.01E+02
44447	1	2.74E-05					
44547	2	1.10E-04					
44632	1	2.74E-05					
44646	9	2.22E-03				2.22E-03	4.50E+02
44647	79	1.71E-01		1.71E-01	5.85E+00	1.71E-01	5.85E+00
44655	3	2.47E-04				2.47E-04	4.05E+03
44747	1	2.74E-05					
46447	3	2.47E-04				2.47E-04	4.05E+03
46646	1	2.74E-05					
54647	1	2.74E-05					
64647	6	9.87E-04				9.87E-04	1.01E+03
66455	1	2.74E-05					
694310	1	2.74E-05					
69434	4	4.39E-04				4.39E-04	2.28E+03
69437	1	2.74E-05					
76353	1	2.74E-05					
76354	3	2.47E-04				2.47E-04	4.05E+03
76454	4	4.39E-04				4.39E-04	2.28E+03
76455	1	2.74E-05					
76554	2	1.10E-04					
76555	8	1.75E-03				1.75E-03	5.70E+02
79333	2	1.10E-04					
79334	1	2.74E-05					
79423	1	2.74E-05					
79424	1	2.74E-05					
79425	1	2.74E-05					
79431	1	2.74E-05					
794310	1	2.74E-05					
79432	2	1.10E-04					
79433	6	9.87E-04				9.87E-04	1.01E+03
79434	17	7.92E-03		7.92E-03	5.59E+00	7.92E-03	1.26E+02
79435	1	2.74E-05					
79437	1	2.74E-05					
79447	2	1.10E-04					
79532	1	2.74E-05					
79534	2	1.10E-04					
79634	2	1.10E-04					
total alleles:	**191**	**1.91E-01** Σp^2 **(CK)**	**5.23**	**1.79E-01** Σp^2 **(CK)**	**5.59**	**1.90E-01** Σp^2 **(CK)**	**5.27**

CHAPTER 3

Lower than expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion

Kristin Ardlie*, Shau Neen Liu-Cordero*, Michael A. Eberle*, Mark Daly, Jeff Barrett, Ellen Winchester, Eric S. Lander and Leonid Kruglyak

* These authors contributed equally to this work

Published as Ardlie *et al.* Lower than expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am. J. Hum. Genet.* 69, 582-589 (2001).

Contributions: Along with Kristin Ardlie, who was an equal partner in this work, I conceived of this project and was responsible for all experimental design, all data collection. I was involved in the data analysis. A large part of the analysis was performed by Mike Eberle and Leonid Kruglyak at FHCRC

Lower than expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion

Kristin Ardlie[1,5,6], Shau Neen Liu-Cordero[1,4,6], Michael A. Eberle[2,6], Mark Daly[1], Jeff Barrett[1], Ellen Winchester[1], Eric S. Lander[1,4] and Leonid Kruglyak[2,3]

1 Whitehead Institute/MIT Center for Genome Research, Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, USA
2 Division of Human Biology and 3 Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
4 Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
5 Current Address: Genomics Collaborative, 99 Erie Street, Cambridge, MA, 02139.
6 These authors contributed equally to this work

Correspondence should be addressed to L.K. (Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, D4-100, Seattle, Washington, 98109 USA. e-mail: leonid@fhcrc.org, phone: 206-667-3120, fax: 206-667-2383)

Understanding the pattern of linkage disequilibrium (LD) in the human genome is important both for successful implementation of disease gene mapping approaches and for inferences about human demographic histories. Previous studies have examined LD between loci within single genes or confined genomic regions, which may not be representative of the genome; between loci separated by large distances, where little LD is seen; or in population groups that differ from one study to the next. We measured LD in a large set of locus pairs distributed throughout the genome, with loci within each pair separated by short distances (on average 124 base pairs). Given current models of the history of the human population, nearly all pairs of loci at such short distances would be expected to show complete LD as a consequence of lack of recombination in the short interval. Contrary to this expectation, a significant fraction of pairs showed incomplete LD. A standard model of recombination applied to this data leads to an estimate of effective human population size N = 110,000. This estimate is an order of magnitude higher than most estimates based on nucleotide diversity. The most likely explanation of this discrepancy is that gene conversion increases the apparent rate of recombination between nearby loci.

INTRODUCTION

With the completion of a reference human genome sequence in sight (International Human Genome Sequencing Consortium 2001), attention is shifting to sequence variation. Most of this variation is in the form of single nucleotide polymorphisms (SNPs). Large numbers of SNPs throughout the human genome have now been discovered and mapped (The International SNP Map Working Group 2001). SNP discovery efforts have been carried out with the belief that these polymorphisms will either turn out to be the causative mutations in human complex disease or can be used as markers to map such mutations by LD (Collins et al. 1997). The success of LD mapping approaches based on a genome-wide map of single nucleotide polymorphisms will depend in large part on the extent of LD between loci. LD can be influenced by a number of factors. These include forces that act on individual genomic regions (such as natural selection and the rates of mutation and recombination) as well as forces that have affected the entire genome (effective population size, population expansions and bottlenecks, population subdivision, and gene flow). Thus patterns of LD are expected to vary from one region of the genome to another as well as among populations.

Previous studies have demonstrated a non-uniform distribution of LD on a chromosomal scale, using estimates of LD based on microsatellite CEPH-Genethon data (Huttley et al. 1999), as well as population differences in the extent of LD (Goddard et al. 2000, Eaves et al. 2000, Taillon-Miller et al. 2000, Kidd et al. 2000). Most current estimates of LD are derived from disparate populations or single genomic regions (Clark et al. 1998,

Tishkoff et al. 2000, Kidd et al. 2000, Fullerton et al. 2000, Nickerson et al. 2000), and are still too few to provide general insights into the patterns and distribution of LD, although data from larger populations and regions are now beginning to emerge (e.g. Dunning et al. 2000, Abecasis et al. 2001). Variation in LD among studies is illustrated by a summary of 5 large single gene studies (Pzreworski et al. 2000), in which 3 show complete LD over at least 2.5kb, while 2 show rapid decline of LD. Thus, even as dense SNPs maps are becoming available, key questions still remain concerning the distribution of LD between SNPs throughout the genome.

We set out to examine LD in a genome-wide collection of locus pairs. For practical reasons, we chose pairs separated by short distances. SNPs are typically detected by comparing the sequences of a short unique stretch of DNA among individuals (such stretches of DNA are commonly known as sequence-tagged-sites or STSs). To examine LD between SNPs, we drew on three large-scale SNP discovery efforts to assemble a collection of 103 STSs that each contained multiple SNPs (a total of 325 SNPs). The STSs are distributed throughout the genome (including chromosomes 1-19, 22 and X). We sequenced these STSs in a globally representative sample of 47 individuals (94 chromosomes) and measured LD between SNPs on the same STS. We report that significantly more SNP pairs show incomplete LD than expected from accepted values of effective human population size and recombination rate, and discuss possible explanations for this observation.

SUBJECTS, MATERIAL, AND METHODS

Samples and Loci

The study sample was comprised of 10 European individuals, 19 individuals from Asia/Pacific, 4 from the Americas, and 14 from Central and North East Africa. Many of these samples overlap with those described in Cargill et al. (1999). One common chimpanzee (*Pan troglodytes*) was also sequenced.

All STS regions were chosen with prior knowledge that they contained at least 2 SNPs. These STS regions, and the SNPs within them, were ascertained in three separate studies. 41 STSs (Group 1) were chosen from the STSs surveyed in Wang et al. (1998). One third of the STSs in that survey were derived from random genomic sequence and two-thirds from the 3'-ends of expressed sequence tags (3'-ESTs), primarily the untranslated regions of genes. These STSs have a size range of 58-430 bp (average of 232 bp). 28 STSs (Group 2) were derived from the genes sequenced by Cargill et al. (1999), and are located primarily within coding sequence. These regions are defined by a primer pair designed to amplify primarily coding sequence from genomic DNA, and range in size from 175-725 bp (average of 436 bp). The remaining 34 regions (Group 3) were derived from a Reduced Representation Shotgun (RRS) library employed by the SNP Consortium (TSC) (Altshuler et al. 2000). These are random genomic regions with a size range of 160-540 bp (average of 354 bp).

Genotyping

All markers were genotyped by ABI sequencing of each STS. Sequences were base-called, assembled, and polymorphic sites were identified by the PolyPhred program with default settings (Nickerson et al. 1997). Results were visually inspected and verified by two observers. Sequencing was performed in the forward direction only; however, extensive confirmatory re-sequencing and data-validation procedures were followed. All SNPs with the minor allele seen only once or twice in the sample, as well as all newly discovered SNPs, were re-sequenced three times to confirm genotypes. Additionally, on average, 22% of all individuals were fully re-sequenced at least 2 times, and genotypes revalidated, for each STS region. We discarded from further analyses those SNPs for which the minor allele was seen only once in our sample (35 SNPs), as well as 2 insertion/deletions.

Analysis of Data

Given the current views on the history of the human population, pairs of SNPs spaced at very short intervals are expected to exist in complete LD. That is, one would expect to observe only three of the four possible haplotypes. The fourth haplotype could arise through breakdown of LD by recombination, recurrent mutation, or gene conversion. We sought to determine the frequency with which all four haplotypes could be observed for nearby loci.

LD is often measured by one of several coefficients that reflect departures of two-locus haplotype frequencies from those expected if the loci were independent (Devlin and Risch 1995, Jorde 2000). These measures are not ideal for the present study, because they are not sensitive to small departures from complete LD that might occur at very short distances, and because they are sensitive to allele frequency. To quantify the breakdown of LD, we used a measure, R_M, which counts the minimum number of recombination events in the history of the sample that are necessary to explain the observed data (Hudson and Kaplan 1985). This measure has previously been used to estimate the rate of recombination in the apolipoprotein AI-CIII-AIV gene cluster (Antonarakis et al. 1988). A recombination event is inferred when all four haplotypes are observed for a pair of loci. For a single STS, R_M is the number of non-overlapping SNP pairs that show all four haplotypes. When tabulating such events, we conservatively considered only locus pairs for which all four haplotypes were unambiguously observed in the data (that is, not hidden in double heterozygotes). For those locus pairs considered most likely to have a hidden fourth haplotype present, haplotypes were resolved empirically using Allele Specific PCR (~10% of STSs). There were no instances in which an additional haplotype was uncovered through Allele Specific PCR.

The inference of a recombination event from the observation of four haplotypes is valid only under the infinite sites model (Kimura 1969), because recurrent mutation can also result in the presence of all four haplotypes. Although most SNPs are thought to arise from unique mutational events, some sites (including CpG dinucleotides and some repetitive sequences) are believed to be highly mutable and subject to recurrent mutation

(Templeton et al. 2000). To obtain a conservative estimate of R_M, we removed all sites that fall into either CpG or mononucleotide run >5 bp categories (85 SNPs or ~30%) from the data set. Including these sites would have raised the observed R_M from 7 to 19.

Coalescent Simulations

Under the standard neutral model (Przeworski et al. 2000), the expected number of recombination events observed in a sample is determined by the parameter $4Nr$, where N is the effective population size and r is the recombination rate per nucleotide per generation. We used the coalescent approach (Hudson 1983) to simulate the history of our sample for different values of $4Nr$. We then calculated the mean and the distribution of R_M.

For each value of $4Nr$, we carried out two sets of simulations. In the first set (type I simulations), random genealogies of our sample were generated for each STS using the coalescent model with recombination (Hudson 1983). A simulated genealogy was accepted if, at the positions where SNPs are observed on the corresponding STS, the trees contained branches such that mutations on those branches would match the number of observations of each allele in the sample. For the accepted genealogies, mutations were placed on these branches, and the resulting genotype observations were added to the simulated data set. If several choices of a branch were available, a single branch was randomly selected with probability proportional to its length. For each value of $4Nr$, 1000 genealogies were generated for each STS. This set of simulations exactly matched

the SNP number, allele frequencies and SNP spacing of each STS. The software used to carry out these simulations is available on our web site.

A concern with this set of simulations is that for each STS, the number of SNPs and their allele frequencies and positions were assigned *a priori* and placed on randomly generated genealogies. In fact, given the size of our sample and levels of human nucleotide diversity, most genomic regions of the size considered here would not be expected to contain multiple SNPs. By choosing regions which do contain multiple SNPs, we are likely to enrich for regions of the genome that have older genealogies with more opportunity for mutation. Such deep genealogies would also allow more opportunity for recombination, potentially biasing the real data set to higher R_M values than predicted by our simulation.

To address this potential bias, we carried out a second set of simulations (type II). These simulations were carried out using the program mksamples distributed by Richard Hudson through his web site (Hudson, 1983). For each value of $4Nr$, we generated random genealogies and placed mutations on these genealogies according to a Poisson process with the rate set by the observed nucleotide diversity in humans ($4N\mu = 8\text{x}10^{-4}$). As in the real data set, we then discarded all STSs that did not contain at least two SNPs with both alleles observed at least twice in the simulated sample. Most genealogies were rejected by these criteria. Only those genealogies with 2, 3, or 4 SNPs and with all SNP alleles observed at least twice were retained for further analysis. We then selected a subset of these genealogies with multiple SNPs that matched the average SNP number,

SNP spacing, and SNP allele frequency distribution of our data set. Specifically, for STSs with 2 SNPs, we generated 10^7 genealogies for segments of 285 bp each, and then selected those with exactly 2 SNPs separated by at least 45 bp (average spacing 125 bp, same as in the real data set). We then binned these simulated STSs according to the geometric average of the two minor allele frequencies. The geometric average was chosen because it is the square root of the frequency of the rarest haplotype under linkage equilibrium. A similar procedure was followed for STSs with 3 and 4 SNPs, except that for STSs with 3 SNPs segment length was 365 bp, the largest distance between two SNPs was set to a minimum of 50 bp, and the smallest distance was set to a minimum of 10 bp. For STSs with 4 SNPs, segment length was 420 bp, the largest distance between two SNPs was set to a maximum of 300 bp, and the smallest distance was set to a minimum of 10 bp. The average distance between SNPs was 130 bp and 121 bp, compared to 132 and 108 bp for the real data, on STSs with 3 and 4 SNPs, respectively. The arithmetic mean of the geometric average of minor allele frequencies for each SNP pair was used to define the frequency bins. We calculated the distribution of R_M by randomly selecting simulated STSs from the binned set such that the number of STSs with 2, 3, or more SNPs, as well as the number in each frequency bin, matched the real data set. The allele frequency distributions of the real and simulated data sets are given in Table 1. The nucleotide diversity for both sets was 2.5×10^{-3}. As expected, this is much higher than for the genome as a whole, because we selected for regions of high polymorphism. We then combined the R_M values for the 68 simulated STSs. This sampling was repeated 1000 times to obtain the distribution of R_M. This set of simulations, together with the ones

including gene conversion (see below) required generating a total of 750 million genealogies, and produced results very similar to those obtained from type I simulations.

To assess the accuracy of matching between the real and simulated data sets, and also to understand the magnitude of the bias introduced by use of random genealogies vs. those conditioned on polymorphism, we carried out type I simulations for simulated data sets generated by type II simulations for values of $4Nr$ between 4×10^{-4} and 9.6×10^{-3}. We found that type I simulations produced estimates of R_M that were around 10% lower than the actual values in the simulated (type II) data, indicating that only a small bias in observed recombination results from greater genealogy depth of regions with higher polymorphism. For example, for $4Nr = 4.8\times10^{-3}$ the type I simulation underestimated the actual value by 11% (expected R_M values for type I and type II simulations were 6.7 and 7.5 for the 68 STSs). Because of the similarity of results from the two types of simulations, we conservatively use the results of type II simulations in Figure 1 and the text.

To simulate the effect of variation in recombination rate, we assumed that the rate is increased by a factor of $1/\alpha$ for a fraction α of the genome, while the rest of the genome does not recombine. Note that the overall rate of recombination per genome is kept fixed at its known value by this assumption. We assigned STSs to groups with recombination rate r/α or 0 with probabilities α and $1-\alpha$, respectively. We then computed the expected number of recombination events. The simulations were carried out for a range of $1/\alpha$ = 1.5 to 20.

Simulations with gene conversion were carried out with mksamples, which implements the model of Wiuf and Hein (2000). We used parameter track_len = 500 for the average conversion tract length in bp, and carried out type II simulations as described above with the ratio f of conversion to recombination events ranging from 0 to 12.

RESULTS

Our final data set consists of 68 STS regions with more than 1 SNP, of which 38 contain 2 SNPs, 19 contain 3 SNPs, 10 contain 4 SNPs, and one contains 5 SNPs (a total of 178 SNPs). Of the 165 overlapping pairs of SNPs on the same STS, 12 unambiguously show all four haplotypes. When the overlaps between pairs on the same STS are taken into account, at least seven obligate recombination events on 6 STSs are needed to explain the observed data (that is, $R_M = 7$). Every event is supported by at least 2 unambiguous observations of each of the four haplotypes. Observations of the rarest haplotype are distributed across samples of different geographic origin. The actual number of recombination events in the history of the sample is likely much higher than 7, because R_M is known to provide an underestimate (Hudson and Kaplan 1985).

We next sought to estimate population parameters consistent with the observation of $R_M = 7$. Under the standard neutral model (Przeworski et al. 2000), the expected number of recombination events observed in a sample is determined by the parameter $4Nr$, where N is the effective population size and r is the recombination rate per nucleotide per

generation. The average recombination rate in humans, 1 cM per Mb, corresponds to $r = 10^{-8}$, and together with the frequently cited value of $N = 10{,}000$ (e.g. Harpending et al. 1998, Harris and Hey 1999) yields $4Nr = 4\text{x}10^{-4}$. We used the coalescent approach (Hudson 1983) to simulate the history of our sample for different values of $4Nr$ between $4\text{x}10^{-4}$ and $9.6\text{x}10^{-3}$ and calculated the mean and the distribution of R_M. With $4Nr = 4\text{x}10^{-4}$, the expected value of R_M is <1, 98% of the replicates give $R_M \bullet 2$, and the highest value observed in 1000 simulated replicates of the data is $R_M = 5$ (this value is observed once). A mean value of $R_M = 7$ is obtained around $4Nr = 4.4\text{x}10^{-3}$ (Figure 1). We can confine $4Nr$ to the range of $2.4\text{x}10^{-3}$-$8.8\text{x}10^{-3}$ with 95% confidence (Figure 1).

DISCUSSION

Our estimate of $4Nr = 4.4\text{x}10^{-3}$, together with a recombination rate $r = 10^{-8}$, leads to a recombination-based estimate of effective human population size $N = 110{,}000$. This estimate is 11-fold higher than the frequently cited value of $N = 10{,}000$ (e.g. Harpending et al. 1998, Harris and Hey 1999). If $r = 10^{-8}$, then $N = 10{,}000$ is clearly excluded by our data. Estimates of effective population size based on sequence diversity rely on the nucleotide mutation rate, which in humans is estimated to be of order $2\text{x}10^{-8}$ per base pair per generation (Drake et al. 1998, Nachman et al. 1998). For a typical nucleotide diversity of ~$8\text{x}10^{-4}$ (Przeworski et al. 2000, Wang et al. 1998, Cargill et al. 1999, Halshuka et al. 1999, The International SNP Map Working Group 2001), the estimate $N = 110{,}000$ would imply a mutation rate of ~$2\text{x}10^{-9}$, an order of magnitude too low. Based on estimates derived from diversity of HLA, mitochondria, and the Y

chromosome, Ayala (1995) argued that the effective long-term human population size is closer to 100,000 than 10,000, which would be consistent with our recombination-based estimate. However, most other recent estimates of N are consistent with a lower value of 10,000. Why are we seeing significantly more apparent recombination events than would be predicted from the generally accepted values of N and r?

While a larger ancestral effective population size is one explanation for our observation of a high R_M, it is not the only one. There are several possible reasons for the discrepancy between observed and predicted values of R_M, some of which we think are unlikely. Our analysis would overestimate the frequency of historical recombination if some of what we consider obligate recombination events were, despite the exclusion of sites known to be highly mutable, the result of recurrent mutation. In principle, two tests could be used to discriminate between recombination and recurrent mutation. First, the probability of recombination increases with distance between two SNPs, while the probability of recurrent mutation does not. The 7 events occur on STSs that are somewhat longer than average (369 bp vs. 330 bp), but this difference is not statistically significant. Second, when three nearby sites are considered, observation of all 4 haplotypes at sites 1 and 2 and at sites 2 and 3 can be explained by a single recurrent mutation at the middle site, but would require the unlikely occurrence of two recombination events. None of the 4 STSs that contain three or more SNPs and show all 4 haplotypes for at least one pair of sites a fall into this category. Two can be explained with either a single recurrent mutation or a single recombination, one can be explained with a single recombination but would require recurrent mutation at two or more sites, and one requires two events of either

type. Once again, these data are not sufficient to provide statistically significant evidence against recurrent mutation, but they are consistent with a recombination-only explanation.

An additional concern is that in choosing regions containing two or more SNPs, we might have chosen regions that have a higher mutation rate (although the studies reporting these SNPs did not find an unexpectedly high fraction of STSs with multiple SNPs (Wang et al. 1998, Cargill et al. 1999, Altshuler at al. 2000)). We examined the human-chimpanzee divergence rates for our regions. 84% of all STSs were successfully amplified and sequenced in the chimpanzee, and overall *Homo-Pan* sequence divergence was 1%. While the mean *Homo-Pan* divergence was lower for Group 2 STSs, which lie primarily in coding regions, this difference was not significant, suggesting no overall mutation rate differences between the groups. Most importantly, the divergence rate is similar to that reported for the genome as a whole (Nachman et al. 1998, Hacia et al. 1999).

Another potential explanation for our data is that recombination across the genome is highly non-uniform, and that our STSs with obligate recombination events fall into high-recombination regions. An extreme version of this model would postulate that the genome consists of non-recombining blocks separated by regions of high recombination. Because the overall rate of recombination per genome is well-established, our results would then suggest that the high-recombination regions constitute ~10% of the genome. We carried out simulations to estimate the effect of such variation in recombination rate on our estimate of $4Nr$. The results showed that for a given $4Nr$, the expected value of

R_M is *lower* when recombination rate is variable, and thus an even higher value of $4Nr$ would be required to explain the observed data.

As described, the SNP markers in the study were derived from three different studies, each of which differed in the total number and the relative population diversity of the samples in which the SNPs were originally discovered. The bias that results from this different ascertainment of SNPs can influence allele frequency and demographic inference. Among the three source groups we surveyed, bias is greatest in groups 1 and 3, where a much smaller sample was used for initial SNP discovery, and least in group 2, because of considerable overlap between the initial discovery sample and the sample used in the present study. We might expect the groups with a smaller discovery sample to contain more common, older SNPs, which are more likely to have experienced recombination events. We therefore examined whether R_M values differed among groups. Of the 7 obligate recombination events, 1 was on an STS from group 1, 2 on STSs from group 2, and 4 on STSs from group 3. These numbers were not significantly different from expectation given the total amount of sequence from each source (6613 bp, 5657 bp, and 10184 bp, respectively), suggesting that we do not see a systematic bias in our measurement of R_M. The specific effects of ascertainment are complex, and are considered for these data in greater detail in another manuscript (Wakeley et al. in preparation).

We believe that the most likely explanation for our observed excess of historical recombination is gene conversion. There is growing recognition that gene conversion

can be a factor in shaping fine-structure patterns of LD. A high rate of conversion events in HLA-DPB1 was reported by Zangenberg et al. (1995). Analyses by Andolfatto and Nordborg (1998), Wiuf and Hein (2000) and Wiuf (2000) demonstrate that at intragenic distances, gene conversion, rather than crossing over, is likely to be the dominant force that breaks up sites, and might account for the demonstrated lack of intralocus associations found in *Drosophila melanogaster*. Gene conversion increases the rate of exchange for closely linked sites, but has negligible effects for more distant sites—as inter-site distance increases, gene conversion events that affect one of the sites become rare compared to crossovers between the sites. Indeed, gene conversion should contribute significantly to the apparent rate of recombination only when the distance between two sites is not appreciably greater than the length of a gene conversion tract (Andofatto and Nordborg 1998).

A rough idea of the effects of gene conversion can be obtained from the following highly simplified model (see also Andofatto and Nordborg 1998, Wiuf and Hein 2000, Wiuf 2000). Assume that H Holliday junctions are formed for each reciprocal crossing over event in a genome, and that each junction is accompanied by a tract of conversion of fixed length L. It is then easy to show that the apparent rate of recombination between sites separated by distances $d < L$ is $2Hdr$. For sites separated by distances $d > L$, the apparent rate of recombination is $dr+Lr(2H-1)$, which is of order dr for $d >> L$ (as would be expected without gene conversion). Note that the apparent rate for $d < L$ is independent of L. The length L of gene conversion tracts is generally thought to be in the range of 350-1000 bp, and thus at the distances between loci considered in this paper, the

apparent rate of recombination should be enhanced by a factor of $2H$. The value of H is not known in humans, but tends to fall in the range of 1-5 for organisms (*Saccharomyces cerevisiae*, *Neurospora crassa*, *Drosophila melanogaster*) in which it has been measured (Foss et al. 1993). A value of H=5 should lower our estimate of $4Nr$ tenfold, bringing it in line with expectation from the generally accepted values of N and r. We tested this prediction by running simulations incorporating the gene conversion model of Wiuf and Hein (2000) with $4Nr = 4\text{x}10^{-4}$. The results indicate that a ratio of conversion events to recombination events between 3 and 10 is consistent with our data, with a ratio of 6 providing the best fit.

A recent review of the literature found $R_M = 55$ in a total of 71,824 bp of sequence from 15 independent regions—a rate of historical recombination events even higher than our observation of $R_M = 7$ in 22,454 bp (Przeworski et al. 2000). The difference could reflect the fact that the analysis of Przeworski et al. included longer regions and did not exclude highly mutable sites. Przeworski and Wall (2001) re-analyzed publicly available polymorphism data for nine loci and found evidence for more recombination than expected from estimates of recombination rates derived from an integration of genetic and physical maps. A model incorporating gene conversion showed a better fit to the data than a model of crossing-over only, but was not sufficient to completely explain the data. These authors considered values H=1 and H=2; a higher value of H could potentially explain their observations.

A final possibility is that a high value of R_M might also reflect the effect of population subdivision and/or admixture. A realistic demographic model for humans is likely to be complex, and population structure has been documented for humans, particularly in African populations (Tishkoff et al. 2000, Wakeley et al. in preparation). A high value of R_M is unlikely under population structure alone, as haplotypes in different subpopulations should have a low chance of recombining with one another (Przeworski and Wall, 2001). However, it could result from recent admixture of formerly subdivided populations, combined with some amount of recombination or gene conversion. Such an example is provided by the *dpp* locus in *Drosophila* (Richter et al. 1997), where a lack of LD at short distances, together with a clade structure of haplotypes, indicated that the population surveyed was a mixture of several divergent haplotypes that had recently recombined, probably through gene conversion, although insufficiently to bring the population to linkage equilibrium.

This study provides a genome-wide look at LD over short distances. We point out that factors such as gene conversion make it difficult to extrapolate properties of LD from short to long distances and probably from one region to another. A more complete characterization of LD in the human genome will thus require a better understanding of molecular processes such as mutation, recombination and conversion, as well as direct genome-wide measurements covering a large range of distances.

ACKNOWLEDGEMENTS

We thank M. Goldis for technical assistance and H. Coller, L. Hartwell, R.C. Lewontin, D. Nickerson, M. Olson, D. Page, M. Przeworski, D. Reich, G. Smith, J. Wakeley, G. Wong, and two anonymous referees for helpful discussions and comments on the manuscript. Supported in part by grants from the NIH to E.S.L and L.K. L.K. is a James S. McDonnell Centennial Fellow.

ELECTRONIC-DATABASE INFORMATION

URLs for data in this article are as follows:

Kruglyak Lab Home Page, http://www.fhcrc.org/labs/kruglyak (for polymorphism data and simulation software)

Hudson Lab Home Page, http://home.uchicago.edu/~rhudson1 (for Richard Hudson's simulation software)

REFERENCES

Abecasis GR, Noguchi E, Heinsmann A, Traherne JA, Bhattacharyya S, Leaves NI, Anderson GG, Zhang Y, Lench NJ, Carey A, Cardon LR, Moffatt MF, Cookson WOC (2001) Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet 68:191-197

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES (2000) A human SNP map generated by reduced representation shotgun sequencing. Nature 407:513-516

Andofatto P, Nordborg M (1998) The effect of gene conversion on intralocus associations. Genetics 148:1397-1399

Antonarakis SE, Oettgen P, Chakravarti A, Halloran SL, Hudson RR, Feisee L, Karathanasis SK (1988) DNA polymorphism haplotypes of the human apolipoprotein APOA1-APOC3-APOA4 gene cluster. Hum Genet 80:265-273

Ayala FJ (1995) The myth of Eve: molecular biology and human origins. Science 270:1930-1936

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R,

Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231-238

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengaird J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Gen 63:595-612

Collins FS, Guyer MS, Chakravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. Science 278:1580-1581

Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311-322

Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148:1667-1686

Dunning AM, Durocher F, Healey CS, Teare MD, McBride SE, Carlomagno F, Xu C-F, Dawson E, Rhodes S, Ueda S, Lai E, Luben RN, Van Rensburg EJ, Mannermaa A, Kataja V, Rennart G, Dunham I, Purvis I, Easton D, Ponder BAJ (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. Am J Hum Genet 67:1544-1554

Eaves IA, Merriman TR, Barber RA, Nutland S, Tuomilehto-Wolf E, Tuomilehto J, Cucca F, Todd JA (2000) The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common diseases. Nat Genet 25:320-323

Foss E, Lande R, Stahl FW, Steiberg CM (1993) Chiasma interference as a function of genetic distance. Genetics 133:681-691

Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengard JH, Salomaa, V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (2000) Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am J Hum Genet 67:881-900

Goddard KA, Hopkins PJ, Hall JM, Witte JS (2000) Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. Am J Hum Genet 66:216-234

Hacia JG, Fan JB, Rydeo O, Jin L, Edgemon K, Ghandour G, Mayer RA, Sun B, Hsie L, Robbins CM, Brody LC, Wang D, Lander ES, Lipshutz R, Fodor SP, Collins FS (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nat Genet 22:164-167

Halushka MK, Fan JB, Bently K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22:239-247

Harpending HC, Batzer, MA, Gurven M, Jorde LB, Rogers AR, Sherry ST (1998) Genetic traces of ancient demography. Proc Natl Acad Sci USA 95:1961-1967

Harris EE, Hey J (1999) X chromosome evidence for ancient human histories. Proc Natl Acad Sci USA 96:3320-3324

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. Theor Pop Biol 23:183-201

Hudson RR, Kaplan NL (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147-164

Huttley GA, Smith MW, Carrington M, O'Brien SJ (1999) A scan for linkage disequilibrium across the human genome. Genetics 152:1711-1722

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

Jorde LB (2000) Linkage disequilibrium and the search for complex disease genes. Genome Res 10:1435-1444

Kidd JR, Pakstis AJ, Zhao H, Lu R-B, Okonofua E, Odunsi A, Grigorenko E, Bonne-Tamir B, Friedlaender J, Schulz LO, Parnas J, Kidd KK (2000) Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, *PAH*, in a global representation of populations. Am J Hum Genet 66:1882-1899

Kimura M (1969) The rate of molecular evolution considered from the standpoint of population genetics. Proc Natl Acad Sci USA 63:1181-1188

Nachman MW, Bauer VL, Crowell SL, Aquadro CF (1998) DNA variability and recombination rates at X-linked loci in humans. Genetics 150:1133-1141

Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res 25:2745-2751

Nickerson DA, Taylor SL, Fullerton SM, Weiss KM, Clark AG, Stengard JH, Salomaa V, Boerwinkle E, Sing CF (2000) Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. Genome Res 10:1532-1545

Przeworski M, Hudson RR, Di Rienzo A (2000) Adjusting the focus on human variation. Trends Genet 16:296-302

Przeworski M, Wall JD (2001) Why is there so little linkage disequilibrium in humans? Genet Res 77:143-151

Richter B, Long M, Lewontin RC, Nitasaka E (1997) Nucleotide variation and conservation at the *dpp* locus, a gene controlling early development in *Drosophila*. Genetics 145:311-323

Taillon-Miller P, Bauer Sardina I, Saccone NL, Putzel J, Laitinen T, Cao A, Kere J, Pilia G, Rice JP, Kwok P-Y (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. Nat Genet 25:324-328

Templeton AR, Clark AG, Weiss KM, Nickerson DA, Boerwinkle E, Sing CF (2000) Recombinational and mutational hotspots within the human lipoprotein lipase gene. Am J Hum Genet 66:69-83

The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928-933

Tishkoff SA, Pakstis AJ, Stoneking M, Kidd JR, Destro-Bisol G, Sanjantila A, Lu R-B, Deinard AS, Sirugo G, Jenkins T, Kidd KK, Clark AG (2000) Short Tandem-Repeat Polymorphism/*Alu* Haplotype Variation at the PLAT Locus: Implications for Modern Human Origins. Am J Hum Genet 67:901-925

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mitten M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077-1082

Wiuf C (2000) A Coalescence approach to gene conversion. Theor Popul Biol 57:357-367

Wiuf C, Hein J (2000) The coalescent with gene conversion. Genetics 155:451-462

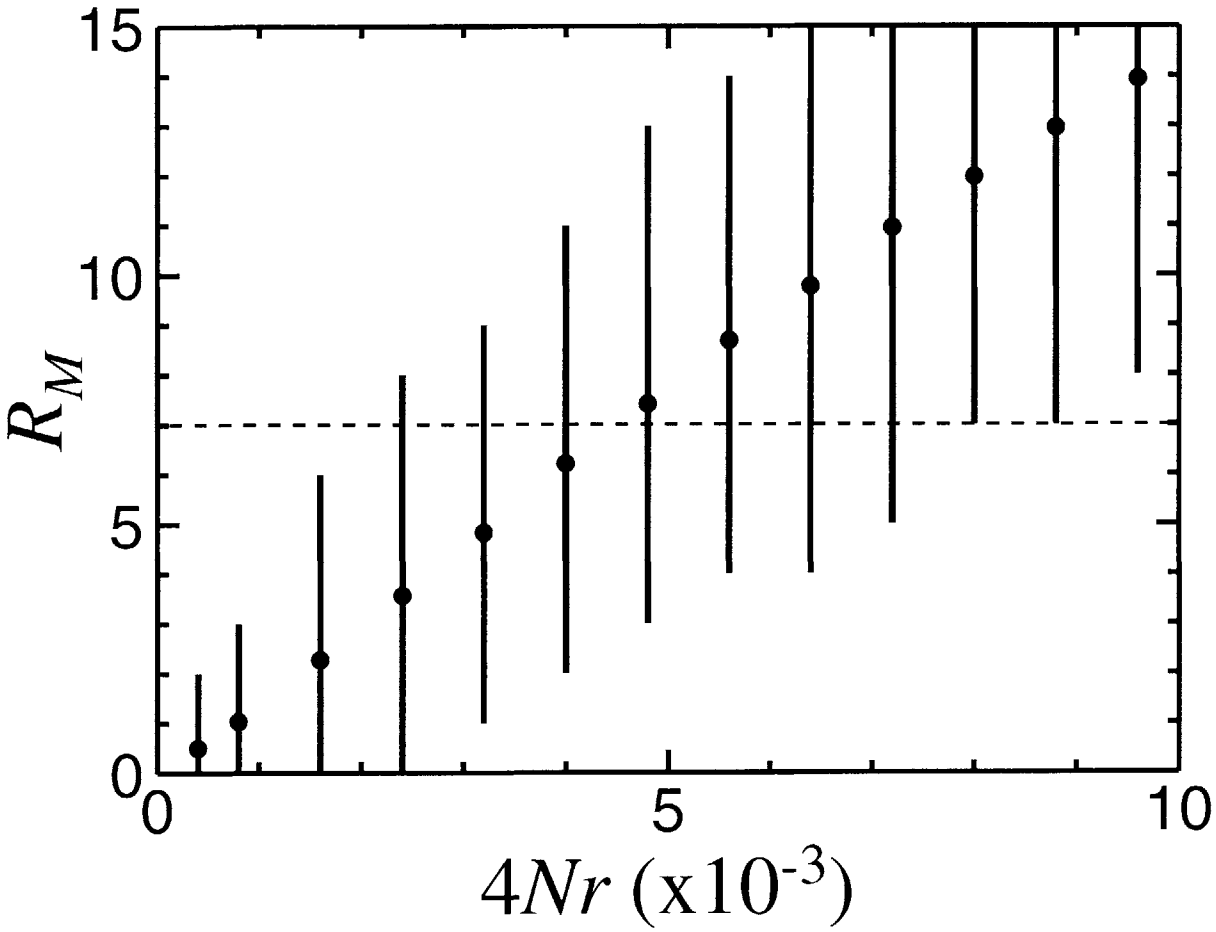
Zangenberg G, Huang M-M, Arnheim N, Erlich H (1995) New HLA-DPB1 alleles generated by interallelic gene conversion detected by analysis of sperm. Nat Genet 10:407-414

TABLE 1. SNP minor allele frequency distributions for the real and simulated data sets

Frequency range	Real Data	Simulated Data
0-0.1	30%	31%
0.1-0.2	14%	16%
0.2-0.3	21%	16%
0.3-0.4	14%	16%
0.4-0.5	21%	21%

FIGURE LEGENDS

Figure 1 Expected number of obligate recombination events (R_M) as a function of population recombination parameter $4Nr$. Dashed line indicates the observed value of R_M = 7. Each solid circle is an average over 1000 simulated replicates of the data. Vertical bars show the values that fall between 2.5% and 97.5% of the simulated distribution


15
10
5
0
0
5
10
R_M
$4Nr$ (x10^{-3})

KEYWORDS

Linkage disequilibrium, gene conversion, recombination, haplotypes, effective population size, single nucleotide polymorphisms, demographic history

CHAPTER 4

The structure of haplotype blocks in the human genome

Stacey B. Gabriel, Stephen F. Schaffner, Huy Nguyen, Jamie M. Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, Shau Neen Liu-Cordero, Charles Rotimi, Adebowale Adeyemo, Richard Cooper, Ryk Ward, Eric S. Lander, Mark J. Daly, and David Altshuler

Published online May 23 2002; 10.1126/science.1069424 (Science Express Reports)

Contributions: I was involved with conception of the project. I was involved with experimental design including numbers and identity of individuals, density of SNPs and size of regions. I decided on the family structure for project; chose individuals and prepared DNA samples. I designed regions for pilot study; participated in some data collection for pilot study.

The structure of haplotype blocks in the human genome

Stacey B. Gabriel[1], Stephen F. Schaffner[1], Huy Nguyen[1], Jamie M. Moore[1], Jessica Roy[1], Brendan Blumenstiel[1], John Higgins[1], Matthew DeFelice[1], Amy Lochner[1], Maura Faggart[1], Shau Neen Liu-Cordero[1,2], Charles Rotimi[3], Adebowale Adeyemo[4], Richard Cooper[5], Ryk Ward[6], Eric S. Lander[1,2], Mark J. Daly[1], and David Altshuler[1,7]

[1]*Whitehead/ MIT Center for Genome Research, Cambridge, MA 02139*
[2]*Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142*
[3]*National Human Genome Center , Howard University, Washington, DC 20059*
[4]*Department of Pediatrics, College of Medicine, University of Ibadan, Ibadan, Nigeria*
[5]*Department of Preventive Medicine and Epidemiology, Loyola University Medical School, Maywood, IL 60143*
[6]*Institute of Biological Anthropology, University of Oxford, Oxford, England OX2 6QS*
[7]*Departments of Genetics and Medicine, Harvard Medical School; Department of Molecular Biology and Diabetes Unit, Massachusetts General Hospital, Boston, MA 02114.*

To whom correspondence should be addressed:* ***altshuler@molbio.mgh.harvard.edu*

Haplotype-based association studies have been proposed as a powerful approach to disease gene mapping, based on linkage disequilibrium (LD) between causal mutations and the ancestral haplotypes on which they arose. Implementing such studies, however, requires information about the general properties (size and diversity) of ancestral haplotypes across the human genome, how these characteristics vary across populations, and the extent to which the local haplotype framework captures the common variation at each region. We report the systematic characterization of haplotype patterns of 51 regions that collectively span •0.4% of the human genome, and compare the results across samples from Africa, Europe and Asia. First, we show that the genome can be objectively parsed into "haplotype blocks" based on patterns of historical recombination across each region. The blocks cover most of the genome and are sizeable, averaging •11 kb in an African (Yoruban) and an African-American sample, and •22 kb in a European and an Asian sample. Within each such block, haplotype diversity is very low: nearly all chromosomes (•90%) match, on average, one of three to five common ancestral haplotypes. Second, we find that both the boundaries of blocks and the specific haplotypes observed are remarkably similar across the population samples examined. Finally, we demonstrate that the majority of all sequence variants track tightly with a haplotype framework that can be defined with only a small number of markers per block. These results provide a foundation for the design of haplotype-based association studies to test common human genetic variation for a role in human disease.

Human genome sequence variation plays a powerful but as yet poorly characterized role in the etiology of common medical conditions, influencing risk of disease, clinical course, and response to therapy. As the vast majority of heterozygosity in the human population is attributable to common variants (those with frequencies >1%), and since the evolutionary selection models for common human diseases (which determine the allele frequencies of causal alleles) are not yet known, one promising approach is to test the universe of common genetic variation for association to medical conditions(1). Moreover, with approximately four million (2-4) of the estimated ten million (5) common single nucleotide polymorphisms (SNPs) already in databases, it is increasingly practical to undertake systematic studies of common genetic variation across a gene or chromosomal region.

In designing and interpreting such studies, it will be necessary to understand the correlations among variant alleles that are observed in the population (termed "linkage disequilibrium"). These correlations arise because each mutation occurs on a given chromosomal haplotype; barring significant mutation or recombination since the most recent common ancestors of the current population, each variant allele will be frozen on the ancestral haplotype on which it arose. Haplotype-based methods have already played a central role in the identification of genes for rare Mendelian diseases (6), and recently, have been applied to the study of common, complex disorders (7).

Many previous studies have examined allelic associations across one or a few gene regions (8-13), concluding that linkage disequilibrium is extremely variable both within and among loci and populations. It was not possible from these studies to draw

general conclusions about the properties of haplotypes in the human genome (14). Recent reports examining a higher density of markers (15-17), however, have revealed a surprisingly simple pattern: blocks of variable length over which only a few common haplotypes are observed, punctuated by a limited number of sites at which it was clear that recombination among these common haplotypes had occurred in the history of the sample. In two regions of the human genome (9, 16), furthermore, it has been demonstrated that dramatic variation in local rates of meiotic recombination determine patterns of linkage disequilibrium, with "hotspots" of recombination coinciding with block boundaries, and "coldspots" representing the regions of low haplotype diversity in between. These studies suggested a model for human haplotype structure (18), but left many critical questions unanswered. First, how much of the human genome exists in such blocks, and what is the size and diversity of haplotypes within blocks? Second, how do these characteristics vary across population samples? Third, can these patterns be parsed using a genome-wide map of common SNPs (2), or will they require direct resequencing (10)? Fourth, and finally, how completely does such a haplotype framework capture common sequence variation within each region? The answers to these questions will determine the design and utility of haplotype-based methods for disease gene discovery, and were the motivation for the current study.

A survey of haplotype blocks across the genome and across populations

To survey haplotype structure, we selected 54 autosomal regions, each with an average size of 250,000 bp, spanning in total 13.4 Mb (•0.4%) of human genome sequence. Regions were evenly spaced across the genome, selected at random subject

only to the availability of contiguous genomic sequence and an average density (in a core region of 150kb) of one candidate TSC SNP every 2kb (19). Genotyping was performed (20) by primer extension of multiplex products and detection by MALDI-TOF mass spectroscopy (21) in 275 individuals sampled from four population groups: 30 parent-offspring trios from Nigeria (Yoruba), 93 members of 12 multigenerational pedigrees of European ancestry (Utah CEPH), 42 unrelated individuals of Japanese and Chinese origin, and 50 unrelated African American samples(22).

We designed assays to 4,532 SNPs (23), of which 3,738 (82%) were successfully genotyped (24)(25). Three of the 54 regions were withheld from further analysis due to inconsistencies in genome assembly and/or evidence that there was a closely related paralogous region making locus-specific PCR difficult (26). In the remaining 51 regions, accuracy of genotype calls was empirically assessed as •99.6%. (27).

There have been varied estimates of the proportion of candidate TSC SNPs that are found to be polymorphic when tested in independent populations. In the large and diverse sample examined here, we find that 89% of the TSC SNPs assayed are polymorphic in at least one of the four population samples, with the proportion polymorphic in each individual sample varying from 70% (Asian) to 86% (African American) (Figure 1a). Although the majority of variants (59%) were observed in all four populations, there are dramatic differences in the allele frequencies of individual SNPs across samples (figure 1b-f), consistent with prior estimates of population differentiation and origin (28).

The size and diversity of haplotype blocks

In order to characterize haplotype blocks across different regions and population samples, it was necessary to create an objective and robust method to define the blocks (29). In the absence of recombination in the history of a sample, observed haplotype diversity is determined by only three factors: the number of ancestral copies represented in the sample, mutation and gene conversion events since those common ancestors, and laboratory errors (assignment of genotypes or map positions). In contrast, where recombination has occurred among the common chromsomes in the history of the sample (termed "historical recombination"), haplotype diversity is increased by the juxtaposition of adjacent ancestral segments. Thus, regions with little evidence for recombination among the common alleles represent the "quantum elements" from which chromosomal diversity is assembled, and form a natural and biological basis for objectively defining haplotype blocks.

The history of recombination between a pair of SNPs can be estimated using the pairwise measure D' (15, 30). In practice, however, the estimate of D' for any given pair of SNPs is highly dependent on sample size, with significant scatter and upwards deviation when small numbers of samples or rare markers are examined. Since our goal was a definition that could be generally applied (that is, not tailored to the specific number of samples and marker characteristics employed), we calculated confidence bounds on D' for each pair of markers (31) and focused on these rather than relying on the maximum likelihood point estimate of D'. For the purposes of defining regions with low levels of historical recombination, we define pairs to be in "strong LD" if the one-

sided upper 95% confidence bound is > 0.98 (that is, consistent with no historical recombination) and the lower bound > 0.7. Conversely, we term "strong evidence of historical recombination" when the upper confidence bound is <0.9 (that is, the data is incompatible with a very low rate of historical recombination). We refer to pairs as "informative" for historical recombination if they fall into one of these two categories. In our study (given the number of chromosomes examined), an average of 87% of all pairs of markers with minor allele frequency > 0.2 were found to be informative. We believe that this method will be robust to variation in the frequencies of the SNPs examined and the sample size examined, since it relies on those pairs for which narrow confidence intervals (that is, precise estimates) have been obtained.

Applying these definitions, we find that a small fraction of very tightly linked informative SNP pairs (separated by distances <1kb) show strong evidence of historical recombination (figure 2a): 14-18% in the African American and Yoruban samples, and 3-6% in the European and Asian samples. In the African and African American samples, the proportion of pairs displaying evidence for historical recombination rises rapidly with distance, increasing to 50% at a separation of •8kb. In the European and Asian samples, in contrast, the fraction of pairs showing strong evidence for recombination rises to 50% at 22kb. As noted previously (12), a difference in the history of recombination among populations is likely attributable to differences in demographic history, since the biological determinants of LD (rates of recombination, mutation, gene conversion) are expected to be constant across groups. These data show for the first time that LD extends to a similar and long extent in Asian as well as European samples.

As our goal was to examine haplotype blocks, we examined the local patterns of D' values across each region (figure 2d-g). Specifically, we defined a haplotype block as a region over which <5% of comparisons among informative SNP pairs show strong evidence of historical recombination. Where many markers were sampled, we simply counted the proportion of pairs with strong evidence of historical recombination, requiring it to be below 5%. Over much of our survey, however, we observed regions in which all of the informative markers showed strong evidence of linkage disequilibrium, but the number of comparisons was insufficient to conclude (simply by counting) that the proportion of such pairs was >95%. By systematically examining the entire dataset, however, we were able to define criteria for accurately identifying scaffolds of two or three markers within which <5% of informative interior pairs show strong evidence of historical recombination (figure 2b, 2c). These criteria (which differed for each population sample (32)) allow us to define blocks even where the marker density is low.

Armed with these criteria, we examined the dataset for haplotype blocks: regions over which <5% of informative pairs show strong evidence of historical recombination. Considering each population sample separately, we identified a total of 928 blocks. Within these blocks, measures of pairwise linkage disequilibrium did not decline appreciably with distance (supplemental figure 1a-c), confirming that these blocks represent regions over which little historical recombination is observed. The span of the blocks averaged 9 kb in the Yoruban and African American samples, and 18kb in the European and Asian samples. The size of each block varied dramatically, however: from <1kb to 94 kb in the African American and Yoruban samples and from <1kb to 173kb in the European and Asian samples. While most of the blocks were small (figure 3a), most

of the sequence spanned by blocks was in large blocks (figure 3b). Specifically, 50% of the sequence in blocks was captured in blocks of 22kb or larger in the Yoruban and African American samples, and in blocks of 44kb or larger in the European and Asian samples.

To extrapolate from these measures to the genome as a whole, it was necessary to disentangle the characteristics of blocks from the spacing of markers used to detect them. Our map consisted of randomly spaced markers (based on their occurrence on the public map), with an average of one marker (with frequency > 0.1) every 7.8 kb across the regions surveyed. The incomplete information in our map leads to two biases. First, in regions in which we had few markers, we are less likely to detect small blocks. Conversely, edge effects cause systematic underestimation of block sizes: a block will typically extend some distance beyond the randomly spaced markers that happen to fall within (and thus, in our survey, define) its boundaries. To estimate the true distribution of block sizes, we performed computer simulations in which block sizes were exponentially distributed (with a specified average size), and markers were randomly spaced (with a mean spacing equal of one every 7.8 kb, the density of high frequency markers (>0.1) in our survey). These simulations provided a good fit to the observed data when the mean size of blocks was set to 11 kb in the Yoruban and African American samples, and 22 kb in the European and Asian samples (Table 1) (33). The model further predicts that the proportion of the human genome spanned by blocks of 10kb or larger is approximately 65% in the Yoruban and African American samples, and 85% in the European and Asian samples.

Having identified blocks according to the criteria above, we were able to examine haplotype diversity with minimal inflation due to historical recombination among the different ancestral segments represented. We note that this block definition—unlike one previously proposed (17)— does not rely upon low haplotype diversity to define blocks. That is, a region could have scant recombination in the genealogy of the samples observed, and yet contain many different common haplotype patterns. Nevertheless, we observe strikingly low diversity of haplotypes as a general feature of regions identified as blocks, with a mean of three to five common (>5%) haplotypes observed in each of the population samples (figure 3c) (34) (35). In addition, we find that these haplotypes can be fully defined with as few as 6-8 randomly chosen common markers. That is, we find that haplotype diversity reaches a plateau where 6-8 markers have been typed, with little evidence for additional common haplotypes where up to 17 markers are included (figure 3c). This demonstrates that low haplotype diversity is not simply an artifact of having examined only a small number of markers, but is a true feature of regions with low rates of historical recombination(36) (37). Haplotype diversity was greatest in the Yoruban and African-American samples, with an average of 5.0 common haplotypes observed. Lower diversity was observed in the European samples (4.2 common haplotypes), and the smallest number of common haplotypes (3.5) observed in the Asian samples. Critically, •90% of all chromosomes match one of these few common haplotypes even where many markers are examined (figure 3d). (38),(39)

Comparison of block boundaries and haplotypes across population samples

The data above demonstrate that most of the human genome consists of sizeable regions over which little evidence for historical recombination is observed, and within which an average of three to five common haplotypes account for 90% of all chromosomes in each population. We next compared these results across the different population samples, examining the location of block boundaries and the identities of the specific haplotypes observed.

To compare block boundaries, we examined pairs of SNPs in sets of two populations. In each population, we asked whether the SNPs were assigned to a single block, or whether they showed evidence of historical recombination. A SNP pair was termed concordant if the assignment was the same in both populations: that is, identified within a single block, or conversely, showing strong evidence of historical recombination. A SNP pair was termed discordant if in the two populations the assignments disagreed (40). The data revealed that the vast majority of SNP pairs (77% - 95%, depending on the population comparison) were concordant across population samples (figure 4a-f). Moreover, where discordance across populations was observed, it was nearly always due to pairs found to display strong evidence of historical recombination in the Yoruban and African American samples, but found in a single block in the European and Asian samples (figure 4a-f). There was great similarity in patterns of historical recombination between the European and Asian samples, with 88% of pairs concordant between these two samples.

We next compared the specific haplotypes observed across the European, Asian, and Yoruban (African) samples. We examined regions spanned by a single block in all samples and, to ensure that haplotypes were well defined, considered only those blocks in which six or more polymorphic markers were obtained (41). Within these blocks, we found that each population sample from contained 3.1 – 4.9 haplotypes that displayed a frequency >5%. The union of these sets, however, contained only 5.3 haplotypes. That is, the specific haplotypes observed in each group were remarkably similar, with 51% (2.7) identified in all three populations, and 72% in two of the three groups. Of the 28% of haplotypes found in a single population sample, furthermore, 90% were found in the Yoruban sample. These data show in a genome-wide data set that three-quarters of haplotypes (>72%) observed in a block are shared across African and non-African samples, with the vast majority of haplotypes found in a single group identified in the Yoruban sample. The similarity in haplotype identities across the European and Asian samples is striking, with an average of only 0.1 haplotypes per block that were unique to either population sample.

The comparison across populations of SNP polymorphism (figure 1) and of recombinant forms and haplotypes (figure 4) all share a common theme. In each case, a greater diversity (of SNPs, of recombinant chromosomes, and of haplotypes) is observed in the African samples, with reduced and highly correlated diversity in the European and Asian samples. These data are strongly supportive of a single "out of Africa" origin (42) for both the European and Asian samples that involved a significant bottleneck, with only a subset of the diversity (of SNPs, of haplotypes, and of recombinant chromosomes) in Africa found in the two non-African populations(11, 12, 43). Since bottlenecks

preferentially effect lower-frequency alleles, this model predicts that the alleles (haplotypes and recombinant chromosomes) present in the African but not the non-African samples would have lower frequencies in Africa than would alleles that are pan-ethnic. We find evidence in support of this hypothesis for both haplotypes and historical recombination data(44).

The strength of correlation between individual variants and haplotype blocks

A major attraction of haplotype based gene mapping is the idea that one can select haplotype tag SNPs ("htSNPs") (10, 17, 45) which capture the common haplotype strucure, and which then can be used to test variation across each region for association to phenotype. No previous study has characterized the extent to which the local haplotype structure conveys information about the common variants contained within each block. In fact, a number of reports (13, 45, 46) have suggested that many SNPs fail to conform to the underlying haplotype structure, and would be overlooked by haplotype based approaches.

To examine this question empirically, we defined a framework of haplotype blocks using a randomly-selected subset of our data (requiring a minimum of 6 markers per block). We then examined the correlation coefficient (r^2) between the remaining SNPs in these blocks and the haplotypes defined by the initial subset of SNPs. Since r^2 values are inversely proportional to the increase in sample-size required in an association study (compared to directly testing each SNP individually), the value of r^2 between the

haplotype framework and these sets of additional SNPs provides a measure of the statistical power and completeness of a haplotype-based association study.

We find that haplotype blocks defined by a few common markers efficiently capture the variation attributable to other markers within their physical span. We measured the maximal r^2 value for each additional SNP (with minor allele frequency > 0.05) when compared to 6 markers used to define the block. The average of these (mean maximal) r^2 values was 0.67 to 0.87. That is, for the average marker, only a small increase in sample size (15-50%) would be needed when using a haplotype-based study rather than a direct association approach. Moreover, we find that within a block, a large majority (77-93%) of all untested markers showed r^2 values greater than 0.5 to the framework haplotypes defined by six markers. This suggests that a minimum of 77% (in the African American samples) and as many as 93% (in the Asian sample) of all other common SNPs within such blocks could be tested with equal or greater statistical power by doubling the sample size of the association study(47). These results directly demonstrate that haplotype blocks can be used to study association to the vast majority of variants within each region with little loss of statistical power.

Discussion

We set out to define human haplotype patterns by typing a dense collection of SNPs across a representative fraction of the human genome, genotyping each in a large number of samples of African, European and Asian ancestry. Our data indicate that haplotype blocks should be considered an objective feature of the human genome —

defined based on the patterns of historical recombination across the human genome. This is in contrast to the interpretation of a study of Chromosome 21 which argued haplotype blocks do not have objective boundaries(17) (48). We show that most of the human genome is composed of such blocks, and estimate that they are typically sizable, averaging 12kb in African samples and 22kb in European and Asian samples. Within each block, a very small number of common haplotypes (three to five) typically capture •90% of all chromosomes in each population. Both the boundaries of blocks and the specific haplotypes observed are shared to a remarkable extent across populations, with the main variation being a subset of alleles (haplotypes and recombinant forms) that are only observed in the Yoruban and African American samples. Finally, blocks defined with a small number of common markers do a quite complete job of capturing the common variation across each locus.

Our results provide a methodological and quantitative foundation for the construction of a haplotype map of the human genome using common SNP markers. Our results show that haplotype blocks can be reliably identified with a modest number of common markers within their span; that is, without complete resequencing of regions of interest. To have confidence that a region is a block, however, requires study in a sufficient number of chromosomes to confidently parse the patterns of historical recombination that underlie common haplotype patterns. Moreover, given the highly variable distribution of block sizes, and the substantial fraction of the genome that exists in large blocks, a hierarchical genotyping strategy will be most efficient. Starting with a lower density of markers (one highly polymorphic marker every 10kb) will capture large blocks; additional polymorphisms can be added where the block pattern and haplotypes

are not yet clearly defined. Our data also clearly indicate that where blocks are small, a greater density of SNPs than are currently available on the public SNP map will clearly be needed. Given the thousand-fold increase in known SNPs over the last three years (2, 49), however, and the availability of additional SNP databases (4), it should be practical to achieve the required marker density across the entire human genome in the near future. Detailed knowledge of human haplotype structure will provide rich information about human history and genome evolution, and provide a foundation for population-based association studies to assess the bulk of human genome sequence variation for a contribution to disease.

Figure Legends

Fig.1. Allele frequency distribution of TSC SNPs across populations. (A) Normalized allele frequency of candidate SNPs. The distribution is normalized to a constant number of chromosomes (n=64 randomly sampled) from the European, African-American, Asian, and Yoroban samples. (B-E) Plots of allele frequency scatter for pairs of populations. The corresponding F_{ST} value is indicated on each plot. (B) Yoruban as compared to European; (C) Asian as compared to European; (D) Yoruban as compared to Asian and (E) Yoroban as compared to African-American. (F) Allele frequencies are correlated between the European and Asian samples when each are compared to the Yoruban sample, supporting a shared "Out of Africa" origin (42).

Fig. 2. Assessment of pairwise linkage disequilibrium across populations. (A) Proportion of informative SNP pairs that display strong evidence for recombination based on confidence intervals on D'. Between 9,860 and 13,980 pairs were examined in each sample. (B,C) Scaffold analysis of Yoruban and African American (B), and European and Asian (C) samples. The y-axis indicates the fraction of internal informative marker pairs displaying strong evidence for recombination when the two-marker criteria, three-marker criteria (32), or no criteria at all are applied to a span of markers. The x-axis indicates the distance between the outermost marker pair; the maximal distance allowed to meet the criteria was determined as described (32). (D-G) Diagram of pairwise D' values for all pairs of markers within region 40A for each population sample: (D)Yoruban; (E) African American; (F) European; (G) Asian. Block diagrams only include SNPs with frequency ≥20% in any given population. Black squares indicate

strong LD (as defined in text); white squares, strong evidence for recombination; gray squares all other uninformative comparisons.

Fig. 3. Block characteristics across populations. (A) Size (kb) distribution of all haplotype blocks found in all populations. (B) Proportion of genome sequence spanned by a block, which has been identified in a block of various size. (D) Summary of haplotype diversity across all blocks. The number of common (•5%) haplotypes per blocks (C) and fraction of all chromosomes accounted for by these haplotypes (D) is plotted as a function of the number of markers typed in each block.

Fig. 4. Comparison of historical recombination across population samples. (A-F) Concordance of block assignments for adjacent SNP pairs, compared across populations. In each plot, the light bars represent the fraction of concordant SNP pairs, and the dark bars the proportion of discordant SNP pairs. Population samples are abbreviated as EU, European sample; AS, Asian sample; AA, African American sample; YR, Yoruban sample.

Table legends

Table 1. Observed and predicted proportion of sequence found in haplotype blocks. Model is based on the best fit to the observed data, and assumes randomly spaced markers with an average density of one every 7.8 kb, and block span an exponentially distributed random variable with a mean size in European sample of 22 kb and of 11 kb in the Yoruban sample. In the model, block boundaries of 2 kb in length are assumed (16).

Supplemental Table 1. Physical location and SNP coverage of clusters. Initial mapping of clusters was made to HG5 (September 00 draft genome sequence). All SNPs were remapped to confirm the relative spacing and intermarker distance using multiple releases of draft genome sequence (HG7, HG8, NCBI draft genome). Several of the original clusters (01, 02, 08, 13, 23, 28, 35, 37, 46, 52) were found to contain sub-regions that were stable internally, but had unstable mapping relative to other parts of the cluster. As such regions could give inaccurate measure of linkage disequilibrium, we subdivided the original (primary) cluster to subclusters containing stretches of SNPS with stable order, orientation and intermarker distance. [b]For each population, the percentage of successful SNP assays is less than the cumulative total of all successful SNP assays (93%). This is because each population sample was assayed in an independent experiment.

Supplemental Figure 1. (A-C) Measures of allelic association of marker pairs versus distance within haplotype blocks, as assessed by the mean value of the correlation

coefficient (r^2) (A); and the mean value of D' (B). (C) Number of pairwise comparisons of markers within blocks, binned by the measure D' between the marker pair. (D) Haplotype frequencies within blocks as estimated by the EM algorithm. The plot represents a comparison of the haplotype frequency based on phased data from full pedigrees versus unphased data from the same individuals, without taking into account family information.

References

1. E. S. Lander, *Science* **274**, 536-9 (1996); F. S. Collins, M. S. Guyer, A. Charkravarti, *Science* **278**, 1580-1 (1997); N. Risch, K. Merikangas, *Science* **273**, 1516-7 (1996).

2. R. Sachidanandam *et al.*, *Nature* **409**, 928-33. (2001).

3. http://snp.cshl.org; http://www.ncbi.nlm.nih.gov.

4. J. C. Venter *et al.*, *Science* **291**, 1304-51. (2001).

5. L. Kruglyak, D. A. Nickerson, *Nat Genet* **27**, 234-6. (2001).

6. E. G. Puffenberger *et al.*, *Cell* **79**, 1257-66. (1994); B. Kerem *et al.*, *Science* **245**, 1073-80. (1989); J. Hastbacka *et al.*, *Nat Genet* **2**, 204-11 (1992).

7. Y. Horikawa *et al.*, *Nat Genet* **26**, 163-75. (2000); J. D. Rioux *et al.*, *Nat Genet* **29**, 223-8. (2001); J. P. Hugot *et al.*, *Nature* **411**, 599-603. (2001); Y. Ogura *et al.*, *Nature* **411**, 603-6. (2001).

8. L. Subrahmanyan, M. A. Eberle, A. G. Clark, L. Kruglyak, D. A. Nickerson, *Am J Hum Genet* **69**, 381-95. (2001); A. R. Templeton, K. M. Weiss, D. A. Nickerson, E. Boerwinkle, C. F. Sing, *Genetics* **156**, 1259-75 (2000); M. J. Rieder, S. L. Taylor, A. G. Clark, D. A. Nickerson, *Nat Genet* **22**, 59-62 (1999); S. A. Tishkoff *et al.*, *Am J Hum Genet* **62**, 1389-402. (1998); J. R. Kidd *et al.*, *Am J Hum Genet* **66**, 1882-99. (2000); P. Taillon-Miller *et al.*, *Nat Genet* **25**, 324-8 (2000); J. C. Stephens *et al.*, *Science* **293**, 489-93. (2001).

9. A. Chakravarti *et al.*, *Am J Hum Genet* **36**, 1239-58. (1984).

10. G. C. Johnson *et al.*, *Nat Genet* **29**, 233-7. (2001).

11. S. A. Tishkoff *et al.*, *Am J Hum Genet* **67**, 901-25. (2000).

12. D. E. Reich *et al.*, *Nature* **411**, 199-204. (2001).

13. A. R. Templeton *et al.*, *Am J Hum Genet* **66**, 69-83 (2000).

14. J. K. Pritchard, M. Przeworski, *Am J Hum Genet* **69**, 1-14. (2001); L. B. Jorde, *Genome Res* **10**, 1435-44. (2000); M. Boehnke, *Nat Genet* **25**, 246-7 (2000).

15. M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, E. S. Lander, *Nat Genet* **29**, 229-232 (2001).

16. A. J. Jeffreys, L. Kauppi, R. Neumann, *Nat Genet* **29**, 217-22. (2001).

17. N. Patil *et al.*, *Science* **294**, 1719-23. (2001).

18. D. B. Goldstein, *Nat Genet* **29**, 109-11. (2001).

19. SNPs from the TSC discovery project were used, as these were identified using a uniform protocol in a multiethnic sample of known composition. A minimum spacing of 500 bp between adjacent SNPs was used to exclude multiple SNPs discovered in the same sequencing read (such SNPs redundantly tag single branches in the genealogy of the region). No other filtering (for example, for repeat content) was applied during SNP selection. A complete description of the characteristics of each region, as well as the details of each SNP tested, are available as Supplemental Table 1

20. Multiplex PCR was performed in five microliter volumes containing 0.1 units of Taq polymerase (Amplitaq Gold, Applied Biosystems), 5 ng genomic DNA, 2.5 pmol of each PCR primer, and 2.5 μmol of dNTP. Thermocycling was at 95 C for 15 minutes followed by 45 cycles of 95 C for 20 s, 56 C for 30s, 72 C for 30 s. Unincorporated dNTPs were deactivated using 0.3U of Shrimp Alkaline Phosphatase (Roche) followed by primer extension using 5.4 pmol of each primer extension probe, 50 μmole of the appropriate dNTP/ddNTP combination, and 0.5 units of Thermosequenase (Amersham Pharmacia). Reactions were cycled at 94 C for 2 minutes, followed by 40 cycles of 94 degrees for 5 s, 50 degrees for 5 s, 72 degrees for 5 s. Following addition of a cation exchange resin to remove residual salt from the reactions, 7 nanoliters of the purified primer extension reaction was loaded onto a matrix pad (3-hydroxypicoloinic acid) of a SpectroCHIP (Sequenom, San Diego, CA). SpectroCHIPs were analyzed using a Bruker Biflex III MALDI-TOF mass spectrometer (SpectroREADER, Sequenom, San Diego, CA) and spectra processed using SpectroTYPER (Sequenom)

21. K. Tang *et al.*, *Proc Natl Acad Sci U S A* **96**, 10016-20. (1999).

22. The European, Asian and African American samples were obtained from the Coriell Cell Repository (http://locus.umdnj.edu/ccr/). The European sample is drawn from the UTAH CEPH pedigree collection. The Asian sample consists of 10 Chinese and 10 Japanese drawn from the Human Variation Panel, and an additional 22 Japanese control samples from the American Diabetes Association GENNID study. The African-American samples constitute the HD50AA diversity panel. Specific sample identifiers are available on The SNP Consortium website (http://snp.cshl.org/allele_frequency_project/panels.html). The Yoruban samples are healthy individuals from a population-based study in Nigeria.

23. Primers and probes were designed in multiplex format (average 4.3-fold multiplexing) using SpectroDESIGNER software (Sequenom, San Diego, CA). A total of 5,283 SNPs were selected from the TSC map, of which assays were successfully designed for 86%; the remaining 14% of SNPs failed primer design. All primers and probe sequences are available at the TSC website

24. Successful genotyping assays were defined as those in which •75% of all genotyping calls were obtained and all quality checks passed (see below). Assays that provided fewer than 75% of genotypes were repeated once in the laboratory and consensus genotypes calculated from the two runs; if not converted into successful assays, a single round of primer redesign and repeat testing was performed. While 75% was used as a minimum threshold, we obtained an average of 94% of all genotypes attempted for each successful SNP.

25. While 82% of assays were successful in at least one population, genotyping success rates in each population range from 72% to 79%. This difference is due the fact that each population was assayed separately in the laboratory, and there is a low rate of laboratory failure in each attempt.

26. We compared the map positions of each SNP across multiple genome builds and independent assemblies (NCBI and UCSC). One of the 54 regions was withheld from analysis due to inconsistencies of relative map positions. (Multiple other regions showed stable map positions at a fine scale, with variable assembly of two or more sub-regions relative to one another. For the purpose of haplotype analysis, such regions were split into sub-regions (see Supplemental Table 1) and analyzed separately.) Two of the 54 regions were withheld because a high proportion of candidate SNPs demonstrated uniformly heterozygous genotypes, indicating the presence of a highly homologous (recently duplicated) region elsewhere in the genome.

27. Hardy-Weinberg equilibrium was evaluated for each population sample, and markers were rejected if they violated H-W equilibrium. Using a threshold of $p<0.01$ (uncorrected for multiple comparison), 1.8% of markers were rejected for violations of H-W equilibrium. This suggests that <1% of markers are in fact out of H-W equilibrium. Two independent tests were used to estimate error rates, providing indistinguishable conclusions. First, we observed 1,068 Mendel errors

in 598,466 polymorphic genotypes examined in multigenerational pedigrees, providing a raw error rate of 0.18%. Since only a subset of genotyping errors will result in a detected Mendel error, we estimate from these data an error rate of •0.4%. In addition, 970 SNPs were assayed more than once in the DNA samples (in total, 394,688 genotypes performed), revealing an independent error rate estimate of 0.4% (1,375 discrepancies identified).

28. L. L. Cavalli-Sforza, P. Menozzi, A. Piazza, *The history and geography of human genes* (Princeton University Press, Princeton, NJ, 1994).

29. K. M. Weiss, A. G. Clark, *Trends Genet* **18**, 19-24. (2002).

30. R. C. Lewontin, *Genetics* **49**, 49-67 (1964).

31. Limits were determined by calculating a probability distribution for D', given the observed two marker genotype data. The upper and lower limits represent the tails of that distribution.

32. The final block identification algorithm was empirically calibrated by the following procedure. Regions were divided into spans of fixed distances (0-50kb), and markers (minor allele frequency >0.1) within each span randomly sampled (2-4 such markers per span). Spans were characterized by the confidence limits on D' between the sampled markers. Spans exceeding minimum thresholds on those confidence intervals were then evaluated by querying all additional informative marker pairs (those not sampled above) for the proportion showing strong evidence of recombination; thresholds were accepted such that < 5% of all internal pairs (averaging over the entire data set) showed strong evidence of recombination. The specific criteria based on these empirical assessments (see figure 2b, c) were as follows. In all cases, the outer-most marker pair was required to be in strong LD with an upper confidence limit (CU) that exceeds 0.98, and a lower confidence limit (CL) that exceeds 0.7. Pairs of markers were required to have confidence bounds of 0.8<CL/0.98<CU, and could span no more than 20 kb in the European and Asian samples and no more than 10 kb in the Yoruban and African-American samples. Runs of three markers were required to have confidence bounds of 0.5<CL/0.98<CU and could span <30 kb in the European and Asian samples(0.75<CL/0.98<CU and <20 kb in the Yoruban and African-American samples). For runs of four or more markers, the fraction of informative pairs in strong LD (0.7<CL/0.98<CU) was simply required to be > 95%; runs of four markers could span on more than 30 kb, while runs of five or more markers were allowed to span any distance.

33. To further confirm that this model applies generally (and is not influenced by the details of our marker spacing), we estimated from the simulations the proportion of pairs at a fixed distance (5kb) that should show evidence of crossing block boundaries (that is, show strong evidence of historical recombination). The model predicts these proportions to be 47% in the Yoruban and African American samples, and 27% in the European and Asian samples. In the empirical data, we

observe 42% and 23% of informative marker pairs separated by 5 kb to show strong evidence of recombination, which offers additional evidence consistent with the model.

34. Haplotype frequencies within blocks were estimated using the expectation-maximization (EM) algorithm of Excoffier and Slatkin. Where pedigree information was available (in the European and Yoruban pedigrees), we used this information to resolve ambiguous phase prior to running the EM. To determine the extent to which pedigree data are needed to determine haplotype frequencies in blocks, we calculated haplotype frequencies in the CEPH and Yoruban samples by two methods. First, we applied the EM without using the data from offspring, and second, we used the full pedigree information to resolve ambiguous phase in the same individuals. We found great consistency of the haplotype frequency estimates in blocks with and without the phase information ($r2>0.99$, supplemental figure 1D). This provides assurance that within blocks, accurate frequency estimates for common haplotypes can be obtained without data from pedigrees. For this reason, haplotype frequency estimations (within blocks of low historical recombination) can be considered accurate even for the two population samples (Asian and African American) for which we do not have pedigree information.

35. L. Excoffier, M. Slatkin, *Mol Biol Evol* **12**, 921-7. (1995).

36. For example, a previous definition for blocks required only that most chromosomes be found in a few common haplotypes. When runs of two or three markers are examined in a small number of samples, however, it is likely this will be the case whether or not there has been recombination in the history of the sample. Such regions — which make up a substantial number of all "blocks" in that survey — do not constitute haplotype blocks in a manner that captures meaningful information about the region.

37. Since the number of markers in a block is correlated with the size of the block, it is possible that the plateau reflects an inverse relationship between block size and haplotype diversity; direct examination of block size and haplotype number reveals this is not the case (data not shown).

38. A low rate of genotyping error is critical to obtaining an accurate measure of haplotype diversity and the proportion in common haplotypes. This is because even a modest 1-2% genotyping error will create many "unique" haplotypes that differ at a single position in a block. In a region with 10 SNPs examined and a 2% error rate: 18% of the chromosomes will contain at least one error, and thus fail to match the few common haplotypes. Thus, direct empirical measures of error rate need to be considered in comparing the complexity of haplotype patterns across studies.

39. Within these blocks, the common haplotypes rarely showed evidence for historical recombination. For example, we performed the four gamete test using

SNPs drawn only from haplotypes with frequency 5% or higher in each block. In only 5% of blocks was one or more violation of the four gamete test observed.

40. To maximize power, these comparisons were made only for SNP pairs spaced five to ten kilobases apart. At shorter distances, nearly all SNP pairs are in a single block, and at greater distances, many SNP pairs are in different blocks.

41. Blocks and haplotypes were identified separately in each population sample, and the results compared for blocks that were physically overlapping in all three samples.

42. R. L. Cann, W. M. Brown, A. C. Wilson, *Genetics* **106**, 479-99. (1984); C. B. Stringer, P. Andrews, *Science* **239**, 1263-8. (1988).

43. D. E. Reich, D. B. Goldstein, *Proc Natl Acad Sci U S A* **95**, 8119-23 (1998); M. Ingman, H. Kaessmann, S. Paabo, U. Gyllensten, *Nature* **408**, 708-13. (2000); S. A. Tishkoff *et al.*, *Science* **271**, 1380-7. (1996).

44. For example, we examined SNP pairs that were in different blocks in the Yoruban samples but in a single block in the European sample. Such pairs had higher D' values in the Yoruban sample (D' = 0.46) than did those that were in different blocks in both population samples (D' = 0.28), indicating a lower frequency to the recombinant forms. We found that the average frequency of all haplotypes in the Yoruban population was 0.21, while those that were found only in the Yoruban sample (but not in the European and Asian samples) were slightly lower, with a mean frequency of 0.16.

45. A. G. Clark *et al.*, *Am J Hum Genet* **63**, 595-612 (1998).

46. S. M. Fullerton *et al.*, *Am J Hum Genet* **67**, 881-900 (2000).

47. The small fraction of SNPs that show r2 values < 0.5 could be attributable to a range of causes. Some may represent branches of the gene tree that are not yet defined with the number of markers employed. Others may be due to gene conversion events or recurrent mutations. We also note that any errors in genotyping or map position would artificially decrease the value of r2, and may make some contribution to the observed results and those reported previously in the literature.

48. The difference between these two interpretations is likely due to the technical impossibility of measuring historical recombination events when a very small number of chromosomes are examined. We note that the example of a haplotype block in figure 2 of Patil et al. shows all four gametes, showing that the "blocks" defined by Patil et al. often contain evidence of historical recombination— and would likely not be considered blocks (based on our definition, or based on haplotype diversity) if examined in a larger number of samples.

49. D. G. Wang *et al.*, *Science* **280**, 1077-82 (1998).

50. This work was supported by funding from The SNP Consortium. We also thank members of the Whitehead Institute Medical and Population Genetics program for helpful discussion; particularly Joel Hirschhorn, David Reich and Nick Patterson for critical reading and comments on the manuscript.

Table 1 : Model versus oberseved distribution of block sizes

	African sample		Non-African sample	
block size	observed % of spanned sequence	predicted % of spanned sequence	observed % of spanned sequence	predicted % of spanned sequence
0-5	12.4	6.3	4.4	1.8
5-10	15.3	15.1	7.4	5.2
10-20	20	31.5	14.9	15.2
20-30	12.8	21.8	16.6	16.6
30-50	22.2	19.1	18	26.9
>50	17.4	6.3	38.7	34.2

Figure 1

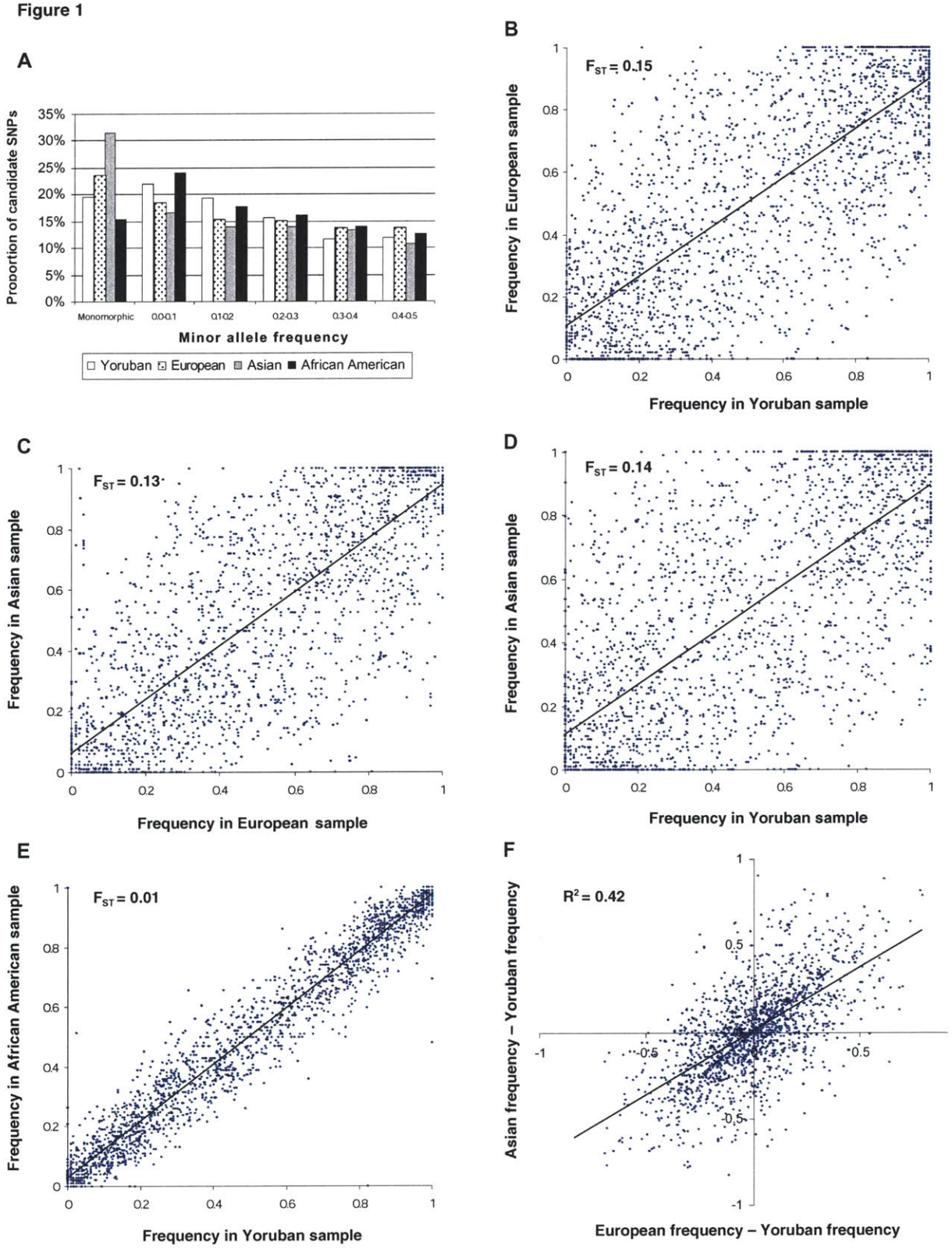

A
Proportion of candidate SNPs
35%
30%
25%
20%
15%
10%
5%
0%
Monomorphic
0.0-0.1
0.1-0.2
0.2-0.3
0.3-0.4
0.4-0.5
Minor allele frequency
Yoruban
European
Asian
African American
B
F_{ST} = 0.15
Frequency in European sample
Frequency in Yoruban sample
C
F_{ST} = 0.13
Frequency in Asian sample
Frequency in European sample
D
F_{ST} = 0.14
Frequency in Asian sample
Frequency in Yoruban sample
E
F_{ST} = 0.01
Frequency in African American sample
Frequency in Yoruban sample
F
R^2 = 0.42
Asian frequency – Yoruban frequency
European frequency – Yoruban frequency

Figure 2

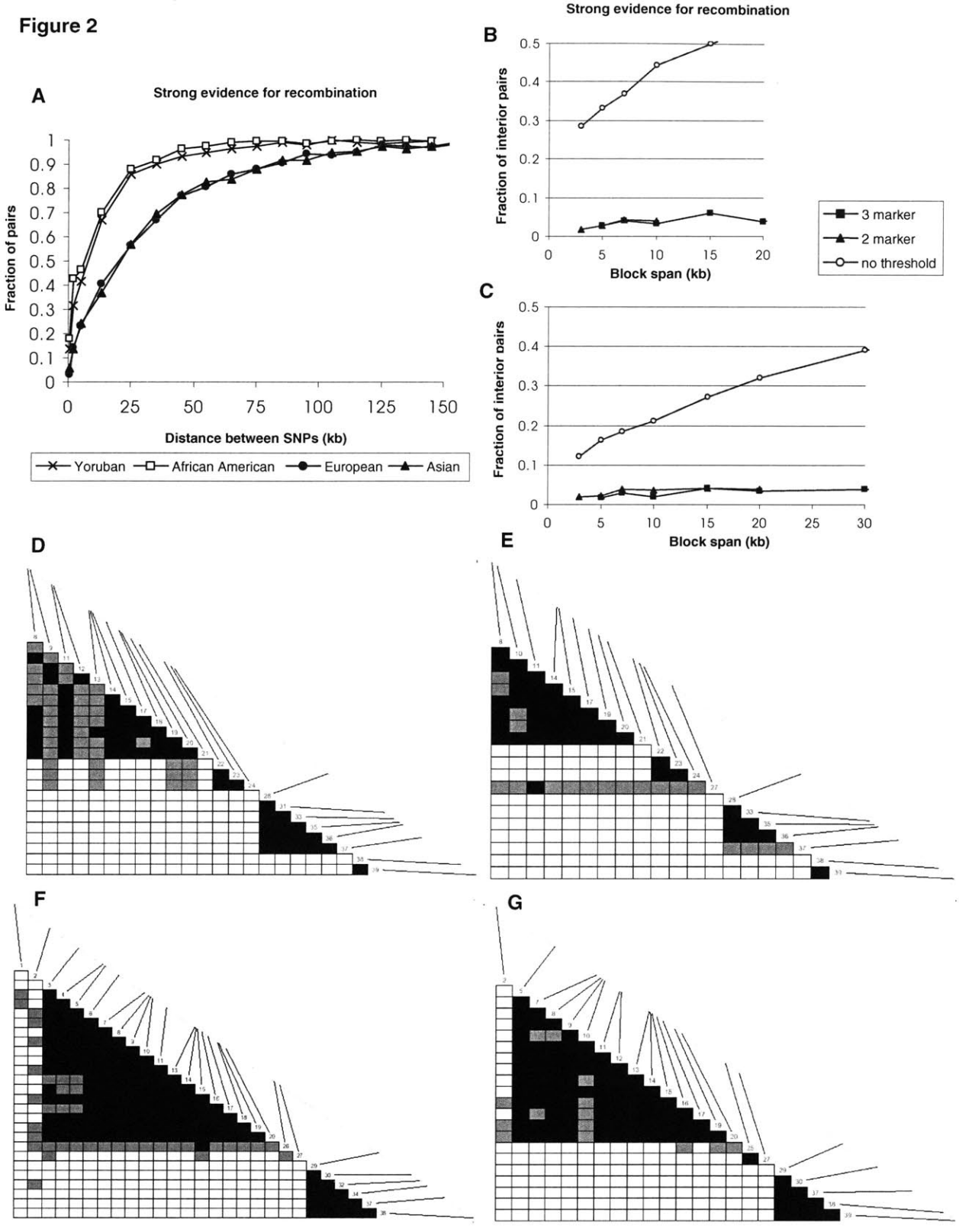

A
Strong evidence for recombination
Fraction of pairs
Distance between SNPs (kb)
Yoruban
African American
European
Asian
B
Strong evidence for recombination
Fraction of interior pairs
Block span (kb)
3 marker
2 marker
no threshold
C
Fraction of interior pairs
Block span (kb)
D
E
F
G

Figure 3

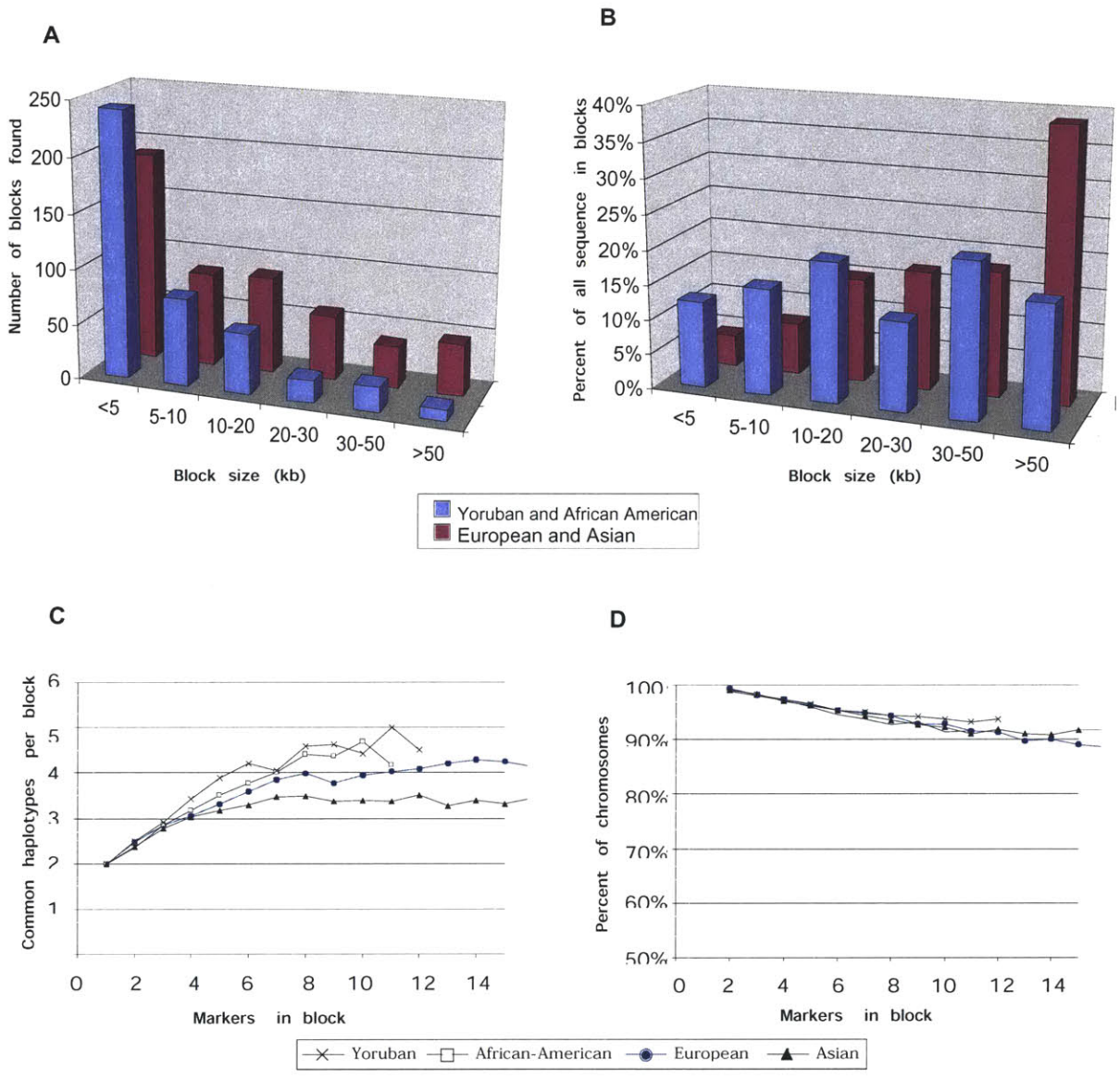
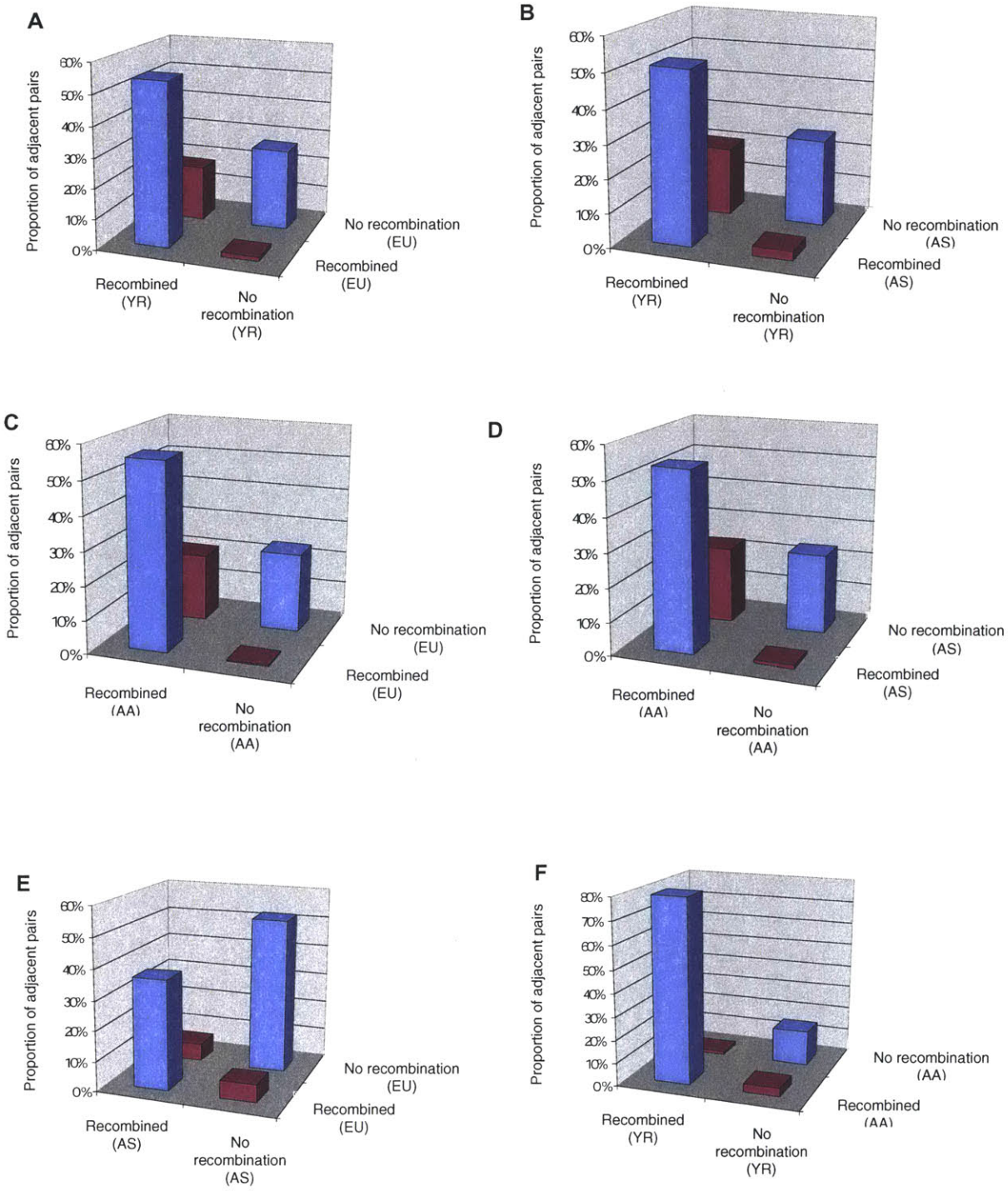
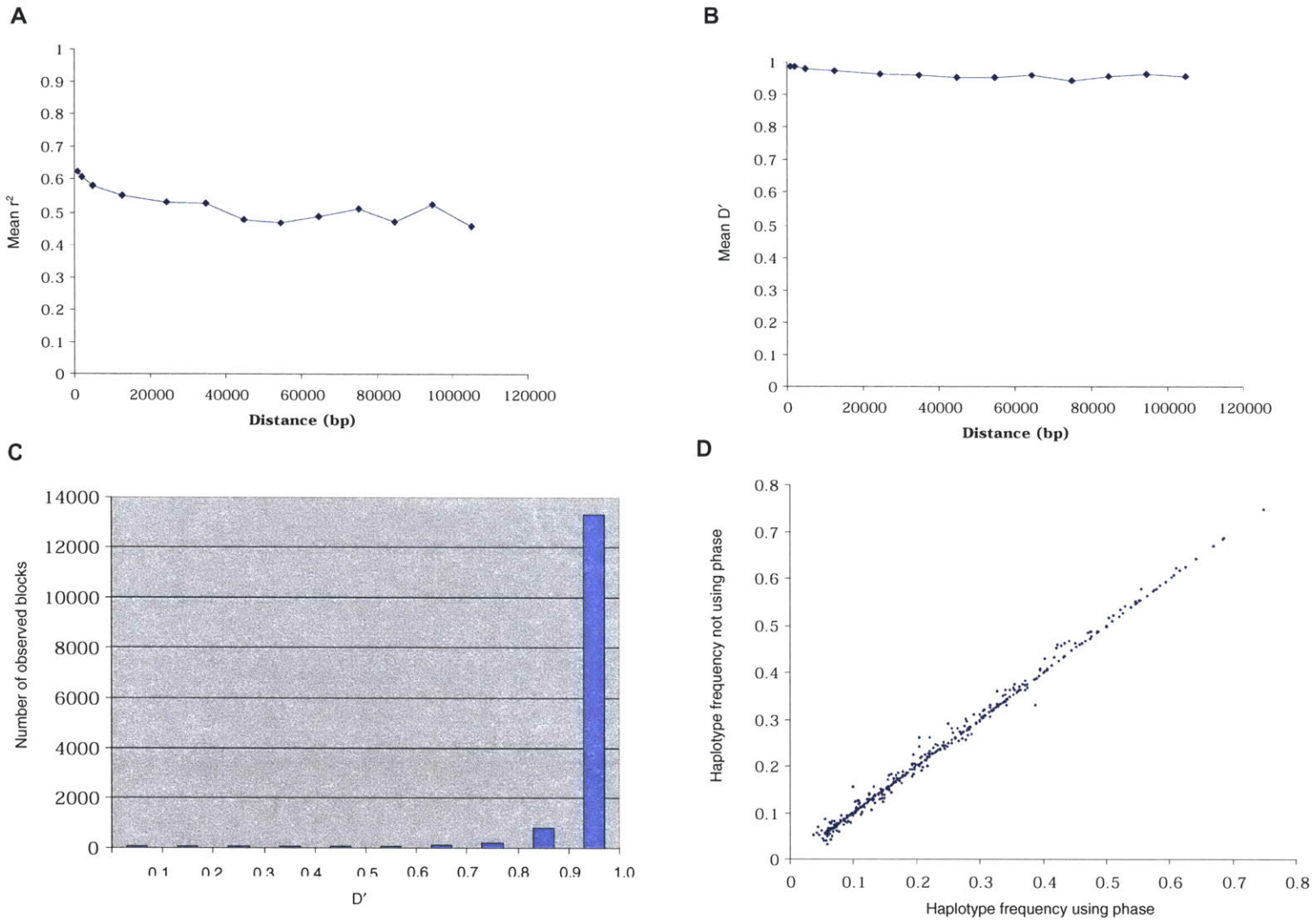

A
B
Number of blocks found
250
200
150
100
50
0
<5
5-10
10-20
20-30
30-50
>50
Block size (kb)
Percent of all sequence in blocks
40%
35%
30%
25%
20%
15%
10%
5%
0%
Yoruban and African American
European and Asian
C
D
Common haplotypes per block
6
5
4
3
2
1
0 2 4 6 8 10 12 14
Markers in block
Percent of chromosomes
100%
90%
80%
70%
60%
50%
Yoruban
African-American
European
Asian

Figure 4


A
Proportion of adjacent pairs
60%
50%
40%
30%
20%
10%
0%
Recombined (YR)
No recombination (YR)
Recombined (EU)
No recombination (EU)
B
Proportion of adjacent pairs
60%
50%
40%
30%
20%
10%
0%
Recombined (YR)
No recombination (YR)
Recombined (AS)
No recombination (AS)
C
Proportion of adjacent pairs
60%
50%
40%
30%
20%
10%
0%
Recombined (AA)
No recombination (AA)
Recombined (EU)
No recombination (EU)
D
Proportion of adjacent pairs
60%
50%
40%
30%
20%
10%
0%
Recombined (AA)
No recombination (AA)
Recombined (AS)
No recombination (AS)
E
Proportion of adjacent pairs
60%
50%
40%
30%
20%
10%
0%
Recombined (AS)
No recombination (AS)
Recombined (EU)
No recombination (EU)
F
Proportion of adjacent pairs
80%
70%
60%
50%
40%
30%
20%
10%
0%
Recombined (YR)
No recombination (YR)
Recombined (AA)
No recombination (AA)

Supplemental Figure 1



Supplemental Table 1

						European sample					Yoruban sample					Asian sample					African American sample				
CHR	Cluster	Primary Cluster [a]	HG0 position		cluster size (kb)	SNPs attempted	Successfully genotyped	Percent successful [b]	Number polymorphic	Density of polymorphisms	SNPs attempted	Successfully genotyped	Percent successful	Number polymorphic	Density of polymorphisms	SNPs attempted	Successfully genotyped	Percent successful	Number polymorphic	Density of polymorphisms	SNPs attempted	Successfully genotyped	Percent successful	Number polymorphic	Density of polymorphisms
1	1a	1	2,920,112	3,076,857	157	57	49	86.0%	39	3.6	58	50	86%	43	3.5	58	43	74.1%	34	4.4	58	41	70.7%	38	3.9
1	1b	1	3,626,300	3,678,525	52	18	15	83.3%	11	4.7	17	16	94%	8	6.4	17	16	94.1%	8	6.4	17	16	94.1%	10	5.1
1	2a	2	4,487,513	4,803,009	315	91	69	75.8%	62	4.1	86	63	73%	56	4.2	86	68	79.1%	56	4.6	86	69	80.2%	62	4.2
1	2b	2	4,496,740	4,520,067	23	15	13	86.7%	13	1.8	14	13	93%	10	2.3	14	14	100.0%	9	2.6	14	14	100.0%	13	1.9
1	3a	3	74,754,000	75,073,226	319	118	98	83.1%	79	4.0	129	80	62%	67	4.8	128	86	67.2%	55	5.8	128	77	60.2%	69	4.6
1	4a	4	77,419,928	77,844,362	424	125	90	72.0%	74	5.7	120	86	71%	75	5.7	119	90	75.6%	57	7.4	119	86	72.3%	77	5.6
1	5a	5	176,405,899	176,790,351	384	85	10	11.8%	5	0.0	85	9	11%	7	0.0	85	5	5.9%	3	0.0	85	8	9.4%	4	0.0
1	6a	6	223,918,761	224,061,673	143	69	52	75.4%	48	3.0	71	53	75%	45	3.0	71	52	73.2%	44	3.2	71	54	76.1%	50	2.9
1	7a	7	226,528,186	226,708,716	181	65	52	80.0%	36	4.8	70	58	83%	47	3.8	70	56	80.0%	39	4.5	70	55	78.6%	46	3.9
1	7b	7	267,669,212	267,713,505	54	18	12	66.7%	12	4.5	18	13	72%	13	4.2	18	14	77.8%	11	4.9	18	13	72.2%	13	4.2
2	8a	8	12,852,989	13,164,396	311	81	58	85.3%	40	3.9	89	52	83%	43	3.8	81	62	76.5%	36	4.3	81	59	72.8%	49	3.3
2	9a	9	19,130,511	19,277,935	147	44	37	84.1%	26	5.7	75	65	87%	53	2.8	75	59	78.7%	36	4.1	75	61	81.3%	47	3.1
2	10a	10	66,248,197	66,568,017	320	72	59	81.9%	48	4.7	65	47	72%	39	4.7	59	43	72.9%	33	5.4	59	40	67.8%	33	5.4
2	11a	11	158,261,352	158,355,714	94	46	38	82.6%	27	3.3	50	40	80%	27	3.4	49	43	87.8%	30	3.1	49	41	83.7%	33	2.7
2	12b	12	158,701,947	158,726,023	24	16	10	62.5%	10	2.4	20	11	55%	10	2.5	22	19	86.4%	18	1.5	22	18	81.8%	17	1.6
2	13a	13	160,098,173	160,246,905	148	22	21	95.5%	14	10.3	83	49	78%	47	3.6	63	54	85.7%	37	4.6	63	53	84.1%	48	3.5
2	14a	14	232,488,231	232,632,052	144	39	32	82.1%	25	5.8	86	60	71%	46	3.7	83	63	75.9%	41	4.1	83	61	73.5%	49	3.5
2	15a	15	233,372,540	233,511,284	139	51	37	72.5%	34	4.1	53	42	79%	32	5.6	53	45	84.9%	32	5.6	53	46	86.8%	34	5.3
3	16a	16	68,962,231	69,001,754	40	14	14	100.0%	13	3.0	14	13	93%	11	3.3	14	13	92.9%	11	3.3	14	12	85.7%	10	3.6
3	16b	16	69,051,642	69,157,633	106	62	53	85.5%	34	3.1	61	52	85%	40	2.6	47	37	78.7%	12	5.6	47	38	80.9%	29	3.5
3	17a	17	132,157,977	132,343,483	186	83	70	84.3%	56	3.2	85	63	74%	53	3.8	82	61	74.4%	44	4.0	82	62	75.6%	57	3.2
3	18a	18	201,062,441	201,348,748	286	91	74	81.3%	69	4.1	93	53	57%	38	7.5	81	63	77.8%	51	5.6	81	58	71.6%	51	5.6
4	19a	19	50,023,246	50,476,591	453	105	76	72.4%	60	4.9	101	58	57%	47	6.3	104	77	74.0%	48	6.1	104	71	68.3%	63	4.7
4	20a	20	126,461,792	126,690,397	229	83	74	89.2%	68	3.4	77	44	57%	40	5.7	86	67	77.9%	58	3.9	86	64	74.4%	58	3.9
4	21a	21	207,161,616	207,306,064	144	54	33	61.1%	21	6.5	48	21	44%	15	7.3	54	32	59.3%	18	7.7	54	31	57.4%	23	6.3
4	21b	21	207,257,278	207,280,582	26	15	9	60.0%	7	3.5	15	6	40%	4	4.8	15	9	60.0%	6	4.1	15	9	60.0%	7	3.2
5	22a	22	63,284,503	63,457,630	173	71	63	88.7%	55	3.1	72	55	76%	51	3.4	72	54	75.0%	46	3.8	72	53	73.6%	49	3.6
5	23a	23	135,471,341	135,646,636	175	91	76	83.5%	51	3.4	93	72	77%	61	2.8	93	75	80.6%	55	3.2	93	74	79.6%	62	2.7
6	24a	24	21,507,393	21,951,398	444	148	100	67.6%	85	5.2	147	101	69%	69	5.0	142	99	69.7%	79	5.6	142	91	64.1%	83	5.4
6	25a	25	88,282,463	88,312,980	31	16	14	87.5%	7	4.0	19	14	74%	10	2.6	19	12	63.2%	6	3.3	19	13	68.4%	10	2.9
6	25b	25	88,340,641	88,395,014	54	21	16	76.2%	14	3.9	23	21	91%	19	2.9	23	14	60.9%	12	4.2	23	15	65.2%	14	3.6
6	26a	26	185,214,680	185,384,850	170	63	57	90.5%	47	3.6	72	64	89%	56	2.9	67	53	79.1%	37	4.3	67	55	82.1%	49	3.9
7	27a	27	50,099,452	50,383,684	284	75	63	84.0%	51	6.3	90	70	78%	59	5.5	77	55	71.4%	48	6.7	77	59	76.6%	50	6.4
7	28a	28	127,720,132	128,184,792	465	113	81	71.7%	65	6.2	119	84	71%	73	5.5	117	79	67.5%	64	6.3	117	78	66.7%	67	6.0
8	29a	29	36,946,115	37,417,430	471	110	94	85.5%	74	3.3	101	78	77%	54	4.6	107	87	81.3%	58	4.3	107	81	75.7%	74	3.3
8	30a	30	106,163,223	106,340,850	178	83	70	84.3%	58	3.1	70	30	43%	25	6.6	55	29	52.7%	17	9.7	55	28	50.9%	23	7.7
9	31a	31	29,116,179	29,291,103	175	33	25	75.8%	17	8.7	32	23	72%	19	9.6	32	25	78.1%	18	10.1	32	23	71.9%	18	10.1
9	31b	31	29,555,787	29,658,244	102	55	46	83.6%	33	3.0	53	35	66%	26	3.7	53	35	66.0%	25	4.0	53	31	58.5%	25	4.0
9	32a	32	113,497,641	113,930,995	433	124	94	75.8%	77	6.0	121	77	64%	58	8.0	120	92	76.7%	68	6.7	120	91	75.8%	80	5.8
10	33a	33	27,036,716	27,229,783	193	71	59	83.1%	51	3.8	72	35	49%	29	6.7	72	57	79.2%	43	4.5	72	50	69.4%	43	4.5
10	33b	33	27,346,240	27,385,275	39	11	11	100.0%	7	3.8	3	1	33%	1	0.0	3	3	100.0%	1	0.0	3	2	66.7%	2	11.2
10	34a	34	108,148,096	108,622,586	474	98	83	83.8%	61	6.5	73	47	64%	38	9.7	74	56	75.7%	36	10.5	74	50	67.6%	42	9.1
11	35a	35	26,986,814	27,154,841	168	60	50	83.3%	40	4.1	72	59	82%	56	3.0	72	52	72.2%	43	3.8	72	44	61.1%	39	4.3
11	36a	36	104,086,142	104,278,329	192	60	53	88.3%	40	4.8	61	52	85%	47	4.2	61	48	78.7%	28	6.9	61	44	72.1%	39	5.0
12	37a	37	32,950,295	33,105,162	155	65	50	76.9%	44	3.5	64	48	75%	41	4.1	65	47	72.3%	36	4.3	65	47	72.3%	41	3.8
12	38a	38	131,383,027	131,717,127	334	92	75	81.5%	67	5.0	91	74	81%	58	5.8	90	77	85.6%	70	4.8	90	76	84.4%	68	4.9
13	39a	39	33,421,472	33,593,592	172	82	73	89.0%	53	3.3	84	58	69%	50	3.4	86	66	77.6%	43	4.0	86	64	75.3%	61	2.8
13	39b	39	33,458,579	33,509,225	51	13	9	69.2%	6	8.4	11	9	82%	8	6.3	11	9	81.8%	5	9.9	11	9	81.8%	8	6.3
13	40a	40	105,682,085	105,984,226	302	103	82	79.6%	74	4.4	99	74	75%	63	5.1	103	80	77.7%	52	5.7	103	80	77.7%	72	4.2
14	41a	41	81,115,396	81,552,062	437	158	127	80.4%	88	5.0	155	116	75%	94	4.6	162	128	79.0%	65	6.6	162	128	79.0%	104	4.1
15	42a	42	35,320,182	35,730,101	410	126	96	76.2%	78	5.3	118	91	77%	80	5.1	126	99	78.6%	77	5.3	126	100	79.4%	92	4.5
15	43a	43	103,020,578	103,177,405	157	55	44	80.0%	28	5.6	61	50	82%	42	3.7	61	46	75.4%	30	5.2	61	46	75.4%	43	4.1
16	44a	44	65,828,511	65,916,227	88	44	39	88.6%	31	2.8	45	41	91%	36	2.4	45	36	80.0%	24	3.7	45	36	80.0%	35	2.5
16	44b	44	65,920,138	66,044,818	125	54	49	90.7%	35	3.4	55	49	89%	42	3.0	55	48	87.3%	32	3.7	55	45	81.8%	40	3.1
17	45a	45	36,100,799	36,275,161	174	77	60	77.9%	51	3.4	77	67	87%	58	3.1	73	57	78.1%	45	4.0	73	51	69.9%	47	3.8
18	46a	46	24,252,610	24,552,651	300	90	78	86.7%	61	4.9	84	64	76%	53	5.3	82	73	89.0%	54	5.1	82	73	89.0%	62	4.5
19	47a	47	7,304,781	7,480,473	176	76	47	61.8%	34	5.0	76	42	55%	32	5.4	74	52	70.3%	36	4.8	74	52	70.3%	40	4.3
20	48a	48	12,459,462	12,637,757	178	68	58	85.3%	45	4.0	76	61	80%	44	4.1	63	56	88.9%	33	5.4	63	55	87.3%	44	4.1
21	49a	49	10,002,212	10,171,621	169	45	25	55.6%	8	17.6	64	27	42%	15	11.4	64	31	48.4%	13	11.2	64	26	40.6%	16	10.7
22	50a	50	24,115,889	24,366,984	251	89	72	80.9%	66	3.8	71	59	83%	57	4.4	71	64	90.1%	51	4.9	71	64	90.1%	62	4.1
22	51a	51	32,043,178	32,297,897	255	72	58	80.6%	53	4.8	51	42	82%	37	6.9	51	44	86.3%	39	6.5	51	45	88.2%	41	6.2
22	52a	52	40,077,996	40,315,218	237	74	52	70.3%	44	5.4	55	34	62%	27	8.7	55	45	81.8%	28	8.5	55	46	83.6%	41	5.8
22	53a	53	46,122,491	46,295,651	173	90	78	86.7%	61	4.3	67	54	81%	39	4.6	67	56	83.6%	46	3.8	67	52	77.6%	45	3.9
22	54a	54	46,131,355	46,386,830	255	82	62	75.6%	54	4.7	67	58	87%	38	6.7	67	59	88.1%	46	5.5	67	59	88.1%	50	5.1
nary statistics					13,473	4398	3444	79%	2745	4.6	4419	3152	72%	2602	4.7	4363	3073	77%	2338	5.1	4363	3193	73%	2771	4.5

CHAPTER 5

Linkage Disequilibrium and Recombination on the X Chromosome Support a Role for Recombination Hotspots

Shau Neen Liu-Cordero, Andrew Kirby, Brendan Blumenstiel, Matthew DeFelice, Stephen F. Shaffner, Stacey Gabriel, Mark Batzer, Eric S. Lander and Mark J. Daly

Contributions: Mark Daly and I conceived of the project and were responsible for the experimental design. Dan Richter retrieved and assembled SNP information from dbSNP. I characterized the Alu insertion polymorphisms and chose and tested all the SNPs. I performed all data collection, except the slides for Sequenom mass spec were spotted and run by Brendan Blumenstiel and Matthew DeFelice. Andrew Kirby, Mark Daly, Steve Shaffner and I performed data processing and analysis.

Introduction

Notwithstanding tremendous effort, massive accumulation of resources, and significant advances in technology, the basic determinants of common disease have continued to elude discovery. Linkage disequilibrium (LD) mapping remains a method that holds a great deal of promise for the mapping and cloning of complex trait mutations. In recent years, a large number of studies have contributed to a dramatic increase in the understanding of the patterns of LD and haplotype diversity in the human genome. A significant accumulation of studies looking in and around genes have demonstrated that pairwise LD is extremely variable across all distances (Abecasis et al., 2001; Ardlie et al., 2002; Bonnen et al., 2000; Clark, 1999; Dunning et al., 2000; Kidd et al., 2000; Reich et al., 2001; Rieder et al., 1999; Stephens et al., 2001; Weiss and Clark, 2002). Even at extremely short distances, pairs of markers fail to show complete LD (Ardlie et al., 2001; Frisse et al., 2001). Although the explanation for this phenomenon is likely to be gene conversion and not meiotic crossover events, it illustrates that this variability in pairwise LD is likely to affect detection of associations for all markers regardless of proximity to a causative mutation. The high level of variation in pairwise LD measures suggests that single markers in LD with a disease mutation will not be sufficient for detection of significant associations for most common diseases.

The power to detect associations relies on the individual characteristics of each marker. Even markers in close proximity to one another often possess different frequencies and population histories. The simple explanation of this variability is that pairwise estimates of LD do not fully describe the underlying haplotype structure. A specific SNP pair is likely to exist on multiple copies of longer haplotypes. The relevant unit of genome structure in the human

population is not an individual SNP, SNP pair, or even a gene. Dense collections of SNPs and haplotypes consisting of larger numbers of markers would provide more information about the fundamental structure of a genomic region.

A number of recent studies using dense sets of markers in specific genomic regions demonstrate that LD is not a monotonic function of distance between markers (Daly et al., 2001; Goldstein, 2001; Jeffreys et al., 2001; Johnson et al., 2001; Patil et al., 2001; Rioux et al., 2001; Taillon-Miller et al., 2000). The extent of LD does not exhibit a simple relationship with any one specific parameter due to the extremely variable local gene genealogies and unique population histories. The results of these studies suggested a new model for the haplotype structure in humans. The genome appears to be subdivided into blocks of extremely limited haplotype diversity. Within these blocks, LD tends to be strong and significant over long stretches of genomic sequence. Interspersed between these blocks are regions exhibiting little or no LD across their length. The lack of LD in these interspersed regions has been postulated to be due to clustered crossover events occurring at recombination hotspots. However, a single recombination event that happened early in the history of a population or other population processes cannot be completely ruled out as the basis for these observations. The exact nature of these apparent hotspots and their generality throughout the genome must be investigated in more depth.

The advent of an extremely dense SNP map throughout the genome has allowed for a study of the generality of the high-resolution haplotype structure found in individual genomic regions (Sachidanandam et al., 2001). The current number of discovered SNPs totals more than 2 million of the approximately 10 million expected common SNPs that exist throughout the genome. The number of SNPs with known exact locations is increasing rapidly as the amount of

finished genome sequence accumulates. Chapter 4 describes a study in which we took advantage of the dense SNP map to confirm the generality of this block-like haplotype structure (Gabriel et al., 2002). Haplotype patterns in 51 regions were systematically characterized in samples from Europe, Africa and Asia. The same phenomenon of haplotype blocks based on patterns of historical recombination was observed across these regions. Haplotype blocks averaged 3-5 common haplotypes across an average block span of 22 kilobases. Haplotype diversity was slightly greater and block size shorter in individuals with recent African ancestry. Even though pairs of markers show a great range of variability in general, this is not true for pairwise comparisons within blocks. Therefore, the same block-like structure emerges when pairwise comparisons are examined in a dense set of markers. These blocks correspond to the regions of limited haplotype diversity. Pairs of markers that span a hotspot show a low level of LD even if they are in very close proximity to one another. It is evident why many studies exhibited variable LD across genes. Block boundaries can fall within a gene, making blocks a more fundamental unit of inheritance in the human population.

The study described in this chapter focuses on genomic regions on the X chromosome, which contain polymorphic Alu insertions in a population of CEPH males. By utilizing the hemizygosity of males for the X chromosome, it is possible to collect unambiguous haplotypes over long stretches. One problem with constructing haplotypes using mother, father, child trios, or even larger families, is that there will be some cases where phase cannot be assigned due to multiple heterozygous genotypes. Therefore it is very difficult to construct haplotypes consisting of many markers. The use of regions surrounding recent polymorphic Alu insertion events provides a way to track specific recombination events. The combination of these two design features, X chromosomes in males and polymorphic Alu insertions allow us to address two main

issues, the patterns of LD and haplotype structure on the X chromosome and the distribution of recombination events.

The X chromosome possesses a unique population history. Studies of polymorphism on the X chromosome reveal a much lower nucleotide diversity than the autosomes (Kaessmann et al., 1999; Nachman et al., 1998; Sachidanandam et al., 2001). One factor that can explain the lower rate of polymorphism on the X chromosome is the lower effective population size (N_e), which 3/4 that of the autosomes since males are hemizygous. Another factor that may contribute is a lower mutation rate (μ). The mutation rate is higher in male than in female meiosis, with the ratio of male to female mutation rate being approximately 1.7 to 1 (Bohossian et al., 2000) . In addition, only 1/3 of the X chromosomes undergo male meiosis. Both of these variables, effective population size and mutation rate, describe the population genetic parameter $\Theta=4N_e\mu$. Taking these factors into consideration, the observed nucleotide heterozygosity (π) on the X chromosome is not that far from what might be expected. The observed value of π, an estimator of the neutral theory population parameter Θ, is $4.69x10^{-4}$ or 61% of the average value for the autosomes (Sachidanandam et al., 2001). It is unclear that a lower polymorphism rate on the X chromosome indicates a greater level of LD than on the autosomes. If polymorphism rate correlates with a shared segment of the genome in the population, a greater extent of LD may be expected on the X chromosome. Long stretches of either high or low sequence diversity were found in a genome wide assessment of human sequence variation in which the distribution of millions of SNPs were analyzed (Reich et al., 2002). The primary determinant of this pattern was shown to be similar gene history and the authors concluded that recombination hotspots contribute significantly to this pattern. Given this observation, it might also be expected that longer LD should exist on the X. Selection pressures on the X chromosome may also suggest

longer tracts of haplotypes and lower haplotype diversity. The description of the general nature of haplotype blocks in the genome contained in Chapter 4 did not include an analysis of the X chromosome. In order to compare the X chromosome to the already existing autosomal data, this chapter describes the patterns of LD and haplotype diversity for six X chromosome regions, each region spanning more than 200 kb.

Another aspect of the haplotype structure that was not addressed in Chapter 4 is the distribution of recombination events and the nature of the presumed recombination hotspots. In particular it is unclear to what extent crossovers are clustered at these hotspots (Petes, 2001; Shiroishi et al., 1995). A random collection of chromosomes from a population will represent a large extent of the genealogical history. Therefore, only "historical" recombination hotspots can be resolved. Specific recombination events between two chromosomes cannot be resolved in this manner. Although the blocks of LD are most likely caused by repeated crossover events at specific hotspots, other events could be responsible such as selection, population substructure, or a lucky crossover event early in the history of a population. The insertion of an Alu repeat onto one specific haplotype in a population provides a unique event in the history of that chromosome. Any differences in haplotypes from that point on in history must be due to mutation events or, more predominantly, recombination events. Therefore, it is possible to limit the section of the genealogical history that is being examined, especially if the insertion occurred relatively late in the history of the population.

The vast majority of Alu elements that exist in the human genome were inserted prior to the expansion of the human population, and are therefore found in all humans. However, there are several younger subfamilies that comprise the majority of polymorphic Alu insertion sites in the human population (Carroll et al., 2001; Roy et al., 2000). The recent introduction of these

subfamilies into the human genome makes them useful for studies within or between human population. There is an inverse correlation between the age of the Alu subfamily and the percentage of polymorphic elements it contains. The Ya5 and Yb8, and Yc1/2 subfamilies have the highest rate of polymorphism and the highest activity with respect to retrotransposition population (Roy-Engel et al., 2001). In this study we chose regions surrounding high frequency polymorphic Alu insertions belonging to these subfamilies. By choosing these most recent Alu insertions, it is possible to increase the chance of finding an insertion polymorphism that occurred within the recent history of our population sample. In achieving this, we could examine a much more restricted portion of the genealogical history, label a specific haplotype permanently and unambiguously, and visualize specific recombination events in the region surrounding the insertion.

Specific recombination events surrounding mutations that arose on a single ancestral chromosome have been seen in a number of rare monogenic diseases (Hastbacka et al., 1992; Hoglund et al., 1995; Richter et al., 1999). However, these are only rare events in very special populations. We wanted to explore the possibility of using recent Alu insertions as a tool to look at actual recombination events in the general population for a set of non affected chromosomes. We assume that haplotype blocks are showing us that recombination has occurred between blocks, but we don't know how much has occurred. It would be advantageous to find a way to quantify the amount of recombination that is actually occurring. Our aim is to demonstrate that recombination is causing hotspots by clustering of crossover events to specific intervals in regions containing recent Alu insertions.

Materials and Methods

Individuals

Males from the CEPH pedigrees were used in this analysis. We chose the males from the total set that would provide the maximum number of independent X chromosomes to genotype. We used 154 males, which for the most part consisted of trios of father, paternal grandfather, and maternal grandfather, none of whom share a common X chromosome. A number of CEPH families are related through a number of individuals. Males who were related from these families were eliminated from the study. A significant number of the CEPH males used in this study are the same or come from the same families as those used in the large haplotype study in Chapter 4.

Polymorphic Alu insertions

The polymorphic Alu insertions were discovered as described in (Roy-Engel et al., 2001) (Carroll et al., 2001). A relatively small number of the polymorphic insertions were discovered on the X chromosome. An even fewer number existed at a high enough frequency in the Caucasian population to be useful. The insertions were genotyped in a panel of CEPH reference DNAs in order to determine allele frequencies if they were polymorphic at all in the CEPH population. Several insertions that had been found to have a high allele frequency in Caucasians were found to be monomorphic in our CEPH population. In addition, insertions which were contained in regions not possessing a high density of discovered SNPs were also excluded from the study. The Alu insertions were amplified under the conditions described in (Roy-Engel et al., 2001) and genotyped on 3% agarose gels.

Marker Selection

SNPs were selected from dbSNP, a comprehensive database of discovered SNPs and their coordinates in the genome sequence. Markers were selected from dbSNP 300 kb regions surrounding the Alu insertions, 150 kb on either side. A total of 442 markers were ordered. A number of markers were eliminated from the study as their positions in the genome sequence were shifted substantially in subsequent releases of the golden path (genome sequence). Approximately equal numbers of SNPs were found by The SNP Consortium (TSC) and discovered in BAC overlap sequence. Primers and probes were designed in multiplex format (average 4.3-fold multiplexing) using SpectroDESIGNER software (Sequenom, San Diego, CA). Markers were eliminated from use in the LD study if they were monomorphic (approx. 23%), contained greater than 1 heterozygote in the male samples (11.3%), or were missing more than 30% of the genotypes (18.1%).

SNP Genotyping

Multiplex PCR was performed in five microliter volumes containing 0.1 units of Taq polymerase (Amplitaq Gold, Applied Biosystems), 5 ng genomic DNA, 2.5 pmol of each PCR primer, and 2.5 μmol of dNTP. Thermocycling was at 95 C for 15 minutes followed by 45 cycles of 95 C for 20 s, 56 C for 30s, 72 C for 30 s. Unincorporated dNTPs were deactivated using 0.3U of Shrimp Alkaline Phosphatase (Roche) followed by primer extension using 5.4 pmol of each primer extension probe, 50 μmole of the appropriate dNTP/ddNTP combination, and 0.5 units of Thermosequenase (Amersham Pharmacia). Reactions were cycled at 94 C for 2 minutes, followed by 40 cycles of 94 degrees for 5 s, 50 degrees for 5 s, 72 degrees for 5 s. Following addition of a cation exchange resin to remove residual salt from the reactions, 7

nanoliters of the purified primer extension reaction was loaded onto a matrix pad (3-hydroxypicoloinic acid) of a SpectroCHIP (Sequenom, San Diego, CA). SpectroCHIPs were analyzed using a Bruker Biflex III MALDI-TOF mass spectrometer (SpectroREADER, Sequenom, San Diego, CA) and spectra processed using SpectroTYPER (Sequenom). 10 individuals were genotyped on two separate source plate, and were therefore genotyped twice. The error rate for these samples was extremely low. There were only 3 discrepant genotypes out of 3251 genotype calls.

Statistical Analysis

Computation of D', analysis of haplotype blocks and haplotype diversity are as described in Chapter 4 (Gabriel et al., 2002).

Results

A survey of LD in six regions on the X chromosome

In order to characterize LD, haplotype structure, and patterns of recombination on the X chromosome, we selected six regions designed to flank polymorphic Alu insertions by 150 kb on either side. SNPs were selected within these regions from dbSNP and included TSC SNPs as well as SNPs discovered in BAC overlaps (as described in the Materials and Methods). We chose a male CEPH population sample consisting of 154 individuals possessing independent X chromosomes. Many of these individuals were the same as used in Chapter 4. Determining SNP genotypes is an easier and robust process when all individuals are hemizygous for all markers. A total of 410 markers were genotyped, but a large number of markers were not included in the

analysis due to monomorphism or extremely low allele frequency, detected heterozygotes, or high assay failure rate. In many cases markers were lost over the course of the experiment due to changes in the golden path coordinates of the human genome sequence. In all, over 60,000 genotypes were obtained covering approximately 1.4 Megabases of total sequence and 198 of 410 total SNPs were informative and possessed high quality assays.

The exact sizes of the regions and the total number of SNPs found for each region are listed in **Table 1**. An average of 33 SNPs covered an average of around 225 kb. Thus, the average SNP density is therefore one SNP approximately every 7 kb. Although this represents less than half the density of the autosomal study described in Chapter 4, based on the results of the autosomal study, this SNP density was determined to be sufficient. In Chapter 4, if fewer SNPs were used, the same number of blocks would have been identified. Furthermore, we expected longer tracts of LD on the X chromosome, due to the lower polymorphism and recombination rates.

Of the six regions we chose to study, only four of the regions contained a polymorphic Alu insertion. However, when considering recombination events surrounding an insertion event, each side of the insertion sites can be considered to be independent. Therefore a total of 8 independent insertion "sides" were available for study. However, only six were included because two of the sides did not contain a sufficient number of SNPs to yield any informative results. The numbers of SNPs on each side of these 4 insertions are listed in **Table 1**. The number of SNPs per independent side ranges from 6 to 31 with an average of approximately 19 SNPs per side. The allele frequencies for the Alu insertions are also listed in **Table 1**. Two of the insertions, Ya5NBC37 and Ya5DP5, possessed allele frequencies of 36.2% and 23.4%,

respectively. Since these frequencies were under 50%, they were the most likely to be the most recent insertions of the set.

Distribution of SNP allele frequencies and monomorphism on the X chromosome

We examined the distribution of allele frequencies for the SNPs in the 6 regions (**Figure 1**). The addition of the low frequency and monomorphic markers brought the total number of markers to 252. Recent unpublished results show a significantly greater level of monomorphism on the X chromosome as opposed to the autosomes. This disparity between autosomes and the X chromosome was more substantial in Caucasians and Asians as opposed to African and African-American populations. The rate of monomorphism was closer to 40% on the X chromosome in Caucasians, whereas it was 22% for the autosomes (D. Altshuler pers comm.). In this study, we did not observe this disparity. The observed frequency of X chromosome monomorphism in this data set was 25%, indistinguishable from the autosomal frequency. In addition, the rate of lower frequency SNPs seems to be suppressed in this sample. The bins containing allele frequencies up to 20% are much lower on the X chromosome than in the autosomal data (**Figure 1**). Accordingly the higher frequency bins contain many more SNPs. This observation may be due to random fluctuations across the X chromosome. Although the autosomal data consisted only of TSC SNPs and not BAC-overlap SNPs, this should not be the cause of an upward bias in allele frequencies in our study. Only SNPs with an allele frequency of greater than 20% were used in both the X chromosome and autosome studies, so this should not affect any of the LD or recombination analyses. We are simply observing a greater proportion of the high frequency SNPs that exist in these X chromosome regions. Finally, we searched for runs of at least 4 monomorphic SNPs in these regions. This analysis showed similar results for the X

chromosome and the autosomes as well. 1.4% of the X chromosome sequence is contained in such runs, as compared to 1.8% in autosomes.

Pairwise linkage disequilibrium and block structure on the X chromosome

To study the historical recombination in the X chromosome, we measured LD between marker pairs in our dense collection of SNPs. We measured LD using the normalized measure D' which represents historical recombination in a pair of polymorphisms. The distribution of mean D' values for pairs of SNPs spaced at varying distances is shown in **Figure 2a**. Low D' values indicate a substantial amount of recombination between the pairs and a high D' is evidence for a small amount of recombination in the history of the alleles. For this analysis, only high frequency SNPs were examined (minor allele frequency $\geq$0.2). A total of 2,043 total pairwise comparisons were made, and the decline in LD closely resembles that of the autosomes, as shown in Chapter 4. **Figure 2b** shows the fraction of pairs showing strong LD. We used the same definition in Chapter 4 for the threshold of "strong LD" and we observe little difference between the X chromosome and the autosomes in the decline of LD with distance.

We also examined D' values across each of the 6 regions on the X chromosome. An example of a table of D' values is shown for region 2 (Ya5DP5) in **Figure 6b**. The regions with the highest level of LD are shown as red squares and low levels of LD are shown as white squares. A block-like structure similar to that found on the autosomes exists on the X chromosome. A total of 24 blocks were found in the 6 regions, with 136 out of a total of 174 markers analyzed found to exist within these blocks. The number of blocks found per region and percent of sequence found in blocks is shown in **Table 1**. The average amount of sequence found in blocks was 38.3% for the six X chromosome regions as opposed to 40.5% for the

autosomes. The average block size was 21 kb as compared to 22 kb for the autosomes. This initial set of data on pairs of markers strongly suggests that the level of LD on the X chromosome is not significantly different than the autosomes.

Haplotype diversity on the X chromosome

We next examined the haplotype diversity within the defined blocks of LD. Only common haplotypes, those with a frequency of ≥5%, were considered. As in Chapter 4, we could not be sure that the density of markers that we used could fully describe the complete haplotype diversity. Therefore, we plotted the mean number of common haplotypes in each block as a function of the number of markers tested. The mean number of common haplotypes levels out at six to eight SNPs typed in a block **(Figure 3a)**. In other words, the maximum number of haplotypes are observed with as few as 6 to 8 markers. Again, the X chromosome appears to level out at a similar rate to the autosomes. However, slightly lower haplotype diversity is observed for Caucasians on the X chromosome as opposed to autosomes. The X chromosome in Caucasian samples appear to have a similar haplotype diversity to an Asian population on the autosomes. There are 4.1 common haplotypes in the Caucasian sample as opposed to 3.7 in Asians and 4.9 in Africans (data not shown) on the autosomes. On the other hand, the CEPH autosomes and X chromosome are indistinguishable when comparing the fraction of chromosomes against the number of markers defining a block **(Figure 3b)**. In order to further characterize the haplotype diversity on the X chromosome, we looked at the haplotype frequency distribution. **Figure 4** shows the haplotype heterozygosity for the X chromosome blocks and autosomal block plotted for increasing numbers of markers in a block. The heterzygosity is not significantly different between the X and the autosomes.

Apart from a slightly lower haplotype diversity on the X chromosome, we were not able to discern any major differences in LD from the autosomes. One of the main differences between the two studies is that only 6 regions were described on the X chromosome as opposed to 51 regions for the autosomes. In addition, the autosomal data possessed a greater SNP density. However, a lower SNP density on the X chromosome would only underestimate the haplotype diversity and would not explain why we did not observe a greater extent of LD on the X chromosome. The LD data and the haplotype structure observed for the X chromosome further supports a role for recombination hotspots creating a block-like structure. In addition, all the data illustrate that the X chromosome and autosomes exhibit a similar level of LD. Although this result may be surprising, on further reflection this observation may be expected in a model where clustered recombination events are the primary determinant of the haplotype block structure of the genome. This point will be addressed further in the Discussion.

Reduced haplotype diversity and distinct recombination events surrounding recent polymorphic Alu insertions

In order to more closely examine the nature of the recombination hotspots, we subdivided the haplotypes by the presence or absence of the Alu insertion. We looked primarily at the two regions which had an insertion frequency under 50%. These regions were likely to have the most recent insertion events because these frequencies are lower than the others. Since the left and right sides of the Alu can be considered independent, these two regions represent 4 entities with respect to examination of specific recombination events. An Alu will insert onto a single chromosomal haplotype. After that point in history, all inserted chromosomes are derived from that ancestral haplotype since an Alu insertion is a permanent event. Since retrotransposition of

Alu repeats is an extremely rare event, it is not likely to occur twice in the same population. As an example, **Figure 5** shows the reduced haplotype diversity of inserted chromosomes for the left side of the Alu insertion in region Ya5NBC37. The frequency of this Alu insertion in our population sample is 36.2%. We used 29 common SNPs to define the haplotypes and only 8 total haplotypes exist over a span of more than 150kb on the inserted chromosomes. The uninserted chromosomes are comprised of 27 different haplotypes, the most frequent of which is only 12.1% and is not present in any of the inserted chromosomes. All 8 haplotypes on the inserted chromosomes can be derived from a single recombination event between the ancestral inserted haplotype and a common uninserted haplotype. The recombinant haplotypes vary in frequency from 1 to 12 occurrences. Each of the recombinants is likely to only have occurred once at various points in the genealogical history and then expanded to varying degrees. There are an average of 6.5 recombinant chromosomes for all 4 independent sides in the inserted chromosomes and an average of 23 haplotypes in the uninserted chromosomes. The haplotype diversity and haplotype patterns strongly suggest that we are witnessing single recombination events at different distances from an Alu repeat that inserted fairly late in the history of our population sample.

In order to examine the distribution of recombination events, we defined the minimum determinable intervals surrounding each recombination event. Ancestral and recombinant haplotypes are shown in **Figure 6a** for both independent sides of Alu Ya5DP5. The exact boundaries of the location of the recombination event could not always be isolated to one SNP. The first point at which the recombinant chromosome diverges from the ancestral haplotype corresponds to the distal limit of the recombination interval. The proximal limit of the recombination interval was determined by comparing the recombinant haplotype to common

uninserted haplotypes. The interval from the distal limit of the recombinant chromosome all the way to the insertion is identical to the ancestral inserted haplotype, and is referred to as the proximal portion of a recombinant chromosome. The point of divergence between this proximal portion of a recombinant chromosome and the corresponding haplotype in a common uninserted chromosomes represents the proximal limit of the recombination interval. Specific recombination events appear to coincide with the hotspots of historical recombination determined by the breaks between blocks in the table of D' values (**Figure 6b**). The intervals containing recombination events are shown as colored boxes surrounding a variable number of SNPs. Haplotype blocks are labeled in red, and D' values and LOD scores are provided for each SNP pair. Interestingly, it appears as if the Alu repeat inserted into a recombination hotspot. Three clusters of crossover events can be seen. Two are immediately on each side of the Alu insertion, and one is closer to the distal end of the left side of the insertion. These three clusters roughly align with two recombination hotspots determined from the collection of uninserted chromosomes. Therefore, we have shown that breakpoints between haplotype blocks appear to be a result of multiple recombination events as opposed to early crossover events or other aspects of population history.

Discussion

Similar levels of LD and haplotype diversity on the X chromosome and autosomes may support a recombination hotspot model

The primary goal of this study was to examine in more detail the nature of recombination hotspots as well as describe the patterns of LD on the X chromosome. Using half the SNP

density in the X chromosome study as compared to the autosomal study described in Chapter 4, we identified a similar block structure and recombination hotspots. Surprisingly, we found no direct evidence that the extent of LD is different on the X chromosome than the autosomes at the level of pairwise comparisons of markers and at the level of haplotype blocks. This observation can actually be seen as further evidence supporting a model of recombination hotspots in the human genome being the primary determinant of haplotype structure. The X chromosome is known to have a lower polymorphism rate, a lower rate of recombination, and a smaller effective population size. However, recombination may be saturated at recombination hotspots on the autosomes. Therefore, a lower rate of recombination on the X chromosome may be enough to break down haplotype structure to the same extent as the autosomes. If crossover events were not clustered, but instead occurred randomly across the chromosome, then it would be more likely that the lower rate of recombination would create longer tracts of LD on the X chromosome. The lower effective population size may have had some effect of lower the haplotype diversity on the X as seen in **Figure 3a**. Since we observe a lower haplotype diversity but a similar extent of LD, the determinants of these two properties may not be completely overlapping.

From a survey of the distribution of SNPs throughout the genome, Reich et al. concluded that a pattern of extended segments of low and high polymorphism rate is primarily caused by patterns of long stretches of similar gene history (Reich et al., 2002). They also concluded that this pattern is best explained by recombination hotspots. From these conclusions, it may be expected that the X, possessing a lower polymorphism rate, would have longer stretches of LD due to a more similar gene history. However, there are so many factors that affect the gene history of the X, and the effect seen by Reich et al. is not strong enough to definitively show that

this is necessary characteristic of the X chromosome. Only six regions were included in this study, so the effect we observed may be due to fluctuation along the X chromosome. We also cannot rule out that the selection of regions containing Alu insertions had an effect on the results. Alu insertions may be more likely to occur at recombination hotspots or may cause specific regions to assume the properties of a hotspot. It will be necessary to look at a larger number of regions as well as the population distribution of LD and haplotype diversity on the X chromosome.

Specific crossover events surrounding recent polymorphic Alu insertions appear to coincide with breaks in haplotype blocks

Haplotype data has been collected from a significant number of genomic regions, and a general picture of haplotype structure has been formed. The genome appears to be segregated into blocks of high LD and low haplotype diversity separated by recombination hotspots. These breaks between blocks are not necessarily due to clustering of crossover events. None of the studies of haplotype structure can discern a difference between historical recombination hotspots, early recombination events, or population history. The evidence certainly favors a model of clustering of crossover events, but cannot definitively demonstrate actual crossover events. In this data set we observed a similar block-like structure as the previous studies. However the presence of recent Alu insertions in our regions allowed us to further examine the root causes of these hotspots. By observing the haplotypes and their frequencies, we were able to determine that we were most likely examining a limited section of the genealogical history. We observed lower haplotype diversity on inserted chromosomes and visualized individual recombination

events. These results indicated, in at least in two examples, that multiple crossover events occurred in several sites that coincided with the sites of historical recombination hotspots.

It may be possible to generalize this method of analysis to other regions not containing Alus. A common haplotype may be used as an "event" after which recombination events occur around that haplotype. The "event" would be the last mutation that created that haplotype. This method may not have the same level of resolution as this Alu insertion study, but it is worth investigating since there are not a large number of these recent polymorphic Alu insertions in the genome.

In summary, the lack of a higher extent of LD on the X chromosome combined with the correlation of hotspots with multiple specific recombination events provide new evidence in support of a recombination hotspot model. The level of LD on the X chromosome also raises new questions about the effects of saturation at recombination hotspots and the effects of the unique population history of the X chromosome.

References

Abecasis, G. R., Noguchi, E., Heinzmann, A., Traherne, J. A., Bhattacharyya, S., Leaves, N. I., Anderson, G. G., Zhang, Y., Lench, N. J., Carey, A., Cardon, L. R., Moffatt, M. F., and Cookson, W. O. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. Am J Hum Genet *68*, 191-197.

Ardlie, K., Liu-Cordero, S. N., Eberle, M. A., Daly, M., Barrett, J., Winchester, E., Lander, E. S., and Kruglyak, L. (2001). Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. Am J Hum Genet *69*, 582-9.

Ardlie, K. G., Kruglyak, L., and Seielstad, M. (2002). Patterns of Linkage Disquilibrium in the Human Genome. Nature Reviews Genetics *In press*.

Bohossian, H. B., Skaletsky, H., and Page, D. C. (2000). Unexpectedly similar rates of nucleotide substitution found in male and female hominids. Nature *406*, 622-5.

Bonnen, P. E., Story, M. D., Ashorn, C. L., Buchholz, T. A., Weil, M. M., and Nelson, D. L. (2000). Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium. Am J Hum Genet *67*, 1437-51.

Carroll, M. L., Roy-Engel, A. M., Nguyen, S. V., Salem, A. H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M., Watkins, W. S., Henke, J., Makalowski, W., Jorde, L.

B., Deininger, P. L., and Batzer, M. A. (2001). Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J Mol Biol *311*, 17-40.

Clark, A. G. (1999). The size distribution of homozygous segments in the human genome. Am J Hum Genet *65*, 1489-92.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. Nat Genet *29*, 229-232.

Dunning, A. M., Durocher, F., Healey, C. S., Teare, M. D., McBride, S. E., Carlomagno, F., Xu, C. F., Dawson, E., Rhodes, S., Ueda, S., Lai, E., Luben, R. N., Van Rensburg, E. J., Mannermaa, A., Kataja, V., Rennart, G., Dunham, I., Purvis, I., Easton, D., and Ponder, B. A. (2000). The extent of linkage disequilibrium in four populations with distinct demographic histories. Am J Hum Genet *67*, 1544-54.

Frisse, L., Hudson, R. R., Bartoszewicz, A., Wall, J. D., Donfack, J., and Di Rienzo, A. (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am J Hum Genet *69*, 831-43.

Gabriel, S., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumensteil, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S., Ward, R., Lander, E., Daly, M., and Altshuler, D. (2002). The sturcture of haplotype blocks in the human genome. Science *submitted.*

Goldstein, D. B. (2001). Islands of linkage disequilibrium. Nat Genet *29*, 109-11.

Hastbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., and Lander, E. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet *2*, 204-11.

Hoglund, P., Sistonen, P., Norio, R., Holmberg, C., Dimberg, A., Gustavson, K. H., de la Chapelle, A., and Kere, J. (1995). Fine mapping of the congenital chloride diarrhea gene by linkage disequilibrium. Am J Hum Genet *57*, 95-102.

Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet *29*, 217-22.

Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C., Clayton, D. G., and Todd, J. A. (2001). Haplotype tagging for the identification of common disease genes. Nat Genet *29*, 233-7.

Kaessmann, H., Heissig, F., von Haeseler, A., and Paabo, S. (1999). DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat Genet *22*, 78-81.

Kidd, J. R., Pakstis, A. J., Zhao, H., Lu, R. B., Okonofua, F. E., Odunsi, A., Grigorenko, E., Tamir, B. B., Friedlaender, J., Schulz, L. O., Parnas, J., and Kidd, K. K. (2000). Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. Am J Hum Genet *66*, 1882-99.

Nachman, M. W., Bauer, V. L., Crowell, S. L., and Aquadro, C. F. (1998). DNA variability and recombination rates at X-linked loci in humans. Genetics *150*, 1133-41.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P., and Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science *294*, 1719-23.

Petes, T. D. (2001). Meiotic recombination hot spots and cold spots. Nat Rev Genet *2*, 360-9.

Reich, D., Schaffner, S., MJ, D., McVean, G., Mullikin, J., Richter, D., Lander, E., and Altshuler, D. (2002). Genome-wide assessment of human genome sequence variation indicates a major role for recombination hotspots. Nature *submitted.*

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001). Linkage disequilibrium in the human genome. Nature *411*, 199-204.

Richter, A., Rioux, J. D., Bouchard, J. P., Mercier, J., Mathieu, J., Ge, B., Poirier, J., Julien, D., Gyapay, G., Weissenbach, J., Hudson, T. J., Melancon, S. B., and Morgan, K. (1999). Location score and haplotype analyses of the locus for autosomal recessive spastic ataxia of Charlevoix-Saguenay, in chromosome region 13q11. Am J Hum Genet *64*, 768-75.

Rieder, M. J., Taylor, S. L., Clark, A. G., and Nickerson, D. A. (1999). Sequence variation in the human angiotensin converting enzyme. Nat Genet *22*, 59-62.

Rioux, J. D., Daly, M. J., Silverberg, M. S., Lindblad, K., Steinhart, H., Cohen, Z., Delmonte, T., Kocher, K., Miller, K., Guschwan, S., Kulbokas, E. J., O'Leary, S., Winchester, E., Dewar, K., Green, T., Stone, V., Chow, C., Cohen, A., Langelier, D., Lapointe, G., Gaudet, D., Faith, J., Branco, N., Bull, S. B., McLeod, R. S., Griffiths, A. M., Bitton, A., Greenberg, G. R., Lander, E. S., Siminovitch, K. A., and Hudson, T. J. (2001). Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet *29*, 223-8.

Roy, A. M., Carroll, M. L., Nguyen, S. V., Salem, A. H., Oldridge, M., Wilkie, A. O., Batzer, M. A., and Deininger, P. L. (2000). Potential gene conversion and source genes for recently integrated Alu elements. Genome Res *10*, 1485-95.

Roy-Engel, A. M., Carroll, M. L., Vogel, E., Garber, R. K., Nguyen, S. V., Salem, A. H., Batzer, M. A., and Deininger, P. L. (2001). Alu insertion polymorphisms for the study of human genomic diversity. Genetics *159*, 279-90.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., and Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928-33.

Shiroishi, T., Koide, T., Yoshino, M., Sagai, T., and Moriwaki, K. (1995). Hotspots of homologous recombination in mouse meiosis. Adv Biophys *31*, 119-32.

Stephens, J. C., Schneider, J. A., Tanguay, D. A., Choi, J., Acharya, T., Stanley, S. E., Jiang, R., Messer, C. J., Chew, A., Han, J. H., Duan, J., Carr, J. L., Lee, M. S., Koshy, B., Kumar, A. M., Zhang, G., Newell, W. R., Windemuth, A., Xu, C., Kalbfleisch, T. S., Shaner, S. L., Arnold, K., Schulz, V., Drysdale, C. M., Nandabalan, K., Judson, R. S., Ruano, G., and Vovis, G. F. (2001). Haplotype variation and linkage disequilibrium in 313 human genes. Science *293*, 489-93.

Taillon-Miller, P., Bauer-Sardina, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. P., and Kwok, P. Y. (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28 [see comments]. Nat Genet *25*, 324-8.

Weiss, K. M., and Clark, A. G. (2002). Linkage disequilibrium and the mapping of complex human traits. Trends Genet *18*, 19-24.

Table 1 Summary of X chromosome regions and SNPs. The details of six genomic regions on the X chromosome are provided. Average SNP density for each region is calculated from the number of SNPs and the total sequence spanned by those SNPs. Only SNPs that passed the quality tests outlined in the Materials and Methods are included. The data on haplotype blocks is also listed. The number of blocks and the proportion of the sequence that exists in blocks are shown for each region. The frequencies of the Alu insertions are also provided.

Figure 1 SNP allele frequency distribution for X chromosome markers. 252 total SNPs were examined in 154 chromosomes from a population of CEPH males. Monomorphic markers were included.

Figure 2 Linkage disequilibrium among pairs of markers on the X chromosome. (A) Mean D' for marker pairs spaced at varying distance. 2,043 total pairwise comparisons were made. (B) Proportion of informative SNP pairs that display evidence for strong LD based on confidence intervals on D'.

Figure 3 Haplotype blocks on the X chromosome compared to autosomes. (A) Haplotype diversity for X chromosome blocks in CEPH samples and autosomal blocks in CEPH and Asian samples. The number of common (≥5%) haplotypes per block. (B) The fraction of all chromosomes accounted for by these common haplotypes is plotted as a function of the number of markers typed in each block.

Figure 4 Haplotype heterozygosity on the X chromosome compared to autosomes. Haplotype heterozygosity for X chromosome and autosome blocks plotted as a function of the number of markers typed in each block.

Figure 5 Haplotype frequencies of inserted and uninserted chromosomes. The left side of the region containing the Ya5NBC37 Alu insertion is described as an example of reduced haplotype diversity in chromosomes with a recent insertion event. The ancestral and all recombinant chromosomes are listed for the insertion chromosomes with their frequencies in our population sample. Only 8 total chromosomes exist for the inserted chromosomes as opposed to 27 for the uninserted chromosomes. The ancestral chromosome and regions derived from the ancestral chromosome are shown in red. The blocks of color surrounding other portions of the chromosomes are provided to illustrate all the differences between the recombinant chromosomes. However, each one of these recombinants can be explained by a single recombination event between the ancestral insertion chromosome and a common uninserted haplotype. The most frequent uninserted haplotype is shown in blue letters and only has a frequency of 12.1%.

Figure 6 Observed crossover events coincide with historical recombination hotspots. The region containing the Ya5DP5 Alu insertion is described (A) The ancestral and recombinant haplotypes for each side of the Alu insertion are listed. The minimal interval that could be determined for each recombination event is shown as different colored boxes. All the SNPs are lined up between (A) and (B). The table of D' values for all pairwise comparisons is shown in (B). Each D' value is shown with its corresponding LOD score. The highest values of D' are shown as red boxes. As the D' value gets lower, the boxes are shown as less red. White boxes represent strong evidence for recombination between the markers. Haplotype blocks are indicated as red lines. Green arrows point to hotspots of historical recombination. Black arrows point from the clustered recombination events to their corresponding historical recombination hotspots.

Table 1

Region	Name	# SNPs	SNPs Left of Alu Insertion	SNPs Right of Alu Insertion	Sequence Spanned (bp)	Averge SNP Density per kb	Blocks Found	Sequence Found in Blocks (bp)	% Sequence Found in Blocks	Alu Insertion Frequency
1	YA5NBC37	37	31	6	246568	6.66	3	104345	42.4	36.2
2	YA5DP5	34	19	15	227017	6.68	6	66483	29.3	23.4
3	YA5NBC98	34	14	20	282042	8.30	6	168467	59.7	72.9
4	YC21RG99	36	16	10	207380	5.76	4	53731	25.9	72.1
5	YB8NBC8	20	20	0	129664	6.48	3	53731	54.7	Monomorphic
6	YA5NBC313	37	31	6	277016	7.49	2	61328	22.1	Monomorphic
	Total	198			1369687	6.92	24	525231	38.3	

Figure 1

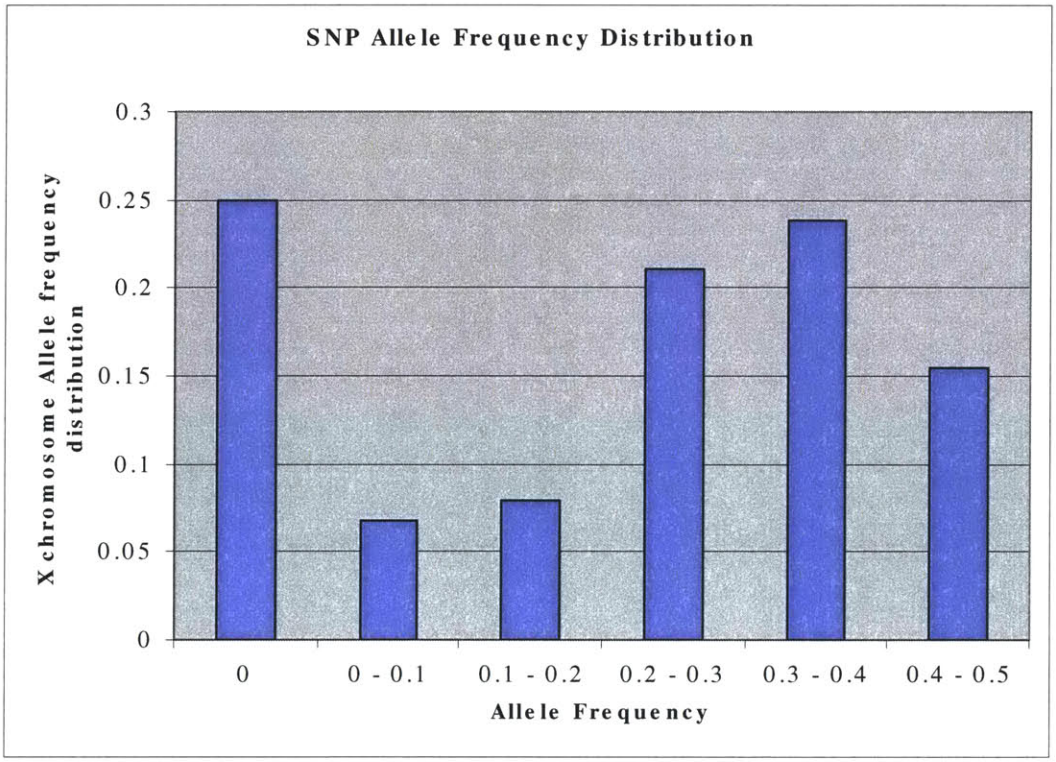
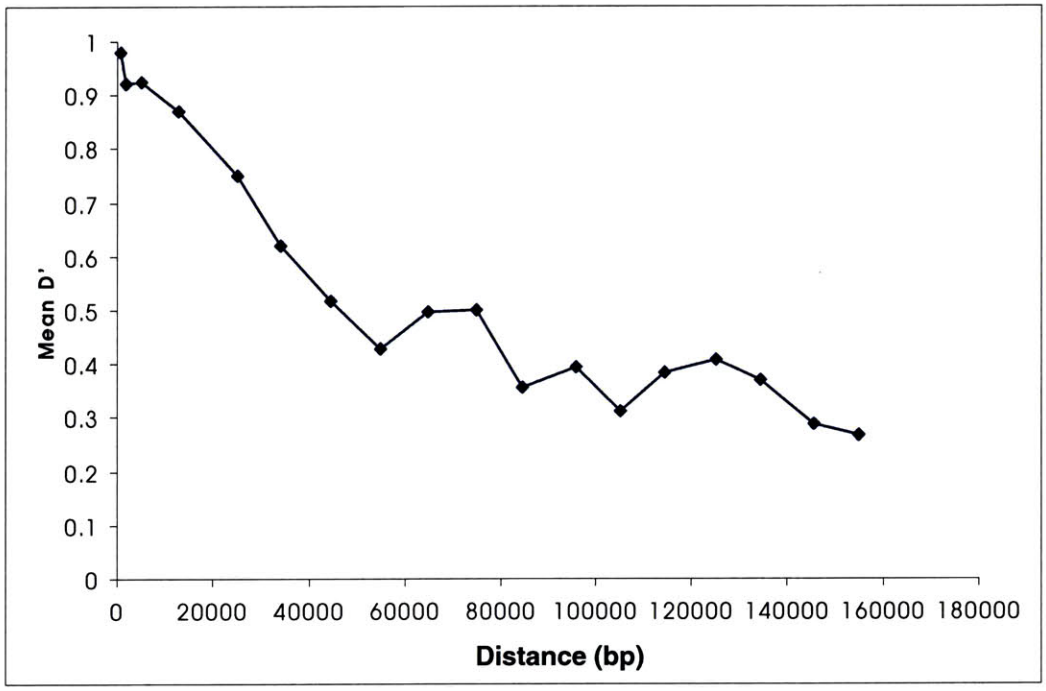

SNP Allele Frequency Distribution
X chromosome Allele frequency distribution
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0 - 0.1
0.1 - 0.2
0.2 - 0.3
0.3 - 0.4
0.4 - 0.5
Allele Frequency

Figure 2

Linkage Disequilibrium on X Chromosome

A



B

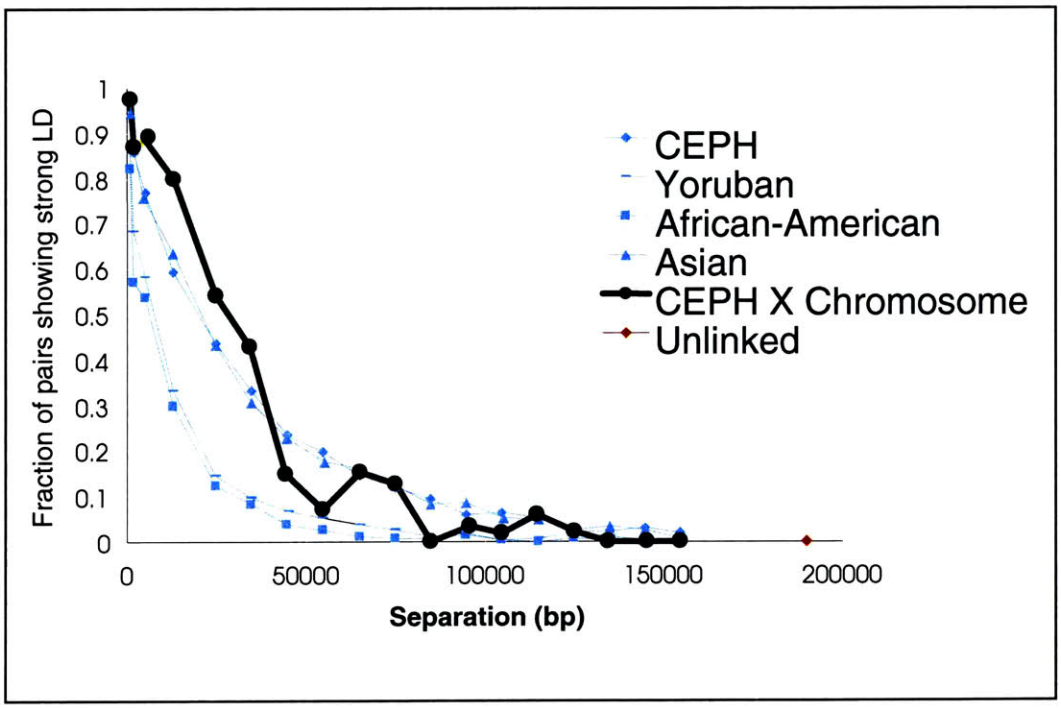
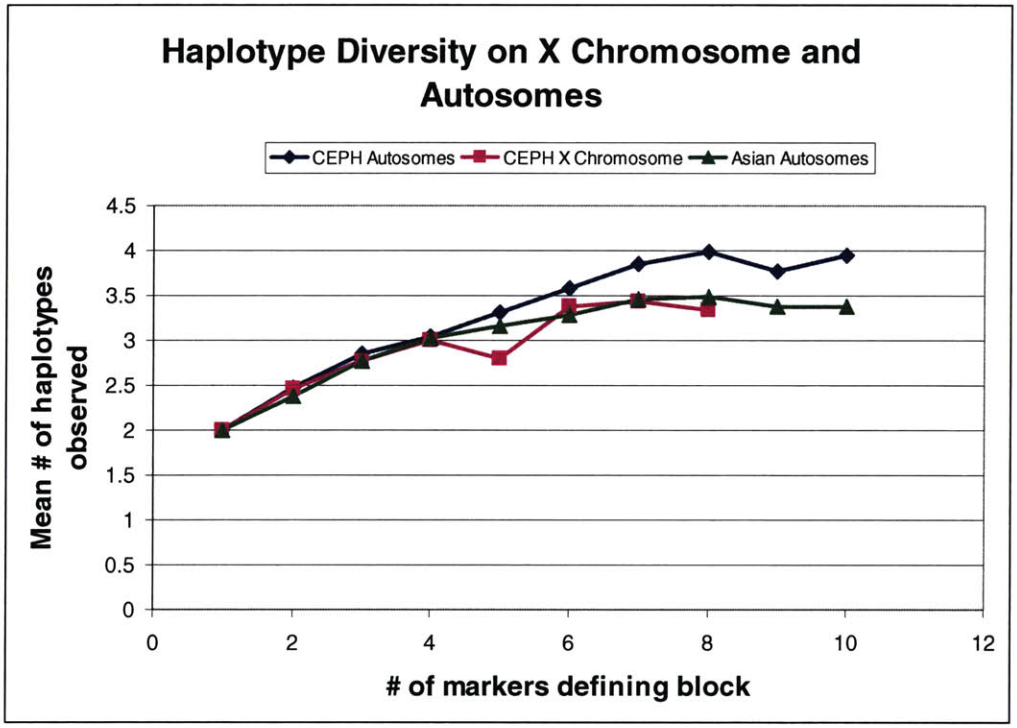


Figure 3

A


Haplotype Diversity on X Chromosome and Autosomes
CEPH Autosomes
CEPH X Chromosome
Asian Autosomes
Mean # of haplotypes observed
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
2
4
6
8
10
12
# of markers defining block

B

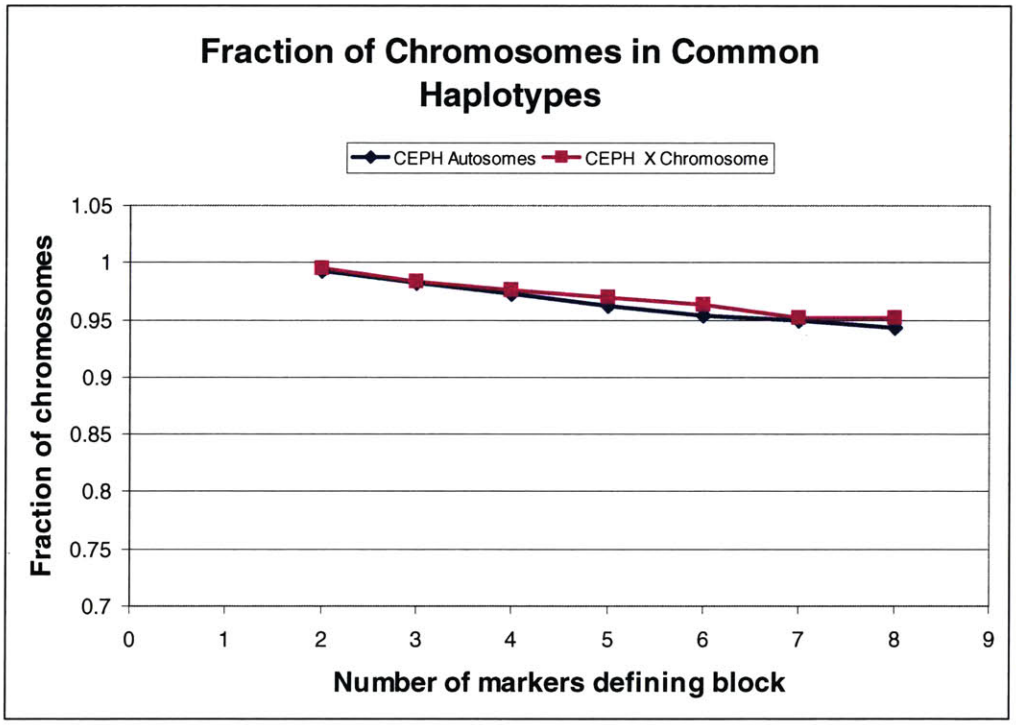

Fraction of Chromosomes in Common Haplotypes
CEPH Autosomes
CEPH X Chromosome
Fraction of chromosomes
1.05
1
0.95
0.9
0.85
0.8
0.75
0.7
0
1
2
3
4
5
6
7
8
9
Number of markers defining block

Figure 4

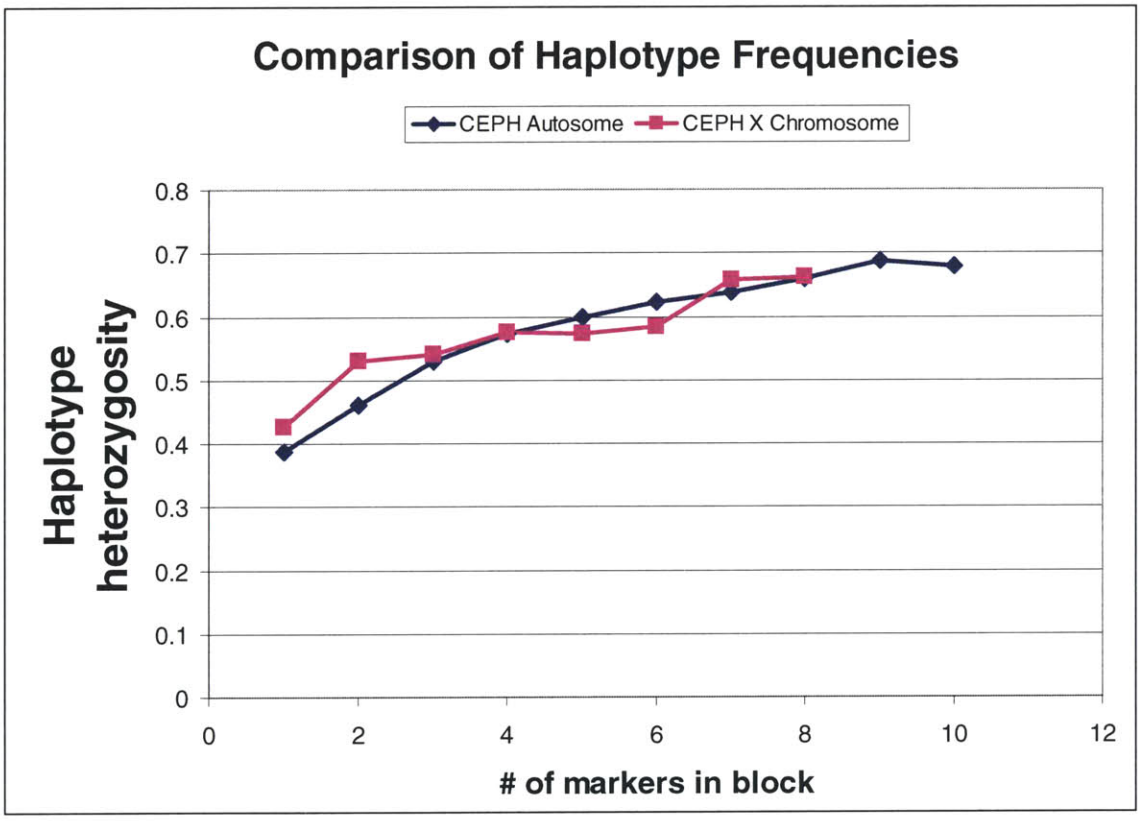

Comparison of Haplotype Frequencies
CEPH Autosome
CEPH X Chromosome
Haplotype heterozygosity
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
12
# of markers in block

Figure 5

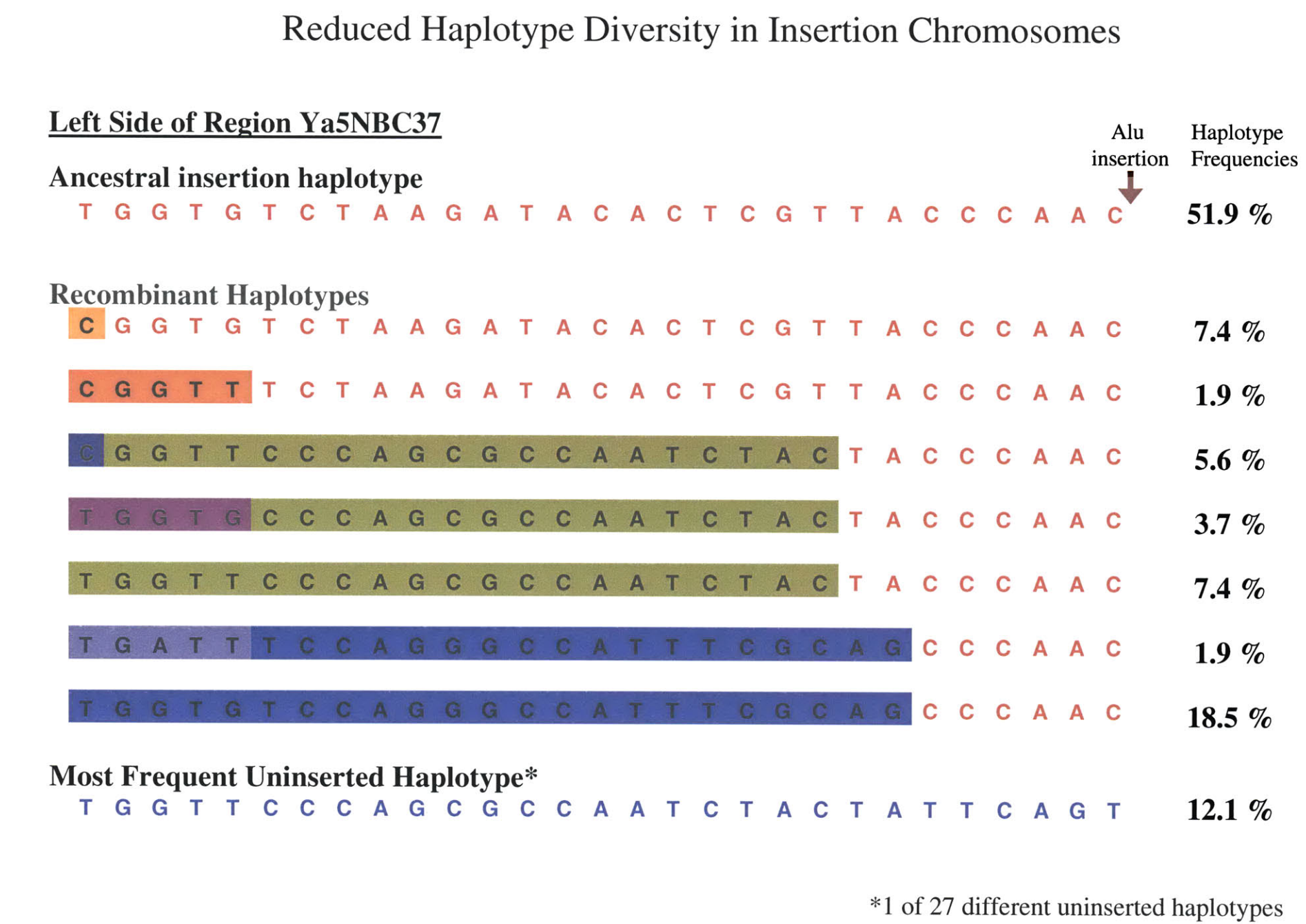

Reduced Haplotype Diversity in Insertion Chromosomes
Left Side of Region Ya5NBC37
Alu insertion
Haplotype Frequencies
Ancestral insertion haplotype
T G G T G T C T A A G A T A C A C T C G T T A C C C A A C
51.9 %
Recombinant Haplotypes
C G G T G T C T A A G A T A C A C T C G T T A C C C A A C
7.4 %
C G G T T T C T A A G A T A C A C T C G T T A C C C A A C
1.9 %
C G G T T C C C A G C G C C A A T C T A C T A C C C A A C
5.6 %
T G G T G C C C A G C G C C A A T C T A C T A C C C A A C
3.7 %
T G G T T C C C A G C G C C A A T C T A C T A C C C A A C
7.4 %
T G A T T T C C A G G G C C A T T T C G C A G C C C A A C
1.9 %
T G G T G T C C A G G G C C A T T T C G C A G C C C A A C
18.5 %
Most Frequent Uninserted Haplotype*
T G G T T C C C A G C G C C A A T C T A C T A T T C A G T
12.1 %
*1 of 27 different uninserted haplotypes

Figure 6

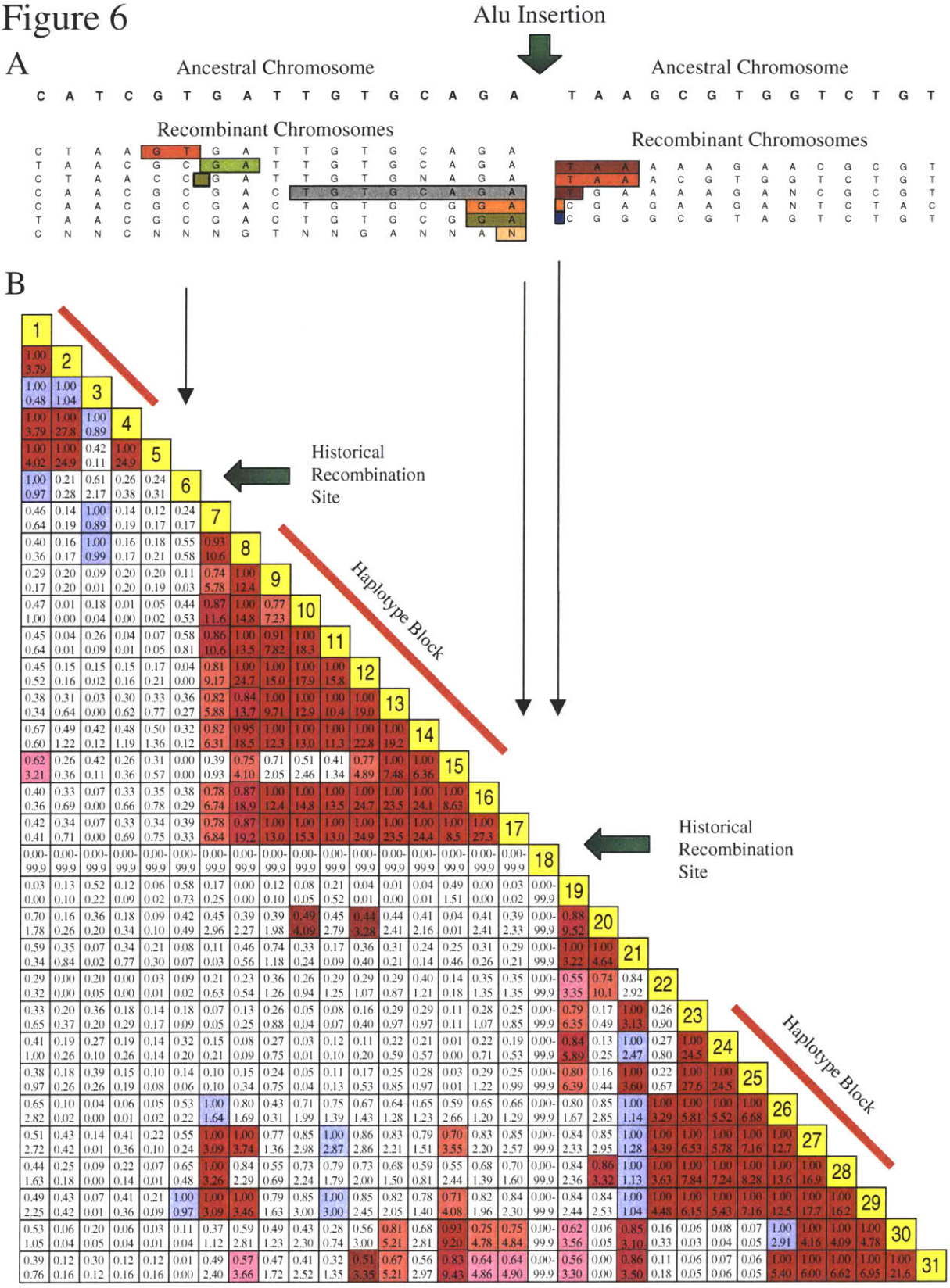


Table of D' Values and LOD Scores for YA5DP5 Region

CHAPTER 6

Perspectives and Future Directions

Looking back

The content of this thesis includes experiments that were performed as many as seven years ago. These experiments were greatly influenced by other studies published and the resources available during the same time periods. The development of ideas surrounding patterns of LD and applications of LD have in many ways imprinted themselves on the structure of this thesis. The chapters were laid out primarily in chronological order, and the changes in ideas that occur between the chapters mimics the changes that occurred in the field in general over the same period of time. The field has drastically changed in the last 10 years. There has been a recent explosion in the number of publications about LD. Only in the last year or two has the scientific community at large begun to exploit the advantages afforded by haplotype data. We knew about the advantages of using haplotypes to examine LD in the first experiments that I performed in this lab, Chapter 2. At the time we saw hints of the patterns of LD that have just recently been fully described. We saw haplotype blocks and limited haplotype diversity within those blocks in the Finnish TCRB data. However, at the time we were not aware of the generality of the results. Although we were aware of the proper way to perform LD studies a long time ago, for a long while it was not possible to perform these studies in large scale due to the lack of polymorphisms and high-throughput genotyping technologies.

The role of resources

As described in the introduction to this thesis, in many ways the research on LD has been guided by the availability of resources. Furthermore, the promise of LD as a mapping tool has contributed to the debate for the creation of large-scale community resources. When microsatellite maps were created, mapping rare monogenic diseases in various isolated founder populations had the most success. During that time I used a denser set of microsatellites provided by the full sequence of the TCRB region. I used this to try to discover an association between a complex trait, multiple sclerosis, and a candidate region. The association study turned out to be negative, but we learned a lot about haplotypes and linkage disequilibrium. TCRB was the subject of debate at the time

and an excellent candidate gene. Accordingly, we also learned that examining single candidate genes one at a time probably was not going to be the most successful method for uncovering causative factors in complex diseases.

The idea that common mutations were most likely going to play an important role in the risk of common diseases became a focus of attention. There was a realization that most large patient samples were ascertained for broad populations and that common variants were not likely to be affected by moderately sized population bottlenecks. Therefore, the focus turned to older globally diverse or broad Caucasian population samples. SNPs became viewed as the preferable markers to work with due to a greater abundance in the genome, a lower mutation rate more amenable to an older population, and a presumably easier method of genotyping. However, only a small number of SNPs had been discovered, so efforts turned to methods for discovering single nucleotide differences. Methods such as resequencing on chips, DHPLC, and brute force sequencing were the major focus at the time. A low resolution SNP genetic map of the genome was created for linkage mapping studies, but was not designed to have a sufficient density for linkage disequilibrium studies. However, high-throughput sequencing was an available technology. Large-scale sequencing efforts were taking place to find all common polymorphisms in huge sets of candidate genes for complex traits. At this time, I chose to sequence a 12-kilobase portion of the X chromosome in 100 Caucasian samples in order to examine LD in a nondescript genomic region and gain haplotype information, as described in Appendix I. The region turned out to be a veritable SNP desert. This study foreshadowed the extremely variable levels of SNP density in general and a low polymorphism rate on the X chromosome, especially in Caucasian samples. In addition, the sequencing study reiterated the limitations of studying single regions. The variability of LD in the genome dictated looking at a more distributed and representative sample of the genome.

Pairwise LD statistics were examined for a number of single gene regions. Extremely variable levels of LD existed even at extremely short distances. We decided to see if this variability extended out to many regions of the genome. Although there was not a dense SNP map at the time, we utilized a fortuitous resource that consisted of multiple SNPs on small STSs used in the creation of the SNP genetic map. We examined

many SNP pairs distributed throughout the genome, described in Chapter 3. We found extremely variable and surprisingly incomplete LD even at these very short distances where one might expect complete LD. Taken alone, this result painted a gloomy picture and showed that even on the very limiting case LD was going to be variable.

We started to realize that an LD curve of pairwise comparisons was possibly not the most informative data. Since the range for any one distance was so large for different regions, applying that information to any single disease region would be extremely difficult. What we really wanted to do is to revisit the multimarker haplotypes that provide so much more detailed information about the structure of the genome. Haplotype structure was examined for a few specific regions implicated in disease. Large scale SNP discovery efforts were making progress. A dense SNP map resource made it possible to select many regions which contained enough SNPs to examine haplotype structure in those regions. In both the single disease regions and a large number of random regions of the genome, a block-like structure began to emerge. Limited haplotype diversity was evident as well. Apparent hotspots of recombination separated these blocks of limited common haplotype diversity. These studies provided a new framework from which to view linkage disequilibrium and the structure of the human genome and the human population.

The role of population geneticists and demographic history

This thesis has not closely addressed the topic of human demographic history. However, this aspect of human genetics is intimately associated with linkage disequilibrium. More specifically, the two topics are the application of LD to mapping human disease genes and the dissection of the history and genetic composition of the human population. Although the focus is somewhat different, these two subjects cannot be easily separated. It is difficult to understand the history of the human population without discerning the shared segments of the genome. Likewise, it is impossible to understand structure of LD in a population sample for which there is no demographic information. Unfortunately, this reality is responsible for the fact that many experiments were designed often had multiple goals. These goals were often intertwined and created results that were confusing and not generally applicable. Population geneticists spent a

great deal of their energies collecting and characterizing extremely well-defined populations but generally did not have the resources to perform large-scale polymorphism discovery or genotyping. Human disease gene mappers studied specific regions in great detail and collected relatively large numbers of markers in these regions, but mostly studied these regions in limited patient populations. It was difficult for both groups to make general claims about the characteristics of LD. Only when researchers began to sequence genomic regions in many individuals from a globally diverse or a general Caucasian population sample did our understanding of the behavior of LD begin to accelerate. These regions were largely candidate genes for complex diseases studied in individuals who did not carry the trait, but regions were also studied that had interesting characteristics due to selection or interesting population histories. It is difficult to collect a globally diverse set of DNA samples which consists of trios or larger families. For this reason, a large number of the haplotype studies that aim to characterize human population history utilized the X chromosome in males. Population geneticists have continued to advocate the importance of expanding LD studies to a wider range of populations. With the accumulation of resources, more attention should become focused on the investigation of LD in a diverse panel of individuals and tracing the pathways of human demographic history.

Where we are now

The experiments contained in this thesis have contributed to a field of study that has developed relatively slowly but is currently on the threshold of enormous explosion. A large body of work from many research groups has brought the study of LD, haplotype structure, and population variation in the human genome to a new level. The remaining portion of this chapter is devoted to summarizing the major finding and current state of the field as well as how these finding branch out into future directions.

Pairwise Linkage disequilbrium

Many years of experiments have brought us to the conclusion that pairwise measures of LD have a more limited utility in describing specific underlying haplotype

structure in the genome than originally thought. Pairwise LD is extremely variable within and between genomic regions of any size and is very dependent on allele frequencies and the individual population histories of the markers. However, a large number of measurements of pairwise LD for different sized regions have been useful for determining average levels of LD across the genome and between populations. The average pairwise LD does show a decrease with distance as far out as 60kb, much further than it was initially believed to be. At the other end of the spectrum, LD is much lower than one would expect at extremely short distances where one would expect almost complete LD. Gene conversion is the most likely explanation for this observation. So the initial predictions, largely created by simulations and population models, differed substantially at both extremes of detectable LD. As described in Chapter 4, observing pairwise LD in a dense set of markers also helped to elucidate that a large part of the genome exists in blocks of LD. However, longer haplotypes were necessary to resolve the limits and diversity of those blocks.

Haplotype blocks and recombination hotspots

A large number of studies have now suggested the general haplotype structure consists of segments with little historical meiotic recombination within. These blocks, defined by clusters of common SNPs, are to be separated by regions where there are very low levels of LD. These regions appear to be hotspots for meiotic recombination, and are not as likely to be due to random genealogical processes or recombination events that occurred very early in the history of the population. Haplotype blocks are around 22 kb in length on average in Caucasians and 4-6 common haplotypes typically comprise more than 90% of all haplotype diversity with in a block. In addition, the X chromosome does not appear to have a longer extent of LD as may be expected. However, this observation is not inconsistent with a model of recombination hotspots being the primary determinant of haplotype structure in human genome.

Population information

The variation in LD between populations has only been loosely described. Broad categories such as Caucasians, Asians, and individuals with recent African ancestry have

been applied to population samples. Haplotype block boundaries are highly similar between populations. Most of the specific common haplotypes are also shared among populations, although the frequencies of these haplotypes tend to differ. Haplotype blocks appear to be slightly smaller and more diverse on average in African or African-American samples. A much more fine resolution of these population differences will be achieved with the genotyping of a larger number of individuals from a more diverse population sample. In addition, the data on haplotype blocks are inconsistent with a constant-sized or expanding population with a uniform recombination rate. These are attributes of many population models. More data of this type will help to solidify the assumptions in human population history.

Abundant resources

The haplotype structure in the genome could only be resolved through genotyping of a dense set of SNPs in a large number of regions. The three main resources that have developed rapidly in recent years are large-scale SNP discovery, genome sequencing, and high-throughput genotyping technologies. Over 2 million SNPs have been discovered and are rapidly being mapped as the complete human genome sequence is constructed. We now have the ability to genotype approximately 50,000 genotypes a day, and that number will be readily increased by scaling up as well as new technologies. These advances will continue to propel our understanding haplotype structure and population history.

The Road Ahead

Haplotype blocks and a LD/haplotype map of the human genome

A great deal of progress has been made in our understanding of the patterns of linkage disequilibrium. However, as is normal for a major discovery and a paradigm shift in a field, more questions have been raised than answered. A few key issues remain. One of the major issues is how to define a haplotype block. LD can occasionally fall off within a block. A block definition is not only important in terms of describing the biology of haplotype structure, but also on a practical level on the relationship between

SNPs and which SNPs would be best to use in disease mapping studies. How frequent in the population does the block have to be and how are will the boundaries be defined. Defining the boundaries can obviously be resolved by using an even more dense set of SNPs. Currently, there are specific thresholds used to determine whether a region is a "block" or not. But I don't believe that a binary categorization is the proper solution. Just in the way that there is variation in LD for pairs of markers, there will be variation between properties of haplotype blocks. I believe there should be a block scoring system instead of a specific cutoff for being a block. A variable rating would better describe the underlying biology. A value of one would be a perfect block with no within-block recombination and a value of zero would be a perfect hotspot, no haplotypes extending past the hotspot. There would be a mean block rating for all the blocks combined, and then cutoffs could be determined based on the rating. I believe that would be the most useful information for researchers who want to use the information to map disease genes. However, this reveals the issue that the best definition may depend on the application.

A haplotype map of the whole genome would prove to be a valuable resource for both the study of human complex traits and human demographic history. Such a resource would require genotyping at least a million SNPs and cataloging the variability of block structure, extent of LD, and haplotype diversity for most of the genome. For disease studies, it would be useful to have the haplotype structure of a region implicated in the disease already assessed before fine-mapping in a patient population. An initial survey of haplotype diversity can give clues about whether or not the characteristics of the region are consistent with those of the disease. A more in-depth examination of haplotype block would also reveal whether or not the boundaries correlate with boundaries of genes. Are shared ancestral segments of the genome the most fundamental unit? Is there a functional difference between very small blocks and extremely long blocks? This may simply represent random genealogical processes, but blocks may also represent a unit of selection. The prospect of addressing all these questions makes it apparent that expanding the scope of haplotype data to the entire genome is an important next stage.

SNPs as haplotype surrogates for simplifying complex traits mapping

In addition to the advantage of pre-characterized genomic regions, a haplotype map of the genome can also significantly reduce the amount of work that an individual researcher would have to do to map a complex trait. The simplicity of haplotype variation in the blocks will significantly increase the power of association studies, thus reducing required sample sizes. Different sets of SNPs within a "perfect" block are essentially equivalent representatives of that portion of the genome. Therefore it is possible to use only the minimum number of SNPs that uniquely define a block as a surrogate for the total haplotype information in a genomic region. This "haplotype tagging" method may increase the likelihood that whole genome searches for LD could be performed. A great deal of time was spent in the field debating about the density of markers necessary to cover the genome adequately for LD mapping. Using haplotype tags will not only reduce the number of markers needed, but more importantly will help ensure that the proper SNPs are used to comprehensively test all independent regions of the genome.

Understanding the structure of human populations and forces that shape LD

As the sequence of the human genome is completed and a large fraction of the common variation in human population is discovered, the most logical next step is to resolve the subsequent levels of population variation. Examining haplotype structure in broad population categories for the whole genome is the most feasible. This may prove to be very close to a sufficient resource in terms of many of the goals for understanding human disease. However, many other questions can be addressed by also focusing on fine scale demographic history and population differences. The issue of isolated founder populations can also be revisited. Do isolated founder populations hold the advantage of greater homogeneity for common haplotypes? This debate has not been entirely settled. By describing the haplotype variation for many different populations, it may eventually be possible to use information from disease studies and specific haplotype information from a population to immediately gain insight on whether or not a particular set of mutations plays a role in the disease etiology in that particular population

The subject of Chapter 5 included experiments on recombination distribution. Using the resources that have been created, we can perform further experiments to

elucidate the forces that have created the observed patterns of diversity, such as population substructure, recombination hotspots, and gene conversion. In Chapter 5, we used recent Alu insertion events as a method to more closely examine the history of recombination. It may be possible to generalize this method by looking at very well defined common haplotype blocks as a unique event in the history of a population. It may be possible to discern specific recombination events using this method.

Analysis tools

Finally, perhaps most importantly, methods for analysis need to advance in step with the accumulation of data. Methods need to be developed to perform the most efficient tests of association between haplotypes and a disease. By using haplotype blocks instead of individual markers in a specific region with association or transmission disequilibrium tests it is possible to obtain a more accurate picture of the breakdown of LD around a mutation. However, for a systematic search across the genome, it will be necessary to perform the minimum number of statistical tests in order to increase the likelihood of observing any significant result. It will also be necessary to create tools that will identify the best SNPs to use as haplotype tags. Analysis tools for identifying blocks and discerning the proper thresholds to use also need to be developed.

Our understanding of LD has both expanded and defined the limits of genetic mapping. In a way, the study of LD has come full circle. The earliest experiments were performed to understand for LD its own sake. Only later was it explored and exploited as a tool for mapping disease mutations. Currently, we have come back to designing experiments simply to understand how the patterns arose. However the patterns of LD are utilized in the future, it will continue to gain attention as an invaluable field of study for medical genetics.

Appendix I

Nucleotide Diversity on the X Chromosomes

Shau Neen Liu-Cordero and Eric S. Lander

Contributions: I was responsible for the conception of the idea for this study, the experimental design, all of the data collection, and all of the data processing and analysis.

Overview and Study Design

The initial goal of this study was to use the X chromosome as a way to examine the extent of linkage disequilibrium. In addition, we sought to exploit the hemizygosity of males to isolate unambiguous haplotypes. It was unclear whether or not a region 12 kb in size would represent the limit of detectable LD in general population. Currently we know from a variety of studies that LD can be seen at much further distances in human population, especially in the Caucasian population. At the time that this experiment was performed very few studies existed which examined nucleotide diversity and the properties of linkage disequilibrium over large contiguous segments of the genome, only one of which examined the X chromosome (Clark et al., 1998; Kaessmann et al., 1999; Rieder et al., 1999). The surprising result of very high nucleotide diversity and low disequilibrium from the lipoprotein lipase (LPL) study indicated that the 9.7 kb of sequence of the LPL gene may be an outlier as opposed to representing the norm for a gene or a genomic region, and further investigation was necessary. There existed many questions that pertained to study design. How should the markers be spaced? Should only high frequency polymorphisms be examined? What is the size of the genomic regions that should be analyzed? Since the diversity in the LPL gene was so high and the level of LD so low even over very short distances, it did not yield much information to help elucidate any strategy for designing LD studies. Based on this, we decided to duplicate the scale of the LPL study on a genomic locus on the X chromosome.

The X chromosome was chosen for the reason that haplotype information can be obtained by simply sequencing in males, who are hemizygous for the X, and alleviates the problems inherent in scoring heterozygous sites in sequence data. In order to isolate a number of polymorphisms with which to study LD, we chose to sequence both strands of a contiguous 12 kb region in 100 Caucasian males, 20 of which were Finnish for comparison with an isolated founder population. The region was amplified as twenty-four overlapping 700 base-pair fragments in each individual and then dye-primer sequencing was performed in both directions. A genomic segment was selected which was devoid of genes, the idea being that there would be less chance for selection in this region, which would could add a confounding factor to analysis. In addition, the region

had a typical GC content of 42% and a typical density of repeats. There is only one other study of extended regions on X chromosome and the purpose of that study was not to look at LD but to address issues pertaining to human population origins and history (Kaessmann et al., 1999).

In population genetic models, the key parameter affecting the balance between mutation and random genetic drift is $3N_e\mu$, for the X chromosome ($4N_e\mu$ for the rest of the nuclear genome). N_e is the effective population size and μ is the mutation rate per nucleotide. Under the neutral theory and infinite sites model, this parameter represents the rate at which these processes of mutation and drift generate and maintain variation in genomic DNA sequence in the absence of the effect of natural selection (Kimura, 1983). The parameter $3N_e\mu$ is related, under equilbrium, to two commonly used measures of nucleotide variation, θ and π. θ is based on the proportion of segregating sites in a sample per nucleotide, and π is the average heterozygosity per site, which is based on the average number of nucleotide differences between two sequences randomly drawn from a sample – the probability that a site will be heterozygous in an individual in a randomly mating population. The values of θ and π are expected to be equal in a population under the infinite sites model. A formal comparison of these two measures, the statistic D, was introduced by Tajima (Tajima, 1989). If the value of the "Tajima's D" statistic is significant, then the null hypothesis of neutrality can be rejected.

Results and Discussion

The results of resequencing the 12 kb region over 100 indivduals were striking. We found only 9 variant sites, with the highest allele frequency being .07, and with most variant alleles present in only one individual (**Table 1**). Based on current estimated values of θ in humans and the fact that the X chromosome has ¾ of the effective population size of the rest of the genome, we would have expected to see on the order of 19 polymorphisms. The corresponding value of nucleotide variation for this region was $\theta=1.57\text{x}10^{-4}$, one-third of other estimates of θ in the human genome (Cargill et al., 1999; Halushka et al., 1999; Li and Sadler, 1991; Wang et al., 1998) and on the very low end of existing X chromosome estimates (Nachman et al., 1998). More surprisingly, the

estimate of heterozygosity for this region was extremely low – both in absolute terms and relative to θ. Specifically, π=3.52x10-5, around one fifth of the value of θ (**Table 2**). Taking into account the effective population size of the X chromosome, the value of π is almost an order of magnitude smaller than the current estimate for the human genome. The extreme difference between π and θ yielded a Tajima's D = –1.91, which is significant enough to reject the assumption of neutrality. So, a region that was initially chosen because it seemed unlikely to be the target of selective pressure appears to be inconsistent with the neutral hypothesis.

With such a paltry collection of polymorphisms in the region, it is impossible to examine LD. Low diversity does not necessarily imply a high level of LD. The results suggest that a better strategy would be to screen many more regions. Several regions could then be examined in great detail. We need to identify regions of sufficient diversity before it is possible to perform a detailed study of LD in those specific regions. Performing detailed studies region by region until one with the proper characteristics surfaces is an extremely cumbersome and inefficient scheme to follow.

The results of this initial study of an X chromosome region prompted us to consider whether or not that region was odd, or was that the norm for the X. In addition, do the estimates of nucleotide variability for the X chromosome differ from other regions of the genome? A number of subsequent studies of the X chromosome have shown an extremely low level of polymorphism on the X chromosome consistent with our results. It is interesting that the original goal of this to see if the LPL gene was an extreme case of high nucleotide diversity and low LD, but in fact we identified a region on the other end of spectrum. This highlights the extreme variation from region to region that has now been seen in a large number of studies.

References

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalayanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q., and Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet *22*, 231-8.

Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E., and Sing, C. F. (1998). Haplotype structure and population genetic inferences from nucleotide- sequence variation in human lipoprotein lipase. Am J Hum Genet *63*, 595-612.

Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R., and Chakravarti, A. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet *22*, 239-47.

Kaessmann, H., Heissig, F., von Haeseler, A., and Paabo, S. (1999). DNA sequence variation in a non-coding region of low recombination on the human X chromosome. Nat Genet *22*, 78-81.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution: Cambridge University Press).

Li, W. H., and Sadler, L. A. (1991). Low nucleotide diversity in man. Genetics *129*, 513-23.

Nachman, M. W., Bauer, V. L., Crowell, S. L., and Aquadro, C. F. (1998). DNA variability and recombination rates at X-linked loci in humans. Genetics *150*, 1133-41.

Rieder, M. J., Taylor, S. L., Clark, A. G., and Nickerson, D. A. (1999). Sequence variation in the human angiotensin converting enzyme. Nat Genet *22*, 59-62.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585-95.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lander, E. S., and et al. (1998). Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome. Science *280*, 1077-82.

Table 1 Summary of discovered polymorphisms

The PCR assays which were sequenced are listed as the GE numbers. The SNP type and location within the assay are listed. F represents a Finnish individual and C represents an individual described as Caucasian. Some repeat polymorphisms are listed that were discovered in the assays. SNPs that were additionally discovered by the method of DHPLC are marked by an X.

Table 2 Calculation of the observed heterozygosity per base pair (π)

The allele frequencies and nucleotide heterozygosities are calculated for each assay. The bottom number represents the overall calculation for all the sequence which is the sum divided by the total amount of sequence in base pairs.

Table 1 Summary of Discovered Polymorphisms

ASSAY	SNPs	VARIANT INDIVIDUALS	COMMENTS	DHPLC discovered SNP
GE1389	724 G/A	F1R, F11FR, F16FR, C31R, C32FR, C33F, C54FR		XXX
GE1390	none			
GE1391	711 C/A	C61FR		
GE1393	none			
GE1394	576 T/G	C31FR		
GE1395	none		CA repeat -polymorphic	
GE1396	none			X
GE1397	484 T/A	C61R		X
GE1398	655 G/C	C62FR		
GE1399	none			
GE1400	none			
GE1401	none			
GE1402	none		CA repeat - no read-through	
GE1403	none			X
GE1405	699 A/*	C56FR		
GE1406	none			
GE1407	none			X
GE1408	none			
GE1409	none			
GE1410	none			X
GE1411	none			
GE1412	none			
GE1413	162 G/A	C51F		
GE1414	324 G/A	C9FR, C10FR, C43FR, C58R, C59FR	PolyT - no read-through	
	490 G/A	C5FR, C39FR		

Table 2 Calculation of Nucleotide Heterozygosity

Polymorphism	Allele frequency	Nucleotide heterozygosity
GE1389	0.07	0.13
GE1391	0.01	0.02
GE1394	0.01	0.02
GE1397	0.01	0.02
GE1398	0.01	0.02
GE1405	0.01	0.02
GE1413	0.01	0.02
GE1414a	0.05	0.10
GE1414b	0.02	0.04
		0.39
	$\pi =$	**3.52E-05**

Appendix II

Supplemental Data to Chapter 3

Distribution of Pairwise Linkage Disequilibrium and Sampling Issues

Authors and contributions are the same as for Chapter 3.

Distribution of Pairwise Linkage Disequilibrium

This appendix serves to provide additional data that was not included in the publication represented by Chapter 3. Chapter 3 focused mainly on observed and expected obligate recombination events, or R_M. To briefly review, we studied 103 STSs that contained 325 SNPs in a globally diverse population sample. These SNPs were very closely linked, with the average distance between SNPs being 124 base pairs. Traditional measurements of LD reflect departures of two-locus haplotype frequencies from unlinked expectations were not included in the publication because they are not sensitive to detecting small departures from complete LD and are sensitive to allele frequency. This Appendix presents the LD distribution and some of the consequences that allele frequencies have on LD statistics.

We initially examined the distribution of pairwise linkage disequilbirum using d^2 as a measurement of LD. This distribution is shown in (**Figure 1**). Given the views on the history of the the human population, pairs of SNPs spaced at very short intervals are expected to exist in complete LD. However, various published results exhibit a large range of LD even at very short distances. In this sample of loci distributed throughout the genome, we observed a large range of LD. It is evident that many of the values indicate near-complete LD, but even at these short distances, some markers behave as if unlinked. In this analysis, we only included markers with a minor allele frequency >0.25. The rationale for this subdivision of the data lies in what was observed when we looked at comparisons between unlinked markers.

Allele Frequency Affects Measurement of Linkage Disequilibrium

We wanted to observe the behavior of this LD statistic in the limiting case where we would expect no LD due to physical linkage. We used the exact same SNPs as the linked data, except the SNPs are compared between the different loci instead of within loci, so there many more possible pairwise comparisons. **Figure 2** illustrates actual and simulated data plotted adjacent to each other. Simulations were performed using the actual allele frequencies of the markers but with varying numbers of simulated chromosomes. In our study we collected data for 92 chromosomes. The blue curve represents simulated data for a sample size equivalent to what we used in the actual data.

It is evident that a large number of aberrantly high values of d^2 are present in a sample size of 92 as opposed to a larger sample size of 1000 represented by the green distribution. The red curve demonstrates the effect of including only markers with higher minor allele frequencies. By looking at markers with a minor allele frequency greater than 0.25, one can largely avoid the sampling problems due to small numbers. The actual data is plotted in the black curve. These data also include markers with a minor allele frequency greater than 0.25. There is a similar effect, but there are still a number of observations showing higher levels of LD. This may be due to a small amount of admixture. The significant number of observations with d^2 between 0.1 and 0.25 in unlinked markers suggests that care should be taken when evaluating those levels of LD as significant in linked data sets, especially when employing small sample sizes or rare alleles. We observe this effect of allele frequency in both the unlinked and the linked data **(Figure 3)**. In the case of unlinked markers, the inacccuracy produces an overestimate of d^2 for low allele frequencies, and in the case of linked markers the value of d^2 tends to be underestimated as lower allele frequencies are included.

Figure 1 Distribution of Pairwise Linkage Disequilibrium

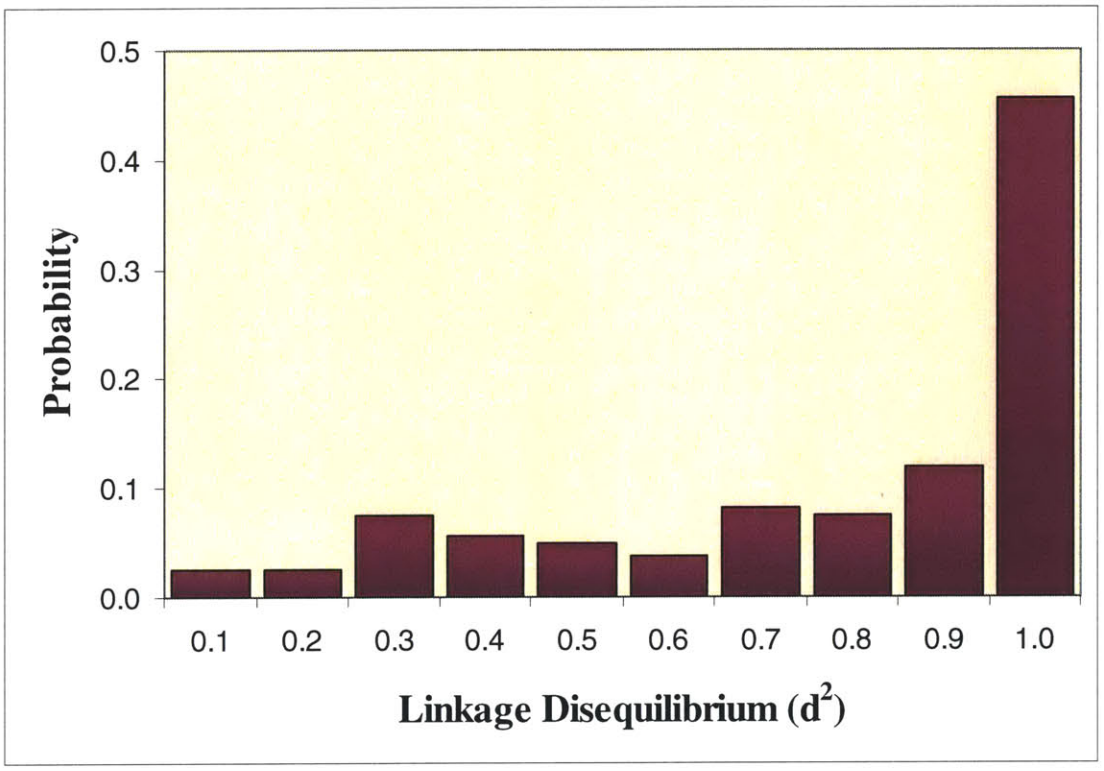
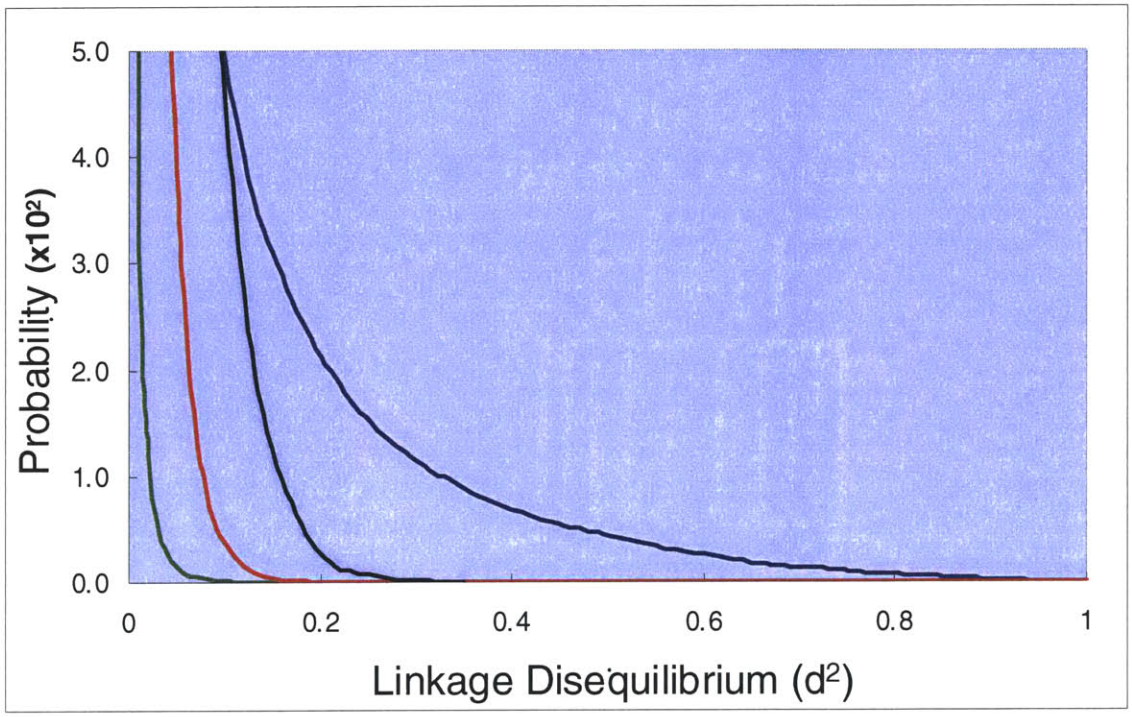


Figure 2 d^2 Provides a Biased Estimate of Linkage Disequilibrium due to Small Sample Size



92 Chromosomes
Avg d^2 = .023 ± .063

1000 Chromosomes
Avg d^2 = .0018 ± .0047

92 Chromosomes
MAF>0.25
Avg d^2 = .012 ± .017

Results: MAF>0.25
Avg d^2 = .025 ± .034

Figure 3 Linkage Disequilibrium as a Function of Allele Frequency

Mean d^2

	Linked Pairs	Unlinked Pairs
All alleles	.35 ± .39	.040 ± .083
MAF > .1	.55 ± .39	.028 ± .039
MAF > .2	.65 ± .34	.027 ± .036
MAF > .3	.74 ± .29	.026 ± .035

Appendix III

The Discovery of Single Nucleotide Polymorphisms and Inferences about Human Demographic History

John Wakeley, Rasmus Nielsen, Shau Neen Liu-Cordero and Kristin Ardlie

Published as Wakeley et al. The discovery of single nucleotide polymorphisms and inferences about human demographic history. Am. J. Hum. Genet. 69, 1332-47 (2001).

Contributions: I was responsible for all data collection, along with Kristin Ardlie who was an equal partner. John Wakeley performed all the analysis.

The discovery of single nucleotide polymorphisms and inferences about human demographic history

John Wakeley[1], Rasmus Nielsen[1*], Shau Neen Liu-Cordero[2,3], and Kristin Ardlie[2**]

[1]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA; [2]Whitehead Institute for Biomedical Research, Cambridge, MA; [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA

*Present affiliation: Department of Biometrics, Cornell University, Ithaca, NY

**Present affiliation: Genomics Collaborative Inc., Cambridge, MA

Running title: SNPs and human history

Corresponding author:

John Wakeley

2102 Biological Laboratories

16 Divinity Ave.

Cambridge, MA 02138

Telephone: (617) 495-1564

FAX: (617) 496-5954

E-mail: wakeley@fas.harvard.edu

Summary

A method of historical inference that accounts for ascertainment bias is developed and applied to single nucleotide polymorphism (SNP) data in humans. The data consist of 95 short fragments of the genome which were selected from three recent SNP surveys to contain at least two polymorphisms in their respective ascertainment samples, and which were then fully resequenced in 47 globally-distributed individuals. Ascertainment bias is the deviation, from what would be observed in a random sample, caused by discovering polymorphisms in small samples or by selecting loci based on levels or patterns of polymorphism. The three SNP surveys from which the present data were derived differ in their protocols for ascertainment and in the size of the samples used for discovery. We implemented a Monte Carlo maximum likelihood method to fit a subdivided population model which includes a possible change in effective size at some time in the past. Incorrectly assuming that ascertainment bias does not exist causes errors in inference, affecting both estimates of migration rates and historical changes in size. Migration rates are overestimated when ascertainment bias is ignored. However, the direction of error in inferences about changes in effective population size (whether the population is inferred to be shrinking or growing) depends on whether the numbers of SNPs per fragment or the SNP allele frequencies are analyzed. We use the acronym SDL for SNP-Discovered Locus in recognition of the genomic discovery-context of SNPs. When ascertainment bias is modeled fully, both the numbers SNPs per SDL and their allele frequencies support a scenario of growth in effective size in the context of a subdivided population. If subdivision is ignored, however, the hypothesis of constant effective population size cannot be rejected. An important conclusion of this work is that SNP data are useful in demographic or other studies only to the extent that their ascertainment can be modeled.

Introduction

Single nucleotide polymorphisms, or SNPs, are the markers of choice for studies of both linkage and historical demography. This is due to the relative abundance of SNPs in the human genome compared with other types of polymorphisms, the efficiency with which they can be assayed, and the ease with which they can be analyzed using the tools of population genetics. It is typically assumed that each SNP is the result of a single mutation event and that different SNPs segregate independently of one another. These assumptions are probably correct much of the time. Then, it is the allele frequencies at SNPs and the distribution of the polymorphisms among sub-populations which can tell us about demographic history. However, SNPs are discovered, and later genotyped, using primer pairs which amplify short fragments of the genome rather than single sites. We refer to these SNP-Discovered Loci as SDLs. Some proportion of SDLs will be found to contain multiple SNPs, especially as the sample sizes from human populations increase. This represents an opportunity to garner more information from polymorphism data, namely the numbers of SNPs per SDL and their joint frequencies in a sample.

The SDL context of SNPs also has important implications for correcting ascertainment bias. The data analyzed below are derived from SDLs discovered in three recent SNP surveys: Wang et al. (1998), Cargill et al. (1999), and Altshuler et al. (2000). The first two reported SDLs which had at least one SNP segregating in a relatively small, geographically-restricted sample and a relatively large, globally-distributed sample, respectively. The third study found polymorphisms in a relatively small, globally-distributed sample, but also introduced a new SNP-discovery protocol, called reduced representation shotgun sequencing, in which it is necessary to impose an upper bound on the number of SNPs per SDL. A large fraction of the 1.42 million SNPs in the high-density SNP map reported recently were discovered using a modified version of this method (The International SNP Map Working Group 2001). In some applications it will be

necessary to model this discovery process. In addition, all of the SDLs studied here contain multiple SNPs because they were originally chosen for a study of genome-wide patterns of linkage disequilibrium (Ardlie et al. 2001) to have at least two SNPs segregating in their respective ascertainment samples. We show that there is substantial information about population history both in the numbers of SNPs per SDL and in the allele frequencies of SNPs. However, the mark of ascertainment bias is different for these two kinds of data. In order to correct properly for ascertainment bias, it is necessary to know the full pattern of polymorphism discovered at an SDL, even if only a single SNP is typed in a later study.

We are concerned with two aspects of human historical demography: population subdivision and changes in effective size over time. Although the human population may be less structured than that of chimpanzees and other close relatives (Kaessmann et al. 1999), it is clear that subdivision has played a role in shaping human polymorphism. There is less agreement about the pattern of changes in the human effective population size (Hawks et al. 2000). The early reports of mitochondrial DNA diversity seemed to indicate a recent large increase in effective size (Cann et al. 1987; Vigilant et al. 1991). When nuclear data became available, the first few data sets appeared to contradict this, showing instead a pattern consistent with a decrease in effective size rather than an increase (Hey 1997). This conclusion was based in part on deviations from the expected frequency distribution of polymorphic sites. Deviations in the frequency spectrum are summarized by Tajima's (1989) statistic, D, which tends to be negative when the effective size of a population increases and positive in when it decreases. A recent survey of available nuclear loci (Przeworski et al. 2000) showed a broad range of D values and concluded that neither a constant effective size nor long-term exponential growth could explain the pattern. Two more recent reports suggests a stronger signature of growth (Stephens et al. 2001; Yu et al. 2001). While humans have certainly increased in number and we might expect to find genetic evidence of this, it is important to keep in mind that census size is not the only

determinant of effective size. In a subdivided population, changes in the rate and pattern of migration can mimic, or obscure, a signature of growth because the effective size of a population is inversely proportional to the migration rate (Wright 1943; Nei & Takahata 1993) and depends on the pattern of migration across the population (Wakeley 2001).

Before describing our model and the effects of ascertainment bias on historical inference, some background for a simpler model is helpful. Expectations about patterns of polymorphism are typically based on the coalescent (Kingman 1982; Hudson 1983a; Tajima 1983), a stochastic model which describes the genealogical history of a sample of DNA sequences. In this model, it is assumed that a sample of size n is taken without replacement and, importantly, without regard to variation in the population. It is also assumed that the population has been of constant effective size over time and not subject to current or historical subdivision. Variation at the genetic locus under study is assumed not to be affected by selection, either directly, or indirectly through linkage to other loci. The standard model also assumes that there is no intra-locus recombination. If these assumptions hold for a sample of DNA sequences from some population, the genealogy of the sample will be a randomly-bifurcating tree with exactly $n-1$ coalescent nodes like the one shown in figure 1. Further, the time during which there were exactly k lineages is exponentially distributed with mean

$$E(t_k) = \frac{2}{k(k-1)} \tag{1}$$

(Watterson 1975; Kingman 1982). These times, t_k, are measured in units of $2N_e$ generations, where N_e is the inbreeding effective size of the population. Equation (1) show that the expected value of t_k is larger when k is smaller, *i.e.* for the more ancient coalescent intervals in the genealogy. The relative branch lengths in the genealogy in figure 1 are those expected from (1).

All the standard predictions of the coalescent, for example in Tavaré (1984), follow

from the two basic results described above: the random-bifurcating structure of genealogies and the exponentially-distributed times to common ancestor events. However, predictions about what should be observed in a sample of genetic data are different depending on the mutation process at the locus under consideration. When the rates of mutation and recombination per site are very low, the infinite sites mutation model without intra-locus recombination is appropriate (Watterson 1975). We use this model below, and exclude the SDLs which show direct evidence of either multiple mutations, recombination, or gene conversion (Ardlie et al. 2001). Under the infinite sites model, there is a one-to-one correspondence between mutations and polymorphic sites in a sample. Considering the genealogy, this means that a polymorphic site which is segregating at frequency i/n in the sample must be the result of a mutation that occurred on a branch of the genealogy which partitions the tips of the tree into two sets, one of size i and one of size $n-i$. The number of mutations that occur on a branch of length T is Poisson distributed with mean $Tl\theta/2$, where l is the length in bp of the locus or SDL, θ is equal to $4N_e u$, and u is the neutral mutation rate per bp per generation.

Inferences about the demographic history of populations are often made by comparing observed data, such as SNP data, to the following prediction of the standard coalescent model with infinite-sites mutation: the expected number of segregating sites at which one base is present in i copies and the other is present in $n-i$ copies in a sample is equal to

$$E(\eta_i) = l\theta\left(\frac{1}{i}+\frac{1}{n-i}\right)/(1+\delta_{i,n-i}) \quad (2)$$

(Tajima 1989; Fu 1995). Because the ancestral state is typically unknown, i ranges from one to $[n/2]$, where $[n/2]$ is the largest integer less than or equal to $n/2$. Thus, $E(\eta_i)$ is the sum of two terms, the expectation for a mutant site pattern, i, and for its complement, $n-i$. To avoid counting the same pattern twice, it must be corrected for the case $i = n-i$ using Kronecker's δ, which is equal to one if $i = n-i$, and zero otherwise. Equation (2)

specifies that singleton polymorphisms ($i = 1$) should be the most abundant, and that the numbers of other kinds of polymorphisms should fall off in a characteristic manner as i increases. If the polymorphic site frequencies in a dataset deviate significantly from this prediction, one or more of the assumptions of the model must be incorrect. Tajima's (1989) D, and the statistics proposed by Fu and Li (1993), will detect deviations in two directions: either too few low-frequency sites, or too many. Figure 2 plots the distributions of Tajimas D among SDLs for the three datasets studied here, and shows that D tends to be positive in two of them. This is the result of an excess of middle-frequency polymorphisms, which we show below is due to ascertainment bias in these two datasets.

Every SDL and SNP has an associated ascertainment sample, the sample in which it was originally discovered. In fact, this is true of any genetic marker. Subsequent genotyping of SNPs is done with different, typically much larger, data samples which may or may not overlap with the ascertainment sample. There are three kinds of samples in this context: samples which are included in the ascertainment study but not in a later data set, samples which are included in both the ascertainment study and a later data set, and samples which are included in a later data set but were not part of the original discovery study. We will refer to the numbers of these ascertainment-only samples, overlap samples, and data-only samples as n_A, n_O, and n_D, respectively. In total, the ascertainment sample is of size $n_A + n_O$ and the data sample is of size $n_D + n_O$. Because the chance that a SNP will be segregating in a small ascertainment sample is higher for middle-frequency polymorphisms than it is for low-frequency polymorphisms, the counts of the two bases segregating in later data samples will tend more toward the middle frequencies than the expectation for random sample given by equation (2). This effect will be exacerbated if a frequency cutoff is used before a SNP is recognized in the ascertainment sample. The bias in frequencies which results from initial screening in a small sample has been described before in other contexts (Ewens et al. 1981; Sherry et al. 1997), and its importance for

human SNPs has recently been emphasized (Nielsen 2000; Kuhner et al. 2000).

Here we describe two further aspects of ascertainment bias: the consequences of choosing uncharacteristically polymorphic loci and the effects of ascertainment bias on the distribution of the number of SNPs per SDL. We are concerned with these phenomena because the data considered here were selected to have at least two SNPs per SDL in the ascertainment sample (Ardlie et al. 2001) and because some of our analyses depend on the distribution of the number of SNPs per SDL. Figures 3 and 4 display simulation results of ascertainment bias under the standard coalescent model. In both figure 3 and figure 4, $l = 400$ bp long SDLs were simulated with $\theta = 0.0005$ per bp, and a data sample size of $n_D = 10$ was assumed. Using Watterson's (1975) result, which is equivalent to the sum of (2) over all i, we find that the expected number of SNPs per SDL, S, is equal to 0.566. Figure 3 shows that the SNP allele frequencies are skewed toward the middle frequencies both (a) when SDLs are required to have SNPs segregating in small ascertainment samples, and (b) when SDLs are selected to contain multiple SNPs. The effect is stronger in the former case than in the latter, but should not be ignored in either case.

The effect shown in figure 3(a) is fairly well known, and follows directly from sampling considerations. It has important consequences for the distribution of mutations over the genealogy of the sample. Mutations which occur during the most recent coalescent interval, t_n, can only be singletons, but mutations which occur on the deeper branches in the genealogy can be segregating at higher frequencies. Thus, by preferentially gathering middle frequency SNPs, more-or-less directly as in figure 3(a), we are also selecting older mutations. The reason this effect is also seen in 3(b), when SDLs are chosen to be highly polymorphic, is that much of the variation in the total length of the genealogy, and thus in the number of SNPs per SDL, is attributable to variation in the length of the longest and most ancient coalescent interval, t_2. Mutations which occur during this interval can be segregating at any frequency in the sample, and thus tend more toward the middle

frequencies than recent mutations.

Figure 4 shows the effects of these same ascertainment processes on one aspect of the distribution of the number of SNPs per SDL: the coefficient of variation of S. Both (a) using smaller ascertainment samples for discovery, and (b) imposing a cutoff for the number of SNPs per SDL cause the coefficient of variation to be smaller than would be observed in a random sample. Imposing a lower bound on S causes this directly, but it is less obvious why the same thing occurs when higher-frequency polymorphisms are selected. Again, the answer is in the placement of the mutations on the genealogy. Using the exponential distribution with mean (1) and considering the Poisson($l\theta/2$) mutation process, it is easy to show that the coefficient of variation of the number of segregating sites at a locus which descend from mutations which occurred during coalescent interval, t_k, is equal to $(\sqrt{k-1+\theta})/\theta$, and is thus smaller for more ancient mutations. It is important to separately consider the effects of ascertainment on the numbers of SNPs per SDL and on the allele frequencies because the consequences for historical inference are different for the two types of data. For example, extreme population growth is known to make sample genealogies star-shaped (Slatkin & Hudson 1991). This results in an excess of singleton polymorphisms due to long external branches, but it also decreases variation in S because most genealogies will tend to be the same size. The effects of milder growth are in this same direction. Population decline reverses these effects, producing an excess of middle-frequency polymorphisms and increasing variation in S among loci. If ascertainment bias is ignored, an analysis of frequency spectra would point towards a shrinking population while an analysis of numbers of SNPs would indicate an expanding population, even though the truth may be that the population has not changed in size.

Materials and Methods

Ascertainment of SDLs

Ardlie et al. (2001) analyzed 106 SDLs chosen from three recent SNP surveys (Wang et al. 1998; Cargill et al. 1999; Altshuler et al. 2000). These were selected on the basis of their having at least two SNPs segregating in the samples used for discovery, and were then fully resequenced in a sample of forty-seven globally-distributed individuals. Consult Ardlie et al. (2001) for a description of these samples. We refer here to the SDLs derived from Wang et al. (1998), Cargill et al. (1999), and Altshuler et al. (2000), as datasets 1, 2, and 3, respectively. Individuals were partitioned into demes, or subpopulations, mostly on the basis of geographic origin, but with some attention to ethnic identity within localities. Table 1 lists these demes and gives the data-only, overlap, and ascertainment-only sample sizes for each dataset. Sample sizes are numbers of chromosomes rather than diploid individuals. The number of chromosomes listed for the CEPH Utah pedigree in dataset 1 (Wang et al. 1998) is odd because the ascertainment sample in that study included a maternal grandmother and her son from family number 1331 (GM07340 and GM07057).

The 106 SDLs studied in Ardlie et al. (2001) included forty-one from the study of Wang et al. (1998), twenty-nine from Cargill et al. (1999), and thirty-six from Altshuler et al. (2000). We excluded four of these SDLs, all from dataset 3, because they were found to have fewer than two SNPs in the ascertainment sample when they were resequenced, and thus did not fit our model of ascertainment bias. In addition, seventeen SDLs were removed – seven from dataset 1, four from dataset 2, and six from dataset 3 – because they showed direct evidence of either recombination or gene conversion (6 SDLs) or of multiple mutations (11 SDLs) (Ardlie et al. 2001). Finally, we excluded one SDL from dataset 3 because it mapped to the X chromosome, and thus has a different effective population size and possibly a different migration pattern than the autosomal SDLs. We also ran all of the analyses with these SDLs included and the results were the same. In sum, our dataset 1 contains thirty-four SDLs and datasets 2 and 3 each contain twenty-five SDLs, all of which appear to fit both our model for ascertainment and the infinite-sites mutation model

without recombination or gene conversion.

The SDLs in dataset 3, which were discovered using the method described in Altshuler et al. (2000), must be treated differently than those in datasets 1 and 2. In this case, the ascertainment sample for each SDL is not identical to the $n_A + n_O$ samples listed in table 1, but rather is a random sample of these taken with replacement. The sizes of these random samples are the "clique sizes" of Altshuler et al. (2000). However, they are not the final ones reported in that paper because the SDLs studied here and in Ardlie et al. (2001) were selected prior to the completion of that study. These clique sizes differ among SDLs, and range from two to six with a mean of three. In order to exclude multi-copy sequences, Altshuler et al. (2000) imposed an upper bound of no more than one SNP per one-hundred bp in a SDL. Thus, in addition to the lower bound of two SNPs, which is true for all three datasets, when analyze dataset 3 we must include an upper bound on the number of SNPs per SDL in the ascertainment sample and take into account the subsampling of the ascertainment sample to form cliques.

A Model of Historical Demography

We used the subdivided population model recently described by Wakeley (2001). This is a generalized version of Wright's (1931) island model, in which the sizes of demes (N), the contributions of each deme to the migrant pool (α), and in the fraction of each deme that is replaced by migrants every generation (m) vary across the population. It is assumed that the number of demes in the population is large relative to the size of the sample under study. Simulation results indicate that the number of demes need only be three or four times the sample size for the large-number-of-demes approximations to hold (Wakeley 1998). The parameters which determine the pattern of genetic variation in a sample are $M = 2Nm$ for each sampled deme and $\theta = 4N_e u$, where N_e is the effective size of the entire population and u is the neutral mutation rate at a locus. The effective size of the population depends on the total number of demes, and the distributions of N, α, and

m among demes. It is important to note that θ in this model is the expected number of nucleotide differences for a pair of sequences from *different* demes. This is a consequence of their being a large number of demes; a randomly chosen pair will almost never be from the same deme.

As in Wakeley (1999) we allow for the possibility of a single, abrupt change in effective size at some time in the past. This could be the result of a change in the total population size, but it could also be caused by changes in the relative sizes of demes, in the relative contributions to the migrant pool, or in the backward migration rates (Wakeley 2001). The large-number-of-demes model is characterized by a short, recent "scattering" phase and a longer, more ancient "collecting" phase (Wakeley 1999). The scattering phase is a stochastic sample-size adjustment which accounts for the tendency of samples from the same deme to be more closely related than samples from different demes. The collecting phase is a Kingman-type coalescent process with effective size N_e. The ancestry of a sample can be described analytically, but is easily simulated and we take this route in modeling ascertainment bias. Genealogies are simulated as follows. First, the scattering phase is performed for each deme's sample using the "Chinese-restaurant" process (Arratia et al. 1992). This is one of several stochastic processes known to produce Ewens' (1972) distribution, which is the appropriate model for the numbers of descendants of the lineages from each deme which enter the collecting phase (Wakeley 1999). Then, conditional on this, the collecting phase for the remaining lineages is a coalescent process, but with a change in effective size at some time in the past. Observed data will depend on M_i, $1 \leq i \leq d$, which are the values of $2Nm$ for each of the d sampled demes, as well as on θ. They will also depend on $Q = N_{eA}/N_e$, the ratio of the ancestral effective size to the current effective size, and on $T = t/(2N_e)$, the time in the past at which the change in effective population size occurred, measured in units of $2N_e$ generations.

Methods of Ancestral Inference

The data have the following structure at each SDL. There are some number of data-only SNPs, S_D, and some number of overlap SNPs, S_O. The other possible type of SNP, ones which were segregating in the ascertainment sample but not in the data sample, S_A, are not directly observed. However, we do have some information about these which we must take into account when we condition on ascertainment. Namely, the sum $S_A + S_O$ must be greater than or equal to two for the SDL to have been selected (Ardlie et al. 2001). For datasets 1 and 2, $S_A + S_O \geq 2$ must be true for the ascertainment sample of $n_A + n_O$ chromosomes listed in table 1. For dataset 3, it must be true in a randomly-chosen ascertainment sample of some smaller size ("clique size"; see above). In addition, for dataset 3 we must also impose the upper bound: $S_A + S_O \leq Z$, where $Z = [l/100]$, *i.e.* that there is no more than one SNP per one hundred bp (Altshuler et al. 2000).

The three categories of SNPs — S_D, S_O, and S_A — are mutually exclusive. Thus, under the infinite sites model, they are generated via mutation on non-overlapping sets of branches in the genealogy of the sample. Figure 1 shows one possible realization of such a genealogy with $n_D = n_O = n_A = 3$, and distinguishes the three possible kinds of branches. Let T_D be the sum of all the solid branches, T_O be the sum of all the short-dashed branches, and T_A be the sum of all the long-dashed branches in the genealogy. Every branch in the genealogy must fall into one of these three categories. Given these values, the numbers of polymorphisms, S_D, S_O, and S_A, are mutually-independent and Poisson-distributed with parameters $T_D\theta/2$, $T_O\theta/2$, and $T_A\theta/2$, respectively. Our analyses depend on this because we calculate likelihoods and other quantities by conditioning on the genealogy of the sample, and averaging values over many simulated genealogies.

In addition to S_D and S_O (and S_A), the full data include the joints frequencies of SNPs among demes and the linkage patterns between SNPs within each SDL. We would

like to use this information to make inferences about the parameters of the model:

$$\Omega = \{\theta, Q, T, M\} \tag{3}$$

where $M = \{M_1, M_2, \ldots, M_{20}\}$. We are most interested in inferences about Q and T, and treat M and θ as nuisance parameters. Ideally, we would like to base our inferences upon $Pr\{Data|\Omega, asc\}$, the likelihood of the full data, given the ascertainment scheme. However, this is computationally infeasible. Instead, we first obtain moment-based estimates of $M = \{M_1, M_2, \ldots, M_{20}\}$ for each of the three datasets based on the the numbers of polymorphisms segregating within each deme. We then use the distribution of the numbers of data-only and overlap SNPs per SDL to make inferences about θ. This step gives information about Q and T as well, because θ is estimated over a grid of (Q, T) values by maximizing $Pr\{S_D, S_O|\theta, Q, T, \widehat{M}, asc\}$. Lastly, fixing both M and θ from these analyses, we use $Pr\{X|\widehat{\theta}, Q, T, \widehat{M}, asc\}$ to make inferences about Q and T, where X is a vector of the frequencies of the less-frequent bases segregating at each SNP on each SDL. We ignore the pattern of linkage between SNPs. As described below, all of these inferences are made conditional on the ascertainment scheme. These procedures are still computationally intensive. It takes several days on a fast workstation to perform all of the analyses described below.

Estimating migration parameters

We estimate the set of demic migration parameters by fitting the expected numbers of SNPs per SDL segregating in each deme to the observed values, conditional on ascertainment. Let S_{Dk} and S_{Ok} be the numbers of segregating sites in deme k for some SDL, and let $S_A^{<k>}$ be the number of SNPs discovered on that SDL which are not segregating in deme k. Thus, $S_A^{<k>}$ includes the ascertainment-only SNPs, S_A, and the overlap SNPs that are not polymorphic in the data sample from deme k. The expected

number of SNPs segregating in the data sample from deme k, given the parameters of the model, and the ascertainment scheme is

$$E[S_{Dk} + S_{Ok} \mid Z \geq S_A^{<k>} + S_{Ok} \geq 2, \theta, M] \tag{4}$$

where $Z = \infty$ for datasets 1 and 2, and $Z = [l/100]$ for dataset 3. Appendix A describes how we compute (4), first by conditioning on the genealogy of the sample then "integrating" over genealogies using simulations. We solve numerically for M and θ by minimizing the difference between (4) and the observed values of S_{Dk} and S_{Ok}. We later discard this estimates of θ in favor of the maximum likelihood estimate described below. However, these moment-based and maximum likelihood estimates of θ were very similar for all three datasets.

The reason Q and T do not appear in equation (4) is that we estimate M only for the case of no change in effective size, $Q = 1$. The parameter T is meaningless in this case. This was done for computational reasons, namely that it is too computationally expensive to estimate M for every value of Q and T. This introduces some error in the results: the likelihood is accurately estimated for $Q = 1$, but will be underestimated for other values of Q (and T). Thus, the direction of error is conservative with respect to the null hypothesis of no change in effective population size.

Estimating θ

Once we have estimated the set of demic migration parameters, M, these are fixed for the rest of the analysis. We calculate the likelihood based on the numbers of SNPs per SDL, conditional on these and on ascertainment:

$$L_S(\theta, Q, T) \;=\; P(S_D, S_O \mid Z \geq S_A + S_O \geq 2, \theta, Q, T, \widehat{M}) \tag{5}$$

Appendix B describes how this quantity is computed. We use (6) to optimize for θ over a grid of paired values of Q and T. The justification for doing this is that most of the information regarding θ is in the number of SNPs per SDL, not in their unrooted allele frequencies (Fu 1994). Thus, our likelihood function, (9) below, is probably close to the true likelihood based on all the data. The values of θ obtained in this step are then fixed, together with the M from before, in computing the likelihood of Q and T using the frequency data.

Joint ML surface estimation for Q and T

If we take $\widehat{\theta}$ to mean the estimates of θ over the grid of Q and T, then the above analysis yields

$$L_S(Q,T) \;=\; P(S_D, S_O \,|\, Z \geq S_A + S_O \geq 2, \widehat{\theta}, Q, T, \widehat{M}) \tag{6}$$

This is the joint likelihood for Q and T based on the distribution of the numbers of SNPs per SDL. We can combine this information with the following likelihood analysis of the SNP frequencies, because the results are independent.

Let the count of the less frequent base at data-only SNP i be $X_D^{(i)}$, and the count of the less frequent base at overlap SNP i be $X_O^{(i)}$. The frequency data at a SDL can be summarized as

$$X \;=\; \left\{X_D^{(1)}, \ldots, X_D^{(S_D)}, X_O^{(1)}, \ldots, X_O^{(S_O)}\right\} \tag{7}$$

Again, we do not keep track of linkage patterns between SNPs, partly because these are genotypic data, but mostly to reduce the computational burden of the calculating the likelihood. The frequency-based likelihood is computed conditional upon the numbers of

SNPs at a SDL:

$$L_X(Q,T) = P(X \mid S_D, S_O, Z \geq S_O + S_A \geq 2, Q, T) \quad (8)$$

Appendix C describes how this is done. We consider the two likelihoods, (6) and (8), to be independent and calculate the overall likelihood of the data as

$$L(Q,T) = L_X(Q,T)L_S(Q,T) \quad (9)$$

In fact, $L_X(Q,T)$ and $L_S(Q,T)$ are not strictly independent because they are both conditional on the estimates of M and because $L_X(Q,T)$ is conditional on the estimates of θ from the optimization of $L_S(Q,T)$.

We also performed all of these analyses without conditioning on ascertainment. This was done by fixing all the lower bounds above at zero and the upper bounds at infinity, by making the ascertainment samples identical to the data samples, and by lumping all polymorphisms into one class: $S = S_D + S_O + S_A$. The next section describes the various effects that ignoring ascertainment bias can have on historical inference. In addition, we ran the analyses under the assumption of no population subdivision by setting every migration parameter equal to 10^4, and compared these results to the more general model.

Results

Our first result is not surprising: θ is overestimated if ascertainment bias is ignored. The values of θ before correcting for ascertainment bias are 0.00224, 0.00122, and 0.0021 for datasets 1, 2, and 3. The corrected values are 0.0010, 0.0008, and 0.0019. For ease of interpretation, these are the values obtained when $Q = 1$, that is when the population has been of constant effective size. Thus, they are not the global maximum likelihood estimates

for the full model, although they do not differ much from them. It is important again to note that under the demographic model used here, and with $Q = 1$, these are equivalent to the expected number of differences per site when two sequences from separate demes are compared. This is different than the average number of pairwise differences in a sample, which would include both within-deme and between-deme comparisons, and would thus be smaller.

Estimates of Migration Parameters

Figure 5 shows that demic migration parameters can be substantially overestimated when ascertainment bias is ignored. The results pictured are those for dataset 2, but the results for datasets 1 and 3 are similar. When SDLs are chosen to be highly polymorphic, the ones obtained are more likely to contain migrants or to be descended from migrants than a random sample. These values of M will remain fixed in most of the analyses below, the exception being the analysis assuming a panmictic population.

Analysis of the Number of SNPs per SDL

Figure 6 plots the likelihood surface for Q and T based on the distributions of the numbers of data-only and overlap SNPs per SDL for each of the three datasets both (a) ignoring ascertainment bias, and (b) modeling ascertainment. The lightest area, bounded by the first contour is the approximate joint 95% confidence region for Q and T, *i.e.* three log-likelihood units from the maximum. Comparing figure 6(a) to 6(b) shows that ignoring ascertainment bias prevents some very unlikely values of Q and T from being rejected, the ones in the lower left which are consistent with a recent increase in effective population size. Figure 6 also shows that the differences between ignoring and modeling ascertainment bias are similar for all three datasets when numbers of SNPs are analyzed.

Because the results in figure 6 are so similar for all three datasets, we combined them in figure 7. When the data are analyzed together and ascertainment bias is ignored (a), a

model of constant effective population size ($Q = 1$) is rejected in favor of one in which the effective size has increased. Correcting for ascertainment bias (b) shows that this result is spurious, and instead reveals a valley in the likelihood surface over much of the same area as the peak in (a). Thus, in the analysis of only the numbers of SNPs per SDL, we cannot reject the hypothesis of no change in effective size ($Q = 1$). The difference between figures 7(a) and 7(b) can be understood by referring back to figure 4, which showed that ascertainment bias decreases variation in the number of SNPs per SDL, thus creating a false signal of population growth.

Analysis of SNP Allele Frequencies

Figure 8 plots the likelihood surface for Q and T based on the allele frequencies at data-only and overlap SNPs for each of the three datasets both (a) ignoring ascertainment bias, and (b) modeling ascertainment. In contrast to the analysis of numbers of SNPs per SDL, the analysis of the frequencies shows great differences in the effects of ascertainment among the three datasets. When ascertainment bias is ignored, datasets 1 and 3 both show a peak in the likelihood surface consistent with a shrinking population. Both dataset 1 and 3 have small ascertainment samples; see table 1. Dataset 2, which has a large ascertainment sample, shows no such peak. As with the tendency for Tajima's D to be positive for datasets 1 and 3 (figure 2), these peaks reflect the overrepresentation of middle-frequency polymorphisms expected from ascertainment bias (*e.g.* see figure 3). When ascertainment bias is modeled properly, as in figure 8(b), all three datasets show the same pattern and none of them reject a constant effective population size. This pattern is similar to that found in the analysis of numbers of SNPs shown in figures 6 and 7, and to the frequency-based surface for dataset 2 (a). That is, the correction of frequencies for ascertainment bias is minor for dataset 2, but is quite striking for datasets 1 and 3.

Combined Analysis with and without Subdivision

Encouraged by the similarity of the results for all three datasets in part (b) of both figure 6 and figure 8, we combined the results of all the analyses according to equation (9). This gives us our best estimate of the demographic history of humans and is shown in figure 9(b). Using just the numbers of SNPs per SDL or just the SNP allele frequencies, it was not possible to reject the hypothesis of no change in effective size, but when all the data are used, a significant signature of population growth emerges. Figure 9(a) shows the corresponding overall picture when it is assumed that the human population is not subdivided. Even if we take ascertainment bias into account, if we ignore population subdivision then we also ignore this apparent signal of population growth in the data. We call this signal apparent because its significance depends on our estimates of M, and we have not properly accounted for variation in these. However, we note that there is also a peak for $Q < 1$ in figure 9(a), which is not visible in the figure because the contours are drawn three log-likelihood units apart. Thus, regardless of our estimates of M, these data support a scenario of population growth, but if we have underestimated M for some reason we may be wrong in calling it significant.

Discussion

Our analysis reveals two very different effects of ascertainment bias: a decrease in among-SDL variation in SNP number and an increase in heterozygosity (allele frequency) within SDLs. The second of these is fairly well known, but the first is not. We have also shown that these two kinds of bias have opposite effects on inferences about historical demography. This is illustrated in figures 3 and 4 for simulated data, and in figures 6 and 8 for polymorphism data from humans. Figure 6 shows close agreement among the three diverse datasets exactly when we expect the effects of ascertainment to be similar for all three. In this analysis of the numbers of SNPs per SDL, when results for the three datasets are pooled to produce figure 7, ascertainment bias introduces a false signal of population

expansion. In contrast, figure 8 shows disagreement among datasets when we expect the magnitude of ascertainment bias to differ, but close agreement when the ascertainment process is included in the likelihood model. In this case, when the frequencies of SNPs are analyzed (figure 8a, datasets 1 and 3), ascertainment bias produces a false signal of population decline. Comparison of these results to figures 3 and 4, and the good agreement among datasets, lends support to the overall picture of human history suggested by figure 9(b).

Wakeley (1999) fit a restricted version of this same demographic model, in which it was assumed that all demes have the same migration parameter, to restriction fragment length polymorphism (RFLP) data from a worldwide sample of humans (Bowcock et al. 1987; Matullo et al. 1994; Poloni et al. 1995). A pattern like that in figure 8(a) for datatsets 1 and 3 was found. While those RFLP data are known to be subject to ascertainment bias (Mountain & Cavalli-Sforza 1994) its contribution to this pattern could not be directly assessed (Wakeley 1999). The present work suggests that the apparent signature of a decrease in effective size observed for the RFLP data in Wakeley (1999) is probably the result of ascertainment bias.

In our computations, we have assumed that recombination and gene conversion do not occur in these short SDLs and that θ does not vary among loci. Both assumptions are false, and a more complete approach would account for this. Our approach was to delete the loci which showed direct evidence of multiple mutations, recombination or gene conversion. Recombination and gene conversion will certainly affect the distribution of the numbers of SNPs per SDL, and could bias the results non-conservatively (Hudson 1983b; Kaplan & Hudson 1985), although the interaction between recombination, ascertainment, demography, and our deletion of recombinant SDLs is difficult to predict. Only 5% of SDLs showed evidence of recombination or gene conversion (Ardlie et al. 2001). As for mutation, there could still be some variation in θ among the SDLs we analyzed. This would result in

greater variation in the number of SNPs per SDL than a constant-size population model would predict. However, this would indicate population decline, which we did not observe (Figures 6b and 6b). The effects of these phenomena on the allele frequencies at SNPs are difficult to predict, but the fact that identical results were obtained whether or not we deleted aberrant SDLs indicates that none of these effects are very strong.

Clearly, the effects of the polymorphism discovery process on later demographic inferences can be quite pronounced. Further, the direction of the bias introduced is not always the same; it depends on which aspect of the data is used for inference. Caution in the design of experiments and in the choice of markers seems indicated. However, our results are also encouraging. If the discovery process is known and ascertainment bias is modeled, then accurate demographic inferences can be made. The present data suggest that both population subdivision and changes in effective population size have been important in human history. Within the limits of our model and our methods of analysis, the data indicate a history of growth in effective size within the context of a subdivided population. The joint 95% confidence region for Q and T, enclosed by the first contour in figure 9(b), is quite broad, consistent with other recent work (Wall & Przeworski 2000), despite the fact that the human population has increased dramatically in census size. Because the effective size of a population depends both on the census size and on the rates and pattern of migration across the population (Wright 1943; Nei & Takahata 1993; Wakeley 2001), studies of historical changes in effective population size must also take subdivision into account. A comparison of figures 9(a) and 9(b) illustrates how population subdivision and growth can be conflated. When subdivision is ignored, the signal of growth in these data is missed. Further, the unexpectedly small observable effect of growth in human genetic data may be due to changes in rates and/or patterns of migration.

Acknowledgments

We thank Eric S. Lander for continuing support and helpful comments on an earlier version of the manuscript. RN and JW were supported by National Science Foundation grant DEB-9815367 to JW. This work was supported in part by grants from the National Institutes of Health to Eric S. Lander.

References

Altshuler D, Pollar VJ, Cowles CR, Etten W JV, Baldwin J, Linton L, Lander ES (2000) A SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407:513–516

Ardlie K, Liu-Cordero SN, Eberle M, Daly M, Barrett J, Winchester E, Lander ES, Krugliak L (2001) Lower than expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. Am J Hum Genet 69:582–589

Arratia R, Barbour AD, Tavaré S (1992) Poisson process approximations for the Ewens sampling formula. Ann Appl Probab 2:519–535

Bowcock AM, Bucci C, Hebert JM, Kidd JR, Kidd KK, Friedlaender JS, Cavalli-Sforza LL (1987) Study of 47 DNA markers in five populations from four continents. Gene Geography 1:47–64

Cann RL, Stoneking M, Wilson AC (1987) Mitochondrial DNA and human evolution. Nature 325:31–36

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daly GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature Genetics 22:231–237

Ewens WJ (1972) The sampling theory of selectively neutral alleles. Theoret Pop Biol 3:87–112

Ewens WJ, Spielman RS, Harris H (1981) Estimation of genetic variation at the DNA level from restriction endonuclease data. Proc Natl Acad Sci, USA 78:3748–3750

Fu X-Y (1994) Estimating effective population size or mutation rate using the frequencies of mutations in various classes in a sample of DNA sequences. Genetics 138:1375–1386

Fu X-Y (1995) Statistical properties of segregating sites. Theoret Pop Biol 48:172–197

Fu X-Y, Li W-H (1993) Statistical tests of neutrality of mutations. Genetics 133:693–709

Hawks J, Hunley K, Lee S-H, Wolpoff M (2000) Population bottlenecks and Pleistocene human evolution. Mol Biol Evol 17:2–22

Hey J (1997) Mitochondrial and nuclear genes present conflicting portraits of human origins. Mol Biol Evol 14:166–172

Hudson RR (1983a) Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217

Hudson RR (1983b) Properties of a neutral allele model with intragenic recombination. Theoret Pop Biol 23:183–201

Kaessmann H, Wiebe V, Pääbo S (1999) Extensive nuclear DNA sequence diversity among chimpanzees. Science 286:1159–1162

Kaplan NL, Hudson RR (1985) The use of sample genealogies for studying a selectively neutral m-loci model with recombination. Theoret Pop Biol 28:382–396

Kingman J FC (1982) On the genealogy of large populations. J Appl Prob 19A:27–43

Kuhner MK, Beerli P, Yamato J, Felsenstein J (2000) The usefulness of single nucletide polymorphism data for estimating population parameters. Genetics 156:439–447

Matullo G, Griffo RM, Mountain JL, Piazza A, Cavalli-Sforza LL (1994) RFLP analysis on a sample from northern Italy. Gene Geography 8:25–34

Mountain JL, Cavalli-Sforza LL (1994) Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. Proc Natl Acad Sci, USA 91:6515–6519

Nei M, Takahata N (1993) Effective population size, genetic diversity, and coalescence time in subdivided populations. J Mol Evol 37:240–244

Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154:931–942

Poloni ES, Excoffier L, Mountain JL, Langaney A, Cavalli-Sforza LL (1995) Nuclear DNA polymorphism in a Mandenka population from Senegal: comparison with eigth other human populations. Ann Hum Genet 59:43–61

Przeworski M, Hudson RR, DiRienzo A (2000) Adjusting the focus on human variation. Trends in Genetics 16:296–302

Sherry ST, Harpending HC, Batzer MA, Stoneking M (1997) *Alu* evolution in human populations: using the coalescent to estimate effective population size. Genetics 147:1977–1982

Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129:555–562

Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, et al. (2001) Haplotype variation and linkage disequlibrium in 313 human genes. Science 293:489–493

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

Tavaré S (1984) Lines-of-descent and genealogical processes, and their application in population genetic models. Theoret Pop Biol 26:119–164

The International SNP Map Working Group (2001) A map of human genome sequence variation containing 142 million single nucleotide polymorphisms. Nature 409:928–933

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wislon AC (1991) African populations and the evolution of human mitochondrial DNA. Science 253:1503–1507

Wakeley J (1998) Segregating sites in Wright's island model. Theoret Pop Biol 53:166–175

Wakeley J (1999) Non-equilibrium migration in human history. Genetics 153:1863–1871

Wakeley J (2001) The coalescent in an island model of population subdivision with variation among demes. Theoret Pop Biol 59:133–144

Wall JD, Przeworski M (2000) When did the human population size start increasing? Genetics 155:1865–1874

Wang DG, Fan J-B, Siao C-J, Berno A, Young P, Sapolsky R, Ghandour G, et al. (1998) Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theoret Pop Biol 7:256–276

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Wright S (1943) Isolation by distance. Genetics 28:114–138

Yu N, Zhao Z, Fu Y-X, Sambuughin N, Ramsay M, Jenkins T, Leskinen E, Patthy L, Jorde LB, Kuromori T, Li W-H (2001) Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol Biol Evol 18:214–222

Appendix A

Let C_k represent the condition: $Z \geq S_A^{<k>} + S_{Ok} \geq 2$. Then, starting from (4) in the text and using the rules for conditional probability, we have

$$E[S_{Dk} + S_{Ok} \,|\, C_k, \theta, M] = \int_\Psi E[S_{Dk} + S_{Ok} \,|\, C_k, \theta, M, G] P(G|C_k, \theta, M) dG \quad (10)$$

$$= \frac{\int_\Psi E[S_{Dk} + S_{Ok} \,|\, C_k, \theta, M, G] P(G, C_k|\theta, M) dG}{P(C_k|\theta, M)} \quad (11)$$

$$= \frac{\int_\Psi E[S_{Dk} + S_{Ok} \,|\, C_k, \theta, M, G] P(C_k|\theta, M, G) P(G|\theta, M) dG}{\int_\Psi P(C_k|\theta, M, G) P(G|\theta, M) dG} \quad (12)$$

In equations (10) through (12) and below, we use Ψ to denote the set of all possible genealogies with branch lengths. This representation suggests that $E[S_{Dk} + S_{Ok} \,|\, C_k, \theta, M]$ can be estimated consistently as

$$\frac{\frac{1}{n}\sum_{i=1}^{n} E[S_{Dk} + S_{Ok} \,|\, C_k, \theta, M, G_i] P(C_k|\theta, M, G_i)}{\frac{1}{n}\sum_{i=1}^{n} P(C_k|\theta, M, G_i)} \quad (13)$$

where G_i is one of n genealogies simulated from $P(G|\theta, M)$.

For each simulated tree, we store T_{Dk}, T_{Ok}, and $T_A^{<k>}$. These are the total length of branches in the genealogy which could give rise to a SNP that is segregating in the data-only sample from deme k, in the overlap sample from deme k, and in the total ascertainment sample but not in deme k, respectively. Branch lengths are measured in units of $2N_e$ generations. Given T_{Dk}, T_{Ok}, and $T_A^{<k>}$, the number of mutations in each of these three classes are independent Poisson random variables with parameters, $T_{Dk}\theta/2$,

$T_{Ok}\theta/2$, and $T_A^{<k>}\theta/2$. Thus, we have

$$E[S_{Dk} + S_{Ok} \mid C_k, \theta, M, G_i] = E[S_{Dk} \mid C_k, \theta, M, G_i] + E[S_{Ok} \mid C_k, \theta, M, G_i] \quad (14)$$

$$= \frac{\theta T_{Dk}}{2} + E[S_{Ok} \mid C_k, \theta, M, G_i] \quad (15)$$

The second term in (15) is calculated by further conditioning on the value of $S_A^{<k>}$:

$$E[S_{Ok} \mid C_k, \theta, M, G_i] = \sum_{j=0}^{Z} E[S_{Ok} \mid Z - j \geq S_{Ok} \geq 2 - j, \theta, M, G_i] P(S_A^{<k>} = j) \quad (16)$$

The expectation on the right is given by

$$E[S_{Ok} \mid Z - j \geq S_{Ok} \geq 2 - j, \theta, M, G_i] = \frac{\sum_{x=2-j}^{Z-j} x P(S_{Ok} = x)}{\sum_{x=2-j}^{Z-j} P(S_{Ok} = x)} \quad (17)$$

and $P(S_{Ok} = x)$ and $(S_A^{<k>} = j)$ are the appropriate Poisson probabilities. Similarly, the term, $P(C_k|\theta, M, G_i)$, in (13) is given by

$$P[Z \geq S_A^{<k>} + S_{Ok} \geq 2, \theta, M, G_i] = \sum_{x=2}^{Z} P(S_A^{<k>} + S_{Ok} = x) \quad (18)$$

and the sum, $S_A^{<k>} + S_{Ok}$, is Poisson distributed with parameter $(T_A^{<k>} + T_{Ok})\theta/2$.

Appendix B

We compute the likelihood, $L_S(\theta, Q, T)$, as follows:

$$L_S(\theta,Q,T) = P(S_D, S_O \,|\, Z \geq S_A + S_O \geq 2, \theta, Q, T, \widehat{M}) \quad (19)$$

$$= \frac{P(S_D, S_O, Z \geq S_A + S_O \geq 2 \,|\, \theta, Q, T, \widehat{M})}{P(Z \geq S_A + S_O \geq 2 \,|\, \theta, Q, T, \widehat{M})} \quad (20)$$

$$\approx \frac{\frac{1}{n}\sum_{i=1}^{n} P(S_D, S_O, Z \geq S_A + S_O \geq 2 \,|\, \theta, Q, T, \widehat{M}, G_i)}{\frac{1}{n}\sum_{i=1}^{n} P(Z \geq S_A + S_O \geq 2 \,|\, \theta, Q, T, \widehat{M}, G_i)} \quad (21)$$

where G_i is a genealogy simulated from $P(G|\theta, Q, T, \widehat{M})$. For each genealogy we store the values of T_D, T_O, and T_A, which are the total lengths of branches which contribute to S_D, S_O, and S_A, respectively. Given the genealogy and therefore these times, S_D, S_O, and S_A are independent Poisson random variables with parameters $T_D\theta/2$, $T_O\theta/2$, and $T_A\theta/2$, respectively. Thus, we have

$$P(Z \geq S_A + S_O \geq 2 \,|\, \theta, Q, T, \widehat{M}, G_i) = \sum_{j=2}^{Z} P(S_D + S_O = j \,|\, \theta, Q, T, \widehat{M}, G_i) \quad (22)$$

Because of independence, the term in the numerator of (21) is given by

$$P(S_D, S_O, Z \geq S_A + S_O \geq 2 \,|\, \theta, Q, T, \widehat{M}, G_i) = P(S_D \,|\, \theta, Q, T, \widehat{M}, G_i) P(S_O \,|\, \theta, Q, T, \widehat{M}, G_i)$$

$$\times P(Z - S_O \geq S_A \geq 2 - S_O \,|\, S_O, \theta, Q, T, \widehat{M}, G_i) \quad (23)$$

The first two terms on the right in (23) are simple Poisson probabilities, and the third term is just the sum of these over a range of values:

$$P(Z - S_O \geq S_A \geq 2 - S_O \,|\, S_O, \theta, Q, T, \widehat{M}, G_i) = \sum_{j=2-S_O}^{Z-S_O} P(S_A = j \,|\, \theta, Q, T, \widehat{M}, G_i) \quad (24)$$

Appendix C

To save space, let C represent the condition: $Z \geq S_A + S_O \geq 2$ and $\Omega^* = \{\widehat{\theta}, Q, T, \widehat{M}\}$. We compute the likelihood as follows:

$$
\begin{aligned}
L_X(\Omega^*) &= P(X|S_D, S_O, C, \Omega^*) && (25)\\
&= \frac{P(X, C|S_D, S_O, \Omega^*)}{P(C|S_D, S_O, \Omega^*)} && (26)\\
&= \frac{P(X, C, S_D, S_O|\Omega^*)}{P(C, S_D, S_O|\Omega^*)} && (27)\\
&= \frac{\int_\Psi P(X, C, S_D, S_O|\Omega^*, G)P(G|\Omega^*)dG}{\int_\Psi P(C, S_D, S_O|\Omega^*, G)P(G|\Omega^*)dG} && (28)\\
&= \frac{\int_\Psi P(X|S_D, S_O, \Omega^*, G)P(C|S_O, \Omega^*, G)P(S_D|\Omega^*, G)P(S_O|\Omega^*, G)P(G|\Omega^*)dG}{\int_\Psi P(C|S_O, \Omega^*, G)P(S_D|\Omega^*, G)P(S_O|\Omega^*, G)P(G|\Omega^*)dG} && (29)\\
&\approx \frac{\frac{1}{n}\sum_{i=1}^{n} P(X|S_D, S_O, \Omega^*, G_i)P(C|S_O, \Omega^*, G_i)P(S_D|\Omega^*, G_i)P(S_O|\Omega^*, G_i)}{\frac{1}{n}\sum_{i=1}^{n} P(C|S_O, \Omega^*, G_i)P(S_D|\Omega^*, G_i)P(S_O|\Omega^*, G_i)} && (30)
\end{aligned}
$$

where, again, Ψ denotes the set of all possible genealogies with branch lengths. The steps above rely upon the fact that conditioning on the genealogy of the sample makes S_D, S_O, and S_A independent and Poisson-distributed with respective parameters, $\theta T_D/2$, $\theta T_O/2$, and $\theta T_A/2$ defined by the genealogy. Again G_i is a genealogy simulated from $P(G|\Omega^*)$. As above, we can compute each of the terms in (30) easily. For example, $P(S_D|\Omega^*, G_i)$ and $P(S_O|\Omega^*, G_i)$ are again simply Poisson probabilities, with parameters $T_D\theta/2$ and $T_O\theta/2$. Also, the term $P(C|S_O, \Omega^*, G_i)$ is identical to expression (24).

Lastly, it follows from the Poisson mutation process that, given a mutation occurs, the place it occurs is uniformly distributed among the branches in the genealogy in proportion

to their lengths. Therefore, we have

$$P(X|S_D, S_O, \Omega^*, G_i) = \prod_{i=1}^{S_D} \frac{t_D^{(i)}}{T_D} \prod_{i=1}^{S_O} \frac{t_O^{(i)}}{T_O} \tag{31}$$

where $t_D^{(i)}$ and $t_O^{(i)}$ are the total length of branches in the genealogy on which a mutation would produce polymorphic site pattern $X_D^{(i)}$ and $X_O^{(i)}$, respectively. The terms $t_D^{(i)}/T_D$ and $t_O^{(i)}/T_O$ in (31) are the probabilities that a mutation which has occurred on the genealogy has occurred on a branch corresponding to the patterns $X_D^{(i)}$ and $X_O^{(i)}$.

	Dataset 1			Dataset 2			Dataset 3		
Deme	n_D	n_O	n_A	n_D	n_O	n_A	n_D	n_O	n_A
Utah – CEPH	6	0	5	0	6	10	6	0	2
Venezuelan – CEPH	2	0	4	0	2	0	2	0	0
Irish	2	0	0	2	0	0	2	0	0
Russian/Adygei	6	0	0	0	6	4	6	0	0
Russian/Zuevsky	4	0	0	0	4	6	2	2	0
Chinese	8	0	0	0	8	2	6	2	0
Cambodian	6	0	0	0	6	0	6	0	0
Melanesian	8	0	0	6	2	0	6	2	0
Japanese	4	0	0	2	2	2	2	2	0
Taiwanese/Ami	6	0	0	6	0	0	6	0	0
Taiwanese/Atayal	4	0	0	4	0	0	4	0	0
South Indian	2	0	0	2	0	0	2	0	0
Amerindian	8	0	0	8	0	0	8	0	2
CAR/Pygmy	6	0	0	6	0	0	6	0	2
Zaire/Pygmy	4	0	0	4	0	0	4	0	0
Sudanese/Dinka	4	0	0	4	0	0	4	0	0
Sudanese/Shilluk	2	0	0	2	0	0	2	0	0
Sudanese/Arab	2	0	0	2	0	0	2	0	0
Ethiopean/Semitic	6	0	0	6	0	0	6	0	0
Lybian/Semitic	4	0	0	4	0	0	4	0	0
Amish – CEPH	0	0	6	0	0	0	0	0	2
African American	0	0	0	0	0	20	0	0	2
French – CEPH	0	0	0	0	0	0	0	0	2
Totals	94	0	15	60	34	46	86	8	12

Table 1: The numbers of data-only (n_D), overlap (n_O), and ascertainment-only (n_A) chromosomes/haplotypes sampled from each deme.

Figure 1. An example genealogy, drawn with branch lengths equal to the coalescent expectations, which shows the structure of the data analyzed here: "D," "O," and "A" are samples which are only in the dataset, samples which are in the dataset and the ascertainment set, and samples which are only in the ascertainment set. Three types of branches are distinguished, corresponding to the three kinds of observable polymorphisms discussed in the text.

Figure 2. The distribution of Tajima's (1989) D among SDLs in each of the three datasets.

Figure 3. The expected numbers of SNPs segregating in different frequencies in a sample of size $n_D + n_O = 10$, relative to the number of singleton polymorphisms. Panel (a) shows the effect of requiring a SDL to have at least one SNP in the first n_O samples drawn from the population, and (b) shows the effects of separating SDLs into classes with different numbers of SNPs with $n_D = 0$. Results are averages over one hundred thousand simulated datasets for a 400 bp long SDL with $\theta = 0.0005$ per bp.

Figure 4. The coefficient of variation of the number of SNPs per SDL a sample of size $n_D + n_O = 10$. Panel (a) assumes $n_D = 0$ and shows the effects of requiring SDLs to have at least k SNPs, and (b) shows the effect of requiring a SDL to have at least one SNP but this must be segregating in the first n_O samples drawn from the population. Results are averages over one hundred thousand simulated datasets for a 400 bp long SDL with $\theta = 0.0005$ per bp.

Figure 5. Estimates of demic migration parameters, $2Nm$, for dataset 2 both with and without modelling ascertainment. For this dataset, five demes had infinite migration rate

estimates when ascertainment was ignored and these are not plotted above.

Figure 6. Likelihood surfaces for Q and T based on the distribution of the numbers of data-only and overlap SNPs per SDL for each of the three datasets, (a) ignoring ascertainment bias and (b) taking it into account.

Figure 7. Combined likelihood surfaces for Q and T based on the distribution of the numbers of data-only and overlap SNPs per SDL for all three datasets, (a) ignoring ascertainment bias and (b) taking it into account.

Figure 8. Likelihood surfaces for Q and T based on the allele frequencies at data-only and overlap SNPs, conditioned on their numbers, for each of the three datasets, (a) ignoring ascertainment bias and (b) taking it into account.

Figure 9. Combined likelihood surfaces for Q and T for all the data, (a) assuming that the population is panmictic, and (b) fitting the subdivided population model described in the text.

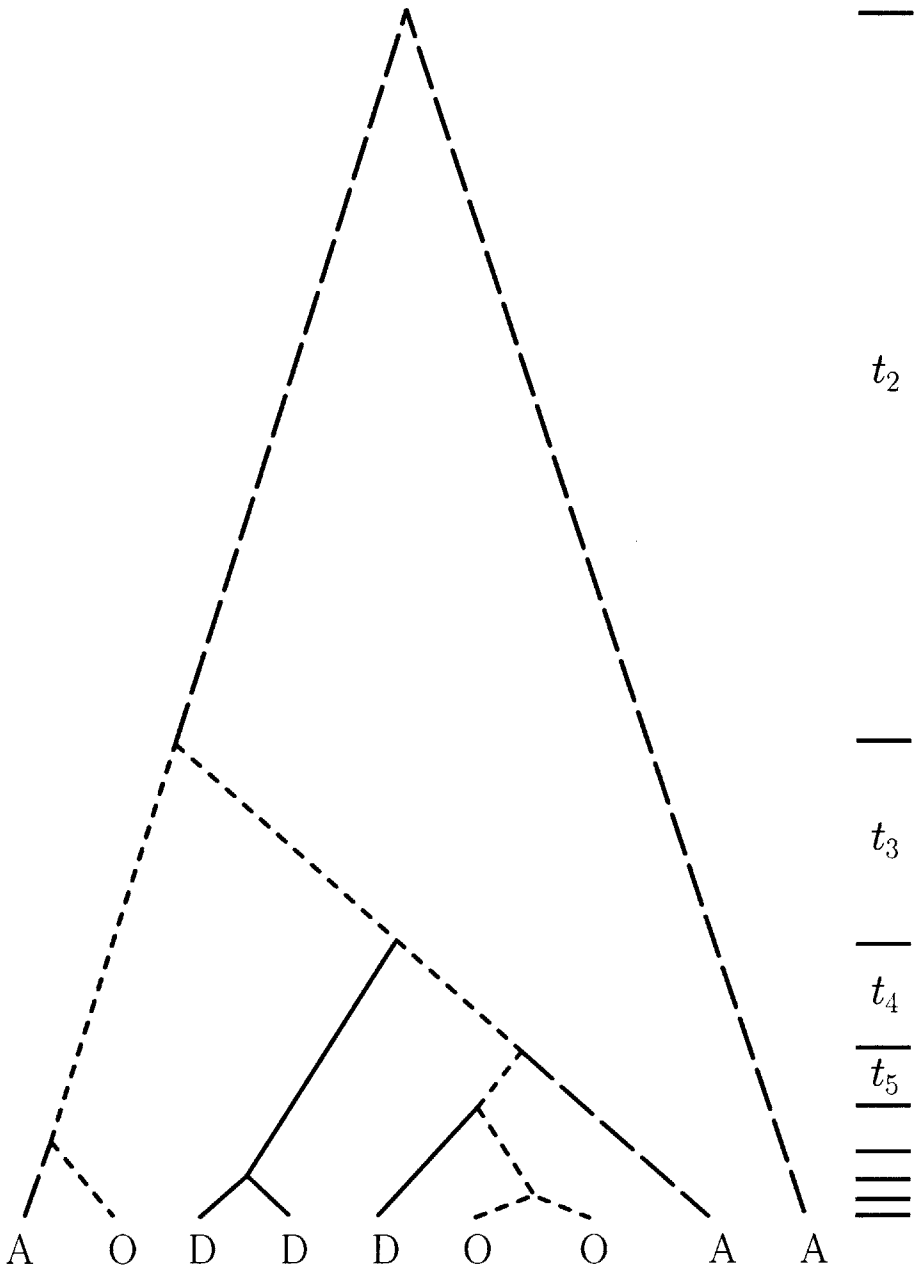


Figure 1: Wakeley et al.

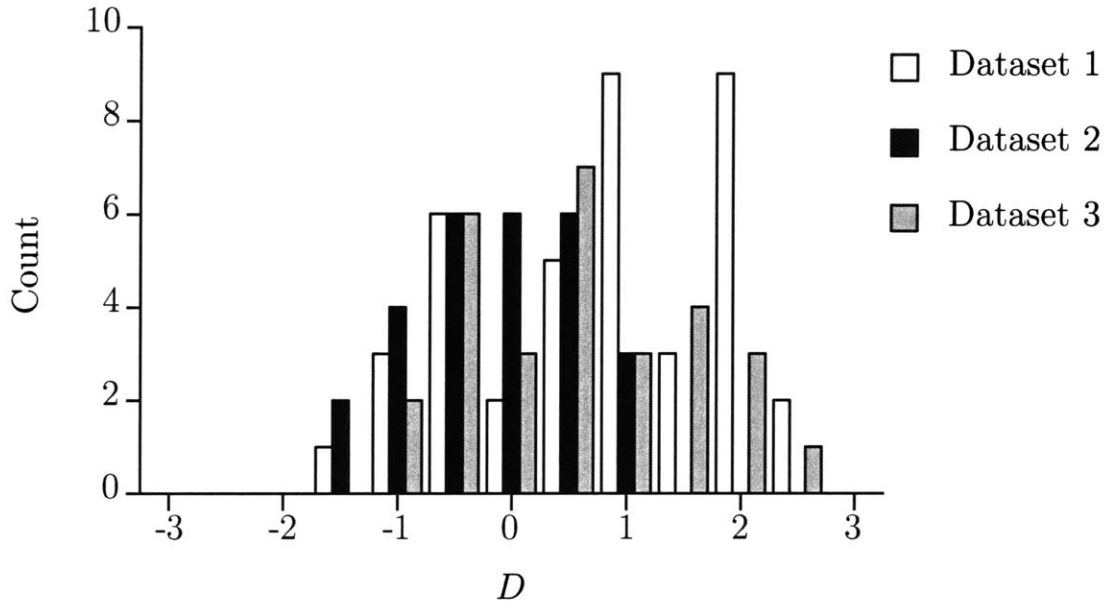


Figure 2: Wakeley et al.

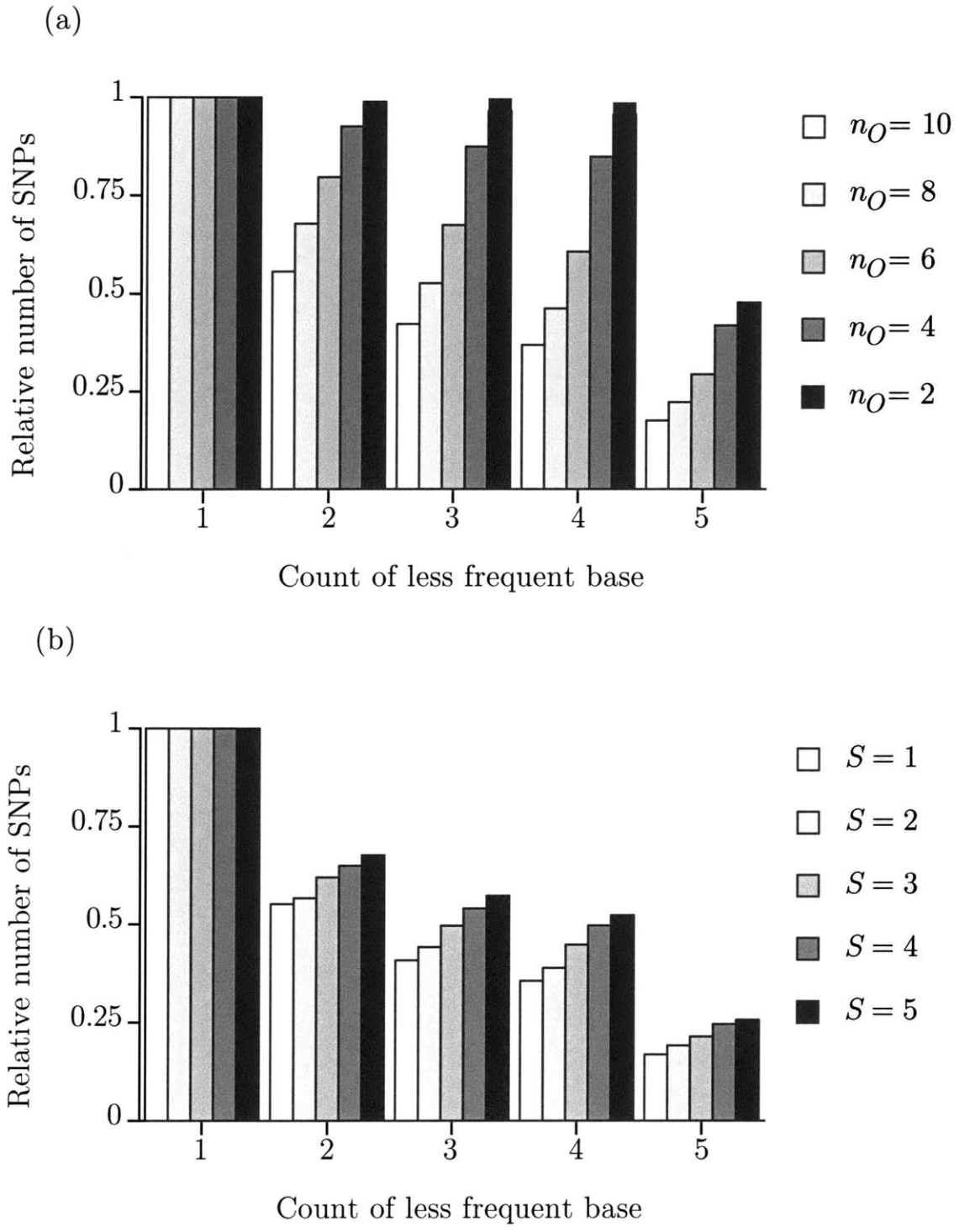


Figure 3: Wakeley et al.

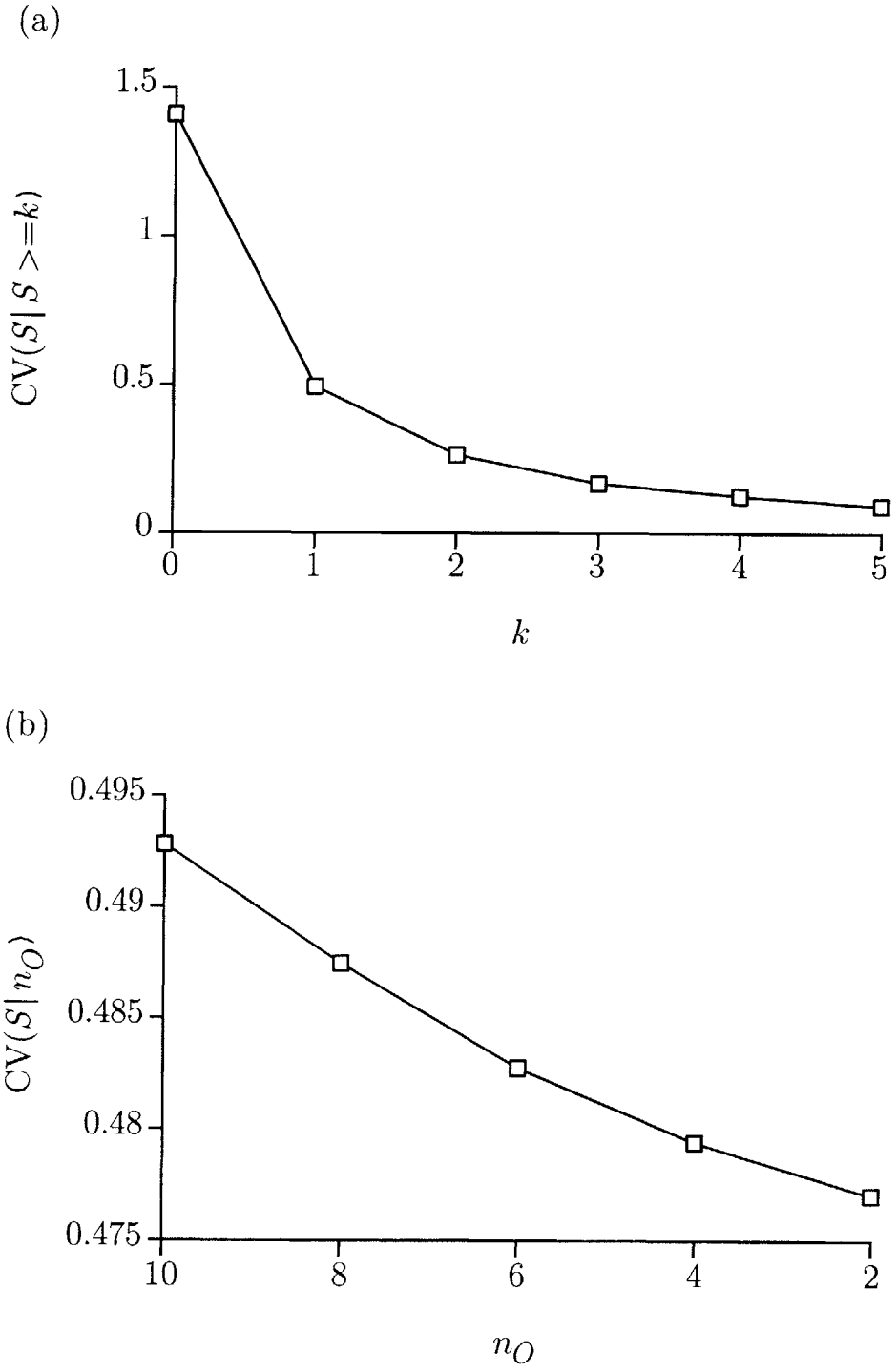


Figure 4: Wakeley et al.

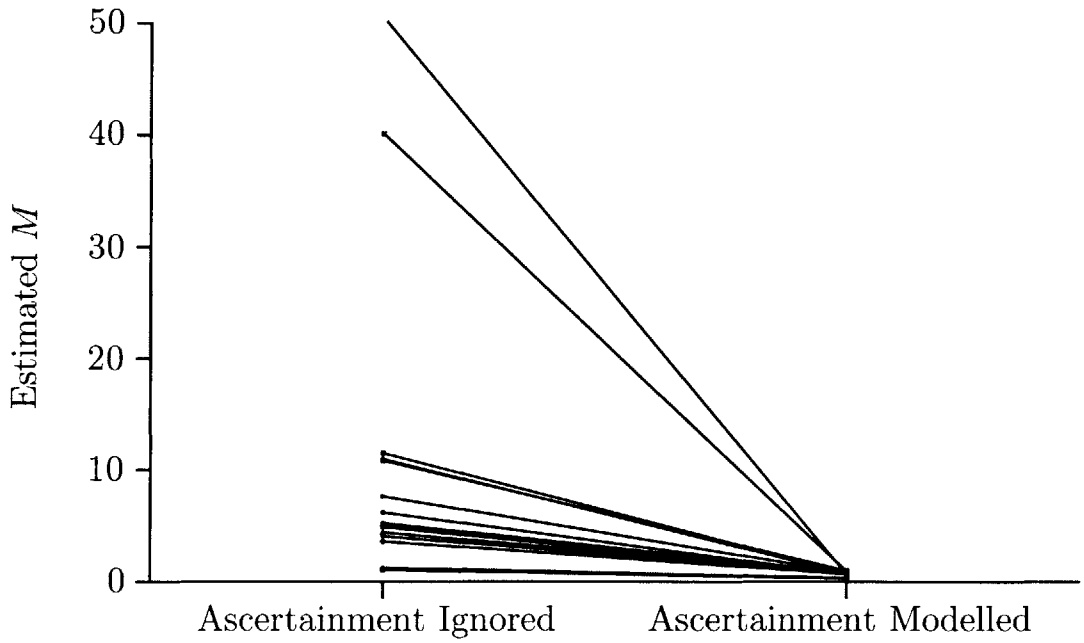


Figure 5: Wakeley et al.

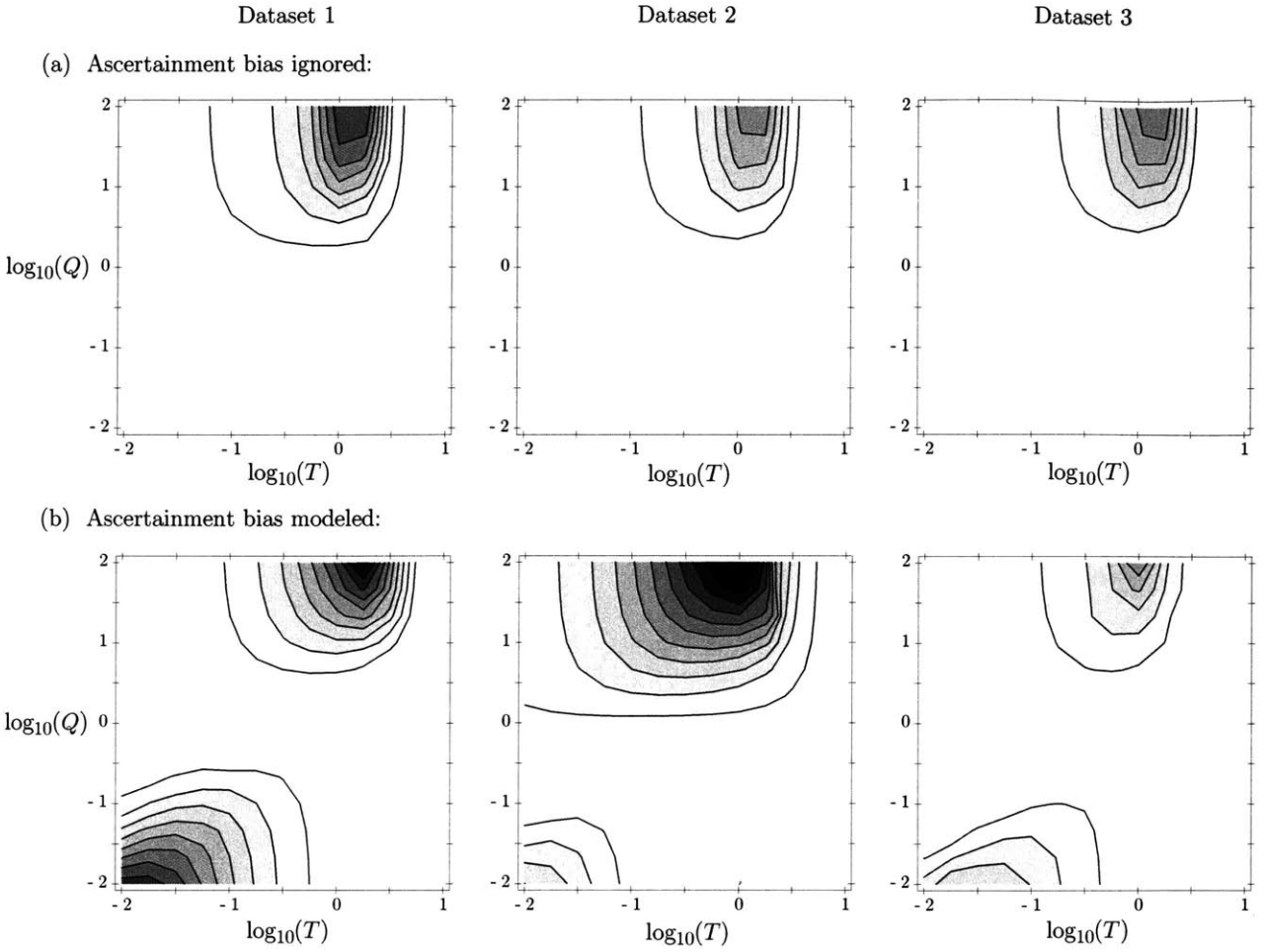
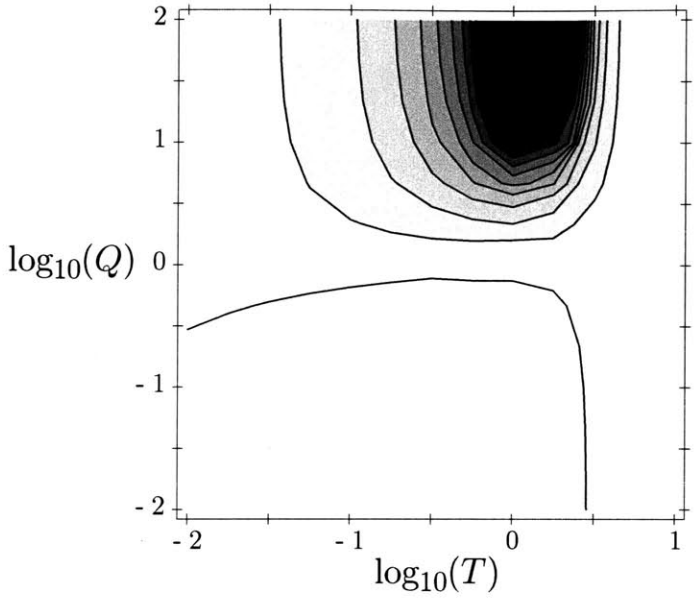


Figure 6: Wakeley et al.

(a) Ascertainment bias ignored:



(b) Ascertainment bias modeled:

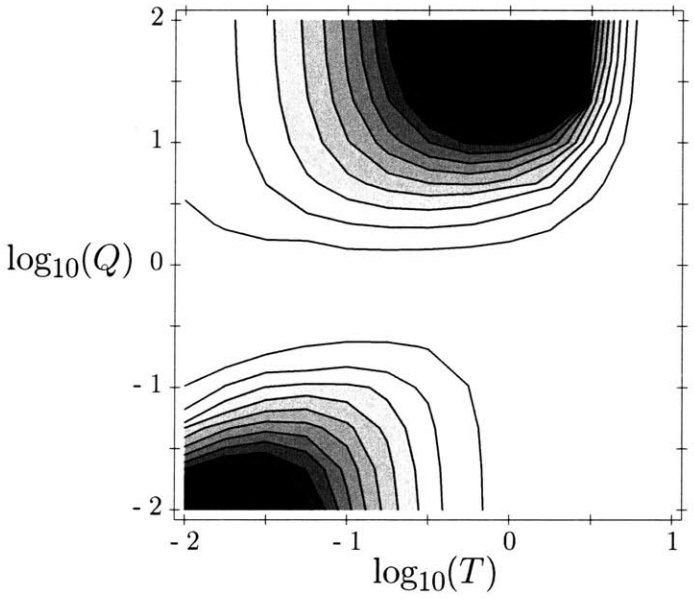


Figure 7: Wakeley et al.

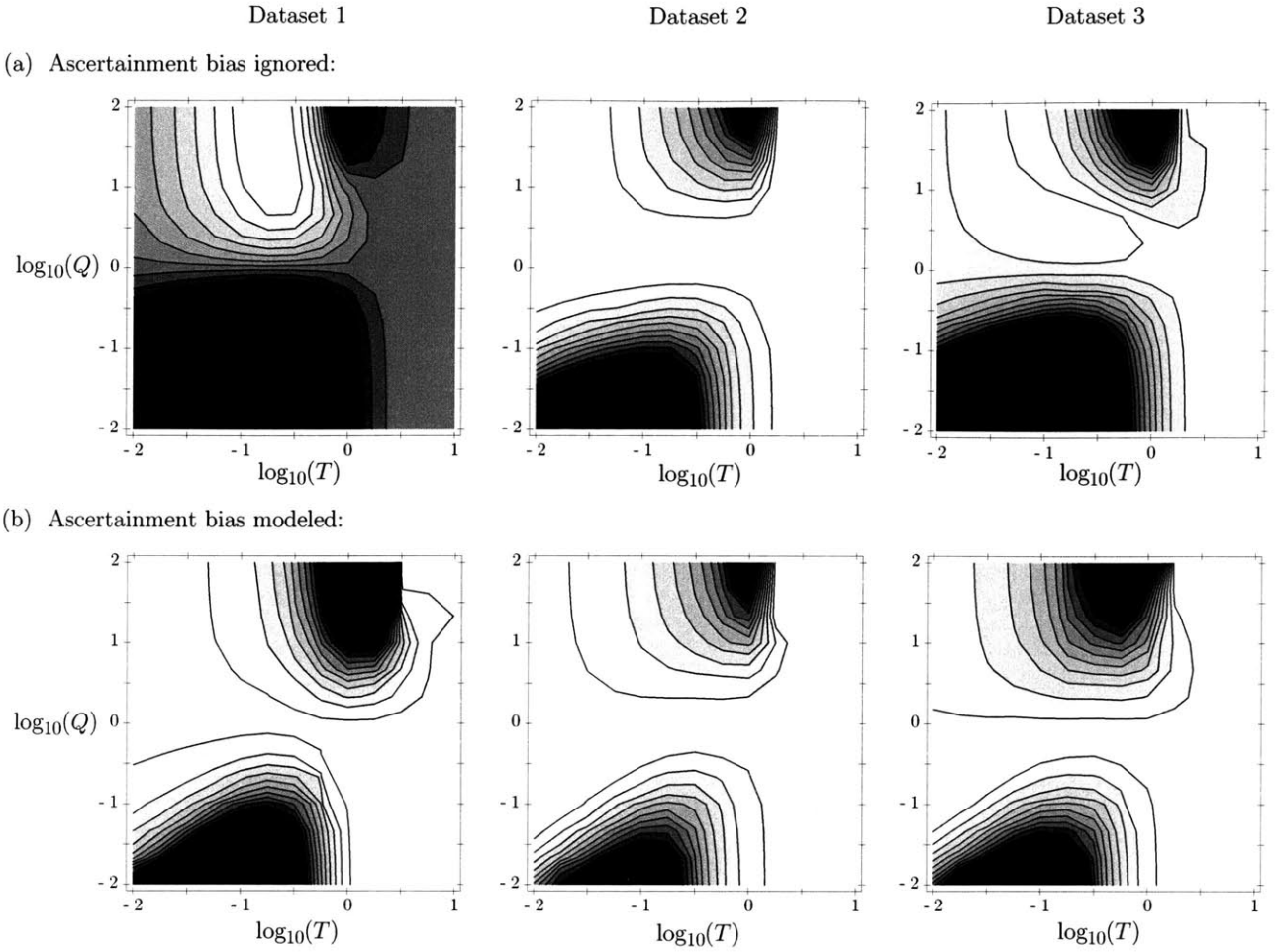
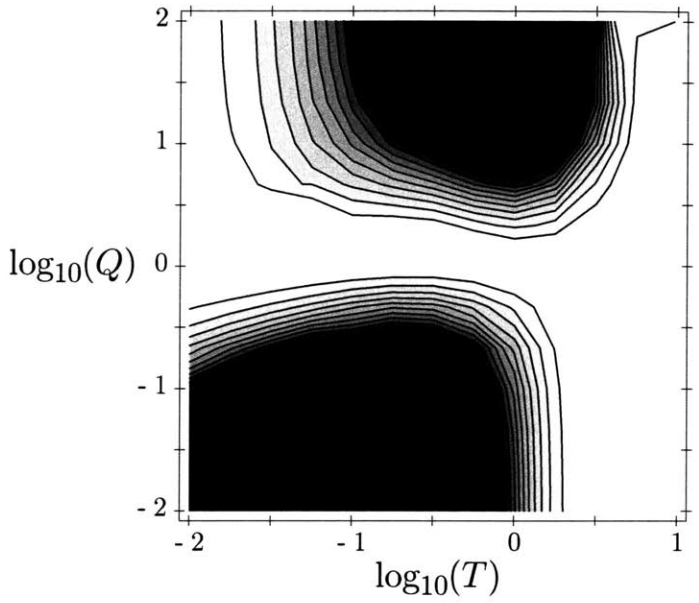


Figure 8: Wakeley et al.

(a) Population subdivision ignored:



(b) Population subdivision modeled:

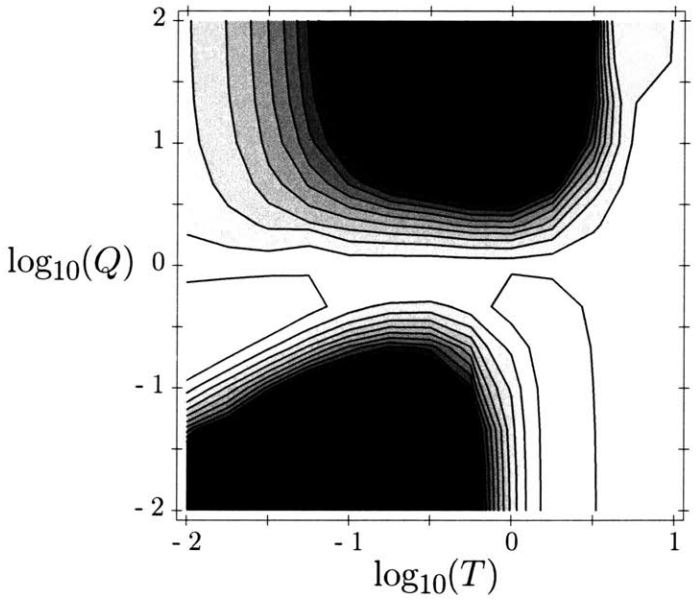


Figure 9: Wakeley et al.

Cover Page

User:	shauneen
Document:	Microsoft Word - Appendix III Cover
Server:	E-650
Time:	05/24/02 08:52:08
Pages requested:	2
Page size:	Letter
Status:	OK