

Aus der Arbeitsgruppe Bioinformatik des  
Max-Delbrück-Centrums für Molekulare Medizin (MDC),  
Berlin-Buch, in Kooperation mit der Medizinischen Fakultät  
der Charité - Universitätsmedizin Berlin

# **The Definition of Multilocus Haplotype Blocks and Common Diseases**

DISSERTATION

zur Erlangung des akademischen Grades  
Doctor rerum medicarum (Dr. rer. medic.)  
im Fach Medizin

vorgelegt der  
Medizinischen Fakultät der  
Charité – Universitätsmedizin Berlin  
Humboldt-Universität zu Berlin

von  
Herrn Dipl.-Math.  
MICHAEL NOTHNAGEL  
geboren am 22.07.1971 in Berlin

Präsident der Humboldt-Universität zu Berlin:  
Prof. Dr. Jürgen Mlynek

Dekan der Medizinischen Fakultät der  
Charité – Universitätsmedizin Berlin:  
Prof. Dr. med. Martin Paul

Gutachter:

1. Univ.Prof. Dr. em. Jens G. Reich
2. Suzanne M. Leal, Ph.D., Associate Professor
3. Prof. Dr. Andreas Ziegler

eingereicht am: 03. März 2004

Datum der Promotion  
(Tag der mündlichen Prüfung): 13. Dezember 2004

## Abstract

Current approaches to haplotype block definition target either absent recombination events or the efficient description of genomic variation. This thesis aims to define blocks of single nucleotide polymorphisms (SNP) as areas of elevated linkage disequilibrium (LD). To this end, a new entropy-based measure for LD between multiple markers/loci, the Normalized Entropy Difference, is developed and is characterized as a multilocus extension of the pairwise measure  $r^2$ . A corresponding algorithm for the block definition is proposed. Its evaluation on a data set of human chromosome 12 from the International Haplotype Map project proves the usefulness of the derived blocks with respect to several features, including their chromosomal coverage and the number and portion of common block haplotypes. The critical role of the SNP density for detectable LD and block structure is demonstrated. The success of association studies in common diseases with block haplotypes serving as multi-allelic markers will depend on whether the Common Variants/Common Diseases (CV/CD) hypothesis holds true for those diseases.

### Keywords:

multilocus linkage disequilibrium, haplotype blocks, common diseases, single nucleotide polymorphisms

## Zusammenfassung

Bisherige Methoden der Haplotyp-Block-Definition zielen entweder auf abwesende Rekombinationsereignisse oder eine effiziente Beschreibung genomischer Variation. Die vorliegende Arbeit definiert Blöcke von Single Nucleotide Polymorphisms (SNP) als Gebiete erhöhten Kopplungsungleichgewichtes (LD). Für dieses Ziel wird ein neues, entropie-basiertes Maß für LD zwischen multiplen Markern/Loci (Normalized Entropy Difference) entwickelt und als eine Multilocus-Erweiterung des paarweisen Maßes  $r^2$  charakterisiert. Ein zugehöriger Algorithmus für die Block-Definition wird vorgeschlagen. Seine Evaluierung an einem Datensatz des menschlichen Chromosoms 12 vom Internationalen Haplotype Map Projekt zeigt die Nützlichkeit der abgeleiteten Blöcke in Hinblick auf verschiedene Eigenschaften, einschließlich ihrer chromosomalen Coverage und der Anzahl sowie des Anteils der häufigen Block-Haplotypen. Der wesentliche Einfluß der SNP-Dichte auf die zu entdeckenden LD- und Blockstrukturen wird demonstriert. Der Erfolg von Assoziationsstudien in komplexen Erkrankungen mit Block-Haplotypen als multiallelischen Markern wird davon abhängen, ob die Common Variants/Common Diseases (CV/CD) Hypothese für solche Erkrankungen erfüllt ist.

### **Schlagwörter:**

Multilocus-Kopplungsungleichgewicht, Haplotyp-Blöcke, Komplexe Erkrankungen, Single Nucleotide Polymorphisms

# Contents

<b>Preface</b>	<b>v</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Genetic background of diseases . . . . .	1
1.1.1 Approaches to statistical gene mapping . . . . .	2
1.1.2 Common diseases and the benefit of haplotypes . . . . .	5
1.2 Haplotypes and linkage disequilibrium . . . . .	8
1.2.1 Estimation of haplotype frequencies . . . . .	8
1.2.2 Pairwise measures for LD . . . . .	9
1.2.3 Multilocus LD measures . . . . .	13
1.3 Methods for the definition of blocks . . . . .	14
1.4 Objective of this thesis . . . . .	16
<b>2 Measure &amp; methods</b>	<b>19</b>
2.1 The concept of entropy . . . . .	19
2.2 The normalized entropy difference $\varepsilon$ . . . . .	20
2.3 Analytical features of $\varepsilon$ . . . . .	21
2.4 An $\varepsilon$ -based block definition algorithm . . . . .	27
2.5 A data simulation algorithm . . . . .	27
<b>3 Applicability of <math>\varepsilon</math></b>	<b>29</b>
3.1 Common haplotypes, coverage, and $\varepsilon$ . . . . .	29
3.1.1 Simulation study design . . . . .	30
3.1.2 Simulation results . . . . .	34
3.2 Applicability of $\varepsilon$ . . . . .	34

3.2.1	Simulation I: A single block . . . . .	35
3.2.2	Simulation II: Large and adjacent blocks . . . . .	38
3.2.3	An established block structure . . . . .	39
<b>4</b>	<b>Block patterns on human chromosome 12</b>	<b>44</b>
4.1	Data set description and objective . . . . .	44
4.2	Analysis of the data set . . . . .	45
4.3	Block lengths and chromosomal coverage . . . . .	47
4.3.1	Lengths and coverage of $\varepsilon$ -defined blocks . . . . .	47
4.3.2	The origin of the block length distribution . . . . .	51
4.4	Haplotypes in $\varepsilon$ -defined blocks . . . . .	52
4.5	Allele frequencies in $\varepsilon$ -defined blocks . . . . .	55
4.6	Pairwise LD measures in $\varepsilon$ -defined blocks . . . . .	55
4.7	Comparison of algorithms . . . . .	58
<b>5</b>	<b>Discussion</b>	<b>64</b>
5.1	The measure $\varepsilon$ . . . . .	64
5.2	The $\varepsilon$ -based block definition algorithm . . . . .	70
5.3	Blocks on human chromosome 12 . . . . .	72
5.4	Implications for medical research and other potential applications . . . . .	78
<b>6</b>	<b>Summary</b>	<b>82</b>
<b>7</b>	<b>Deutsche Zusammenfassung</b>	<b>86</b>
	<b>Abbreviations</b>	<b>91</b>
	<b>Bibliography</b>	<b>93</b>

# List of Figures

1.1	Schematic example of LD between two SNPs . . . . .	9
1.2	$D'$ as an indicator for missing haplotypes . . . . .	12
2.1	$\varepsilon$ 's dependence on the numbers of loci and haplotypes . . . . .	24
2.2	Comparison of $r^2$ , $\Delta S$ , and $\varepsilon$ . . . . .	26
3.1	Effect of small errors in $p$ on $-p \log p$ . . . . .	30
3.2	Simulation I: $\varepsilon$ values . . . . .	37
3.3	Simulation II: $\varepsilon$ and pairwise LD values . . . . .	40
3.4	$\varepsilon$ and pairwise LD values for Daly et al. (2001) . . . . .	42
4.1	Baylor HapMap: Pairwise LD values . . . . .	46
4.2	Baylor HapMap: $\varepsilon$ values . . . . .	48
4.3	Baylor HapMap: $\varepsilon$ -based block definition . . . . .	49
4.4	Baylor HapMap: Distributions of physical block length and window size . . . . .	53
4.5	Baylor HapMap: SNP allele frequency distribution in blocks .	56
4.6	Baylor HapMap: Correlations between $\varepsilon$ and $ D' /r^2$ . . . . .	57
4.7	Baylor HapMap: Comparison of block definitions . . . . .	60
4.8	Baylor HapMap: SNP allele distribution in blocks derived from $ D' /r^2$ . . . . .	63

# List of Tables

1.1	Table of block definition algorithms . . . . .	15
1.2	Physical block lengths in the literature . . . . .	16
3.1	Average bias of $\varepsilon_{cmn}$ for twice as many rare than common haplotypes . . . . .	32
3.2	Average bias of $\varepsilon_{cmn}$ for a total of 20 haplotypes . . . . .	33
3.3	Simulation I: percentage of accurate detections . . . . .	36
4.1	Baylor HapMap: Statistics for $\varepsilon$ -defined blocks . . . . .	50
4.2	Baylor HapMap: Concordance of block length and window size distributions . . . . .	52
4.3	Baylor HapMap: Common haplotypes in $\varepsilon$ -defined blocks . . .	54
4.4	Baylor HapMap: Correlations between $\varepsilon$ and $r^2/ D' $ . . . . .	58
4.5	Baylor HapMap: Block statistics for pairwise LD measures . .	61
4.6	Baylor HapMap: Concordance of SNP inclusion in blocks . . .	62

# Preface

Statistical genetics has seen its rise from a very specialized field to a large scientific area within the last 30 years. It combines the disciplines of medicine, biology, statistics, and computer science to find and map genetic causes of diseases in human and other organisms. Each of these areas is rapidly evolving; so is statistical genetics. First papers on haplotypes blocks appeared in 2001, whereas the frequency of published articles investigating this phenomenon changed from monthly to almost weekly in 2003.

Haplotype blocks are an interesting subject, with a number of possible applications. However, the existing methods for their definition deliver inconsistent and sometimes contradicting results and are in constant debate. The major drawback of these methods is their lack of direct assessment of multilocus LD, which is important for the mapping of disease genes in common but complex diseases, such as lipid disorders, hypertension, or Alzheimer's. In this thesis, I develop a multilocus LD measure and a block definition algorithm that is based on it and demonstrate their usefulness. Their features are investigated and compared with existing measures and methods. The application to a whole chromosome data set allows for an in-depth evaluation of them and also for some general conclusions about the nature of haplotype blocks in the human genome. The proposed LD measure could be beneficial in other applications as well.

I would like to thank a number of people for their assistance and help. First of all, I want to thank my wife, Ute Hölting, who encouraged me to work on my thesis, even when she was 10,000 kilometers away on a remote continent. I want to thank Prof. Jens G. Reich and Dr. Klaus Rohde for

their supervision and advice which made this thesis possible. I am indebted to Prof. Jürg Ott and Prof. Suzanne M. Leal for giving me the opportunity to work in their groups and for the helpful discussions, comments, and questions that very much improved my understanding of statistical genetics and the quality of this thesis. I want to thank Prof. Herbert Schuster for drawing my attention to genetic disease studies and for making my visit to New York possible. I would like to thank Prof. Richard A. Gibbs for providing the very valuable data set of human chromosome 12 ahead of public release and Prof. Gerard te Meerman for thoughtful comments on this thesis. Very special thanks go to my friends Dr. Claudia Iserhot and Dr. Boris Jerchow for not hesitating to take a first and a second look at a completely unreadable draft. Last but not least I want to thank my colleagues from the Bioinformatics department at the MDC, Berlin, from the Laboratory of Statistical Genetics at Rockefeller University, New York, and from the Laboratory of Molecular and Human Genetics at Baylor College, Houston, for collaboration, discussions, and companionship. And as the very last person on my list, thanks go to the unknown man behind the bar at *Java Girl* on 66th St. for the best unflavored double-size, double-shot macchiato in town.

*Berlin, February 29, 2004*

Michael Nothnagel

# Chapter 1

## Introduction

### 1.1 Genetic background of diseases

A familial aggregation of many diseases in humans has long been observed and their heritability has been suspected. However, statistical descriptions of heredity and inferences about its biological basis are only some 150 years old [38]. Diseases with a genetic component, like other phenotypic traits, are usually distinguished as being either Mendelian or complex. *Mendelian traits* are characterized by well-defined phenotypes, one or two genetic disease loci with high penetrance, a small phenocopy rate and usually small susceptibility allele frequencies. This clear genotype-phenotype relation results in a clear pattern of inheritance. Mendelian diseases are usually rare in the population.

*Complex traits* show a less clear relationship between genotype and phenotype due to two or more of the following characteristics: ill-defined phenotypes, incomplete penetrance, high phenocopy rate, genetic heterogeneity, oligogenic or polygenic inheritance, epistasis, mitochondrial inheritance, imprinting, and an often large contribution of environmental influences [85]. Unfortunately, most common diseases in humans resemble complex traits. Examples are hypertension, lipid metabolism disorders, some forms of Alzheimer's disease, and depression. Mendelian and complex traits differ only to the extent of these problems. An individual's genetically determined response to medication is the trait of interest in pharmacogenetics [136, 95].

### 1.1.1 Approaches to statistical gene mapping

How can loci which influence a trait or a disease be mapped on the genome? In animals, plants, and bacteria, breeding techniques are routinely used to create individuals with a defined genotype at one or more loci, for example knock-out mice. This approach cannot be used with humans for ethical reasons. Statistical methods have been developed to circumvent this problem.

#### Genetic markers

Since the trait-affecting gene is a priori unknown, all methods use *genetic markers*. These are variations of the genome that can be genotyped at reasonable cost and time. *Microsatellites* and *single nucleotide polymorphisms* (SNPs) are markers that are in general use today [130]. Microsatellites (or: short tandem repeats, STR) are a special form of frequent repeats of short DNA sequences (minisatellites). They are useful due to their widespread distribution throughout the genome and their large number of alleles. A measure for allelic diversity is the heterozygosity,  $H = 1 - \sum_{i=1}^{n_a} p_i^2$  where  $p_i$  and  $n_a$  denote the frequency of the  $i$ -th allele and the total number of alleles in the population [130].  $H$  exceeds 0.7 for a high portion of microsatellites, making them very informative for linkage analysis (see below). SNPs are usually bi-allelic and, thus, show a low heterozygosity, but have the advantage of low mutation rates and low genotyping costs for large-scale genotyping through automation [50]. By 2001, more than 2.1 million [40, 41] SNPs had already been identified throughout the genome. By November 2003, this figure had jumped to 5.7 million [29], forming a huge source of genetic markers. This makes them suitable to carry out genome-wide association studies [123] (see below). SNPs with a minor allele frequency above 0.1 are called *common*, whereas the other SNPs are called *rare*, although this threshold varies in the literature. This thesis will focus on SNPs as markers of choice.

#### Linkage disequilibrium and recombination

The basis for statistical gene mapping in humans is the phenomenon of *linkage disequilibrium* (LD) or *allelic association*. An individual's chromosomal

*genotype* consists of two *haplotypes*, one derived from the maternal gamete and the other from the paternal one [130]. In a narrower sense, a haplotype is the allelic combination of the chromosomal loci under consideration. LD is the non-random association between marker alleles on the same haplotype, i.e. it denotes their stochastic dependence.

How does LD come into existence? Mutations always occur in an already existing haplotype. The new allele will at first be found only in combination with the other loci's alleles in this haplotype. If crossing-over or gene conversion events occur between two loci during meiosis [16, 140], their alleles are newly combined (*recombination*). Over time, repeated recombinations will undermine the strength of LD. In general, the probability of a recombination increases with growing distance between two loci. It is measured by the *recombination fraction*  $\theta$ , i.e. the probability that a gamete will be recombinant with respect to two loci [106, 130]. There will be no recombination between loci for  $\theta = 0$  (*complete linkage*), whereas  $\theta = 0.5$  denotes *unlinked*, independent loci. Selection and population bottlenecks lead to a reduced diversity of haplotypes and can thereby strengthen LD. Genetic drift, i.e. the random changes in the haplotype frequencies due to the sampling of haplotypes inherited from one generation to the next, can also strengthen LD. Thus, the disequilibrium state of a genomic region depends on many, often unknown, factors, including population history and the size and structure of the genomic region. LD can arise from other sources, e.g. population admixture and sample substructure, due to the Wahlund effect [57]. Depending on their history, different populations exhibit different amounts of LD [120]. Populations that have gone through bottlenecks in their population history, e.g. by migration or founder effects, typically show longer ranging LD and less variation.

### **Statistical gene mapping methods**

Statistical gene mapping assumes that a marker showing evidence of affecting a phenotype is itself a variation in a causative gene or is in LD with such a variation. Methods can be differentiated by their type of analysis (linkage

analysis vs. association analysis [39]), by their type of assumptions (model-based vs. model-free), by the trait phenotype (qualitative vs. quantitative), or by their data basis (family-based vs. population-based [104]).

**Linkage Analysis.** Linkage analysis traces the co-segregation of a phenotype and markers with arbitrary alleles in families to detect recombination events.  $\theta$  should be smaller than 0.5 for markers close to a causative locus. Parametric methods ("lod-score analysis" [106]) explicitly model the disease's mode of inheritance,  $\theta$ , and other parameters. A likelihood-ratio (LR) test is used to test whether a model that employs the maximum-likelihood estimate [43] of  $\theta$ ,  $\hat{\theta} = \operatorname{argmax}_{\theta \in [0, \frac{1}{2})} L(\theta)$ , can explain the observations in the families significantly better than an unlinked-loci model:

$$\operatorname{lod}(\hat{\theta}) = \log_{10} \frac{L_{\hat{\theta}}}{L_{\theta=0.5}} . \quad (1.1)$$

A lod-score of 3 or above for single loci [99] or of 3.3 or above for whole-genome scans [105, 84, 130] corresponds to a significance level of  $\alpha = 0.05$ . Wrong specification of the mode of inheritance can lead to a dramatic loss of power [26]. Model-free methods make implicit or local assumptions instead of an explicit or global modelling of the mode of inheritance. They test for an excess of allele-sharing among relatives with similar phenotypes, using Pearson's  $\chi^2$  test, a mean test, and other tests [130, 155, 82, 77]. Haseman & Elston methods [58, 44, 20, 161] test for negative slopes in linear regression models on genetic markers. Variance component (VC) methods [17] also include environmental and other factors in the regression and compare them by LR tests. Twins are especially useful in this analysis to differentiate the contributions of these factors [93, 129].

**Association analysis.** For short distances between marker and gene locus, large sample sizes are necessary to detect any recombination. Linkage analysis therefore loses resolution. Association analysis ("LD mapping") indirectly tracks historical recombination events and can boost the resolution in the fine-mapping of disease genes: In general, LD decreases more rapidly

with increasing distance between loci. Marker alleles in strong LD with a susceptibility allele will show an association with the phenotype. For a case-control design as a straightforward model-free approach, marker alleles and affection status can be tested in a contingency table for significant departures from an odds-ratio of 1 for a particular allele or the relative risk using a  $\chi^2$  test or Fisher's exact test [130, 4]. Extensions of the VC methods [45] and of the lod-score method [49] can simultaneously test for linkage and association. To protect against spurious LD due to factors like population admixture or sample stratification, cases and controls can be matched for strata, e.g. sub-populations, or by design [19], as is done with the transmission-disequilibrium-test (TDT [126, 135]) and its successors [6, 134, 118, 76], e.g. the family-based association tests (FBAT [62]). The methods of Genomic Control [34, 36, 35, 10] and Structured Association [114, 115, 116, 112] have been developed to correct for confounding sources of LD. Model-based association methods model the cell probabilities of the contingency table under a supposed mode of disease inheritance. Models allowing for LD or not are compared by using a LR test [147].

### 1.1.2 Common diseases and the benefit of haplotypes

**The genetic architecture of common diseases.** Common diseases are often complex traits that suffer from reduced penetrances, increased heterogeneity, and other factors. Successful association studies identify increased susceptibility allele frequencies in affected individuals. Risch & Merikangas [123] noted that the power of statistical gene mapping methods depends on the allelic spectrum of the susceptibility mutations. Using deterministic models from classical population genetics, Reich & Lander [121] predicted that if a high overall frequency of susceptibility alleles exists, then these alleles are few but common. Thus, common diseases would result from combinations of common variants where each single variant is only of modest effect on the trait (*CD/CV hypothesis* [22]). Susceptibility loci are then subject to only mild selection, perhaps even with a selective advantage for heterozygous individuals. The assumption of little locus heterogeneity is critical for the

success of association studies [133]. Some researchers have suggested that association analysis is futile in complex traits, asserting that genes generally do not act additively, but in a multiplicative manner through complex networks of genes and/or environmental factors [61, 142]. Single gene effects are then usually too small to be detected by association analysis. However, important additive effects can be expected, in particular if the CD/CV hypothesis holds [27].

Recent studies found evidence for the CD/CV hypothesis in several diseases, for example the APOE gene in Alzheimer's disease [121, 89]. Pritchard & Cox [117] doubt this evidence for several reasons. They describe the evidence as "low-hanging fruits", caused by higher penetrances and the like, and describe the CD/CV hypothesis as a best-case scenario, rejecting the presented sample of diseases as too small and too biased to draw general conclusions. Furthermore, Pritchard [111] predicted extensive allelic heterogeneity at many loci, using stochastic multiplicative disease models. Population genetics theory predicts that common alleles are old, so only weak selection has acted against them or there may have been a heterozygous advantage. Thus, a number of recombination events have occurred and weakened LD around a gene locus in the past. Consequently, association studies lose statistical power and require increasing sample sizes that quickly become infeasible [48]. Higher marker densities could attenuate this problem.

Selection is compatible with high frequencies of causative alleles, if common diseases are rather young. Once, perhaps, advantageous allele combinations might have become disease risk factors under changing environmental conditions. An example is the increasing rate of type II diabetes and adipositas in industrialized countries due to the over-supply of food. If the susceptibility alleles are old, then a large proportion of the human population will possess them. Variants detected in one population could then be generalized to and tested for in almost every other population. The observation of common allele markers in a region critically depends on the mutation rate and also on the ascertainment scheme for the markers [111, 117].

It is sometimes possible to map the genes in common diseases, which are characterized by a somewhat ill-defined phenotype, by differentiating sub-

phenotypes, where each of them follows a classical Mendelian pattern. The low heterogeneity in each sub-phenotype then allows for the application of classical statistical gene mapping approaches.

**The benefit of haplotypes.** A new mutation arises on a particular haplotype that is only shortened by recombination events. The strength of the correlation between this haplotype and the mutation depends on the haplotype frequency during the event and the succeeding population history. Instead of using single SNPs to test for association, several approaches that utilize SNP haplotypes have been proposed [25].

First, haplotypes can simply serve as multi-allelic markers. They combine the advantages of SNPs with an increase in marker heterozygosity and, therefore, informativeness for gene mapping, when compared to single SNP markers [97, 78, 163, 79, 162, 160, 137]. Association for all the SNPs can then be tested simultaneously by using the haplotypes without much loss of power; it might even ease the multiple testing problem, since correlations between the different markers are implicitly modelled [46, 30, 69]. If a small number of haplotypes describes most of the genetic variation of the included SNPs, these haplotypes can be differentiated by *haplotype-tagging SNPs* (ht-SNPs [69]) where one SNP allele is unique to a particular haplotype. The use of htSNPs can also reduce genotyping costs and, thus, enables a study of larger populations for an equal amount of funding. Second, clusters of similar haplotypes could be enriched in the case group, or the haplotypes that contain a susceptibility mutation could be excessively shared among cases. Methods exploiting these phenomena have been suggested [83, 86, 13, 91]. Third, haplotypic structure can be modelled in a log-linear regression model [25]. Finally, some effects may seem to be haplotype-specific [68] and, thus, haplotypes need to be considered.

**The observation of haplotype blocks.** There is an ongoing debate about the pattern of LD in the human genome. Recent findings indicate a structure with regions of high LD and with limited numbers of haplotypes, interspersed by regions of low LD [46, 30, 31, 108, 32, 1, 120, 144, 18]. The

latter can be due to either high recombination rates (“hot spot”) or high rates of gene conversion in that region [47, 143, 122, 28, 67, 51, 119]. This structure of cold and hot spots of recombination has already been confirmed by physical evidence in the MHC and SHOX genes [66, 94]. Block patterns might also occur stochastically [154, 144]. Selection action on intra-genomic variation and on the distribution of genes in the genome could be a possible explanation for the variation in the recombination rate [107].

Blocks as regions of elevated LD can provide haplotypes to be used as genetic markers and delimit regions where htSNPs can reasonably be defined. They could also provide information on the spacing of SNPs in association studies, i.e. where SNPs should be considered and where not. To assist these objectives, the human haplotype map (HapMap) project [29, 69, 46, 30, 32] is now underway.

## **1.2 Haplotypes and linkage disequilibrium**

### **1.2.1 Estimation of haplotype frequencies**

One way to infer haplotypes and their frequencies is to physically observe them, as has been done by Perlegen Sciences, Inc. [108]. Direct sequencing is still time-consuming and comparatively expensive, but might become a standard technique in near future. Until then, haplotypes and their frequencies need to be estimated from the genotypic data.

There are several groups of estimation methods. In their method, Clark and others [24, 53, 153] aim to maximize the number of resolved haplotypes. Other methods use an Expectation-Maximization (EM) algorithm to find the haplotype set that maximizes the posterior probability of a given genotype set [42, 59, 90, 23]. These Maximum-Likelihood (ML) methods are often combined with Markov Chain Monte-Carlo (MCMC) techniques, e.g. Gibbs sampling, for more efficient sampling that result in a faster frequency estimation and thereby enable the processing of longer sequences [124, 138]. These methods are today’s established standard. Bayesian methods also incorporate prior knowledge on the haplotypes [100, 88, 153].

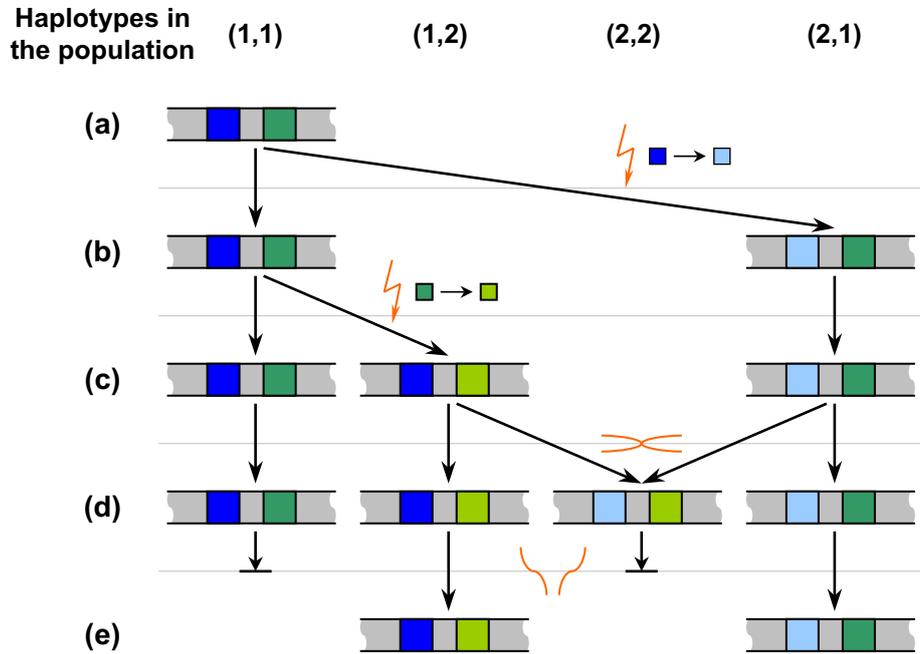


Figure 1.1: **Schematic example of LD between two SNPs.** (a) Nucleotide sequence with no variation at two considered loci (*blue* and *green*). (b) A mutation of the blue nucleotide creates a SNP in the population. LD cannot be assessed due to missing variation at the second nucleotide. (c) A second SNP emerges due to a mutation of the green locus. LD is, thus, initially complete ( $|D'| = 1$ ,  $r^2 < 1$ ). (d) The fourth haplotype (2,2) is created by a recombination between the haplotypes (1,2) and (2,1) ( $|D'| < 1$ ,  $r^2 < 1$ ). A mutation with the same result is extremely unlikely. Repeated recombinations will lower LD between the SNPs, population bottlenecks, admixture, and genetic drift will strengthen it. (e) Only two complementary haplotypes out of four are left in the population due to bottlenecks and genetic drift ( $|D'| = 1$ ,  $r^2 = 1$ ).

## 1.2.2 Pairwise measures for LD

A number of measures for the strength of LD have been proposed. They can be broadly differentiated by their ability to consider exactly two loci or more than two loci at a time. There is a vast amount of literature on the matter of LD measures [33, 64, 52, 87, 70, 8, 113, 60, 98], however, the commonly used measures are limited to LD between two loci.

To formally introduce pairwise LD measures, consider two bi-allelic loci,

possessing alleles 1 and 2 each. Let  $p_{ij}$  denote the probability of haplotype  $(i, j)$ , i.e. locus 1 exhibits the allele  $i$  and locus 2 the allele  $j$ . Let  $p_{i\cdot}$ ,  $p_{\cdot j}$  denote the marginal (or single) frequencies of alleles  $i$  and  $j$  at loci 1 and 2, respectively. These probabilities can be arranged in a contingency table:

	1	2	$\Sigma$
1	$p_{11}$	$p_{12}$	$p_{1\cdot}$
2	$p_{21}$	$p_{22}$	$p_{2\cdot}$
$\Sigma$	$p_{\cdot 1}$	$p_{\cdot 2}$	1

Under linkage equilibrium, the expected haplotype frequencies are the product of the marginal allele frequencies:  $p_{ij} = p_{i\cdot}p_{\cdot j}$ . The deviation from the expectation for this particular haplotype is measured by:

$$D_{ij} = p_{ij} - p_{i\cdot}p_{\cdot j} \quad (i, j = 1, 2). \quad (1.2)$$

For two bi-allelic loci, the absolute value of the deviation is the same for all four haplotypes:  $D_{ij} = (-1)^{i+j}D$  where  $D = p_{11} - p_{1\cdot}p_{\cdot 1}$ . Thus, the deviation for one haplotype describes the other three as well. Under LD, the allele probability distribution for a particular marker conditional on another marker allele differs from the marginal distribution. Knowledge of LD, therefore, complements the information provided by the markers' marginal allele frequencies. To allow for comparisons between different pairs of loci,  $D$  can be standardized to  $[-1, 1]$  or  $[0, 1]$  in several ways [33], among them the commonly used measures  $D'$  and  $r^2$  (also denoted as  $\Delta^2$ ):

$$D' = \begin{cases} \frac{D}{\min(p_{1\cdot}p_{\cdot 2}, p_{\cdot 1}p_{2\cdot})} : D > 0 \\ \frac{D}{\min(p_{1\cdot}p_{\cdot 1}, p_{2\cdot}p_{2\cdot})} : D < 0 \end{cases}, \quad r^2 = \frac{D^2}{p_{1\cdot}p_{2\cdot}p_{\cdot 1}p_{\cdot 2}}. \quad (1.3)$$

The state of  $|D'| = 1$  is called *complete LD*, whereas *perfect LD* is present if  $r^2 = 1$ . The latter situation can only occur as a consequence of population bottlenecks and genetic drift. Figure 1.1 illustrates the occurrence of LD

between two SNPs. Since LD depends on the age of the SNP-creating mutations, the population history, genetic drift, the recombination fraction, and other factors, it is highly variable even between close loci. Other pairwise measures have been proposed [33, 64, 87].

Although these measures are useful to assess pairwise LD, they cannot consider more than two loci and, thus, are blind to simultaneous associations between alleles of more than two loci. Furthermore, the measure  $D'$  is not suitable for differentiating different degrees of LD. It equals  $\pm 1$  if at least one haplotype is missing [8]. Missing haplotypes are more probable for rare SNP alleles and for multiple SNP sequences than for short sequences of common SNPs. Also, for small to moderate sample sizes, estimates of  $D'$  can exhibit a considerable upward bias [148, 146]. Even if  $D'$  is estimated to be below 1, it might be strongly biased. So  $D'$  is rather an indicator for missing haplotypes, perhaps due to absent recombination events, than a reliable measure of LD. See figure 1.2 for an illustration. The strength of LD between a trait locus and a marker, measured by  $r$ , is indirectly proportional to the power of finding an association [146]. This is not the case for  $D'$ .  $r^2$  is, thus, a measure of choice in disease association studies [8].

Morton et al. [98] describe LD by parametrically modelling an association probability (or recombination probability), using population genetics theory with regard to recombination fraction, effective population size, and other factors. It is, however, not clear how useful and robust the proposed measure  $\rho$  is if one or more of the assumptions is not fulfilled, namely the neutrality and population history assumptions. The SNP detection strategy can result in an ascertainment bias of the LD estimate [5]. This effect will not be discussed further here.

**Extent of LD in human populations.** For some applications, e.g. the spacing of SNP markers in association studies, it is important to know how far LD extends in a region. The rate of LD decay between two loci can, under some assumptions, be described by the recombination fraction  $\theta$  [130]:

$$D_g = (1 - \theta)^g (p_{ij}^0 - p_i \cdot p_j), \quad (1.4)$$

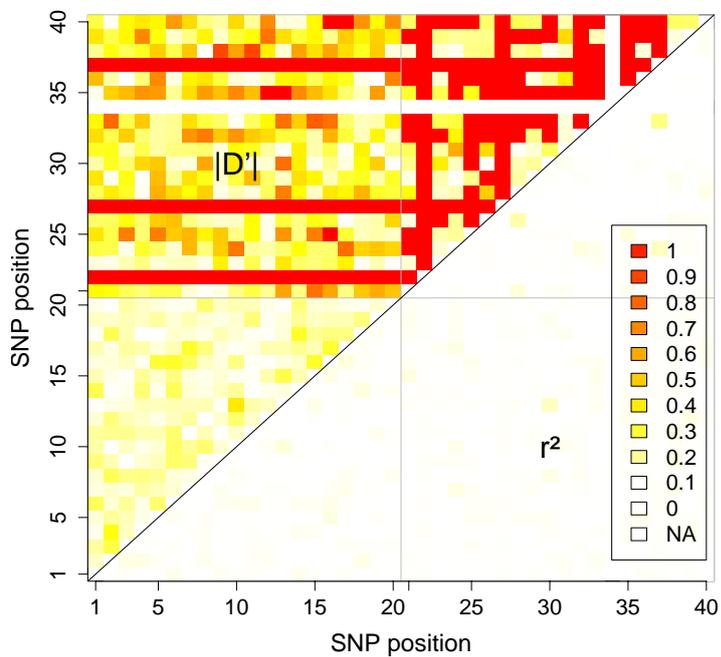


Figure 1.2:  $D'$  as an indicator for missing haplotypes. 200 haplotypes of 40 SNPs each were generated using SNaP (see section 2.5), assuming linkage equilibrium between all SNPs. SNPs 1–20 were common (minor allele frequencies  $p = 0.4 - 0.5$ ) whereas SNPs 21–40 were rare ( $p = 0.01 - 0.1$ ). Despite equilibrium,  $D'$  values (upper left triangle) often signal complete LD between rare SNPs due to missing haplotypes and often strong LD for pairs of common and rare SNPs.  $r^2$  (lower right triangle) is not affected by the allele frequencies, with values close to 0.

where  $D_g$  denotes the deviation (1.2) of the haplotype frequency from its equilibrium frequency after  $g$  generations, given an initial haplotype frequency

$p_{ij}^0$  at generation 0. For example,  $D_g$  roughly equals 0 after 10 generations for  $\theta = 0.5$ , whereas for  $\theta \leq 0.008$ , it is still considerably larger than 0 after 1000 generations [130].

Kruglyak [81] predicted useful LD would not be found in humans beyond 30 kb on average for the general population, using coalescent model simulations and assuming selective neutrality. Further distance predictions were 3 kb for outbred populations and 100 kb for isolated populations and rare SNPs. However, these estimates appear to be too pessimistic. Recent studies found LD to often extend over longer distances [8]. Pairwise  $r^2$  values above 0.4 in an outbred European population were found for distances greater than 100 kb [145].  $|D'|$  values equal to 1 were found for pairs 45 kb or further apart [137], and the average distance for markers with values close to 1 was estimated to be above 100 kb [120]. The extent of LD and the haplotype patterns can vary strongly between different populations [8, 128, 120, 37].

### 1.2.3 Multilocus LD measures

Pairwise LD measures miss at least some multilocus LD information. One way to compensate for this limitation is to consider all pairwise LD values between multiple loci. The problem still remains how to combine this information to multilocus LD descriptions. For example, Hedrick [60] proposed a weighted sum of pairwise  $|D'|$  values to describe multilocus LD but this quantity remains difficult to interpret [9]. Measures have been developed for a direct multilocus LD description. Sabatti & Risch [127] measured the agreement between pairs of markers by haplotype homozygosity and suggest its use especially for highly polymorphic markers. Bennett [14] and Weir [156] proposed measures for three or more loci. Unfortunately, these are haplotype-specific and describe exclusively higher order effects. Lower order terms and the marginal allele frequencies can also constrain these effects [149]. Models of the extent of historical recombination in a region, using population genetics theory and coalescent models, have also been proposed for measuring LD [113, 8, 75]. This approach assumes selective neutrality; the population history has to be modelled and estimated from genotypic data.

This parametric approach is computationally challenging and might be misleading if the region is subject to strong selection or if the population model is wrong.

### 1.3 Methods for the definition of blocks

The term *block* has been used to describe different objects with differing objectives [151]. Blocks were defined in order to define haplotypes in association studies, to reduce genotyping costs, or to delimit boundaries for candidate genes [21]. Depending on the objective, blocks were either defined as “islands” of high LD in a “sea” of low LD or as a segmentation of a genomic region into disjointed, adjacent blocks. The principal criterion for block detection is either a combination of pairwise  $D'$  values, some haplotype diversity criterion, or the coincidence of block boundaries with known recombination hot spots.

The existing algorithms are categorized in table 1.1. Gabriel et al. [46] and others [168, 109, 132, 150] look for genomic regions with no substantial amount of recombination. They use  $D'$  or its confidence intervals in order to find evidence of recombination. Blocks are defined as regions where only a limited proportion of SNP pairs shows strong evidence of recombination. Daly et al. [30] search for regions of low haplotypic diversity by comparing the observed haplotypic heterozygosity with the expectation under linkage equilibrium in sliding windows. They then estimate a haplotype transition probability  $\Theta$  (“historical recombination rate”) for fixed (“ancient”) haplotypes by using a Hidden-Markov model, where  $1 - \Theta = D'$ . Dawson et al. [32] use both  $D'$  and a reduced-haplotype-diversity criterion ( $\leq 5$  haplotypes provide  $\geq 75\%$  frequency coverage). Wang et al. [154] define blocks as regions, where all pairs of SNPs exhibit complete LD ( $|D'| = 1$ ) or where at least one of the four possible haplotypes has a frequency below 0.01. Surprisingly, no methods based on  $r^2$  have been published so far.

Other approaches focus on the partition of a genomic region. Patil et al. [108] use a rough haplotypic diversity criterion, requiring sample-size dependent common haplotypes to provide a certain amount of frequency coverage.

## Approaches and criteria in existing block definition algorithms

<i>Algorithm</i>	<i>Approach</i>		<i>Criterion</i>		<i>htSNPs</i>
	<i>Island</i>	<i>Partition</i>	<i>Pairwise LD</i>	<i>Haplotypes</i>	
Daly [30]	x		x	x	
Gabriel [46]	x		x		
Zhu [168]	x		x		
Phillips [109]	x		x		
Twells [150]	x		x		
Shifman [132]	x		x		
Dawson [32]	x		x	x	
Wang [154]	x		x		
Patil [108]		x		x	x
Zhang [164][165]		x		x	x
Anderson [7]		x	x	x	
Mannila [92]	(x)	x	x	x	

Table 1.1: Tabulation of the objectives and used criteria of existing block definition algorithms. Methods usually search for regions of high LD (island approach) or for a segmentation of the sequence into blocks (partition approach). Island methods are predominantly  $D'$ -based, whereas partition methods apply haplotype diversity criteria and often aim for the definition of htSNPs.

Blocks would maximize the ratio of the number of SNPs in a window by the number of common haplotypes in that window. Zhang et al. [164, 165] and others [92] formalize this approach by a dynamic programming (DP) algorithm. Anderson & Novembre [7] combine measures of haplotypic block diversity and LD decay between blocks to find an optimal partition, employing the minimal description length principle. Both approaches require a parameterization of the block structure. Also, multilocus LD interaction is only described by the occurrence of a few haplotypes. Partition algorithms have the drawback that a large number of blocks merely contain a single SNP. More severe is the algorithms' prerequisite of frequency estimates for haplotypes of potentially very long size. Since haplotype frequencies need to be estimated and the sample size limits the length of haplotypes that can reliably be estimated, this also limits the maximum block size.

## Block lengths found in previous studies

<i>Authors</i>	<i>Chromosome</i>	<i>Block lengths [kb]</i>
Gabriel [46]	various	1-173
Zhu [168]	1, 3, 17	1-45
Daly [30]	5	3-92
Jeffreys [66]	6	-100
Twells [150]	11	37-110
Haiman [54]	15	13-50
Phillips [109]	19	1-153 (-338), 38% physical coverage
Patil [108]	21	-115 ( $\varnothing$ 7.8)
Dawson [32]	22	-804, 41.8% SNP coverage
May [94]	X/Y	-3

Table 1.2: Physical block lengths and, if provided, the block coverage of the chromosome with regard to included SNPs and included physical distances as found in previous studies.

Johnson et al. [69] introduced the concept of htSNPs as a means of reducing genotyping efforts. More algorithms exclusively following this approach have been proposed [72, 139]. Meng et al. [96] sought to select genetic markers for association analysis, using  $D'$ . Zhang & Jin [166] developed the HAPLO-BLOCKFINDER program, which implements several algorithms.

**Block lengths found in previous studies.** A number of studies have already been carried out which look for haplotype blocks. Table 1.2 lists the block lengths that were found in some of these studies and also the chromosomal coverage provided by the blocks, when whole chromosomes were investigated. Physical block lengths vary greatly, from 1 kb through 804 kb.

## 1.4 Objective of this thesis

The detection of haplotype blocks in the human genome is a recent discovery, and the methods for their definition are still under development and open debate. Applications of these blocks include at least three objectives. First, blocks with a number of common haplotypes can be used as multi-

allelic markers with higher heterozygosity in disease association studies to improve statistical power. Second, the definition of htSNPs in blocks can reduce genotyping efforts in medical studies, while nearly the same amount of genetic variation is described. Blocks also suggest where SNPs should be spaced denser or sparser in genomewide studies. Third, blocks are also an interesting feature of human genomic structure in itself. They contradict the long-held assumption that recombination events occur with uniform probability along the genome. Some blocks of elevated LD might coincide with non-coding regions that are functionally relevant and preserved by selection. Those regions could be detected by between-population or between-species comparisons.

The preceding sections have shown how existing block definition algorithms are based on two methods: Either are pairwise  $D'$  values with differing lower limits used to detect regions of little or no recombination, or blocks are defined by employing some haplotypic diversity criterion, where a small number of common haplotypes provide high chromosomal frequency coverage. Neither of these methods describes multilocus LD directly.  $D'$  basically describes absent haplotypes that are often due to missing recombination. But even if recombination events have occurred in the past, they do not reduce association power at once but gradually. There will often be useful LD in a region for disease mapping due to other sources, e.g. population bottlenecks, that would be wasted by using  $D'$ . So far, no methods have utilized  $r^2$  for the block definition despite its direct relation to association power. Furthermore, pairwise measures might miss multilocus LD information, and it is not entirely clear how they could optimally be combined for block detection algorithms. Chromosomal coverage methods look for regions of low diversity, which will often, but not necessarily, detect regions of high LD.

The objectives of this thesis are, therefore, to propose, first, a new multilocus LD measure and, second, a new block definition algorithm that is based on this measure, and to thoroughly investigate the features of both. To this end, the thesis is structured as follows: Chapter 2 introduces the new multilocus LD measure, which is not based on population model assump-

tions and does not share the indicator-like behavior of  $D'$ . The measure's analytical features are investigated, and it is compared to existing measures. The measure is then utilized in a haplotype block definition algorithm. The succeeding section outlines the employed data simulation program. Chapter 3 investigates the potential robustness of the measure in the concept of common haplotypes, using simulated data sets. The ability of the measure to reasonably describe LD and its applicability are demonstrated on both, simulated data sets and in a real-world data set with an established block structure. Chapter 4 applies the proposed block detection algorithm to a data set of the whole human chromosome 12. The resulting blocks are characterized, and the influence of the algorithm's control parameters on the blocks is investigated. This application also allows for some substantial conclusions about the nature of haplotype blocks in the human genome. The proposed algorithm is compared with others that are based on pairwise LD measures. The results and their implications are discussed in chapter 5. Chapter 6 summarizes this thesis.

# Chapter 2

## Measure & methods

### 2.1 The concept of entropy

In physics and information theory, *entropy*,  $S$ , describes the non-order or the degree of (non-)structure of a system [131, 12]. It is defined as

$$S = - \sum_i p_i \log p_i, \quad (2.1)$$

where the  $p_i$ 's denote the probabilities of the different states that the system can assume and  $\log$  denotes the *logarithmus naturalis* (other bases for the logarithm could be used). The sum includes all (possible) states of the system. Missing states ( $p_i = 0$ ) do not contribute to  $S$ , since  $0 \log 0 = 0$  by definition.  $S$  achieves its maximum if all states are equally probable. In this case, the system exhibits the lowest degree of structure, and only the minimum amount of information about the system's actual state is available a priori. An observation of the system will then provide a maximum gain in information.  $S$  equals 0 if there is only one state. Then the system's state is exactly known, and no further information can be gained about it. If two systems are considered, their joint entropy reaches its maximum if the systems are independent of each other [131]. The joint entropy then equals the sum of the single-system entropies.

The concept of entropy has already been applied to haplotypes and to

SNPs. Jawaheer et al. [65] looked at deviations from the uniform distribution of haplotypes that consisted of three markers. Judson and others [71, 2, 3] proposed the choice of SNP subsets that explain most of the haplotypic variation by maximizing their entropy. Hampe et al. [55] aim to select most informative SNPs for association analysis.

## 2.2 The normalized entropy difference $\varepsilon$

By applying the concept of entropy to genetic data, a sequence of two or more loci is considered a system. The possible haplotypes represent the states of this system. Consider  $m$  bi-allelic loci, e.g. SNPs. Their sequence can assume  $2^m$  haplotypes  $(a_1^i, \dots, a_m^i) \in \{1, 2\}^m$  of which  $n$  are assumed to be present. The entropy is used to measure the non-order of the observed loci sequence:

$$S_B = - \sum_{i=1}^n p_i \log p_i, \quad (2.2)$$

where  $p_i = p_{a_1^i, \dots, a_m^i}$  denotes the frequency of haplotype  $i$ . Under the hypothesis of linkage equilibrium,  $p_i$  can be expressed as the product of the marginal allele frequencies at the loci:

$$q_i = q_{a_1^i, \dots, a_m^i} = \prod_{k=1}^m p^{(k)} \mathbb{1}_{\{a_k^i=1\}} (1 - p^{(k)}) \mathbb{1}_{\{a_k^i=2\}}, \quad (2.3)$$

where  $q_i = q_{a_1^i, \dots, a_m^i}$  denotes the frequency of the  $i$ -th haplotype,  $a_k^i$  denotes the allele of the  $k$ -th SNP position ( $k \in \{1, \dots, m\}$ ) at haplotype  $i$ ,  $p^{(k)}$  denotes the frequency of allele 1 at the  $k$ -th SNP, and  $\mathbb{1}_{\{x\}}$  equals 1 if  $x$  is true and zero otherwise. The entropy that would be expected under linkage equilibrium conditional on the SNP marginal frequencies is then:

$$S_E = - \sum_{i=1}^{2^m} q_i \log q_i. \quad (2.4)$$

The term (2.4) is also the maximum entropy the sequence can assume if the marginal allele frequencies are held constant. This is easily proven by

considering each SNP as an independent system and the SNP sequence as the union of these systems.

Deviations from the equilibrium state represent a gain in structure or of a priori information about the loci sequence. This information complements the information provided by the marginal frequencies. Deviations will result in a decreased entropy compared to the equilibrium case. The difference between expected and observed entropy,

$$\Delta S = S_E - S_B, \quad (2.5)$$

is thereby a measure for the sequence's deviation from its linkage equilibrium state. The term  $\Delta S$  coincides with a term that was developed by Zhao et al. [167] using likelihood theory. In analogy to the normalization of  $D$  (1.2),  $\Delta S$  is scaled to  $[0, 1)$  by  $S_E$  to allow for comparisons between different sets of loci and is denoted by  $\varepsilon$ :

$$\varepsilon = \frac{\Delta S}{S_E} = 1 - \frac{S_B}{S_E}. \quad (2.6)$$

This new measure for LD is called the *Normalized Entropy Difference* [103]. In the following, the number of incorporated loci,  $m$ , will be called the *window size*, while  $\varepsilon_m$  denotes the corresponding value of  $\varepsilon$ .

## 2.3 Analytical features of $\varepsilon$

### Assessment of LD significance

Consider the sample size  $N = \sum_{i=1}^{2^m} n_i$ , where  $n_i = n_{a_1^i, \dots, a_m^i}$  represents the number of occurrences of the  $i$ -th haplotype in the sample ( $p_i = \frac{n_i}{N}$ ). We further define the haplotype frequency deviations from the expectation under equilibrium,  $\delta_i = \delta_{a_1^i, \dots, a_m^i} = p_i - q_i$ , with the  $q_i$ 's defined as in (2.3). Note that  $\sum_{i=1}^{2^m} \delta_i = 0$ . The likelihood of a sequence is designated by  $L_B$  for the observation and by  $L_E$  under the assumption of linkage equilibrium.  $L_E$  is completely determined by the  $m$  marginal allele frequencies;  $L_B$  is determined by  $2^m - 1$  haplotype frequencies.

A first result is the following equation:

$$\Delta S = \frac{1}{N} \log \frac{L_B}{L_E}. \quad (2.7)$$

To prove equation (2.7), it is sufficient to show that  $S_B = -\frac{1}{N} \log L_B$  and  $S_E = -\frac{1}{N} \log L_E$ :

$$\begin{aligned} \frac{1}{N} \log L_B &= \frac{1}{N} \sum_{i=1}^{2^m} n_i \log p_i \\ &= \sum_{i=1}^{2^m} (q_i + \delta_i) \log (q_i + \delta_i) = -S_B \end{aligned} \quad (2.8)$$

$$\begin{aligned} \frac{1}{N} \log L_E &= \sum_{i=1}^{2^m} (q_i + \delta_i) \log q_i \\ &= \underbrace{\sum_{i=1}^{2^m} q_i \log q_i}_{= -S_E} + \underbrace{\sum_{i=1}^{2^m} \delta_i \log q_i}_{= (*)} \end{aligned} \quad (2.9)$$

It remains to show that  $(*)$  equals zero:

$$\begin{aligned} (*) &= \sum_{i=1}^{2^m} \delta_{a_1^i, \dots, a_m^i} \log q_{a_1^i, \dots, a_m^i} \\ &= \sum_{i=1}^{2^m} \sum_{k=1}^m \delta_{a_1^i, \dots, a_m^i} \mathbb{1}_{\{a_k^i=1\}} \log p(k) + \\ &\quad \sum_{i=1}^{2^m} \sum_{k=1}^m \delta_{a_1^i, \dots, a_m^i} \mathbb{1}_{\{a_k^i=2\}} \log(1 - p(k)) \\ &= \sum_{k=1}^m \log p(k) \underbrace{\left( \sum_{i=1}^{2^m} \mathbb{1}_{\{a_k^i=1\}} \delta_{a_1^i, \dots, a_m^i} \right)}_{= (**)} + \end{aligned}$$

$$\sum_{k=1}^m \log(1 - p(k)) \underbrace{\left( \sum_{i=1}^{2^m} \mathbb{1}_{\{a_k^i=2\}} \delta_{a_1^i, \dots, a_m^i} \right)}_{= (***)} \quad (2.10)$$

Thus, only the  $\delta_i$ 's of those haplotypes that possess allele 1 at the  $k$ -th locus contribute to the sum subsequently multiplied by  $\log p(k)$ . To show that this sum  $(**)$  equals 0 for all  $k$ , we substitute  $p_i - q_i$  for  $\delta_i$  and then find that both sums assume the marginal frequency  $p(k)$ :

$$\begin{aligned} (***) &= \sum_{i=1}^{2^m} \mathbb{1}_{\{a_k^i=1\}} p_{a_1^i, \dots, a_m^i} - p(k) \sum_{i=1}^{2^m} \prod_{l=1, l \neq k}^m p(l)^{\mathbb{1}_{\{a_l^i=1\}}} (1 - p(l))^{\mathbb{1}_{\{a_l^i=2\}}} \\ &= p(k) - p(k) \prod_{l=1, l \neq k}^m (p(l) + 1 - p(l)) = 0 \end{aligned} \quad (2.11)$$

$(***)$  is resolved in the same way. Thus,  $(*) = 0$  and equation (2.7) holds.  $\square$

Since the log-likelihood ratio is approximately  $\chi^2$  distributed,  $2 \log \frac{L_B}{L_E} \sim \chi^2$  [157, ch. 13], we are now able to state that

$$2N \Delta S \sim \chi_{2^m - (m+1)}^2 \quad (2.12)$$

approximately holds.  $2N \Delta S$  can, therefore, be used to test whether the haplotype frequencies significantly differ from the expectation under linkage equilibrium. Since this is an asymptotic test, sample sizes must not be too small.

### Incorporation of several loci

By definition, the measure  $\varepsilon$  allows for the incorporation of an unlimited number of loci. However, this number is limited by the sample size in practice. Some haplotypes of rare frequency that are present in the population may be missing in the sample due to the limited sample size. This can result in an upwardly biased LD estimate. Also, the need to estimate haplotype frequencies limits the possible number of SNPs that can be incorporated.

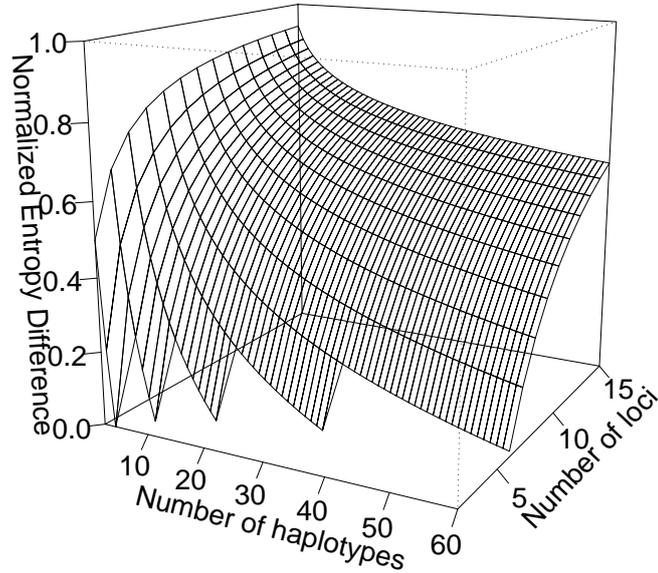


Figure 2.1: **Influence of the numbers of loci and haplotypes on  $\varepsilon$  for the case of equal haplotype frequencies.** Increasing numbers of haplotypes for a constant number of loci result in decreased values of  $\varepsilon$ , as do decreasing numbers of loci, when the number of haplotypes is held constant. The marginal allele frequencies are not held constant in this illustration. The measure  $\varepsilon$  changes greatly for small numbers of loci and for small numbers of haplotypes. A large part of real-world data is expected to fall into this area.

Usual study settings will, thus, allow for the calculation of  $\varepsilon$  for sequences of 8 to 12 SNP loci at most, corresponding to 256 to 4096 possible haplotypes. Section 2.4 proposes a method to describe LD for longer loci sequences.

SNPs that were found to be bi-allelic in one sample are often found to be mono-allelic in another sample. A mono-allelic SNP has the effect of lowering the effective number,  $m$ , of considered SNPs in the calculation of  $\varepsilon$  by 1.

## Haplotypes and their frequency pattern

The measure  $\varepsilon$  is sensitive to both the number of haplotypes that are observed and their frequencies.  $\varepsilon$  equals 0 if and only if a sequence is in its equilibrium state, i.e. the haplotype frequencies are completely defined by the marginal allele frequencies.  $\varepsilon$  increases with decreasing numbers of haplotypes present at the sequence and also increases with deviations from their equilibrium frequencies.  $\varepsilon$  distinguishes between various degrees of LD beyond the absence of more than one haplotype. It is easily proven that for exactly two haplotypes,  $\varepsilon$  assumes the window-depending value of  $\frac{m-1}{m}$ , which is always smaller than 1. Figure 2.1 illustrates the dependence of  $\varepsilon$  on the numbers of loci and occurring haplotypes for the case of equal haplotype frequencies.

## Approximate equality of $\Delta S$ and $r^2$ for two SNPs

At first glance, the rationale behind the entropy-based LD measure  $\varepsilon$  seems rather remote from usual approaches. However, the values of  $\Delta S$  and  $\frac{1}{2}r^2$  are close to each other for two bi-allelic loci, if the marginal allele frequencies are not too close to 0 or 1:

$$\Delta S \approx \frac{1}{2}r^2. \quad (2.13)$$

Figure 2.2 illustrates the very similar behavior of  $r^2$  and  $\Delta S$  and also  $\varepsilon$  for the case of two SNPs. To prove equation (2.13), consider two bi-allelic loci with alleles 1 and 2 each. We use a similar notation as before: let the haplotype frequencies be denoted by  $p_{ij}$ , the marginal allele frequencies by  $p_i$  and  $p_j$ ;  $D = p_{11} - p_1 \cdot p_{.1}$ ,  $D_{ij} = (-1)^{i+j}D$ .  $\Delta S$  is expressed in terms of the log-likelihood ratio using equation (2.7). The approximation (2.13) is done by power series expansion to second order, using  $\log(1+x) = \sum_{i=1}^{\infty} (-1)^{i+1} \frac{x^i}{i}$ . For relations between  $r^2$  and the  $\chi^2$  distribution see [156, p. 137].

$$\Delta S = \frac{1}{N} \log \frac{L_B}{L_E}$$

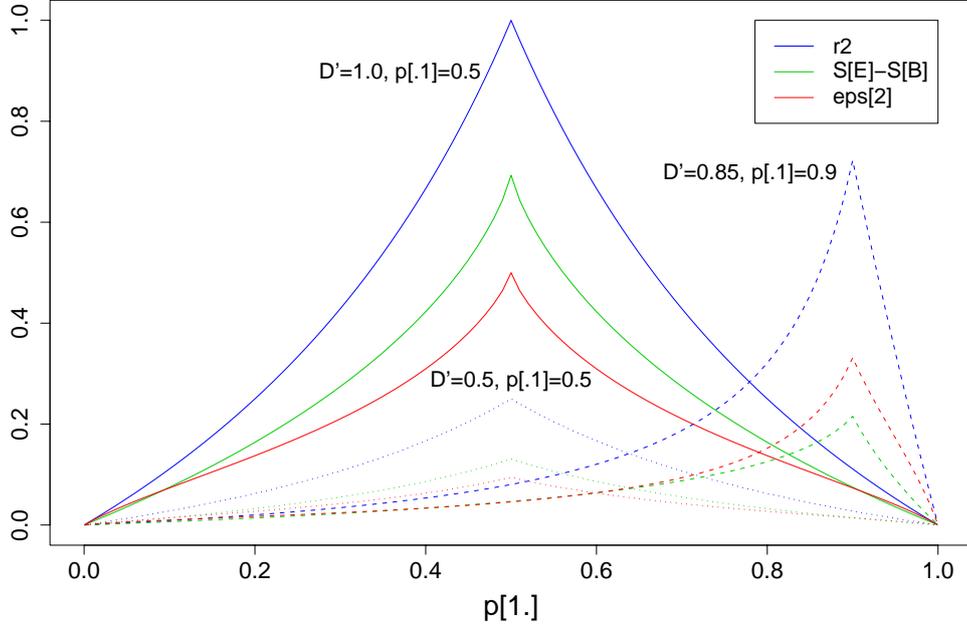


Figure 2.2: **Comparison of  $r^2$ ,  $\Delta S$ , and  $\varepsilon$  for two SNPs following a work by Hedrick [60].**  $p_{1.}$ ,  $p_{.1}$  denote the marginal frequencies of allele 1 at the SNP loci. In three considered pairs,  $D'$  and  $p_{.1}$  are fixed to some value, whereas  $p_{1.}$  varies from 0 to 1. Haplotype frequencies are calculated from  $p_{1.}$ ,  $p_{.1}$ , and  $D'$ , thus, allowing for the calculation of other LD measures. The value of  $D'$  controls the height of the curves, whereas the value of  $p_{.1}$  influences the location of the peak. The curves of  $r^2$  ( $r2$ ),  $\Delta S$  ( $S[E]-S[B]$ ), and  $\varepsilon_2$  ( $eps[2]$ ) show a remarkably similar shape and peak at the same location. For extreme marginal allele frequencies,  $\varepsilon_2$  and  $\Delta S$  diverge from  $r^2$ .

$$\begin{aligned}
&= \sum_{i,j=1}^2 (p_i p_j + D_{ij}) \log \left( 1 + \frac{D_{ij}}{p_i p_j} \right) \\
&\approx \sum_{i,j=1}^2 \left[ \left( (-1)^{i+j} D - \frac{1}{2} \frac{D^2}{p_i p_j} \right) + \frac{D^2}{p_i p_j} \right] \quad (2.14) \\
&= \frac{1}{2} \frac{D^2}{p_{1.} p_{.1} p_{2.} p_{.2}} = \frac{r^2}{2}
\end{aligned}$$

$D$  is confined to  $[\max(-p_{1.p.1}, -p_{2.p.2}), \min(p_{1.p.2}, p_{2.p.1})] \subseteq [-0.25, 0.25]$  [57, p. 51]. If  $|D_{ij}| < p_{i.p.j}$ , approximation (2.14) holds. For very high or very low marginal frequencies, the approximation becomes bad.  $\square$

The value of  $\Delta S$  is indirectly related to the increase of the required sample size to achieve a certain power in disease association mapping [113] as does  $r^2$  for the two-loci case [146, 8].

## 2.4 An $\varepsilon$ -based block definition algorithm

The measure  $\varepsilon$  directly describes LD and is sensitive to both the number and the frequency pattern of the present haplotypes. In analogy to  $D'$ -based methods, an  $\varepsilon$ -based block definition algorithm could search for contiguous regions of high LD, separated by regions of decreased LD. Since the sample size limits the window length that  $\varepsilon$  can be used with, a number of overlapping windows of smaller size could be used instead. This does not exactly describe multilocus LD for long sequences, but will approximate it.

### An $\varepsilon$ -based block definition algorithm.

1. Choose a *window size*  $m \in \{2, 3, \dots\}$ , i.e. the number of SNPs to be used in the calculation of  $\varepsilon$ .
2. Choose a *threshold*  $t \in [0, 1)$  for  $\varepsilon_m$ .
3. Use a sliding window of size  $m$  along the SNP sequence; define *haplotype blocks* or *regions of elevated LD* as contiguous windows for which  $\varepsilon_m$  does not drop below the chosen threshold.

Window size and threshold represent the algorithm's control parameters. The algorithm will deliver blocks of high LD and follows the island approach.

## 2.5 A data simulation algorithm

A software program, SNaP [102], was developed in C [73] to generate data sets that follow the two observations from section 1.1.2, namely SNP sequences

with a discrete pattern of blocks and a limited haplotypic diversity within these blocks. The program assumes a simple model to this end:

1. The sequence is composed of a series of independent blocks of one or more SNPs each.
2. For each block, a set of one or more haplotypes and corresponding frequencies is specified.
3. Block haplotypes are sampled independently from the other blocks. The block haplotypes are concatenated to form the sequence haplotype.

Thus, SNPs from different blocks are always in linkage equilibrium, whereas SNPs from the same block might exhibit medium or strong LD, depending on the specified haplotypes. This model is only suitable to simulate clear block patterns; it cannot explicitly model LD within a block and LD decay at its borders which remains an open problem and has been resolved only for very small sequences [158]. A suitable choice of block haplotypes can partially compensate for this drawback.

To simulate individuals, the program first checks if a haplotype set was specified for each block. If not, it is randomly generated according to a block-specific number. In this case, each possible haplotype has the same probability to be included in the set. The algorithm does, therefore, not follow an evolutionary model for the haplotypes. Subsequently, block haplotypes are sampled independently for each block from the sets and then concatenated to form the sequence haplotype. Two sampled sequence haplotypes form an individual's genotype (dual-haplotype). The program allows for the generation of an unlimited number of individuals and nuclear families with a random or fixed number of children in the families. Optionally, a quantitative trait phenotype or a disease affection status can randomly be assigned to the individual conditional on his or her genotype at one or more loci. The program allows for two sampling schemes, added genotyping errors, removal of causative SNPs, and other features. Except for case-control sampling, each locus is in Hardy-Weinberg equilibrium due to the independent sampling of haplotypes implemented.

# Chapter 3

## Applicability of $\varepsilon$

Equipped with a multilocus LD measure and a block definition algorithm from chapter 2, the following objectives are pursued in this chapter:

1. Investigation of the influence of the common haplotypes concept on the calculation of  $\varepsilon$ .
2. Demonstration of the ability of the measure  $\varepsilon$  to describe LD and detect block structures.

To this end, simulated as well as real-world data sets were analyzed.

### 3.1 Common haplotypes, coverage, and $\varepsilon$

In real-world applications, haplotype frequency estimation is always prone to error. This is particularly true for rare haplotypes. Each haplotype contributes with the term  $-p \log p$  to the entropy term (2.1), where  $p$  denotes the haplotype frequency. Unfortunately, small deviations from the true haplotype frequency can cause large deviations in (2.1) for small  $p$ ; see figure 3.1 for an illustration. The effect of those small deviations for rare haplotypes on  $\varepsilon$  is not obvious and difficult to tackle analytically. It would be advantageous if  $\varepsilon$  would not overly depend on small frequencies. Then only haplotypes with frequencies greater than a certain threshold (*common haplotypes*) needed to be considered for the reliable calculation of  $\varepsilon$  under the

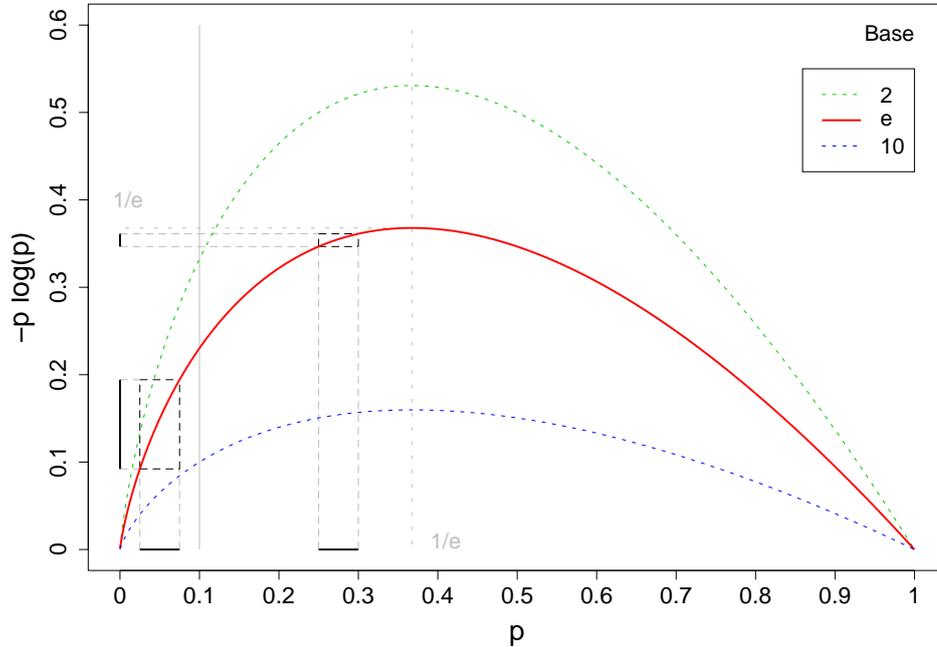


Figure 3.1: **Graph of  $-p \log p$  for three different bases.** Errors in the haplotype frequency estimation only mildly affect the entropy for moderate frequencies but strongly for small ones, e.g. frequencies below a threshold of 0.1 (solid grey line). The graph illustrates the changes in  $-p \log p$  for a frequency error of 0.05.

condition that the common haplotypes provide sufficient *frequency coverage*, i.e. the sum of their frequencies is greater than some threshold. This would result in a more robust estimation of  $\varepsilon$  in real-world applications.

### 3.1.1 Simulation study design

A simulation study was carried out to answer the following questions:

1. Are rare haplotypes negligible for the calculation of  $\varepsilon$ ?
2. How should  $\varepsilon$  be estimated if only common haplotypes are used?
3. Which frequency coverage and definition threshold for the common haplotypes does a reliable estimation of  $\varepsilon$  require?

The following parameters with a potential influence on the calculation of  $\varepsilon$  were considered:

Number of common haplotypes (cHT):	2, 3, 4, 5
Threshold for cHT definition:	0.2, 0.15, 0.1, 0.05, 0.01
cHT's frequency coverage:	60%, 70%, 80%, 90%
Total number of haplotypes:	20 <i>or</i> three times the number of cHT
Frequency pattern of cHT:	<i>Inequality Pattern</i> All but one cHT's frequencies equal the threshold frequency <i>or</i> <i>Equality Pattern</i> All cHT have equal frequencies.
Number of SNPs in sequence:	3, 6, 8

Rare haplotypes were assumed to have equal frequencies. Every possible combination of parameter values was considered (factorial design). SNPs were required to be strictly bi-allelic. For each point in this parameter grid, 1000 data sets with 1000 haplotypes each were randomly and independently generated using SNaP (see section 2.5). Each possible sequence haplotype had the same probability of being included in the limited set from which the data set haplotypes were sampled. For each data set, the measure  $\varepsilon$  was estimated in three different ways:

- $\varepsilon_{all}$ : All haplotypes were used.
- $\varepsilon_{lve}$ : Only the common haplotypes, with their frequencies unchanged, were used.
- $\varepsilon_{cmn}$ : Only the common haplotypes, with their frequencies re-scaled so that their sum equalled 1, were used.

Perl [152] and R<sup>1</sup> scripts were developed for processing, data handling, and the statistical analysis.

---

<sup>1</sup>For more details on this OpenSource S-PLUS clone see <http://www.r-project.org/>.

### Bias of $\hat{\varepsilon}$ when only common haplotypes are used

<i>Common Haplotypes</i>		<i>Frequency Pattern &amp; Coverage</i>			
<i>Number</i>	<i>Threshold</i>	<i>Inequality Pattern</i>		<i>Equality Pattern</i>	
		70%	90%	70%	90%
2	0.05	–	0.03±0.18	–	0.03±0.12
		–	-0.02◦ 0.14	–	0.02◦ 0.09
		–	-0.61◁ 0.24	–	-0.47◁ 0.12
	0.10	0.05±0.21	0.03±0.16	0.06±0.18	0.02±0.13
		-0.02◦ 0.18	0.00◦ 0.12	0.02◦ 0.17	0.01◦ 0.09
		-0.60◁ 0.29	-0.56◁ 0.19	-0.56◁ 0.22	-0.47◁ 0.12
	0.20	0.07±0.18	0.03±0.14	0.07±0.18	0.03±0.12
		0.01◦ 0.18	0.00◦ 0.11	0.02◦ 0.17	0.02◦ 0.09
		-0.57◁ 0.26	-0.50◁ 0.14	-0.55◁ 0.22	-0.47◁ 0.12
3	0.05	–	0.08±0.08	–	0.07±0.04
		–	0.06◦ 0.13	–	0.06◦ 0.09
		–	-0.44◁ 0.22	–	-0.21◁ 0.11
	0.10	0.14±0.09	0.08±0.07	0.13±0.09	0.07±0.04
		0.10◦ 0.19	0.06◦ 0.11	0.10◦ 0.18	0.06◦ 0.09
		-0.38◁ 0.26	-0.36◁ 0.15	-0.33◁ 0.22	-0.23◁ 0.11
	0.20	0.13±0.08	0.07±0.05	0.13±0.08	0.07±0.04
		0.11◦ 0.18	0.06◦ 0.10	0.11◦ 0.18	0.06◦ 0.09
		-0.29◁ 0.24	-0.28◁ 0.12	-0.29◁ 0.23	-0.20◁ 0.11
4	0.05	0.16±0.07	0.10±0.05	0.16±0.05	0.08±0.02
		0.12◦ 0.20	0.08◦ 0.13	0.14◦ 0.19	0.08◦ 0.10
		-0.22◁ 0.28	-0.28◁ 0.17	-0.17◁ 0.22	-0.08◁ 0.11
	0.10	0.16±0.06	0.09±0.03	0.16±0.05	0.08±0.02
		0.14◦ 0.20	0.08◦ 0.11	0.14◦ 0.19	0.07◦ 0.10
		-0.30◁ 0.25	-0.18◁ 0.14	-0.29◁ 0.22	-0.10◁ 0.11
	0.20	–	0.08±0.02	–	0.08±0.03
		–	0.08◦ 0.10	–	0.08◦ 0.10
		–	-0.08◁ 0.11	–	-0.23◁ 0.11
5	0.05	0.16±0.05	0.10±0.03	0.16±0.04	0.08±0.02
		0.14◦ 0.20	0.09◦ 0.12	0.15◦ 0.19	0.08◦ 0.10
		-0.25◁ 0.28	-0.05◁ 0.15	-0.16◁ 0.22	-0.03◁ 0.11
	0.10	0.16±0.04	0.09±0.02	0.16±0.04	0.09±0.01
		0.15◦ 0.19	0.08◦ 0.10	0.15◦ 0.19	0.08◦ 0.09
		-0.12◁ 0.22	-0.12◁ 0.12	-0.01◁ 0.22	-0.02◁ 0.10

Table 3.1: Mean±SD, 1.◦3. quartile, and min.◁max. of  $\varepsilon_{cmn} - \varepsilon_{all}$  for 6 SNPs, three times as many haplotypes in total as common haplotypes, and both haplotype frequency patterns. Results for 60% and 80% coverage and for thresholds 0.01 and 0.15 fit into the trend and were therefore omitted from the presentation.

### Bias of $\hat{\varepsilon}$ when only common haplotypes are used

<i>Common Haplotypes</i>		<i>Frequency Pattern &amp; Coverage</i>			
<i>Number</i>	<i>Threshold</i>	<i>Inequality Pattern</i>		<i>Equality Pattern</i>	
		70%	90%	70%	90%
2	0.05	0.19±0.20	0.11±0.18	0.19±0.17	0.08±0.12
		0.12◦ 0.32	0.06◦ 0.22	0.15◦ 0.29	0.07◦ 0.14
		-0.44◁ 0.42	-0.45◁ 0.28	-0.37◁ 0.33	-0.34◁ 0.16
	0.10	0.20±0.19	0.10±0.15	0.19±0.16	0.08±0.11
		0.13◦ 0.32	0.08◦ 0.19	0.15◦ 0.29	0.07◦ 0.14
		-0.40◁ 0.39	-0.42◁ 0.23	-0.38◁ 0.34	-0.33◁ 0.16
	0.20	0.18±0.19	0.09±0.13	0.18±0.18	0.08±0.12
		0.14◦ 0.30	0.07◦ 0.16	0.14◦ 0.29	0.07◦ 0.14
		-0.38◁ 0.34	-0.37◁ 0.19	-0.37◁ 0.33	-0.35◁ 0.15
3	0.05	0.22±0.10	0.14±0.07	0.21±0.08	0.10±0.04
		0.19◦ 0.28	0.12◦ 0.18	0.20◦ 0.26	0.10◦ 0.12
		-0.35◁ 0.38	-0.32◁ 0.28	-0.20◁ 0.29	-0.15◁ 0.13
	0.10	0.22±0.09	0.12±0.06	0.21±0.08	0.10±0.04
		0.20◦ 0.27	0.11◦ 0.15	0.20◦ 0.26	0.10◦ 0.12
		-0.30◁ 0.32	-0.29◁ 0.18	-0.18◁ 0.29	-0.13◁ 0.14
	0.20	0.21±0.08	0.10±0.04	0.22±0.07	0.10±0.04
		0.20◦ 0.26	0.10◦ 0.13	0.20◦ 0.26	0.10◦ 0.12
		-0.21◁ 0.29	-0.20◁ 0.14	-0.17◁ 0.29	-0.13◁ 0.13
4	0.05	0.22±0.06	0.13±0.04	0.21±0.04	0.10±0.02
		0.18◦ 0.26	0.12◦ 0.15	0.19◦ 0.24	0.10◦ 0.11
		-0.27◁ 0.32	-0.23◁ 0.18	-0.04◁ 0.26	-0.07◁ 0.12
	0.10	0.21±0.06	0.11±0.03	0.20±0.05	0.10±0.02
		0.19◦ 0.24	0.10◦ 0.13	0.19◦ 0.24	0.10◦ 0.11
		-0.23◁ 0.29	-0.08◁ 0.15	-0.10◁ 0.26	-0.08◁ 0.12
	0.20	–	0.10±0.02	–	0.10±0.02
		–	0.10◦ 0.11	–	0.10◦ 0.11
		–	-0.09◁ 0.12	–	-0.16◁ 0.12
5	0.05	0.20±0.05	0.12±0.03	0.20±0.03	0.10±0.02
		0.18◦ 0.24	0.10◦ 0.14	0.18◦ 0.22	0.09◦ 0.10
		-0.07◁ 0.29	-0.04◁ 0.16	-0.04◁ 0.24	-0.03◁ 0.11
	0.10	0.20±0.04	0.10±0.02	0.19±0.04	0.10±0.01
		0.18◦ 0.22	0.10◦ 0.11	0.18◦ 0.22	0.09◦ 0.10
		-0.04◁ 0.25	-0.02◁ 0.13	-0.03◁ 0.24	-0.01◁ 0.11

Table 3.2: Mean±SD, 1.◦3. quartile, and min.◁max. of  $\varepsilon_{cmn} - \varepsilon_{all}$  for 6 SNPs, 20 haplotypes in total, and both haplotype frequency patterns. Results for 60% and 80% coverage and for thresholds 0.01 and 0.15 fit into the trend and were therefore omitted from the presentation.

### 3.1.2 Simulation results

Statistics on the difference  $\varepsilon_{cmn} - \varepsilon_{all}$  were calculated to assess the potential bias of the estimate  $\hat{\varepsilon}$ , when  $\varepsilon_{cmn}$  is used. The calculation of  $\varepsilon_{lve}$  did not lead to meaningful results with values above 1 and below 0 (data not shown). It is, therefore, excluded from the presentation. Tables 3.1–3.2 on pages 32f list the statistics on  $\varepsilon_{cmn} - \varepsilon_{all}$  for the case of 6 SNPs. Results for 3 and 8 SNPs were similar and, therefore, omitted from the presentation.

These tables show that a persistent bias occurs in the calculation of  $\varepsilon$ , when only common haplotypes are used. This bias increases with the number of rare haplotypes that are present and decreases with increasing coverage by the common haplotypes. The variance of  $\varepsilon_{cmn} - \varepsilon_{all}$  decreases with a growing number of common haplotypes. Besides, the number of common haplotypes and their frequency pattern have a minor influence on this bias. The inequality pattern, combined with low thresholds for common haplotypes, causes a slightly higher bias than the equality pattern does. The bias becomes smaller, when more SNPs are considered (data not shown).

The measure  $\varepsilon$  is sensitive for rare haplotypes, and some information is missed if their frequencies are not used in the calculation. However, for 90% coverage by the common haplotypes, the bias of  $\varepsilon_{cmn}$  is confined to 0–0.15 or 0.20 in most cases, depending on the number of rare haplotypes. For 70% coverage, the bias can assume values up to 0.3. Some rare cases show extreme positive as well as negative bias. Closer inspection of these cases reveals that those SNPs, which become mono-allelic if only the common haplotypes are considered, are an important source for the variance of the bias. In cases with negative bias, 2-3 SNPs became mono-allelic in common haplotypes, whereas in cases with strong positive bias, all SNPs remained bi-allelic.

## 3.2 Applicability of $\varepsilon$

This section will demonstrate  $\varepsilon$ 's ability to reasonably describe LD and block structures in two simulation studies and by application to a previously analyzed data set with an established block structure.

### 3.2.1 Simulation I: A single block

#### Objective of analysis and simulation study design

If an LD block is present in a SNP sequence, a reasonable multilocus LD measure should assume high values over this block for windows that exactly match the block location and its size. It should assume lower values for windows of different size or different location than the block. To investigate if  $\varepsilon$  shows such a behavior, multiple data sets were simulated using the SNaP program (see section 2.5). In an ideal scenario, all data sets had a single haplotype block and four neighbouring SNPs in linkage equilibrium on either side in common. The simulated data sets differed in the number of SNPs within the block and in the number of observed haplotypes at the block as well as in their frequency pattern. The single SNP allele frequencies were set to values of 0.4, 0.5, 0.2, 0.3 (left side of the block) and 0.2, 0.1, 0.3, 0.1 (right side). Block sizes ranged from 2 to 10 SNPs and numbers of observed haplotypes from 2 to 8. To take a possible influence of the haplotype frequency pattern in the block on  $\varepsilon$  into account, four different models were chosen:

- Model **e**: All occurring haplotypes are equally probable.
- Model **1r**: There is one major haplotype (frequency  $p_i = 0.7$ ), all additional haplotypes have equal frequencies.
- Model **2r**: There are two major haplotypes ( $p_i = 0.35$  each), all additional haplotypes have equal frequencies.
- Model **2**: There are two groups of haplotypes of approximately equal size. Within each group, haplotypes have equal frequencies. One group's frequencies sum up to 0.7, the sum in the other group is 0.3.

For each possible combination of values for block size, number of observed haplotypes, and frequency model, 100 replicates with 1000 haplotypes each were randomly generated. Window sizes 2–10 were used for  $\varepsilon$ .

**Percentage of replications where  $\varepsilon$  assumes its maximum over an LD block with a block-matching window**

<i>Block</i>		<i>Block Haplotypes Frequency Pattern</i>							
<i># SNPs</i>	<i># HT</i>	<b>e</b>		<b>1r</b>		<b>2r</b>		<b>2</b>	
2	2	32	(0.50)	32	(0.50)	32	(0.50)	32	(0.50)
	3	100	(0.14)	63	(0.21)	100	(0.14)	63	(0.21)
	4	0	(-)	22	(0.07)	2	(0.08)	2	(0.08)
3	2	65	(0.54)	65	(0.54)	65	(0.54)	65	(0.54)
	3	67	(0.25)	56	(0.36)	67	(0.25)	56	(0.36)
	4	87	(0.26)	75	(0.29)	87	(0.26)	87	(0.26)
	6	69	(0.12)	86	(0.19)	81	(0.18)	71	(0.14)
	8	0	(-)	100	(0.13)	43	(0.17)	0	(-)
4	2	73	(0.58)	73	(0.58)	73	(0.58)	73	(0.58)
	3	74	(0.42)	79	(0.45)	74	(0.42)	79	(0.45)
	4	69	(0.38)	77	(0.43)	75	(0.40)	75	(0.40)
	6	84	(0.29)	88	(0.37)	82	(0.34)	87	(0.31)
	8	89	(0.21)	96	(0.31)	86	(0.29)	91	(0.23)
6	2	90	(0.66)	90	(0.66)	90	(0.66)	90	(0.66)
	3	79	(0.60)	84	(0.62)	80	(0.61)	84	(0.62)
	4	79	(0.57)	84	(0.59)	81	(0.58)	81	(0.58)
	6	94	(0.51)	93	(0.55)	89	(0.54)	91	(0.52)
	8	98	(0.45)	98	(0.52)	93	(0.51)	95	(0.47)
8	2	98	(0.72)	98	(0.72)	98	(0.72)	98	(0.72)
	3	93	(0.71)	92	(0.72)	92	(0.71)	92	(0.72)
	4	91	(0.68)	95	(0.70)	95	(0.69)	95	(0.69)
	6	95	(0.63)	95	(0.66)	98	(0.65)	96	(0.63)
	8	99	(0.59)	98	(0.64)	99	(0.63)	98	(0.60)
10	2	100	(0.78)	100	(0.78)	100	(0.78)	100	(0.78)
	3	99	(0.76)	99	(0.77)	99	(0.76)	99	(0.77)
	4	98	(0.74)	98	(0.75)	97	(0.75)	97	(0.75)
	6	98	(0.70)	99	(0.73)	99	(0.72)	99	(0.70)
	8	100	(0.67)	99	(0.71)	100	(0.70)	100	(0.68)

Table 3.3: Percentage of replications, where  $\varepsilon$  assumes its maximum value over an LD block when window size and location match the block (simulation I). The number of SNPs in the block (*# SNP*), the number of haplotypes at the block (*# HT*), and the block haplotype frequency model varied in the simulation (see text). The mean values of  $\varepsilon$  in those replications, where it assumed its maximum over the block, are given in parenthesis.

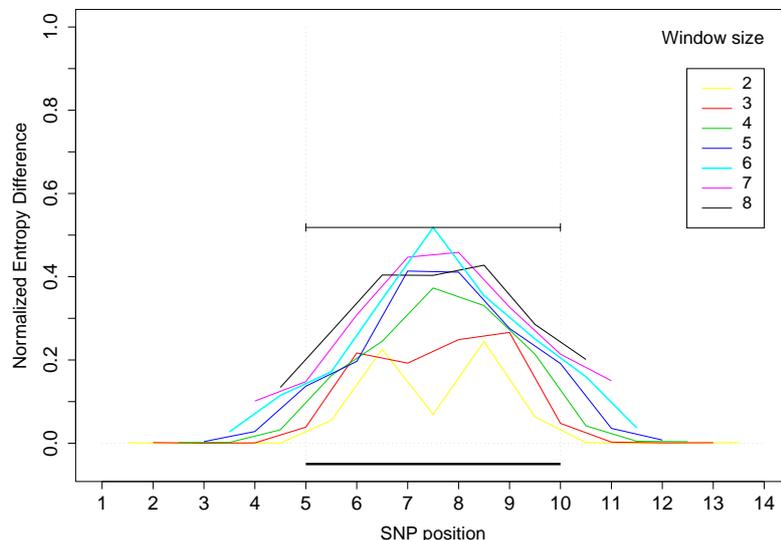


Figure 3.2: **Typical pattern of  $\varepsilon$  values for sliding windows of sizes 2–8 along a simulated SNP sequence (simulation I).** The sequence contained an LD block of 6 SNPs at positions 5–10 (depicted by the bold line below the graph). Seven haplotypes that follow frequency pattern **2** (see text) were present in the block. The maximum value for  $\varepsilon$  is assumed for window size 6 (cyan line) at the block-matching location. The horizontal line on top of the peak depicts this maximizing window with its corresponding value of  $\varepsilon$ .

## Simulation results

Table 3.3 on page 36 lists the percentage of replications where  $\varepsilon$  assumes its maximum value over the LD block with regard to window size and location. If a block included four or more SNPs, the results for the different numbers of loci and the various haplotype patterns were approximately the same (data not shown). Data sets generated with haplotypes that consisted of five and seven loci fit very well into the general trend (data not shown).

The proportion of correct maxima is often small for the smallest window size of 2, but constantly much higher for moderate sizes. It is 0 for four equally probable haplotypes composed of two SNPs and for eight equally probable haplotypes composed of three SNPs, because these loci are in linkage equilibrium. For moderate to larger window sizes,  $\varepsilon$  assumes its maxi-

imum for the correct window size and location in most replications, predominantly above 90%. For these sizes, the frequency pattern of the haplotypes does not influence the measure's ability to detect block structure.

Figure 3.2 illustrates a typical case in the application of  $\varepsilon$ . Growing window sizes increase the value of  $\varepsilon$ , until it assumes its maximum over the block if the correct block size is used. Then the value of  $\varepsilon$  decreases with even larger window sizes.  $\varepsilon$  approaches 0 outside the block, since linkage equilibrium was assumed between the SNPs in this area in the simulation. The figure also illustrates an effect of larger window sizes. For windows of moderate or large size that overlap a block border, e.g. a window that includes the SNP positions 3–7, the value of  $\varepsilon$  does not drop to 0, since there is still some LD between at least some SNPs in the window. Thus, larger windows show a *smoothing effect*. This effect becomes stronger with growing sizes.

### 3.2.2 Simulation II: Large and adjacent blocks

#### Objective and simulation design

To investigate the behavior of  $\varepsilon$  in a more complex, but still somewhat idealized situation, 1000 random haplotypes of a sequence that contained four independent blocks and several single SNPs were generated using SNaP. Three situations were modelled in the sequence:

1. *Large block*: A block, B1, of 20 SNPs (positions 4–23) with eight haplotypes of unequal frequencies was simulated.  $\varepsilon$  cannot be estimated with this window size in real-world applications due to sample size limitations (see section 2.3).
2. *Two adjacent blocks*: To assess the measure's ability to separate two blocks that have no single SNP between them, two independent blocks, B2 and B3, were simulated. These adjacent blocks were composed of five and four SNPs at positions 28–32 and 33–36, with four and three haplotypes of unequal frequencies, respectively.

3. *Single block*: A fourth block, B4, composed of six SNPs at positions 42–47 with four haplotypes of unequal frequencies was also generated.

These three units were separated by interspersed single SNPs of differing allele frequencies that were in linkage equilibrium to each other and to the blocks. Window sizes 2–8 were used with  $\varepsilon$ .

### Simulation results

Figure 3.3 depicts the  $\varepsilon$  values and pairwise LD values for a typical outcome in the simulations. All window sizes signal strong LD over block B4, but  $\varepsilon$  assumes its maximum for the correct window size and location. The adjacent blocks B2 and B3 are identified as two separate blocks:  $\varepsilon$  drops sharply between the blocks for moderate window sizes, although not completely to 0 for sizes greater than 2 due to the smoothing effect.  $\varepsilon$  indicates strong LD for larger window sizes (6 to 8), but does not reach the peak levels of the smaller correct window sizes.  $\varepsilon$  is constantly high over block B1 and increases with window size. Although  $\varepsilon$  is not directly applicable to a block of size 20 due to sample size limitations, it strongly indicates LD along the block for moderate window sizes.

All four blocks can also be recognized by using the pairwise LD measures  $D'$  and  $r^2$  (see figure 3.3): B2 and B3 exhibit complete LD in almost all pairs while B1 and B4 show a more complex structure. The sharp block borders are due to the assumption of linkage equilibrium between the interspersed SNPs and the SNPs in blocks.

### 3.2.3 An established block structure

#### Data set description and objective of analysis

Among the first investigators to search for a discrete haplotype block structure are Daly et al. [30]. They analyzed 103 common SNPs (minor allele frequency greater than 5%) from a sequence of 500 kb on chromosome 5q31, resulting in an average distance of 4.9 kb between adjacent SNPs. The data set included 129 trios (516 independent chromosomes) of European descent.

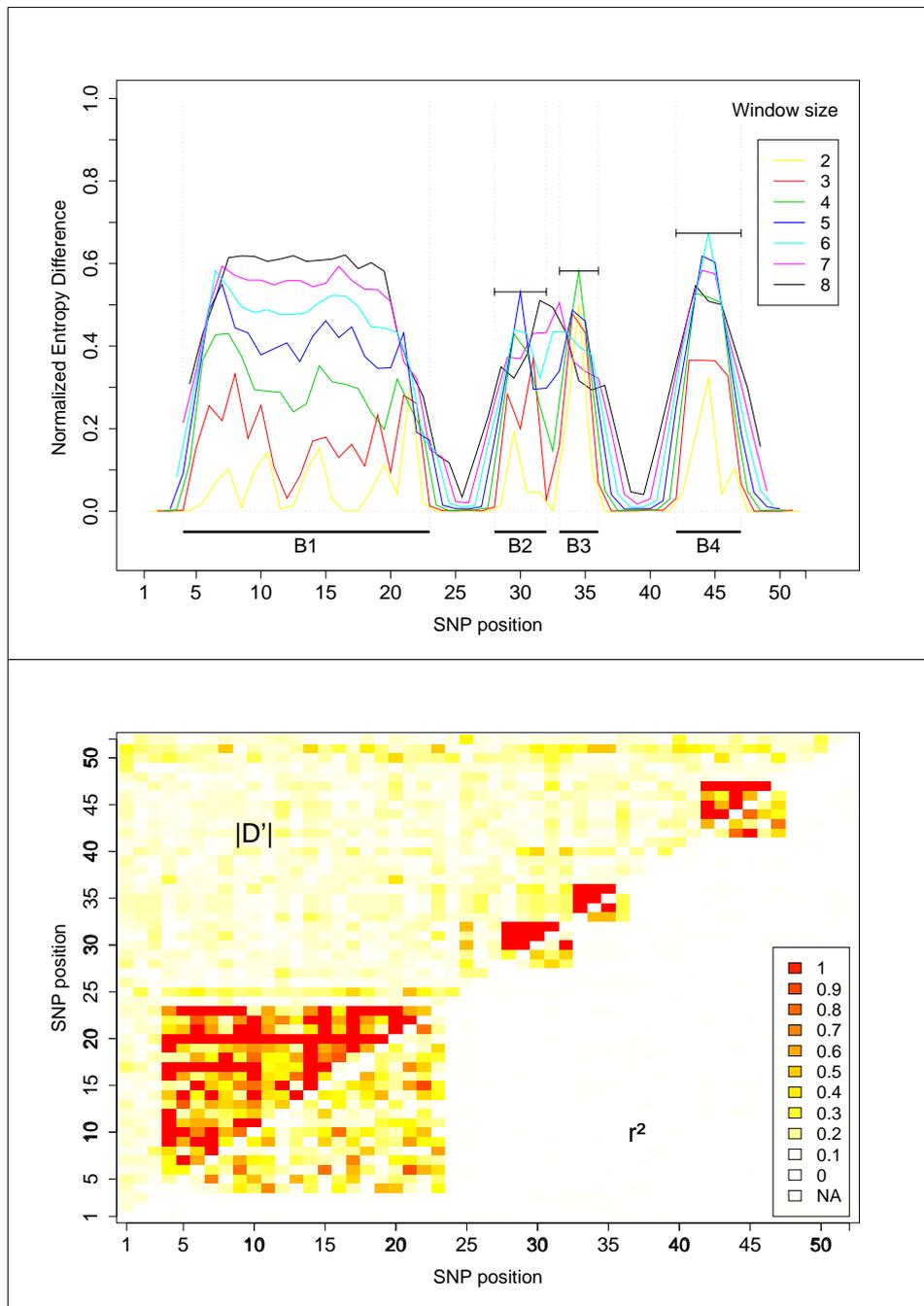


Figure 3.3: **Typical LD pattern in simulation II.** Top:  $\varepsilon$  values of sliding windows for a simulation II data set. Bold lines below the graph depict the four blocks generated. Horizontal lines within the graph depict the windows that assumed the local maxima of  $\varepsilon$  (B2–B4) and the maximizing windows at the block borders (B1). Bottom: Pairwise  $|D'|$  values (upper left triangle) and  $r^2$  values (lower right triangle) for all SNP pairs from the same data set. Both measures indicate strong LD within the blocks.

The authors found several blocks, where only 2–4 haplotypes accounted for over 90% of the frequencies at these blocks. The blocks were described as being separated by regions with higher historical recombination frequencies  $\Theta$  (see section 1.3). The blocks were composed of five to eleven SNPs, with the exception of one block that included 31 SNPs.

$\varepsilon$  was calculated in this data set to test if it would detect an elsewhere established block structure. The genotypic data set was obtained from the web site<sup>2</sup>. Haplotype frequencies were estimated from the trios using an EM-based Maximum-Likelihood approach [124]; genotype information from the children was only used to infer the parents' haplotypes.  $\varepsilon$  was calculated for window sizes between 2 and 8. For comparison, all pairwise values of  $D'$  and  $r^2$  were also calculated.

### Analysis results

Figure 3.4 illustrates the values of  $\varepsilon$  for sliding windows of sizes 2–8 along the SNP sequence and the pairwise LD values  $|D'|$  and  $r^2$  [103]. 99 SNPs within blocks were specified in Daly et al. [30, fig. 2], but another four SNPs fell between the blocks and were not specified. Thus, the blocks could not securely be matched with the exact SNP positions in the publicly available data set.

In general, the whole region exhibits strong levels of LD, presumably due to the dense marker spacing.  $\varepsilon$  clearly indicates two blocks of sizes 8 and 5 on the left edge of the sequence (positions 1–8 and 10–14) and a 3-SNP block (99–101) with very strong LD or an extended 5-SNP block (99–103) with weaker LD on the right edge. These blocks coincide with the ones specified in Daly et al. [30, fig. 2] and appear as bold disjointed triangles in the pairwise LD matrix (fig. 3.4, bottom). The borders of these blocks are sharp. The sites of more frequent recombination around these blocks from Daly et al. [30, fig. 2] correspond to the drops of  $\varepsilon$  (top) and to the sharp triangle borders (bottom) in figure 3.4.

$\varepsilon$  is constantly high over the middle part of the sequence, indicating strong

---

<sup>2</sup><http://www-genome.wi.mit.edu/humgen/IBD5/>

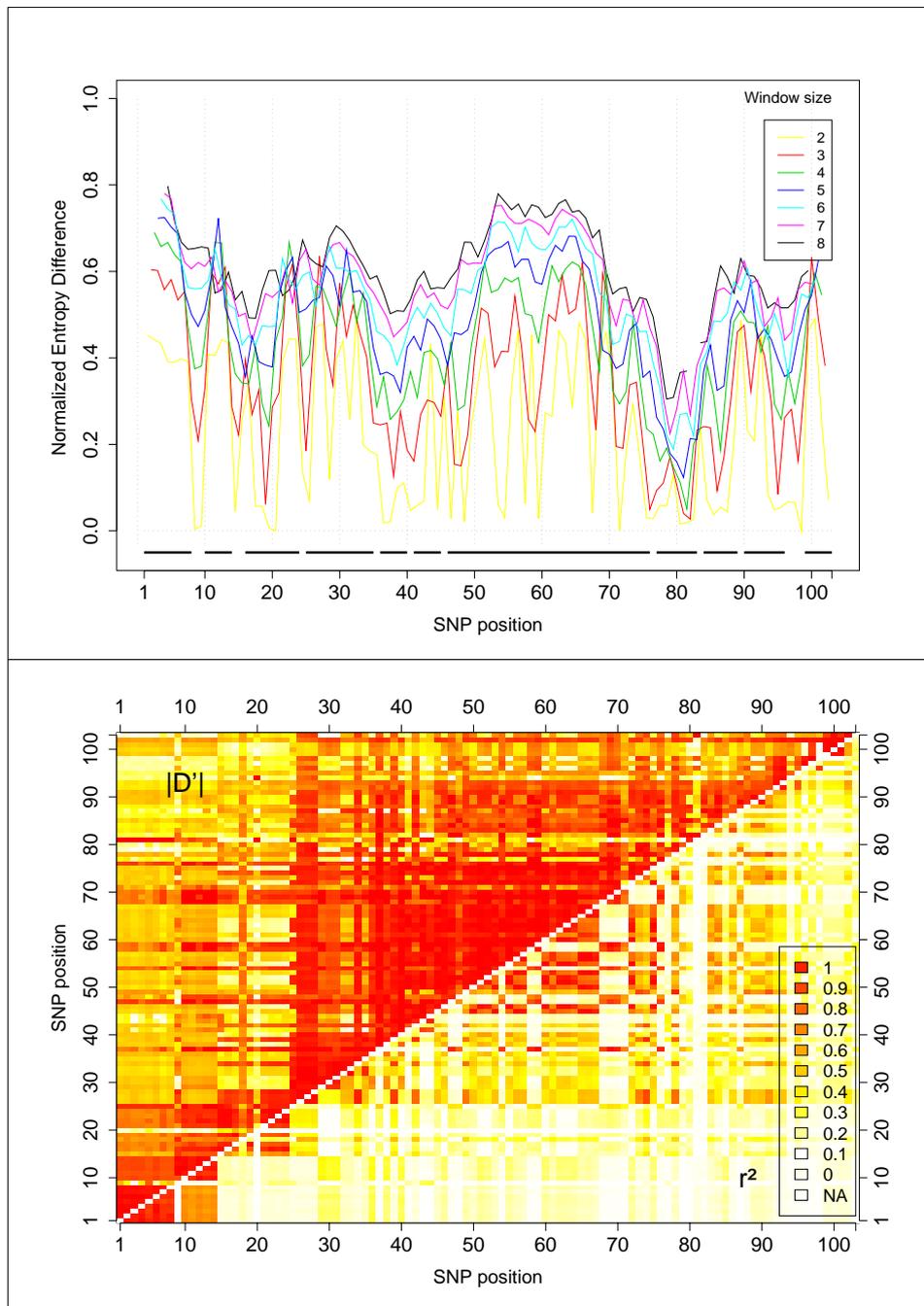


Figure 3.4: LD values for the Daly et al. [30] data set. Top:  $\varepsilon$  values for sliding windows of sizes 2–8. One  $\varepsilon_8$  value is missing due to haplotype estimation problems. The presumed locations of blocks from Daly et al. [30, fig. 2] are depicted by bold lines below the graph. Bottom:  $|D'|$  values (upper left triangle) and  $r^2$  values (lower right triangle) of all SNP pairs. The SNP positions do not represent physical distances.

long-ranging LD, while also hinting at some substructures with even higher LD within this area. The drop in the value of  $\varepsilon$  at SNP loci 81 and 82 strongly suggests linkage equilibrium in this regions. This region corresponds to the gap between blocks 7 and 8 stated in Daly et al. [30, fig. 2] and to the bright lines (low LD) at positions 81/82 in the pairwise matrix (fig. 3.4, bottom). The positions 9, 15, 20, 25, 36, 46, 59, 71, and 95–98 show similar drops, but to a lesser extent. Some of them roughly coincide with the block borders from Daly et al. [30, fig. 2], namely 25, 36, 46 and 81.

However, the genomic region between SNPs 16 and 80 cannot be described as a series of incoherent blocks. This is not surprising, since this region contains a haplotype with a frequency of 0.38 [30]; this major haplotype almost guarantees strong LD for every considered window size.  $\varepsilon$  recognized substructures between SNP loci 26 and 35 and between loci 47 and 75 with even stronger LD that correspond to blocks 4 and 7 in Daly et al. [30, fig. 2]. These results are confirmed by the pairwise LD matrix (fig. 3.4, bottom).

The proposed algorithm from section 2.4 was also applied to the data set. Small window sizes resulted in high variation of  $\varepsilon$  and fragmented the sequence into small blocks. Large window sizes suffered from over-smoothing, resulting either in blocks resembling almost the whole sequence for smaller thresholds or in very few extended blocks for high thresholds. For moderate window sizes (4–5) and thresholds (0.4–0.6), the resulting blocks partially coincide with those defined in Daly et al. [30], namely blocks of SNPs 1–8, 9/10–14, and 98–103. The influence of window size and threshold on the block definition is investigated more thoroughly in chapter 4.

# Chapter 4

## Block patterns on human chromosome 12

### 4.1 Data set description and objective

As part of the international efforts to create a human haplotype map<sup>1</sup> that are now underway, chromosome 12 was sequenced in 30 CEPH (European American) trios by the Human Genome Sequencing Center (HGSC) of the Baylor College of Medicine<sup>2</sup> [29] in collaboration with ParAllele<sup>3</sup>, using molecular inversion probe genotyping technology [56]. Given the size of chromosome 12 of about 132 Mb<sup>4</sup>, a remarkable average density of one SNP per 15.6 kb has been achieved by June 2003. The data set was kindly provided by Richard A. Gibbs of HGSC.

Chapter 3 demonstrated that  $\varepsilon$  reasonably reflects LD and block structure. The chromosome 12 data set provides the opportunity to look for block structures on the scale of a whole chromosome of medium size. In this chapter, the proposed  $\varepsilon$ -based block definition algorithm will be investigated. In

---

<sup>1</sup>The International Haplotype Map (HapMap) Project (<http://www.hapmap.org/>)

<sup>2</sup>Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, U.S.A., <http://www.bcm.tmc.edu/>

<sup>3</sup>ParAllele BioScience Inc., 384 Oyster Point Blvd., South San Francisco, CA 94080, U.S.A., <http://www.p-gene.com/>

<sup>4</sup>By November 2003, the Ensemble web page at <http://www.ensembl.org/> stated the length of chromosome 12 to be 132,018,379 bp.

particular, the analysis will yield answers to the following questions:

- Is the SNP coverage dense enough to deliver reasonable results?
- Is the algorithm that was proposed in section 2.4 capable of defining reasonable block structures?
- How do window size and threshold influence the block definition? Which parameters are suitable?
- How are block length, genome coverage, block haplotype diversity and other block quality measures influenced by the control parameters?
- How do the results of an  $\varepsilon$ -based algorithm relate to other LD measures and algorithms based on them, with an emphasis on  $D'$ ?

## 4.2 Analysis of the data set

The analysis is based on the Human NCBI<sup>5</sup> Build 31 HapMap Fix 12 from Baylor College, batches 1 and 2, of chromosome 12 from June 2003. This release contained 8,475 SNP genotypes. The SNPs were checked for deviations from Hardy-Weinberg equilibrium. 6,815 of these SNPs were found to be bi-allelic in the sample and considered for further analysis (80.4% of the original set). The SNPs were further required to have at least 70% known alleles to protect against strongly biased allele frequency estimates and to avoid problems during haplotype frequency estimation due to missing data. Finally, 3,567 SNPs were included in the subsequent analysis (42.1% of the original set), providing an average, but non-uniform resolution of 37.0 kb.

Values of  $\varepsilon$  for sliding windows of sizes 2–9 and pairwise values of  $D'$  and  $r^2$  were calculated using self-developed C programs and Perl scripts. Haplotype frequencies in a considered window were estimated using an EM-based algorithm for trios, where the children's genotypes are only used to infer the parents' haplotypes [124]. Families with inconsistent genotypes were excluded from the estimation in that particular window. Blocks were defined

---

<sup>5</sup>National Center for Biotechnology Information

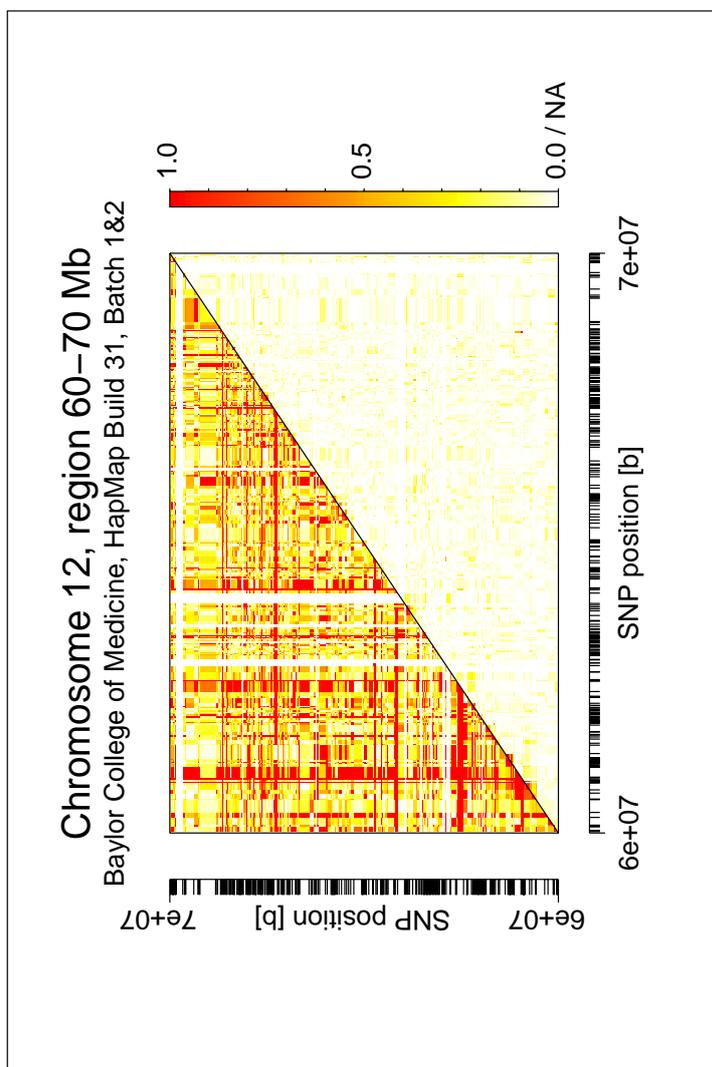


Figure 4.1: **Pairwise LD values in the example region.** Pairwise  $|D'|$  values (upper left triangle) and  $r^2$  values (lower right triangle) for SNP positions in the region 60–70 Mb (black lines below and left of the graph). This figure shows some striking similarities with figure 1.2 on page 12, namely the occurrence of a number of long red bars in the  $D'$  triangle. Again,  $r^2$  is not affected by missing haplotypes and assumes higher values, with few exceptions, only close to the diagonal. This is in clear contrast to  $D'$ .

according to the algorithm from section 2.4 using thresholds 0.1, 0.2, ..., 0.9. Perl scripts were developed for data handling, processing, block definition, and rudimentary statistics. R scripts were developed for the statistical analysis and for the presented figures.

**Example region.** An example region of SNPs with physical positions between 60 and 70 Mb was chosen as a persistent illustration of the different analysis steps. This region roughly corresponds to the regions 12q14.1-3 and 12q15; the filtered sample contained 295 SNPs in this region. Figure 4.1 depicts the pairwise values of  $|D'|$  and  $r^2$  for the example region as the classical approach to LD description. Figures 4.2 and 4.3 illustrate the typical pattern of calculated  $\varepsilon$  values and defined blocks in the example region. The analysis was carried out on the whole chromosome 12 data set, unless otherwise noted.

## 4.3 Block lengths and chromosomal coverage

### 4.3.1 Lengths and coverage of $\varepsilon$ -defined blocks

Table 4.1 contains physical length statistics of the blocks that were defined by the  $\varepsilon$ -based algorithm. The lengths vary from 2 kb to over 4 Mb. Physical length and chromosomal coverage, i.e. the proportion of the chromosome included in blocks, increase with larger window sizes and lower thresholds. 1–5% of chromosome 12 exhibits strongest LD.

$\varepsilon$  values for window sizes 2–3 are very variable and result in a fragmentation of the sequence into short blocks. For window sizes 7 and above, the smoothing problem of  $\varepsilon$  becomes serious: For low thresholds, blocks become unreasonably large and contain regions of very low LD. For more stringent thresholds, coverage decreases rapidly and only a few regions of very strong LD are defined as blocks. Also, the minimum number of SNPs in a block is limited by the used window size. Window sizes of 4–6 with thresholds between 0.4 and 0.6 are a good compromise between multilocus LD assessment and protection against too much variability and over-smoothing. The

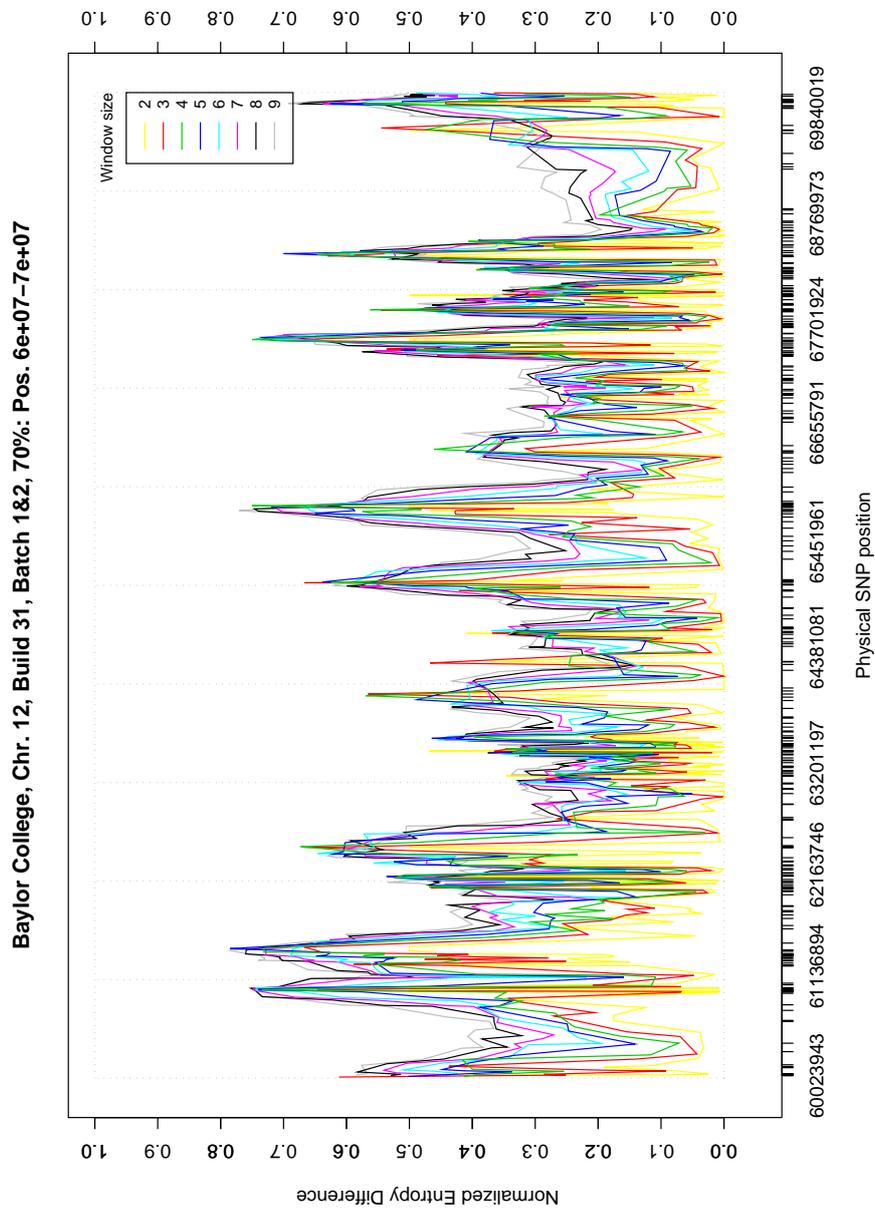


Figure 4.2:  $\varepsilon$  LD profiles.  $\varepsilon$  LD profiles for window sizes 2–9 for SNPs in the example region (60–70 Mb, black lines below the graph). There is much variability in this region, especially for small windows. Some  $\varepsilon$  peaks coincide with triangles of higher  $r^2$  values in figure 4.1.

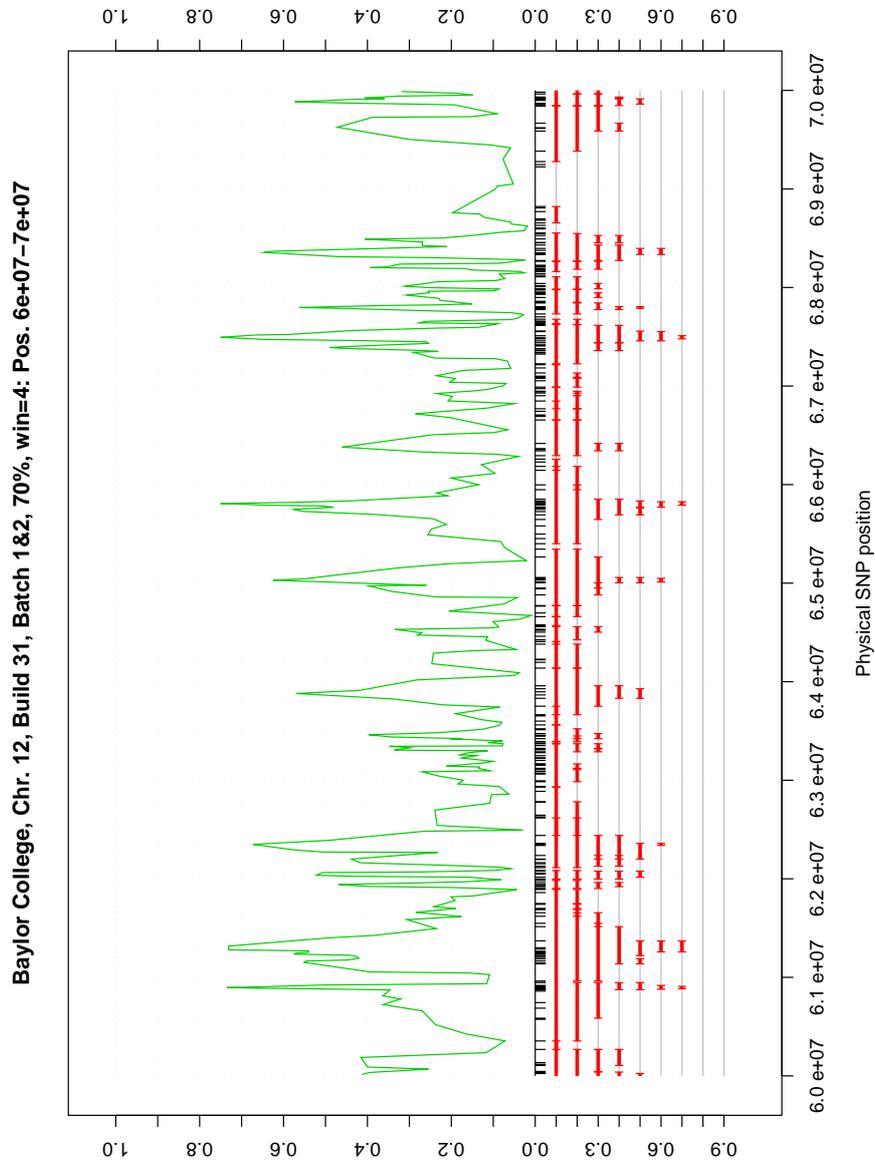


Figure 4.3:  $\epsilon$ -based block definition. Blocks (red lines) that are defined by  $\epsilon_4$  (green line) and thresholds 0.1–0.9 in the example region (60–70 Mb). SNP positions are marked by small black lines in the middle of the graph.  $\epsilon$  values increase in both directions along the ordinate from the middle black line that denotes 0. Low thresholds result in long blocks in partially low LD.

## Lengths and coverage of $\varepsilon$ -defined blocks on chromosome 12

<i>Threshold</i>	<i>Window size</i>				
	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>
0.10	47/22	197/124	–	–	–
	2<4412	12<5720	–	–	–
	25.2%	71.41%	–	–	–
0.20	28/16	128/81	250/180	413/301	653/500
	2<351	6<5407	14<5825	36<5919	56<7752
	12.2%	47.4%	69.5%	85.4%	92.2%
0.30	23/12	91/59	172/116	268/186	355/269
	2<262	6<720	12<5376	22<5862	28<5986
	7.7%	27.7%	46.4%	63.4%	72.1%
0.40	20/10	59/42	144/79	194/140	277/199
	2<245	5<382	11<4638	25<5369	39<5793
	4.7%	11.5%	25.8%	35.5%	44.9%
0.50	16/9	45/32	86/63	164/97	223/150
	2<148	5<215	12<558	21<4638	24<5369
	1.2%	4.1%	9.3%	17.2%	23.4%
0.60	–	46/31	71/47	106/83	207/126
	–	97<174	13<382	22<392	32<4638
	–	1.8%	2.9%	5.1%	11.1%
0.70	–	–	58/40	82/62	102/88
	–	–	19<148	21<215	28<262
	–	–	0.6%	0.9%	1.3%
0.80	–	–	–	25/25	21/21
	–	–	–	25<25	21<21
	–	–	–	< 0.1%	< 0.1%

Table 4.1: Mean/median and minimum<maximum of the physical block length [kb], complemented by the chromosomal physical coverage provided by these blocks. Blocks were defined using the  $\varepsilon$ -based algorithm (section 2.4) with varying window sizes and thresholds. A threshold of 0.1 for moderate and large window sizes resulted in nearly complete coverage and many overlapping blocks.

block lengths for these parameters vary from 10 kb to again over 4 Mb, with medians ranging from 31 kb to about 200 kb, and a physical chromosomal coverage between 2% and 36%. In general, these block lengths fit into the

picture found in previous studies (see p. 16) but the chromosomal coverage is lower. This could be due to the minimum number of SNPs per block required by the used window size. Thus, the higher coverage in Dawson et al. [32] and Phillips et al. [109] would be due to a large proportion of 2- and 3-SNP blocks. This is further investigated in section 4.7.

Correlation coefficients between the blockwise mean values of  $\varepsilon$  and the physical block lengths were calculated separately for each combination of window size and threshold. The coefficients varied from -0.13 to 0.44, but were usually confined to [0.1, 0.3]. Thus, the strength of LD had only a mild effect on the block length within the threshold-defined groups.

### 4.3.2 The origin of the block length distribution

$\varepsilon$  and block length are not strongly correlated, once the blocks are defined by  $\varepsilon$ . Thus, the question about the primary origin for the block length distribution within the threshold-defined groups arises. Figure 4.4 illustrates the block length distributions for two different window sizes and various thresholds in the block definition. For window size 2, the pattern of a highly skewed distribution of the block length confirms previous studies that used pairwise LD measures [46, 132, 109]. For larger window sizes, the skewness of the distribution decreases (see fig. 4.4). How can this pattern be explained?

The black lines in figure 4.4 illustrate the distribution of the physical size of windows with two and five SNPs, respectively. These curves have very similar shapes, when compared to the block length distributions. This indicates that the SNP distance distribution is the primary origin of the block length patterns. To describe this concordance more formally, the Bhattacharyya concordance measure [15] for two probability distributions with densities  $f_1$ ,  $f_2$  was used. The measure is defined as

$$D_B = \int_a^b \sqrt{f_1(x)f_2(x)}dx \quad (4.1)$$

and assumes 1 if the distributions are identical and 0 if the distributions do not overlap. Table 4.2 lists the comparison results for various window sizes and thresholds. For medium range thresholds that are considered useful

## Concordance of block length and SNP distance distributions

<i>Threshold</i>	<i>Window size</i>				
	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>
0.1	0.991 (716)	0.869 (483)	0.726 (278)	0.500 (140)	0.256 (55)
0.2	0.992 (575)	0.948 (496)	0.869 (393)	0.797 (302)	0.697 (211)
0.3	0.984 (445)	0.976 (406)	0.956 (378)	0.927 (337)	0.886 (304)
0.4	0.975 (307)	0.985 (260)	0.982 (272)	0.970 (256)	0.949 (232)
0.5	0.935 (99)	0.952 (123)	0.952 (146)	0.952 (142)	0.954 (145)
0.6	– (0)	0.952 (51)	0.896 (55)	0.941 (66)	0.927 (74)
0.7	– (0)	– (0)	0.854 (14)	0.791 (14)	0.754 (17)

Table 4.2: Concordance of the block length distributions and the SNP distance distribution on chromosome 12 as assessed by the Bhattacharyya measure  $D_B$  (4.1), using intervals (bins) of size 25 kb. The number of blocks is given in parentheses. For medium window sizes (4-6) and thresholds (0.4-0.6) the distributions are nearly identical. Deviations for high thresholds are explained by the insufficient number of blocks to reliably estimate the empirical block length distribution.

(see section 4.3.1), both distributions are nearly identical. The distributions deviate from one another for larger window sizes and extreme thresholds. For low thresholds, the resulting “blocks” do not resemble LD blocks.

## 4.4 Haplotypes in $\varepsilon$ -defined blocks

To investigate the influence of  $\varepsilon$  on the number and the pattern of haplotypes in the blocks, the haplotype frequencies within the blocks were estimated. Blocks that were defined by window sizes 2–6 and thresholds between 0.2–0.7 were considered. Due to computer memory and sample size limitations, this

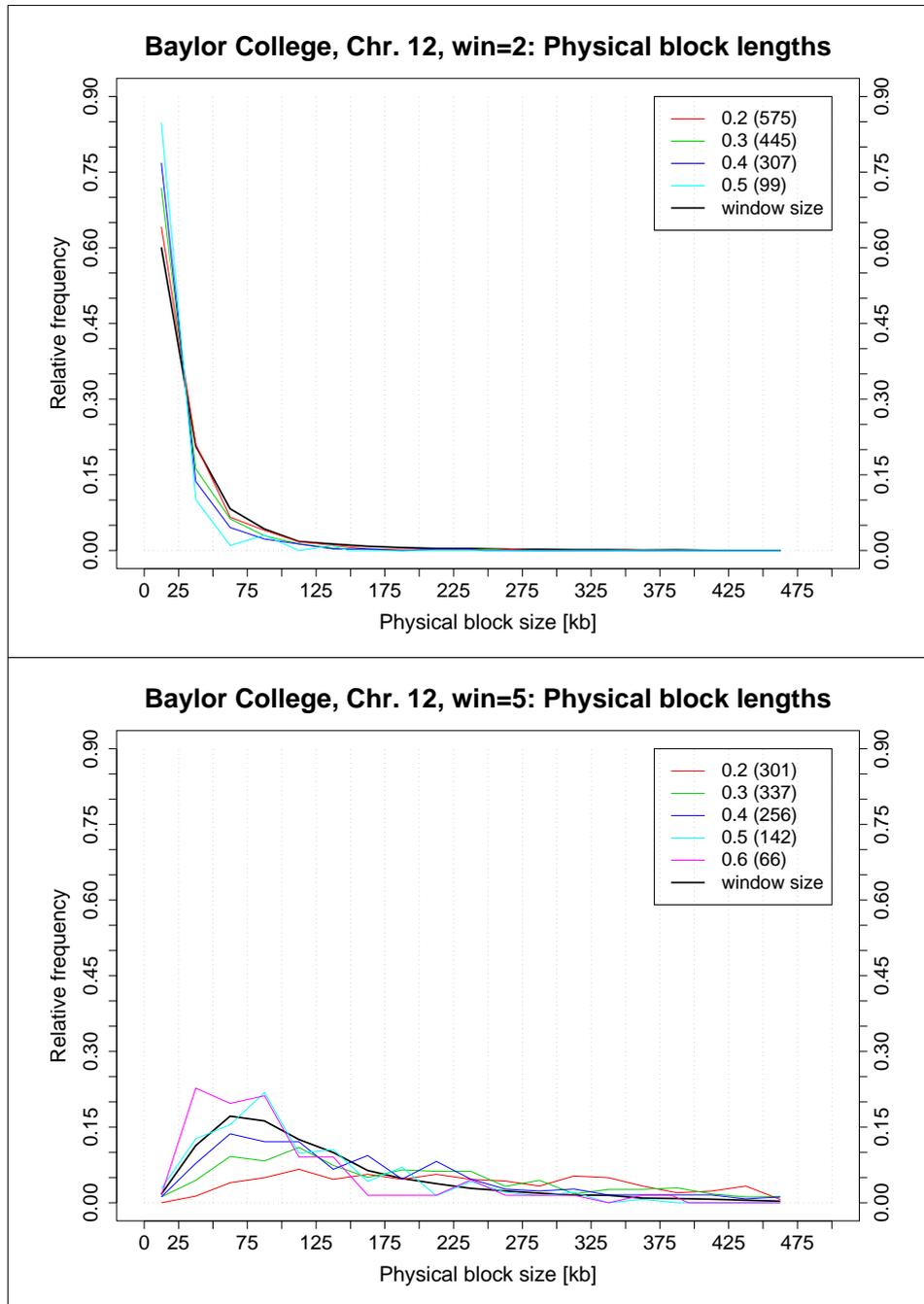


Figure 4.4: **Distributions of physical block lengths and SNP distances.** Histogram of the distributions of the physical block lengths (colored) and of the physical sizes of the sliding windows (black) for window sizes of two SNPs (top) and five SNPs (bottom) that were used with different thresholds in the block definition. The number of detected blocks is given in parentheses. Although not identical, the distributions of window size and block length are very similar in both cases. Lines instead of bars were chosen in the histogram to enable the simultaneous illustration of a number of distributions.

### Common haplotypes in $\varepsilon$ -defined blocks

$\varepsilon$		$t = 0.05$		$t = 0.10$		$t = 0.20$	
<i>Win.</i>	<i>Thr.</i>	<i>HT</i>	<i>Cov.[%]</i>	<i>HT</i>	<i>Cov.[%]</i>	<i>HT</i>	<i>Cov.[%]</i>
2	0.1	3.0/3	97.0/98.3	2.5/3	93.4/95.7	1.8/2	83.5/84.9
	0.3	2.1/2	98.1/99.0	1.9/2	96.3/97.3	1.7/2	93.7/95.3
	0.5	2.0/2	100.0/100.0	2.0/2	99.7/100.0	1.9/2	98.2/100.0
3	0.2	4.4/4	91.8/94.0	3.1/3	82.3/84.4	1.7/2	64.0/66.6
	0.4	3.0/3	96.4/97.2	2.4/2.5	92.5/93.5	1.8/2	82.9/83.6
	0.6	2.0/2	99.2/99.2	1.9/2	98.5/99.2	1.7/2	95.4/98.3
4	0.2	5.4/5	76.8/80.5	2.9/3	60.6/65.1	1.5/1	46.8/45.8
	0.4	4.0/4	92.5/94.4	2.9/3	84.7/86.6	1.8/2	68.9/71.1
	0.5	3.2/3	95.3/95.8	2.6/3	90.9/92.2	1.8/2	80.1/80.3
	0.6	2.6/3	96.8/97.5	2.2/2	94.5/94.1	1.8/2	88.4/89.2
	0.7	2.0/2	99.0/99.2	2.0/2	99.0/99.2	1.8/2	95.5/98.7
5	0.2	5.4/5.5	61.9/63.6	2.4/2	44.7/42.0	1.2/1	36.7/28.3
	0.4	4.9/5	85.5/87.4	3.1/3	72.9/75.4	1.6/2	53.9/53.2
	0.5	4.0/4	91.6/93.1	2.8/3	83.7/85.8	1.8/2	69.3/70.7
	0.6	3.1/3	94.3/95.2	2.7/3	90.8/91.9	1.9/2	79.8/79.2
	0.7	2.3/2	95.8/97.0	2.1/2	94.0/93.7	1.8/2	89.7/91.1
6	0.2	4.3/4	45.6/42.8	2.1/2	38.8/39.3	1.3/1	40.5/38.0
	0.4	5.2/5	77.3/80.2	2.8/3	61.3/62.9	1.4/1	45.5/43.3
	0.5	4.7/5	86.5/88.3	3.1/3	75.8/78.3	1.7/2	55.7/56.5
	0.6	3.9/4	91.9/92.9	2.9/3	84.9/85.7	1.8/2	69.3/70.7
	0.8	2.0/2	97.4/97.4	2.0/2	97.4/97.4	2.0/2	97.4/97.4

Table 4.3: Common haplotypes in  $\varepsilon$ -defined blocks on chromosome 12. Blocks were defined by various window sizes (*Win.*) and thresholds (*Thr.*) for  $\varepsilon$ . Listed are the mean/median number of haplotypes (*HT*) with frequencies above threshold  $t$  and the mean/median frequency coverage provided by them (*Cov.*).

was only feasible for blocks of 12 SNPs or less. This is no strong limitation, since most blocks had fewer markers.

Table 4.3 lists the average numbers of common haplotypes above certain frequency thresholds in  $\varepsilon$ -defined blocks and the coverage they provide. Results for thresholds 0.3 and 0.7 fit very well into the trend and were omitted from the presentation. For larger window sizes and low block-defining thresholds, the common haplotypes provide only modest coverage, thereby

indicating the existence of a number of low-frequency haplotypes at the block. More stringent thresholds for larger window sizes, roughly along the line 4–0.4, 5–0.5, 6–0.6, result in more than 80% coverage by three or less haplotypes with frequencies above 0.1 on average and five or less for frequencies above 0.05. The number of common haplotypes in these blocks varied from 1 to 6. For highest thresholds per window size, e.g. 0.7 for window size 4, there are usually two haplotypes that explain nearly all variation at the block.

## 4.5 Allele frequencies in $\varepsilon$ -defined blocks

The minor allele frequency distribution of all SNPs in the sample was estimated and compared to those of SNPs that were included in  $\varepsilon$ -defined blocks. The result is illustrated in figure 4.5. The allele frequency distribution in the sample is similar to previous studies [109].

About 35% of the SNPs that are included in blocks have minor allele frequencies above 0.4, when higher block thresholds are used. Thus, common SNPs are enriched in  $\varepsilon$ -defined blocks with high LD, when compared to the sample distributions. The enrichment becomes higher for increasing LD. Still, some rare SNPs are also included, but to a lesser extent than in the sample. For thresholds  $\geq 0.5$ , about 90% of the SNPs are common, with minor allele frequencies above 0.1.

## 4.6 Pairwise LD measures in $\varepsilon$ -defined blocks

Half of the algorithms previously proposed for the definition of block are based on the pairwise measure  $D'$ . It was, therefore, investigated, how  $\varepsilon$ -defined blocks relate to this measure and also to  $r^2$ . To this end, the correlations of the mean values and the median values of  $|D'|$  and  $r^2$  with the mean  $\varepsilon$  values were calculated, separately for each used window size and threshold.

Figure 4.6 illustrates the striking contrast between the pairwise measures for window size 4. While the blockwise mean values of  $r^2$  and  $\varepsilon$  sometimes show a very strong correlation, the mean values of  $|D'|$  and  $\varepsilon$  are only weakly

**Baylor College, Chromosome 12, SNP allele distribution**

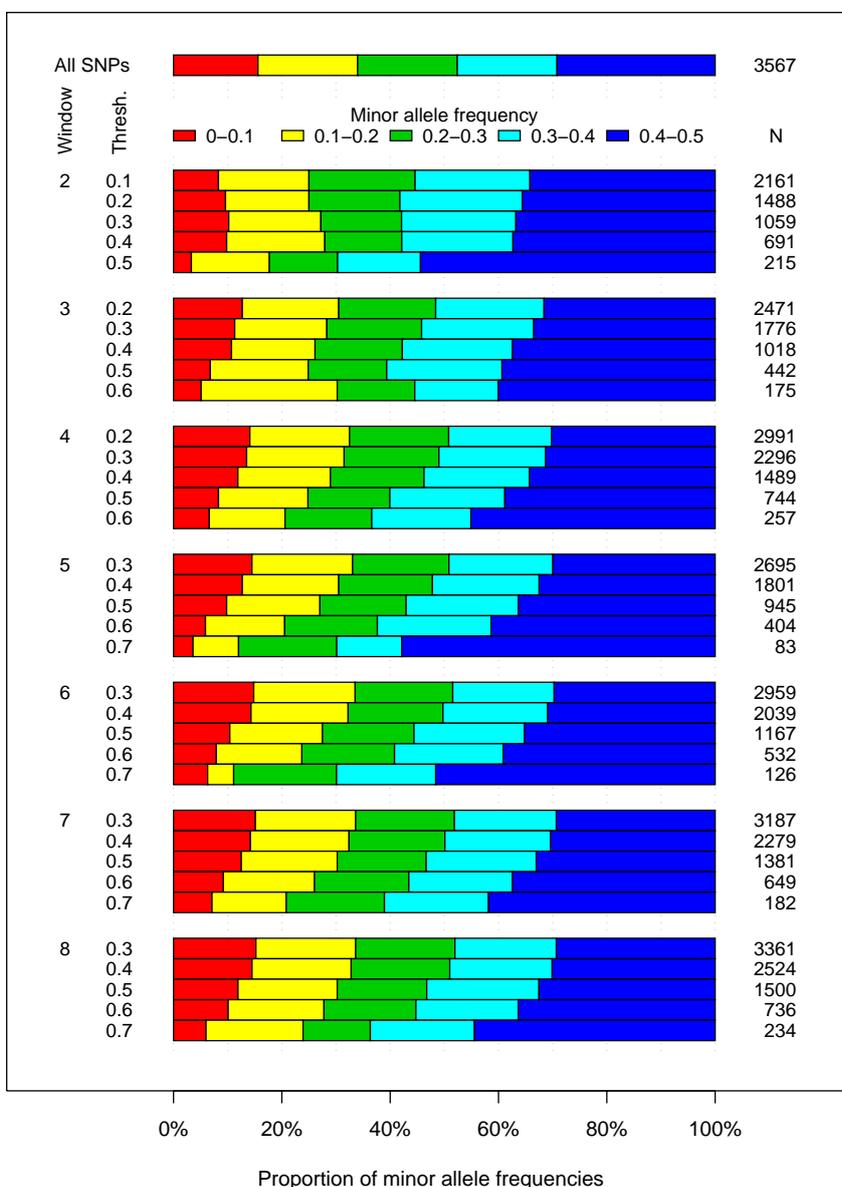


Figure 4.5: Minor allele frequency distribution for all SNPs in the sample and for SNPs included in  $\epsilon$ -defined blocks. Columns on the left specify the window size and threshold used for the block definition, whereas the numbers of SNPs included in those blocks are given on the right. Blocks of higher LD show a clear trend of enriching very common SNPs.

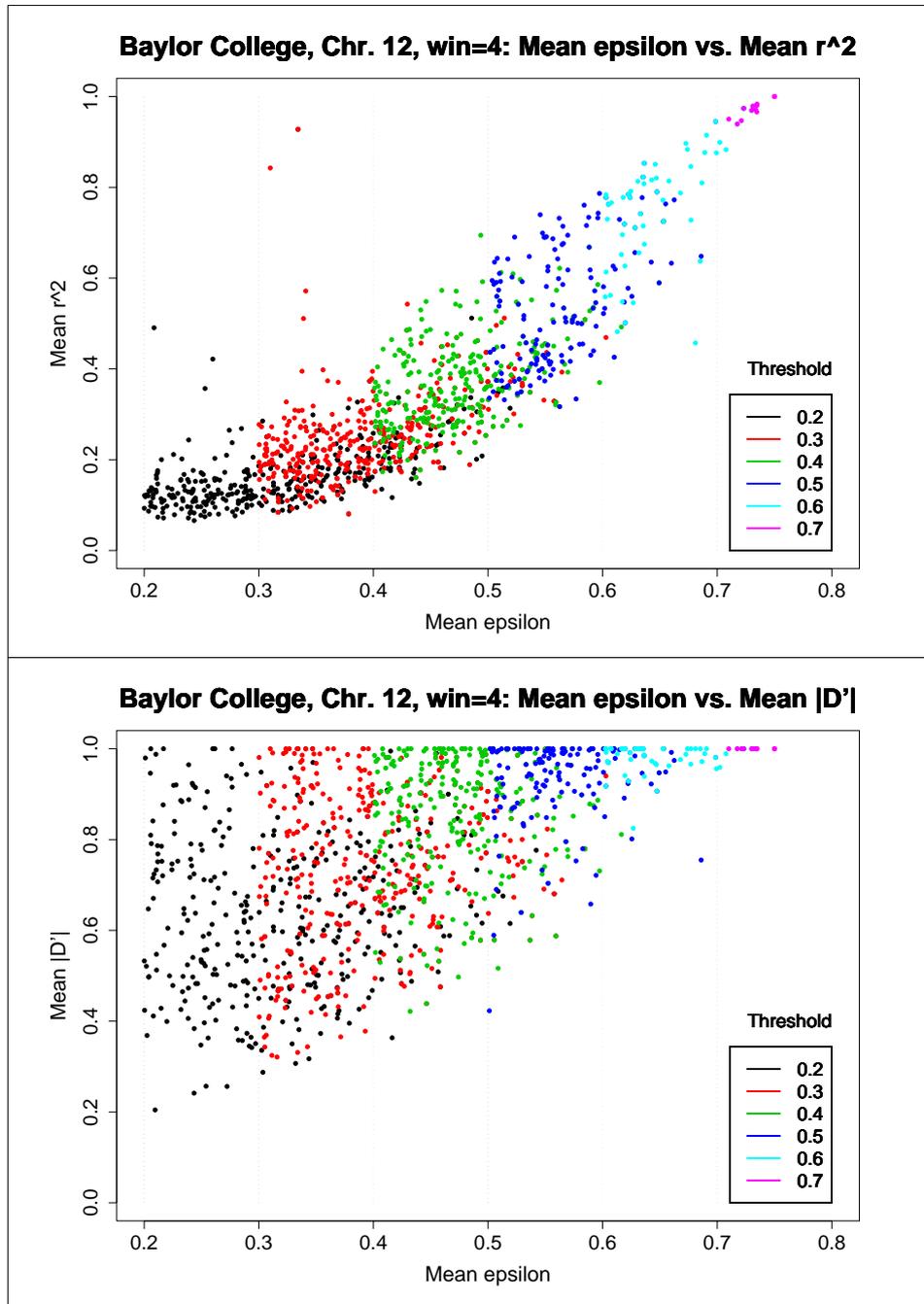


Figure 4.6: **Plot of  $\varepsilon$  vs.  $r^2$  and  $\varepsilon$  vs.  $|D'|$  in  $\varepsilon$ -defined blocks.** Mean  $\varepsilon$  values vs. mean  $r^2$  values (top) and mean  $|D'|$  values (bottom), respectively, in blocks defined by various thresholds for  $\varepsilon_4$ . There is strong correlation between the mean values of  $\varepsilon_4$  and  $r^2$  within the threshold-defined groups (colored dotted lines), but almost none between values of  $\varepsilon$  and  $|D'|$ .

### Correlation between $\varepsilon$ and $r^2/|D'|$ in $\varepsilon$ -defined blocks

$r^2/ D' $ Threshold	Window size				
	2	3	4	5	6
0.2	0.96/ 0.40	0.56/ 0.09	0.55/ 0.11	0.48/ 0.06	0.55/-0.00
0.3	0.97/ 0.49	0.62/ 0.22	0.50/ 0.19	0.48/ 0.09	0.48/ 0.08
0.4	0.98/ 0.41	0.73/ 0.25	0.54/ 0.09	0.47/ 0.11	0.39/ 0.00
0.5	-/-	0.91/ 0.27	0.59/ 0.18	0.37/ 0.13	0.33/-0.07
0.6	-/-	0.97/ 0.42	0.66/ 0.14	0.45/ 0.10	0.31/ 0.13
0.7	-/-	-/-	0.93/-	0.38/ 0.58	0.11/ 0.24

Table 4.4: Correlations between the mean values of  $\varepsilon$  and the mean values of  $r^2$  and  $|D'|$ , respectively, with regard to window size and threshold in the block definition. The strong correlation between  $\varepsilon$  and  $r^2$  weakens with larger window sizes.

correlated; the mean of  $\varepsilon$  roughly serves as a lower boundary for the mean of  $|D'|$ .

Table 4.4 lists the mean value correlations for most pairs of window sizes and thresholds. The correlation between the mean values of  $\varepsilon$  and  $|D'|$  is considerable only for highest thresholds for  $\varepsilon$ , e.g. window size 5 and threshold 0.7. The mean values of  $\varepsilon$  and  $r^2$  are often highly correlated within the window size and threshold defined subgroups. This correlation becomes weaker for larger window sizes. The correlations are weaker for blockwise median values (data not shown).

## 4.7 Comparison of algorithms

To compare the outcome of the proposed algorithm from section 2.4 to that of other methods and to emphasize the different objectives of  $D'$ -based methods (recombination) and those based on  $\varepsilon$  and  $r^2$  (LD), various algorithms were applied to the data set.

**Applied algorithms.** Half of the existing algorithms are based on  $D'$  and use differing thresholds for the minimum of all pairwise  $|D'|$  values or for the confidence intervals of it. To represent these pairwise methods, a minimum

$|D'|$  algorithm was applied with thresholds 0.2, 0.8, and 1.0. This method considers a set of successive SNPs to be a block, as long as  $|D'|$  does not drop below the specified threshold for each pair of these SNPs. So far, no methods have been published that are based on  $r^2$ . To look at the potential of those methods, algorithms similar to the above defined  $|D'|$ -based ones were applied. They require the minimum  $r^2$  or the mean  $r^2$  value of all SNPs of a block to be above a certain threshold. Thresholds of 0.1, 0.3, and 0.5 were considered.

The  $\varepsilon$ -based algorithm from section 2.4 was applied using window sizes 4–6 and thresholds 0.2–0.6. Pairwise LD values were calculated using self-developed C programs and an EM-based haplotype estimation algorithm [124]. The block definition algorithms were implemented in R.

**Results.** Figure 4.7 illustrates the resulting blocks for the different block methods in the example region. The outcome for  $D'$  – a virtual segmentation into many small blocks – differs strongly from the other LD measures' outcome. Both  $r^2$ -based methods yield the more similar results to the  $\varepsilon$ -based approach. Blocks defined by  $r^2$  are often smaller; some blocks are completely missing in either way. "3+" denotes the additional requirement of at least 3 SNPs per block. This considerably reduces the number of blocks for the pairwise methods, especially for  $D'$ . Patterns for  $r^2$  and  $\varepsilon_4$  with higher thresholds look similar under this condition.

Table 4.5 presents the block lengths and chromosomal coverage yielded from the various methods. The very high SNP coverage of 72.7% for  $|D'| = 1.0$  contrasts with only 35.9% physical coverage and is due to the non-uniform distribution of the SNPs on the chromosome. It essentially means that for two arbitrarily chosen adjacent SNPs, at least one out of the four haplotypes is missing with a probability of about 0.7. In general, pairwise methods utilizing  $D'$  and  $r^2$  find many more blocks than the  $\varepsilon$ -based algorithm, but these blocks are smaller and contain fewer SNPs. Depending on the thresholds used, the physical coverage is comparable. If at least three SNPs are required to be in a block, the picture changes dramatically. The number of blocks and their coverage is reduced by roughly 50%, whereas the physical block length

Baylor College, Chr. 12, Build 31, Batch 1&2, 60–70 Mb: Block definition results

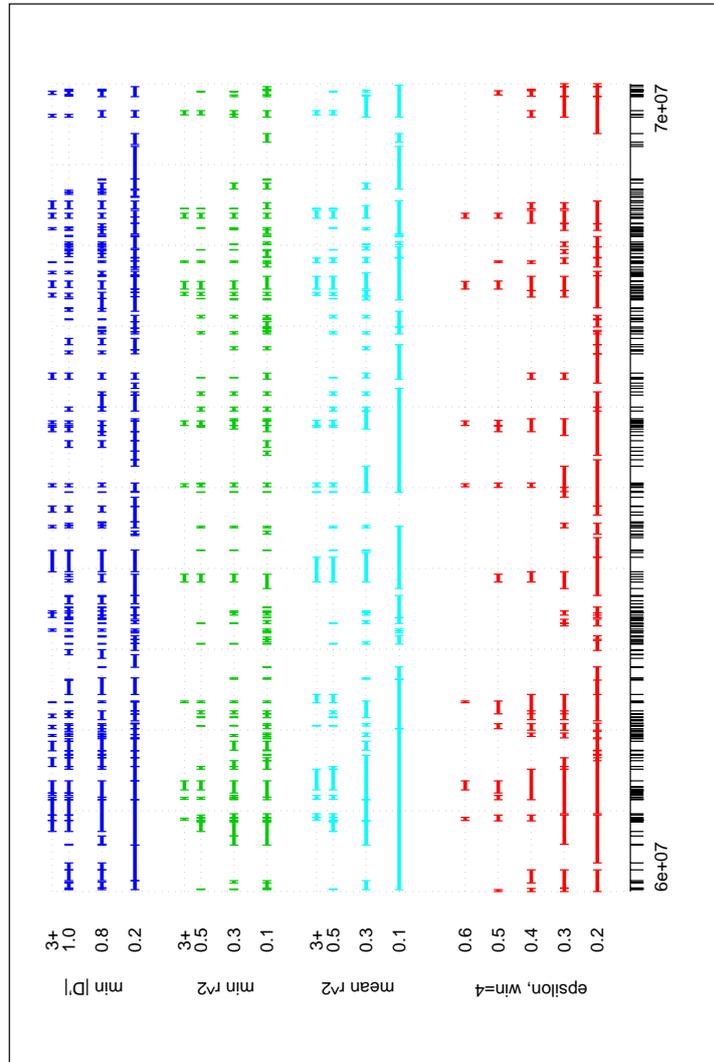


Figure 4.7: **Illustration of blocks derived from  $|D'|$ ,  $r^2$ , and  $\epsilon$ .** Block definition results, using methods based on  $\epsilon$ ,  $|D'|$ , and  $r^2$  and various thresholds (see text) for SNP positions in the example region (60–70 Mb, black lines below the graph). 3+ designates the complementary condition that blocks are required to contain at least 3 SNPs to be considered.

**Length and coverage of blocks derived from  $D'$ ,  $r^2$ , and  $\varepsilon$**

<i>Method</i>	<i>Blocks</i>	<i>SNPs per block</i>			<i>Physical block length [kb]</i>		
		<i>Mean</i>	<i>Med.</i>	<i>Coverage</i>	<i>Mean</i>	<i>Med.</i>	<i>Coverage</i>
min $ D'  = 1.0$	1078	2.63	2	72.7%	44.4	24.2	35.9%
3+	421	3.65	3	42.0%	69.7	43.1	22.0%
min $ D'  \geq 0.8$	1133	2.98	2	85.0%	55.3	31.5	47.0%
3+	543	4.06	4	59.3%	80.5	52.3	32.8%
min $ D'  \geq 0.2$	936	4.21	4	98.2%	115.8	73.0	81.3%
3+	703	4.96	4	90.2%	137.0	90.9	72.2%
min $r^2 \geq 0.5$	563	2.42	2	37.7%	27.2	14.6	11.5%
3+	151	3.61	3	15.3%	53.4	35.1	6.1%
min $r^2 \geq 0.3$	740	2.59	2	52.2%	40.9	19.3	22.7%
3+	239	3.86	3	25.8%	80.2	41.5	14.4%
min $r^2 \geq 0.1$	563	2.83	2	74.2%	50.8	25.9	38.2%
3+	151	4.09	4	45.8%	83.0	47.4	25.3%
mean $r^2 \geq 0.5$	486	3.04	2	41.8%	54.3	18.9	19.8%
3+	190	4.69	4	25.3%	113.3	54.0	16.2%
mean $r^2 \geq 0.3$	510	4.38	3	63.6%	108.1	46.1	41.4%
3+	327	5.74	4	53.6%	152.5	81.1	37.4%
mean $r^2 \geq 0.1$	240	13.23	9	92.3%	447.8	277.1	80.6%
3+	203	15.29	11	90.6%	525.0	346.9	80.0%
$\varepsilon_4 \geq 0.6$	55	4.78	4	7.2%	71.2	47.4	2.9%
$\varepsilon_4 \geq 0.5$	146	5.23	5	20.9%	85.9	62.8	9.3%
$\varepsilon_4 \geq 0.4$	272	5.81	5	41.7%	144.2	79.2	25.8%
$\varepsilon_4 \geq 0.3$	378	6.61	6	64.4%	171.6	116.5	46.4%
$\varepsilon_4 \geq 0.2$	393	8.36	7	83.9%	249.9	180.5	69.5%

Table 4.5: Mean & median (Med.) of the number of SNPs per block and the physical block length [kb], supplemented by the chromosomal block coverage (Cov.). Blocks were defined by pairwise methods utilizing min  $|D'|$ , min  $r^2$ , and mean  $r^2$  and by the  $\varepsilon_4$ -based method with various thresholds. “3+” denotes the additional requirement of at least 3 SNPs per block.

approximately doubles. Thus, the majority of blocks that are defined by the pairwise methods are actually just pairs.

Pairwise and  $\varepsilon$ -based methods differ in the number of blocks and their lengths. The question then arises if the longer blocks that are defined by  $\varepsilon$  are simply mergers of blocks that are delivered by the pairwise methods,

### Concordance of SNPs in blocks derived from $D'$ , $r^2$ , and $\varepsilon$

$pair \rightarrow \varepsilon/\varepsilon \rightarrow pair$ [%]		$\varepsilon_4 \geq$				
<i>Method</i>	<i>SNPs</i>	0.2	0.3	0.4	0.5	0.6
		2991	2296	1489	744	257
min $ D'  = 1.0$	2592	88.3/76.6	72.0/81.3	49.6/86.3	26.1/91.0	9.7/97.7
3+	1623	91.7/49.7	80.5/56.9	61.7/67.3	35.4/77.2	14.3/90.3
min $ D'  \geq 0.8$	3033	87.2/88.4	69.3/91.5	46.7/95.2	24.1/98.4	8.5/100.0
3+	2218	91.9/68.1	77.8/75.1	57.8/86.1	31.8/94.9	11.5/98.8
min $ D'  \geq 0.2$	3502	84.3/98.7	65.0/99.1	42.4/99.7	21.2/99.9	7.3/100.0
3+	3279	85.8/94.1	67.2/95.9	44.4/97.9	22.4/98.9	7.8/99.6
min $r^2 \geq 0.5$	1346	99.0/44.5	89.6/52.5	70.0/63.3	43.7/79.0	17.4/91.1
3+	574	99.8/19.2	98.8/24.7	92.0/35.5	73.2/56.5	39.2/87.5
min $r^2 \geq 0.3$	1861	96.5/60.0	83.9/68.0	62.1/77.6	35.8/89.5	13.4/96.9
3+	972	99.3/32.3	96.3/40.8	83.7/54.7	58.2/76.1	25.4/96.1
min $r^2 \geq 0.1$	2645	90.5/80.0	73.5/84.7	50.9/90.4	27.1/96.2	9.5/98.1
3+	1695	97.2/55.1	87.9/64.9	67.6/77.0	39.5/89.9	14.7/97.3
mean $r^2 \geq 0.5$	1490	99.1/49.3	90.1/58.4	70.3/70.4	42.8/85.8	16.2/94.2
3+	929	99.9/31.0	97.6/39.5	86.7/54.1	61.6/76.9	25.9/93.8
mean $r^2 \geq 0.3$	2270	96.2/73.0	82.2/81.3	58.5/89.1	31.4/95.7	11.2/99.2
3+	1952	98.2/64.1	88.1/74.9	65.2/85.4	36.2/94.9	13.1/99.2
mean $r^2 \geq 0.1$	3294	87.2/96.0	67.8/97.2	44.7/98.8	22.4/99.3	7.8/100.0
3+	3233	88.2/95.3	68.6/96.6	45.4/98.5	22.8/99.1	7.9/99.6

Table 4.6: Concordance of SNP inclusion in blocks by different methods. The percentage of SNPs included in blocks by one method that were also included by the other method (pairwise  $\rightarrow \varepsilon/\varepsilon \rightarrow$  pairwise). “3+” denotes the additional requirement of at least 3 SNPs per blocks. “*SNPs*” states the total number of SNPs included by a particular method and threshold.

or if these blocks do not coincide. To this end, the proportions of SNPs included in blocks by a pairwise method that were also included by the  $\varepsilon$ -based method and vice versa were calculated. The results for window size 4 are listed in table 4.6. Results for size 5 were similar.

Most SNPs included by  $\varepsilon$  are also found by  $D'$ , but less than 50% the other way around. This is not surprising, given the high proportion of SNPs that are in complete LD with at least one adjoining SNP ( $> 70\%$ ). 60–70% percent of the SNPs in  $r^2$ -defined blocks for thresholds  $> 0.3$  are also found

### Baylor College, Chromosome 12, SNP allele distribution

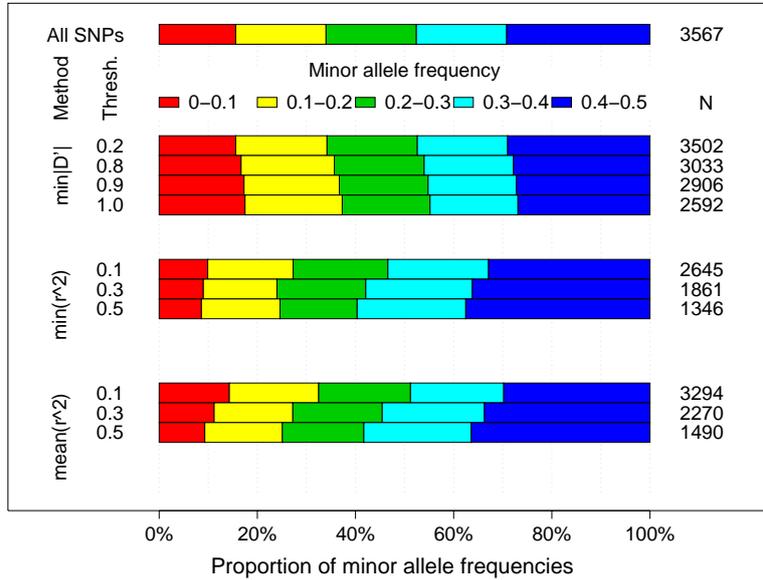


Figure 4.8: **Minor allele frequency distribution for all SNPs in the sample and for SNPs included in blocks derived from  $|D'|$  and  $r^2$ .** Columns on the left specify the pairwise LD measure and the threshold used for the block definition, whereas the number of SNPs included in those blocks is given on the right. Blocks of higher  $r^2$  show a clear tendency to enrich very common SNPs.  $|D'|$ -based blocks show an opposite tendency to include rarer SNPs.

by  $\varepsilon_4 \geq 0.4$ . On the reverse, for medium window sizes of  $\varepsilon$ , most of the block SNPs belong to regions with elevated values of  $r^2$ . If again three SNPs per block are required, these figures change.  $|D'| = 1.0$  and  $\varepsilon_4 \geq 0.4$  only agree on about two thirds of the SNPs, despite approximately similar SNP coverage rates (cf. table 4.5). For higher  $\varepsilon$  thresholds,  $D'$  includes a higher proportion of SNPs. Most of the SNPs included by both  $r^2$  methods with threshold 0.5 are also included by  $\varepsilon_4 \geq 0.4$ .

Figure 4.8 illustrates the SNP minor allele frequency distributions in blocks derived from pairwise LD measures.  $r^2$  enriches very common SNPs in blocks, similar to  $\varepsilon$ -derived blocks (see p. 56). The result for  $D'$  is quite the opposite. The allelic proportions remain nearly the same, with a slight tendency to include more rare SNPs.

# Chapter 5

## Discussion

### 5.1 The measure $\varepsilon$

**Assumptions & Features.** Linkage disequilibrium (LD) between loci is defined as the deviation of the haplotype frequencies from their expectation under independence (linkage equilibrium). There are a number of useful pairwise LD measures, but multilocus measures have been traditionally scarce. The proposed measure  $\varepsilon$  utilizes the long-established concept of entropy by considering a number of loci as a system and the haplotypes at these loci as the possible states of the system. Like  $D$ ,  $\varepsilon$  compares the observed state to the expected one under equilibrium. In analogy to the common pairwise measures  $D'$  and  $r^2$ , it normalizes the difference to allow for comparison between different sequences.  $\varepsilon$  can be interpreted as the relative gain in structure of a loci sequence due to LD.

In theory, an unlimited number of loci can be incorporated into  $\varepsilon$ 's LD assessment. Thus,  $\varepsilon$  is a true multilocus measure for LD and overcomes a major drawback of pairwise measures that are blind against multilocus LD. The problem with pairwise measures — that of how to summarize their values for longer marker sequences to approximate multilocus LD — vanishes. However, the number of loci is limited in applications by the necessity to estimate haplotype frequencies in the population, either by direct sequencing or by estimating them from phase-unknown genotypic data. Rare haplotypes

often contain information on past recombination events. Not including this information, whether by small sample sizes that often miss rare haplotypes or by setting lower frequency limits to protect against errors, can lead to inflated estimates of LD, as will be discussed later. Usually study settings will enable  $\varepsilon$  to consider 8–12 loci at the most.

Although the number of loci that can be considered at one time by  $\varepsilon$  is limited, longer-ranging LD can be approximated by the use of sliding windows. Also, the separation of high order LD effects from medium order effects would require impossibly large sample sizes.  $\varepsilon$  provides convenient profiles to describe LD along loci sequences. When very long loci sequences are considered, the increasing number of marker pairs may cause problems in the calculation and combination of pairwise LD measures. The  $\varepsilon$  measure operates sequentially and is, therefore, not affected by an exponentially increasing number of pairs. On the other hand,  $\varepsilon$  is blind against LD between markers that are more distant than the specified window. More sophisticated algorithms that allow for the exclusion of in-between markers are needed for assessing LD in sequences of non-adjacent SNPs. These methods are not as straightforward as the pairwise value matrix approach due to the higher number of loci being considered.

$\varepsilon$  is sensitive to the number of haplotypes and their frequency pattern. Since  $\varepsilon$  targets LD, it can distinguish between different degrees of LD and does not share the indicator-like behavior of  $D'$  for missing haplotypes that are often, but not always, due to missing recombination events.  $\varepsilon$  cannot equal 1, so there is no equivalent to “perfect” or “complete” LD. Common pairwise LD measures only focus on allelic combinations instead of on a sequence of loci. For more than two loci or alleles, the ability of  $D$  (1.2) to simultaneously describe all haplotypic deviations from equilibrium is lost. Higher-order deviation terms [14, 156] are haplotype-specific and need yet to be combined into a single LD expression.  $\varepsilon$  preserves  $D$ 's feature with regard to a sequence loci by incorporating all haplotypic deviations into a single expression for the allelic structure. It is, therefore, sequence-oriented, treating the set of haplotypes as one unit, rather than each haplotype separately.

$\varepsilon$  measures LD, conditional on the marginal frequencies that are observed

in the sample. It does not make model assumptions about population history, selective neutrality of the loci under consideration, and other parameters. Thus,  $\varepsilon$  uses an empiric, non-parametric approach. This approach is in concordance with  $D$ -based pairwise measures, but in contrast to proposed parametric LD measures [98, 113, 8]. Those measures require the modelling of the population history and the genetic region under consideration, the incorporation of recombination rates and other parameters into the model, and the estimation of the model parameters from the sample data and sometimes other sources. It is not yet clear how the violation of one or more of these assumptions, e.g. due to a disease susceptibility locus under selection pressure, will affect LD assessment. Population histories are often not known in detail and are, therefore, difficult to quantify. For example, consider the often unknown proportion of a population that fell victim to famines and other catastrophes at certain points in its history, or the degree of admixture and migration in borderland regions, in immigration societies, or during war-time. A non-parametric measure, although possibly missing information on the particular population history, might turn out to be more robust in its LD assessment.

Population bottlenecks are likely to decrease the variation in a sequence under consideration. The reduced number of haplotypes will lead to an increase in LD. Population admixture presumably increases multilocus LD due to deviations from Hardy-Weinberg equilibrium, but this effect and its implications for  $\varepsilon$  need further investigation, since deviations are not necessarily the same for multiple markers and might have opposing effects. The pairwise measures, and  $D'$  in particular, show considerable variance in their LD assessment.  $\varepsilon$  could show lower variance, due to its joint consideration of not only two, but usually more, loci that have potentially different histories and allele frequencies, and due to its smoothing effect. The potentially lower variance of  $\varepsilon$  needs further investigation in the future.

$\varepsilon$  assumes the sample to be representative of a single population. If this assumption is violated, e.g. due to sample stratification, then  $\varepsilon$  will indicate possibly inflated LD. Appropriate sampling schemes that are designed to protect against spurious LD in real-world applications need to be employed.

Only a fixed number of loci are considered for the calculation of  $\varepsilon$ ; no physical distance information whatsoever is incorporated into this calculation. However, since LD itself is highly variable over physical distances, depending on the region and the population under investigation, such an incorporation is only useful with a parametric approach in which this factor is explicitly modelled.

The multilocus measure proposed by Sabatti & Risch [127] compares single marker homozygosities to the haplotypes' homozygosity. The rationale is to measure similarity between marker alleles on the same haplotype. To formulate an LD measure, they normalize the difference of the homozygosity values to  $[-1, 1]$  by assessing the extreme values of the difference in a computationally demanding approach. This measure is similar to  $D'$ ,  $r^2$ , and  $\varepsilon$  in that it normalizes a difference between an observed and an expected value under independence, while assuming the marginal frequencies to be fixed. It also yields a single expression for LD. However, the measure is difficult to interpret and quickly becomes infeasible to calculate with increasing numbers of loci. Haplotype diversity criteria from various proposed block detection algorithms provide only an indirect assessment of multilocus LD.

As has been shown in section 2.3, the entropy difference  $\Delta S$  provides an approximate  $\chi^2$  test for a significant deviation of the haplotype frequencies from equilibrium. Thus, the pair  $(\varepsilon, \Delta S)$  allows for the simultaneous assessment of strength and significance of LD.  $\varepsilon$  and  $r^2$  share a number of similarities, namely their ability to distinguish different degrees of LD, the simultaneous assessment of strength and significance of LD, their ties to likelihood theory, the approximate coincidence of  $\Delta S$  and  $r^2$  for two bi-allelic loci with intermediate allele frequencies, and the high correlation between their mean values in  $\varepsilon$ -defined blocks. Therefore,  $\varepsilon$  can be considered a generalization of  $r^2$  to haplotypes of more than two bi-allelic loci. This relationship is even more advantageous, as the value of  $\Delta S$  indirectly relates to the increase of the required sample size to achieve a certain power in disease association mapping [113], as does  $r^2$  for the two-loci case [146, 8]. For extreme marginal allele frequencies,  $\varepsilon$  diverges from  $r^2$ . Also, the correlation becomes weaker for longer sequences.  $\varepsilon$  summarizes the information from the pairwise  $r^2$  to a

large extent, while additionally adding multilocus LD information. An extension of  $\varepsilon$  to multi-allelic markers that might be useful in future applications is straightforward.

**Common haplotypes.** The simulation of data sets allows for the control of hidden, unobserved factors in the data. Methods of analysis can be evaluated under a wide range of conditions by varying the parameters, e.g. the sample size, in the data sets. The results can give insight into a method’s behavior in real-world applications. However, simulated data sets are almost always too pure in comparison to real data sets due to the limited set of assumptions and the lack of hidden factors. Therefore, the interpretation of results from simulation studies requires caution.

The simulation studies from section 3.1, where only common haplotypes are employed in the estimation of  $\varepsilon$  and where their frequencies are re-scaled so that they sum to 1 ( $\varepsilon_{cmn}$ ), demonstrated a persistent bias in the estimates of  $\varepsilon$ , when compared to  $\varepsilon_{all}$ , where all haplotypes were used. The positive bias that was found in most cases was predictable. For most replications in the simulations, the bias of  $\varepsilon_{cmn} - \varepsilon_{all}$  was confined to 0–0.15 (few rare haplotypes) or 0.05–0.2 (many rare haplotypes) when the common haplotypes provided 90% frequency coverage. So, while all haplotype frequencies should be used for the calculation if possible,  $\varepsilon$  could still be used with some reliability, when only common haplotypes that provide 90% coverage are known. Rare haplotypes contain information on past recombination events that lowered LD in the specified region. Not including this information in the estimation will usually result in inflated estimates of LD.

Rather surprising is the observation that  $\varepsilon_{cmn}$  is sometimes *negatively* biased, i.e. it estimates LD to be lower than  $\varepsilon_{all}$ . This was the case for a small proportion of the replications. Closer inspection of these replications revealed that the bias of  $\varepsilon_{cmn}$  is caused by the interplay of the number of rare haplotypes and the number of SNPs that become mono-allelic if only common haplotypes are considered. A mono-allelic SNP lowers the effective number of SNPs considered for the calculation of  $\varepsilon$  by 1 and, thus, decreases the value of  $\varepsilon_{cmn}$ . SNPs with rare alleles have a similar effect. Mono-allelic

SNPs are more likely to occur with fewer common haplotypes. This explains the decreasing variance of  $\varepsilon_{cmn} - \varepsilon_{all}$  with a growing number of common haplotypes.

The data simulations allowed no further assumptions about the haplotypes in those blocks; each possible haplotype had the same probability of being included in the pool from which the data sets were sampled. Thus, haplotypes were unrelated. This can result in artificially low values of LD in the blocks, since haplotypes evolve from one another if no population migration or admixture are considered. In real-world populations, rare haplotypes often differ from the other haplotypes only by a single rare SNP allele, whereas the simulated data sets tend to have more diverse haplotypes. In real-world data sets, the number of SNPs that are mono-allelic in common haplotypes can become larger. The window-size reducing effect in the calculation of  $\varepsilon$  is then strengthened. Thus, this study considered a worst-case scenario. The bias of  $\varepsilon_{cmn}$  can be predicted to be smaller in real data sets.

Data simulation programs utilizing the coalescent model [63, 110] were also considered. This model is based on the insight that, under a selectively neutral model, the process of neutral mutations can be separated from the genealogical process to explain an observed sequence [101, 125]. More parameters, such as bottlenecks, population growth, and recombination, can be incorporated into the model. Unfortunately, tests with a larger number of parameter settings often yielded unreasonably high values of LD when compared to those that were found in real-world data sets, with complete LD between almost all markers. Also, simulating haplotype blocks with a fixed number of SNPs that are embedded in a sequence of additional SNPs that are in linkage equilibrium was not possible with the programs, nor was a priori specification of the number of haplotypes. Therefore, the coalescent model was determined to be inappropriate for this application and was not further considered.

**LD assessment.** The application of  $\varepsilon$  to both simulated and real-world data sets demonstrated that the measure reasonably describes LD and block structures. Under ideal conditions, where a sequence of loci in LD was sur-

rounded by additional loci in linkage equilibrium,  $\varepsilon$  assumed its maximum value for the window size matching the block with regard to length and location for most replications. Larger windows showed a smoothing effect, i.e. an only partial drop of  $\varepsilon$ , when the window included loci in linkage equilibrium. The effect became stronger for larger windows.  $\varepsilon$  used with smaller window sizes usually separated adjacent blocks with no SNPs between them. Larger windows detected LD but could not separate between the different blocks due to the smoothing. This makes larger window sizes of 8 SNPs or above impractical for use in sliding windows, although they are still useful for the assessment of LD in those sequences. Over longer blocks, where  $\varepsilon$  could not be applied due to sample size limitations, it persistently indicated high LD values due to the nearly constant haplotype frequencies in the smaller windows. Although not describing LD in the block directly,  $\varepsilon$  approximated it with smaller window sizes. Feasible sample sizes do not allow differentiation of LD effects of high order from those of lower order, so almost all the information on LD is captured by the smaller windows.

Finally,  $\varepsilon$  re-detected several blocks of an established block structure in a previously described data set [30]. These blocks were characterized by sharp borders. Some other block borders were not supported due to reduced values of  $\varepsilon$ . The region under consideration showed elevated to high values of LD and an embedded sub-region with even higher values. This substructure was described by  $\varepsilon$ , but not by the  $D'$ -based method [30]. The method used by Daly et al. [30] aimed at recombination events and therefore used  $D'$  in the block definition. The differing block patterns between both measures can be explained by their differing objectives.

## 5.2 The $\varepsilon$ -based block definition algorithm

The existing block definition algorithms form two groups that pursue different objectives. Chromosomal coverage (CC) methods seek to describe a maximum amount of the genetic variation by a minimum number of haplotypes. Therefore, they result in a partitioning of the region, the chromosome, or the genome under investigation into blocks. They do not, however, look

for regions of elevated LD; this is a mere side effect of the approach. Methods that pursue the latter objective rather differentiate between regions of high and low LD (island approach). Regions of high LD are marked as blocks, whereas other regions remain unmarked.

The  $\varepsilon$  measure directly describes LD and is sensitive to both, the number of present haplotypes and their frequency pattern. Thus, it combines basic ideas behind both existing block definition approaches, but specifically targets LD. The  $\varepsilon$ -based algorithm proposed in section 2.4 follows the island approach: in analogy to existing  $D'$ -based methods, it defines blocks as contiguous regions of increased LD, separated by regions of decreased LD. The  $\varepsilon$ -based algorithm delivers regions of higher structure with regard to LD. It additionally allows for description of embedded substructures, as seen in its application to the data set from Daly et al. [30], where the SNP spacing in a region was too dense to be separated into clear disjointed blocks.

The  $\varepsilon$ -based approach is broader than the CC approach; it describes a fundamental feature of a region or a chromosome that can be utilized for more than one purpose. Regions of low LD will either have rare SNPs or many rare haplotypes. CC methods will result in very short, often single-SNP blocks in those regions. The definition of haplotype-tagging SNPs (htSNPs) within these blocks is futile. CC methods also require the estimation of potentially very long haplotypes. But the sample size limits the length of haplotypes, whose frequencies can reliably be estimated, and therefore also the maximum number of SNPs in a block that CC methods can detect. In contrast,  $\varepsilon$ -defined blocks require a minimum number of SNPs per block due to the size of the window used, but can contain a theoretically unlimited number of SNPs. For sparser SNP maps, the minimum number of SNPs in a block is a limitation, since LD between two SNPs will often pass undetected for a window size of 4, even if they are relatively distant. For higher resolutions, this problem will vanish.

For sparse densities of SNPs covering the genomic region under investigation, a clear block-like pattern, with sharp borders that are caused by steep drops of  $\varepsilon$ , is likely to occur. This pattern will vanish when the SNP map becomes denser, and will give way to a more gradual LD decay, even close

to recombination or gene conversion hot spots. Very strict block algorithms that exclude any recombination [154] will result in very small blocks. The optimal criterion for the block definition is a trade-off between short, but rather “pure” blocks — i.e. with few haplotypes and little genomic coverage — and longer, but rather “impure” blocks — i.e. with smaller degrees of LD and more haplotypes, but also higher coverage. The choice of window size and threshold in the proposed algorithm and the interpretation of the resulting blocks will depend on the objective of the analysis.

The  $\varepsilon$ -based algorithm has two major drawbacks. Due to the inevitable smoothing effect of window sizes larger than 2, the  $\varepsilon$ -based algorithm yields longer blocks on average and often includes SNPs at the block edges that are only in weak LD with SNPs in the block core, particularly for low thresholds. Also, the use of longer windows can result in overlapping blocks. To make the block assignment of each SNP unique, the algorithm can be refined by implementing a *max-cut* feature. For overlapping blocks, the block with the highest mean value of  $\varepsilon$  will be considered most important and accepted. The SNPs from the block are removed from the SNP sequence, and the sequence is divided into two independent subsequences. The algorithm iteratively defines blocks of highest  $\varepsilon$  values in the remaining sub-sequences, removes their SNPs, and creates more subsequences, until there are no more overlapping blocks. Pairwise measures, either  $r^2$  or  $\varepsilon_2$ , can be considered to remove SNPs from the block edges, for example if all pairwise values with such a SNP drop below a certain threshold. This approach could, however, also discard existing multilocus LD information and should be investigated further.

### 5.3 Blocks on human chromosome 12

The  $\varepsilon$ -based algorithm has been applied to a whole chromosome data set. After filtering, the analyzed chromosome 12 data set provided an average density of 37.0 kb between adjacent SNPs. This density is high for a whole chromosome, but rather sparse when compared to recent studies that focussed on particular regions and provided resolutions from 1 kb through 15 kb [8, 21]. However, a large number of blocks could be defined, and LD

along the chromosome could be described in impressive detail. This is presumably due to the non-uniform, clustered distribution of the SNPs. An average SNP density of about 40 or 50 kb is a lower limit for reasonable block detection. A considerable loss in detected blocks occurred when only 80% of randomly chosen SNPs from the original set were used in the block definition; the loss was dramatic for 50% or less (data not shown).

A high number of combinations of window sizes and threshold values has been investigated. The data set analysis demonstrated that  $\varepsilon$  is very variable, when used with small window sizes of 2 or 3 SNPs. Block definition based on these sizes resulted in fragmentation into small, predominantly pairwise blocks, even in regions of high LD. On the other hand, window sizes of 7 or greater suffered from over-smoothing. Only high thresholds then delivered reasonable blocks of elevated LD, but missed many smaller regions of high LD or regions with intermediate levels of LD in consequence. Moderate window sizes of 4–6 and thresholds in the range of 0.4–0.6 yielded the best results with regard to several block characteristics. These parameter values represent a good compromise between multilocus LD utilization in the block definition and protection against too much variability and over-smoothing. Blocks that were defined in the suggested range for window size and threshold, i.e. 4–6 and 0.4–0.6, contained on average 3 or less haplotypes with frequencies above 0.1 and 5 or less above 0.05 that provided 80% or more frequency coverage. This finding allows for two conclusions: First, a large proportion of these blocks would be embedded in or coincide with blocks detected by CC methods. Since CC methods do not take LD into account, the opposite is not necessarily the case. Second, the low numbers of common haplotypes providing a high coverage make  $\varepsilon$ -defined blocks useful for both, the utilization of their haplotypes as multi-allelic markers and the definition of htSNPs. Thus, the  $\varepsilon$ -defined blocks are useful for two major objectives in the definition of blocks. For highest thresholds, e.g. 0.7 for window size 4, there are usually two haplotypes that explain nearly all variation at the block.

The block lengths from the  $\varepsilon$ -based algorithm fit well into the picture found in other studies [32, 109]. For window size 2, the same pattern of a

highly skewed distribution of the block length, as in previous studies using pairwise LD measures [46, 132, 109], was observed. However, the coverage was lower for larger window sizes. This is due to the minimum SNP number in blocks, conditional on the window size. Thus, the higher coverage in other studies would be due to a large proportion of pairwise SNP blocks. The application of  $D'$ - and  $r^2$ -based methods confirms this hypothesis, as will be discussed later. The  $\varepsilon$ -based algorithm finds fewer but longer blocks than methods based on pairwise measures.

For increasing LD thresholds, the average block length became much shorter and the chromosomal coverage decreased dramatically. This reflects the general decay of LD for increasing distances. Further analysis of the block length distribution demonstrated that it is only marginally influenced by the strength of LD within threshold-defined groups, but almost completely by the marker distance distribution. The SNP spacing in a study then basically determines the block lengths to be found within these groups. High LD will be found over short and long distances, but to a lesser extent for the latter.

Blocks that were defined in the suggested range for window size and threshold, i.e. 4–6 and 0.4–0.6, contained on average 3 or less haplotypes with frequencies above 0.1 and 5 or less above 0.05 that provided 80% or more frequency coverage. This finding allows for two conclusions: First, a large proportion of these blocks would be embedded in or coincide with blocks detected by CC methods. Since CC methods do not take LD into account, the opposite is not necessarily the case. Second, the low numbers of common haplotypes providing a high coverage make  $\varepsilon$ -defined blocks useful for both, the utilization of their haplotypes as multi-allelic markers and the definition of htSNPs. Thus, the  $\varepsilon$ -defined blocks are useful for two major objectives in the definition of blocks. For highest thresholds, e.g. 0.7 for window size 4, there are usually two haplotypes that explain nearly all variation at the block.

SNPs with rare alleles show a similar effect to mono-allelic SNPs that lower the effective window size and thereby the value of  $\varepsilon$ . In reverse, blocks in high LD can be expected to show a large proportion of common SNPs with high-frequency minor alleles. The findings in section 4.5 confirm this

hypothesis. Stronger LD, as measured by  $\varepsilon$ , leads to an enrichment of very common SNPs, whereas the proportion of rare SNPs declines. This tendency is not surprising. The occurrence of only a few common haplotypes in  $\varepsilon$ -defined blocks is more unlikely and the deviation from linkage equilibrium is, therefore, stronger for common SNPs than for rare SNPs.

Strong correlations were found between the mean values of  $\varepsilon$  and  $r^2$  in the block groups defined by window size and threshold. This correlation became weaker for larger window sizes and stronger for higher thresholds. The correlation between the mean values of  $\varepsilon$  and  $|D'|$  was generally found to be weak. These observations allow for two conclusions: First, the weak correlation between  $|D'|$  and  $\varepsilon$  can be explained by the different objectives the measures are suitable for. While  $\varepsilon$  measures LD,  $D'$  basically indicates missing haplotypes, which might be due to absent recombination events in the past. Only the highest thresholds for  $\varepsilon$  will inevitably deliver regions with little recombination. In those cases, e.g. window size 5 and threshold 0.7, correlation between  $\varepsilon$  and  $|D'|$  is high. Second, the correlation between  $\varepsilon$  and  $r^2$  supports the suggestion of  $\varepsilon$  as a multilocus generalization of  $r^2$ . It also indicates that  $\varepsilon$  captures the information contained in a matrix of pairwise  $r^2$  values and conveniently summarizes it. This relationship is less clear for the block-wise median  $r^2$  values. A possible explanation is that LD will usually be strong in the block core and decay towards the edges. The number of pairs in which at least one SNP is located at a block edge is much larger than the number of pairs in which both SNPs belong to the block core. Thus, the median value of LD will be lower than the mean, and correlation becomes weaker.

Section 4.7 compared the  $\varepsilon$ -based algorithm with methods based on the pairwise LD measures  $D'$  and  $r^2$ . All three methods pursue the definition of blocks, but with regard to different objectives. The blocks detected by these methods partially, but not completely, overlap. This is due to the incomplete correlation between LD and recombination, since other factors, such as population bottlenecks, influence LD as well.  $D'$ -based algorithms basically segment the chromosome into many small blocks and include more

SNPs in them, but provide only a slightly higher physical coverage than the other methods. Some 50% of the blocks contained exactly two SNPs. Blocks also showed an enrichment of rare SNPs. This phenomenon is due to the non-uniform SNP distribution and to the dense SNP spacing. Each SNP has a probability of about 0.7 that at least one haplotype with an adjacent SNP is missing. This can be due to absent recombination between the densely spaced markers in the past, but also to population bottlenecks in the past – or the sample was too small to include the haplotypes. The usefulness of those pairwise blocks is questionable, since haplotypes in these blocks are not beneficial, nor is the definition of htSNPs in this case.

If at least three SNPs are required per block, the number of blocks, their coverage, and the number of SNPs included in blocks drops by about 50%. Depending on the thresholds used,  $D'$ - and  $\varepsilon$ -defined blocks provide similar coverage, but  $D'$  results in more and smaller blocks, with regard to both the number of included SNPs and the physical block length. Despite the same coverage, the methods include partially different sets of SNPs in blocks and agree on them for only about 2/3 for  $|D'| = 1.0$  and  $\varepsilon_4 \geq 0.4$ . This demonstrates the different objectives of the algorithms and of the measures on which they are based, i.e. recombination vs. LD. Despite many recombination events in the past, regions can be strong in LD due to other factors, e.g. population bottlenecks, selection, and genetic drift.

Strategies employing the mean and the minimum of the pairwise  $r^2$  values yielded blocks in regions similar to  $\varepsilon$ . Again, there was a considerable proportion of pairwise blocks. The requirement of three or more SNPs per blocks greatly reduced the number of blocks for the minimum strategy, but to a much lesser extent for the mean strategy. If blocks of three or more SNPs and higher LD thresholds were considered, i.e.  $\min r^2 \geq 0.3$  or  $\text{mean } r^2 \geq 0.5$ , then most of their SNPs were contained in  $\varepsilon_4$ -defined blocks with a threshold of 0.4. Comparing these two methods,  $r^2$  resulted in twice as many blocks as  $\varepsilon$  did, but with an equal number of included SNPs. The concordance of included SNPs between both methods is 2/3, but over 85% of the SNPs that are included in  $r^2$ -derived blocks with at least three SNP are also included in  $\varepsilon$ -derived blocks. Therefore,  $\varepsilon$  summarizes the mean values of  $r^2$

for longer blocks.  $r^2$  also detected some blocks consisting of two or three SNPs that cannot be detected by  $\varepsilon_4$ , but missed some blocks with strong multilocus, but only mild two-locus interaction. The concordance between the algorithms again supports the notion of  $\varepsilon$  as a generalization of  $r^2$ .

**Conclusions.** The analysis of the chromosome 12 data set strongly suggests that there are only a few clear distinct haplotype blocks. The detected blocks depend on the measure, the method, and the control parameter values that are used in their definition. The phrase *block* creates the illusion that a chromosome can be segmented into clear disjointed blocks, which is not true. Phrases such as *regions of elevated LD* should be used instead to avoid such misunderstandings. Different aims of analysis will require that different parameter values be chosen and will result in different “blocks”. LD is influenced by many factors, such as the age of mutations, population history and admixture, genetic drift, and selection. Thus, there is presumably no single explanation for the occurrence of these distinguished regions.

The chromosome 12 data set was sampled from North American families of European descent [29]. Since LD patterns depend on the history of the population under consideration, the findings of this analysis do not necessarily apply to populations from other geographical regions. The commonly accepted hypothesis on the origin of human populations describes the colonization of the earth by humans basically as a series of repeated migrations, bottleneck events, and admixture events that started from a small population in north-eastern Africa, first within and then out of Africa, and that later continued in other parts of the world (out-of-Africa hypothesis [141]), although this hypothesis is still subject to some debate [159]. The migrating populations went through bottlenecks that reduced the genetic diversity and eventually led to higher LD values on average. Haplotype blocks in genomes from many African populations such as Nigerians, who are supposed to have had few bottlenecks in their history [74], can, thus, be expected to be shorter in general, as is the extent of LD in those populations [132]. Asian populations might show block lengths similar to Europeans, but a proportion of these blocks might be located at different positions. The approach of

the HapMap project [29] which samples all the major human populations is therefore justified to describe the variation in LD and the haplotype patterns in the whole human population. Coinciding block patterns between different populations could point to hidden factors, such as selection, as a reason for this coincidence and assist in the fine-mapping of diseases or in the search of preserved genomic regions.

## 5.4 Implications for medical research and other potential applications

Two major objectives in the definition of haplotype blocks are the potential use of the block haplotypes as multi-allelic markers with increased heterozygosity in disease association studies and the capture of a maximum amount of genetic variation by a minimum number of markers to considerably lower genotyping costs in medical studies. The basis for both objectives is the observation that markers are often not independent, but show allelic associations (LD) in the populations. Existing block definition methods do not describe LD directly.  $D'$ -based methods merely look for recombination events. Absence of recombination will often yield strong LD, but enough LD often remains in the region to be exploited for gene mapping, even after some recombination events have occurred, due to other sources of LD. CC methods aim to describe longer sections of DNA by a limited number of haplotypes. Although this can be useful for the definition of htSNPs, those regions do not necessarily have to be strong in LD if the SNPs are rare, thereby decreasing the statistical power for gene mapping. The  $\varepsilon$ -based method is distinct from both,  $D'$ -based methods and chromosomal coverage methods, as it directly looks for strong LD. However, since all methods describe different aspects of the same phenomenon, they will get to similar blocks if these aspects coincide, as was demonstrated for the chromosome 12 data set.

The proposed  $\varepsilon$ -based algorithm is easy-to-use and yields regions of elevated LD (“LD blocks”) with a very limited number of common haplotypes that provide high frequency coverage and an increasing proportion of common

SNPs with growing LD. Thus, these blocks are suited for two main applications of haplotypes in statistical gene mapping, i.e. as multi-allelic markers and for the definition of htSNPs. Haplotypes cannot be more frequent than their rarest SNP allele. Since haplotypes are only useful for disease mapping if they occur with notable frequency in samples, it is often recommended to exclusively use common SNPs for haplotype construction. The enrichment of common SNPs makes filtering the data set for rare SNPs ahead of the analysis unnecessary.

The question arises, under which circumstances  $\varepsilon$ -defined regions of elevated LD and their haplotypes are actually useful in the mapping of disease genes. The success of association studies in mapping a trait-influencing gene depends critically on the population history and the age and the frequency of the causative mutation and the marker mutations. If the causative mutation is younger than the surrounding LD structure, it has occurred on a single haplotype, and only few recombination events have disturbed this correlation. The strength of this correlation depends on the haplotype frequency, when the mutation has occurred: the strength grows with lower frequency. The mutation will then be in very strong LD with all markers in the region, and the haplotypes from the LD block will serve as very useful markers. If, however, the mutation is older than the LD structure, then the mutation is likely not to be unique to a single haplotype, but to occur on several.

The Common Diseases/Common Variants (CD/CV) hypothesis [22] assumes the disease-causing mutations to be common and rather old. Due to the age of these mutations, LD around them can be expected to have substantially decreased. For these mutations to be found in regions of elevated LD and to be present in only one or two haplotypes that are detectable by haplotype analysis, very dense marker maps need to be used or other forces than mutation and recombination need to have acted, e.g. population bottlenecks and genetic drift. The out-of-Africa hypothesis on the origin of modern humans assumes bottlenecks to have repeatedly occurred due to migration and, therefore, supports the suggested usefulness of haplotypes. The common haplotypes in  $\varepsilon$ -defined blocks are good candidates to be used in haplotype-based association studies, if the CD/CV hypothesis holds. Com-

mon haplotypes will also be useful, if a mutation recently occurred in an existing LD structure and the mutation-carrying haplotype's frequency was amplified by the repeated migration and bottlenecks events. If, however, the allelic spectrum of common diseases turns out to predominantly consist of many rare mutations, as was predicted in a simulation study [111], and if factors such as bottlenecks and migrations would have only a minor impact on LD during the human evolution, then the use of common haplotypes is futile. Both predictions have their point; presumably none of them will account for all common diseases. Thus, the use of haplotypes in the statistical mapping of complex disease genes will often, but not necessarily, increase statistical power. Haplotypes can only attenuate the problems that are due to often complex relations between particular genes and a disease.

Many genomewide association studies for various diseases are now planned or underway. In the initial phase of these studies, the number of required SNPs for detecting regions of high LD and for defining haplotypes needs to be assessed. The human genome contains about  $3 \cdot 10^9$  base pairs. The use of 100,000 SNPs, which will be available on a single Affymetrix GeneChip by 2004<sup>1</sup>, would result in an average distance (resolution) between two adjacent SNPs of  $\frac{3 \cdot 10^9}{1 \cdot 10^5} = 30$  kb. However, when using chromosome 12 as a model chromosome, a considerable portion of SNPs will be lost to filtering due to mono-allelism and errors in the sample (20% and  $> 50\%$ , respectively, for the chromosome 12 data set), thereby lowering the resolution to 36–60 kb or more. The median block size was found to be  $\approx 80$  kb for  $\varepsilon_4 \geq 0.4$  and less ( $\approx 30$  kb) for  $D'$ -based algorithms. Therefore, a considerable portion of “interesting” regions could be missed. Studies that utilize 200,000–300,000 SNPs are reasonable for genome-wide haplotype block detection and association studies. This is consistent with the prediction of 500,000 required SNPs by Kruglyak [81] and others [71], albeit more optimistic.

The next question is then, where these SNPs should be located on the genome. Genomic regions of high LD show little variation in the number of haplotypes and, thus, fewer SNPs are needed to capture the diversity of this

---

<sup>1</sup>see <http://www.affymetrix.com/> for details

region. On the other hand, regions in a state close to linkage equilibrium will require every possible SNP to be included in the study.  $\varepsilon$  LD profiles could provide a convenient means for the spacing of SNPs in whole-genom scans.

**Other possible applications.** The  $\varepsilon$  measure's potential is not limited to association studies. LD is an interesting feature of the genome in itself. More biological phenomena, e.g. regulatory regions [80] or the organization of chromosomes, might to be found to match the phenomenon of LD. Regions of extremely high or low LD might be subject to selection pressure and could resemble functionally relevant chromosomal regions. A comparative study of populations or species could reveal preserved regions, within as well as outside of coding regions. LD might turn out to be helpful in the prediction of these phenomena. For this kind of application, a multi-allelic marker extension of  $\varepsilon$  that allows for the use of markers other than SNPs is required.

Differences and similarities with regard to LD structure between populations or species could be utilized to infer their history and divergence. For example, populations that diverged more recently from each other can be expected to show more similar LD patterns. Thus, their patterns can be expected to cluster according to the time since the bifurcation event when using an appropriate distance measure and assuming a tree-like evolution of populations. This is certainly an area of future investigation that can complement other approaches, such as sequence comparisons.

Phenotypic variation can only be explained, apart from environmental factors and chance, by genetic variation. LD is an important aspect of this variation and might correlate with other aspects of the genome. So even in completely sequenced genomes, an  $\varepsilon$  LD profile might summarize or condense important aspects of the genetic variation and assist in the analysis of these genomes.

# Chapter 6

## Summary

Common diseases in humans usually resemble complex traits, where the clear relationship between genetic causes and phenotypic expression, that can be found in Mendelian traits, is disturbed by numerous factors. Classical methods for statistical gene mapping that succeed for Mendelian traits lose power when applied to common diseases, and often deliver non-significant or non-reproducible results. Recent observations indicate a structure of the human genome, where regions of elevated linkage disequilibrium (LD) are interspersed by regions of low LD, creating haplotype “blocks”. Haplotypes of these blocks can be used to improve the power of statistical mapping methods and to reduce work efforts, thus enabling larger genetic disease studies for an equal amount of funding.

Existing methods either use the pairwise LD measure  $D'$  or haplotype diversity criteria for the definition of blocks. Both kinds of criteria do not describe LD directly.  $D'$  targets missing recombination, not LD. Its usage can result in a loss of information on association, since LD usually decays only gradually and is influenced by other factors, such as selection, population bottlenecks, and genetic drift. Furthermore, pairwise LD measures are blind against multilocus LD interactions. Existing multilocus measures are either haplotype-specific or computationally challenging and have not yet been employed for the definition of haplotype blocks.

The proposed new measure, the Normalized Entropy Difference  $\varepsilon$ , allows

for a direct, multilocus assessment of LD, conditional on the single SNP allele frequencies in the sample. It applies the established concept of entropy as used to describing a probabilistic system to sequences of genetic markers, where haplotypes of these markers represent the system's states. Analogous to classical pairwise measures, such as  $r^2$  and  $D'$ , it normalizes the difference between the observed state and the expected state under linkage equilibrium.  $\varepsilon$  yields an expression on the LD state of a sequence, not a particular haplotype. A related quantity,  $\Delta S$ , can be used for testing the significance of the haplotype frequency deviations from the expectation.  $\varepsilon$  is sensitive to both, the number of present haplotypes and their frequencies, and does not share the indicator-like behavior of  $D'$  for missing haplotypes. The theoretically unlimited number of loci that can be incorporated in  $\varepsilon$  is limited by the sample size in real-world applications.  $\varepsilon$ , like  $D$ -based measures, uses an empiric, non-parametric approach. It does not make model assumptions on population history, selective neutrality of the loci under consideration, or other parameters. Furthermore, it preserves the feature of a single expression, jointly describing all haplotype frequency deviations, from the case of two bi-allelic loci.  $\varepsilon$  has analytical and statistical bounds to the pairwise measure  $r^2$  and can be regarded as a multilocus extension of it.  $\varepsilon$  profiles provide a convenient means to describe LD along loci sequences.

A persistent bias occurs when only common haplotypes are used in the calculation of  $\varepsilon$ . This bias is caused by the interplay of information loss on recombination events by excluding rare haplotypes and the number of SNPs that become mono-allelic in the common haplotypes. A worst-case scenario simulation study found the bias to be confined to 0–0.20 in most cases. When used with small window sizes of 2 or 3 SNPs,  $\varepsilon$  shows great variability, whereas larger window sizes, i.e. 7 or greater, show a strong tendency of smoothing. Sizes 4–6 are a good compromise between multilocus LD assessment and protection against too much variability and over-smoothing. For simulated clear LD block patterns,  $\varepsilon$  almost always has its maximum for the correct, i.e. block-matching, window size and location.  $\varepsilon$  can usually differentiate adjoining blocks and also indicates longer LD structures. It partially confirmed an established block structure in a real-world data set;

differences in the blocks positions can be explained with the differing aims of the measures  $\varepsilon$  and  $D'$ .

The proposed  $\varepsilon$ -based algorithm defines haplotype blocks as regions of contiguous windows, where LD does not drop below a certain threshold. It parallels the island approach of existing,  $D'$ -based block definition methods by allocating SNPs in regions of elevated LD to blocks, while the other SNPs remain unallocated. The algorithm additionally allows the description of embedded substructures by using differing thresholds. This can be important in samples of very densely spaced markers with high background levels of LD. A required minimal number of SNPs per block, the possible inclusion of SNPs at the block edges that are only in low LD with the other SNPs of the block, and the possible overlapping of blocks are the drawbacks of the algorithm. Refinements of the algorithm can resolve the latter two.

The application to a data set of the whole human chromosome 12 allows for an evaluation and characterization of the blocks delivered by the  $\varepsilon$ -based algorithm. The average distance of 37.0 kb between to adjoining SNPs in the filtered sample yields substantial results, although even more complex LD structures can be expected to be found at a higher resolution. Window sizes of 4–6 and thresholds of 0.4–0.6 yield the best results with regard to several block characteristics. The found physical block lengths confirm findings of previous studies, but the blocks provide a lower chromosomal coverage. This is due to the window-size induced minimal number of SNPs in the blocks. Blocks defined in the suggested parameter range contains three or less haplotypes with frequencies above 0.1 on average that provide 80% or higher frequency coverage. Thus, a number of these blocks would have been found by chromosomal coverage methods or would have been embedded in those blocks. Common SNPs are enriched in  $\varepsilon$ -defined blocks. Within the parameter-defined groups,  $r^2$  shows a strong correlation with  $\varepsilon$ , although decreasing with increasing window size. This supports the suggestion of  $\varepsilon$  as a multilocus extension of  $r^2$ . The correlation between  $\varepsilon$  and  $|D'|$  in these groups is weak, except for very high thresholds, and supports the claim of differing objectives of these measures.

The direct comparison of block methods based on the three measure  $\varepsilon$ ,  $r^2$ ,

and  $D'$  yielded two findings: First, some 50% of the blocks defined by pairwise methods contain exactly two SNPs, whereas  $\varepsilon$ -defined blocks include more SNPs. The  $\varepsilon$ -based algorithm finds fewer but longer blocks than pairwise methods do. Second, the sets of SNPs included in blocks by these methods only partially overlap, since  $\varepsilon$  and  $r^2$  target LD, whereas  $D'$  targets absent recombination.

The application of the algorithms to the data set also revealed that the resulting blocks very strongly depend on the chosen parameters, with only a few regions with sharp, block-like borders. The misleading term *block* should, therefore, be avoided in future applications. The threshold of choice is a trade-off between shorter blocks with fewer haplotypes and lower chromosomal coverage or longer blocks with more haplotypes and lower LD, but higher coverage. The choice of the parameters in the algorithm ultimately depends on the aim of analysis. Since the data was sampled from an European-American population, the findings cannot necessarily be generalized to other populations, e.g. populations of African descent.

The proposed LD measure and the block definition algorithm deliver regions in elevated LD and with only a limited number of haplotypes. Thus, they meet two major objectives in the definition of haplotype blocks: the use of haplotypes as multi-allelic markers in association studies and the effective description of genetic variation. These haplotypes can potentially improve the power of gene mapping studies, depending on the age of the disease mutation and the surrounding LD structure and of the population history, including bottlenecks and admixture. Considering chromosome 12 as a model chromosome, genomewide genetic studies of human diseases should consider the use of 200,000 to 300,000 SNPs to effectively exploit the advantages of haplotypes in the analysis.  $\varepsilon$  profiles could be utilized in the effective spacing of the SNPs along the chromosome. The genome's feature of LD has appeal beyond its use in association studies. Non-coding regions in high LD might prove functionally relevant and could be detected by inter-population or inter-species comparisons. Matching biological phenomena of possible relevance to diseases with LD could improve their predictability. The proposed measure  $\varepsilon$  might prove to be a useful tool in these and other investigations.

# Chapter 7

## Deutsche Zusammenfassung

Komplexe Erkrankungen stellen gewöhnlich komplexe Phänotypen dar, bei denen eine klare Beziehung zwischen genetischen Ursachen und phänotypischer Ausprägung, wie sie bei Mendelschen Phänotypen gefunden wird, durch oft zahlreiche Faktoren gestört ist. Statistische Methoden für die Genkartierung, die mit Erfolg auf Mendelsche Traits angewendet wurden, verlieren dadurch statistische Power, wenn sie auf komplexe Erkrankungen angewendet werden. Ihre Ergebnisse sind dann oft nicht signifikant oder nicht reproduzierbar. Neuere Studien scheinen das Vorliegen einer Blockstruktur im menschlichen Genom bezüglich des Kopplungsungleichgewichts (LD) zu belegen. Haplotypen dieser Blöcke können die Power klassischer Methoden erhöhen und den Aufwand genetischer Krankheitsstudien verringern.

Existierende Methoden benutzen das paarweise LD-Maß  $D'$  oder Haplotyp-Diversitätskriterien, um Haplotypblöcke zu definieren. Beide Kriterien beschreiben LD nicht direkt.  $D'$  zeigt im wesentlichen fehlende Haplotypen anstelle von LD an, häufig aufgrund nicht stattgefundener Rekombination. Seine Verwendung kann einen Verlust an Assoziationsinformation nach sich ziehen, da LD gewöhnlich nur graduell abfällt und von weiteren Faktoren, z.B. Selektion und Populationsgeschichte, abhängt. Paarweise LD-Maße sind blind für Multilocus-Interaktionen; existierende Multilocus-Maße sind entweder haplotyp-spezifisch oder sehr aufwendig in der Berechnung und wurden bisher nicht für die Definition von Blöcken eingesetzt.

Das vorgestellte neue Maß, die Normalisierte Entropiedifferenz  $\varepsilon$ , ermöglicht eine direkte Beschreibung von Multilocus-LD für gegebene SNP-Allelfrequenzen in der Stichprobe. Es überträgt das etablierte Konzept der Entropie zur Beschreibung eines probabilistischen Systems auf eine Sequenz von genetischen Markern, bei der die Haplotypen die Zustände des Systems repräsentieren. In Analogie zu klassischen paarweisen Maßen, z.B.  $r^2$  und  $D'$ , normalisiert  $\varepsilon$  die Differenz zwischen beobachtetem Zustand und erwartetem unter Kopplungsgleichgewicht.  $\varepsilon$  liefert einen Ausdruck für den LD-Status einer Sequenz, nicht eines speziellen Haplotypen. Eine verwandte Größe,  $\Delta S$ , ermöglicht einen Test auf signifikante Abweichungen der Haplotypfrequenzen von der Erwartung.  $\varepsilon$  wird durch die Anzahl der vorkommenden Haplotypen und ihre Frequenzen beeinflusst; es teilt nicht das Indikatorverhalten für fehlende Haplotypen mit  $D'$ . Die Stichprobengröße beschränkt die theoretisch unbegrenzte Anzahl von Markern, die mit  $\varepsilon$  beschrieben werden können (Fenstergröße), in praktischen Anwendungen.  $\varepsilon$  nutzt, ähnlich zu  $r^2$  und  $D'$ , einen empirischen, nichtparametrischen Ansatz. Es werden keine Modellannahmen über Populationsgeschichte, Selektion und andere Parameter gemacht.  $\varepsilon$  erhält die Eigenschaft einer beschreibenden Größe für alle Abweichungen vom Fall zweier biallelischer Loci und kann als Multilocus-Erweiterung von  $r^2$  angesehen werden.  $\varepsilon$ -Profile sind eine handliche Methode, um LD entlang von Sequenzen zu beschreiben.

Ein beständiger Bias tritt auf, falls nur häufige Haplotypen für die Schätzung von  $\varepsilon$  verwendet werden. Er entsteht durch das Zusammenspiel von Verlust an Information über Rekombinationsereignisse durch den Ausschluß seltener Haplotypen und von SNPs, die bei ausschließlicher Betrachtung häufiger Haplotypen monoallelisch werden. Simulationen ungünstigster Fälle schätzen den Bias auf 0–0,20.  $\varepsilon$  zeigt große Variabilität mit kleinen Fenstergrößen von 2–3 SNPs und einen starken Glättungseffekt mit 7 oder mehr SNPs. Fenstergrößen zwischen 4 und 6 stellen einen guten Kompromiß zwischen der Abschätzung des Multilocus-LD und dem Schutz gegen zu starke Variabilität und zu starke Glättung dar. Bei simulierten klaren LD-Blockmustern wird  $\varepsilon$  fast immer maximal über dem Block mit einer blockgleichen Fenstergröße; es kann oft benachbarte Blöcke unterscheiden und längere LD-

Strukturen anzeigen. Es bestätigt in Teilen eine publizierte Blockstruktur in einem realen Datensatz; Unterschiede in den Blöcken können durch die unterschiedlichen Ziele von  $\varepsilon$  und  $D'$  erklärt werden.

Der vorgestellte Algorithmus definiert Haplotypblöcke als Regionen von zusammenhängenden Fenstern, in denen  $\varepsilon$  nicht unter einen bestimmten Schwellenwert fällt. Dieser Ansatz gleicht dem Inselansatz existierender,  $D'$ -basierter Methoden, indem er SNPs in Regionen mit erhöhtem LD Blöcken zuordnet, während für die anderen SNPs keine Zuordnung erfolgt. Der Algorithmus kann außerdem eingebettete Substrukturen mit verschiedenen Schwellenwerten beschreiben, etwa im Falle dicht verteilter Marker mit hohem Hintergrund-LD. Nachteile des Algorithmus sind die fensterinduzierte Minimalanzahl von SNPs pro Block, der mögliche Einschluß von SNPs an den Blockrändern in nur geringem LD mit den anderen Block-SNPs, und die mögliche Überlappung von Blöcken. Die beiden letzteren können durch Verfeinerungen des Algorithmus aufgehoben werden.

Die Anwendung des  $\varepsilon$ -basierten Algorithmus auf einen Datensatz des gesamten menschlichen Chromosoms 12 erlaubt die Evaluierung und Charakterisierung der resultierenden Blöcke. Ein mittlerer Abstand von 37,0 kb zwischen benachbarten SNPs liefert substantielle Resultate, wobei noch komplexere LD-Strukturen für höhere SNP-Dichten erwartet werden können. Fenster von 4–6 SNPs und Schwellenwerte zwischen 0,4–0,6 liefern die besten Resultate bezüglich verschiedener Blockcharakteristika. Die gefundenen physikalischen Blocklängen bestätigen frühere Studien; die chromosomale Abdeckung (Coverage) ist allerdings geringer. Dies ist eine Folge der Minimalanzahl von SNPs pro Block. Blöcke, die mit den empfohlenen Parametern definiert werden, enthalten durchschnittlich drei oder weniger Haplotypen mit Frequenzen über 0,1 und über 80% Frequenzabdeckung. Viele dieser Blöcke würden so von Chromosomal-Coverage-Methoden gefunden oder in deren Blöcke eingebettet sein. SNPs mit zwei häufigen Allelen sind überdurchschnittlich häufig in diesen Blöcke enthalten. Innerhalb der parameterdefinierten Gruppen zeigt  $r^2$  eine starke Korrelation mit  $\varepsilon$ , die sich zwar mit zunehmender Fenstergröße abschwächt, jedoch die Notation von  $\varepsilon$  als einer Multilocus-Erweiterung von  $r^2$  unterstützt.  $D'$  zeigt eine nur schwache Kor-

relation mit  $\varepsilon$  in diesen Gruppen, mit Ausnahme höchster Schwellenwerte, und unterstützt die Feststellung der unterschiedlichen Ziele beider Maße.

Der direkte Vergleich von Blockmethoden, die auf  $\varepsilon$ ,  $r^2$ , and  $D'$  basieren, liefert zwei Beobachtungen: Erstens enthalten ungefähr 50% der durch paarweise Maße definierten Blöcke genau zwei SNPs, während  $\varepsilon$  zu Blöcken mit mehr SNPs führt. Der  $\varepsilon$ -basierte Algorithmus findet weniger, aber längere Blöcke. Zweitens stimmen die Mengen der in Blöcke eingeschlossenen SNPs nur teilweise für die Maße überein, da  $\varepsilon$  und  $r^2$  LD beschreiben, während  $D'$  auf abwesende Rekombinationen abzielt.

Die Analyse des Datensatzes zeigt auch, daß die resultierenden Blöcke wesentlich von der Wahl der Parameter abhängen und nur wenige Regionen scharfe, blockartige Grenzen zeigen. Der irreführende Ausdruck *Block* sollte daher in zukünftigen Anwendungen vermieden werden. Die Wahl des Schwellenwertes ist eine Abwägung zwischen kürzeren Blöcken mit weniger Haplotypen und geringerer Coverage und längeren Blöcken mit mehr Haplotypen und geringerem LD, aber höherer Coverage. Sie ist daher abhängig vom Analyseziel. Die vorliegende Analyse stützt sich auf eine Stichprobe einer europäisch-amerikanischen Population und kann daher nicht ohne weiteres auf andere Populationen, z.B. afrikanischer Abstammung, übertragen werden.

Das vorgeschlagene LD-Maß und der darauf basierende Algorithmus liefern Regionen in erhöhtem LD und mit einer begrenzten Anzahl von Haplotypen. Sie erfüllen damit zwei Ziele, die mit der Definition von Haplotypblöcken verfolgt werden, nämlich die Benutzung von Haplotypen als multiallelische Marker in Assoziationsstudien und die effektive Beschreibung genetischer Variation. Diese Haplotypen können potentiell die statistische Power in Genkartierungsstudien erhöhen, in Abhängigkeit vom Alter der Krankheitsmutation, der umliegenden LD-Struktur und der Populationsgeschichte. Die Verallgemeinerung der Ergebnisse von Chromosom 12 führt zu der Empfehlung, 200.000–300.000 SNPs in genomweiten Krankheitsstudien zu betrachten, um die Vorteile von Haplotypanalysen auszuschöpfen.  $\varepsilon$ -Profile könnten die effektive Platzierung der SNPs auf dem Genom unterstützen. LD als Eigenschaft des Genoms hat eine Anziehungskraft, die über Assoziationsstudien hinausgeht. Nichtkodierende Regionen in hohem LD könnten

sich als funktionell relevant herausstellen und durch vergleichende Studien von Populationen oder Spezies gefunden werden. Das vorgeschlagene Maß  $\varepsilon$  hat das Potential zu einem hilfreichen Werkzeug in diesen und anderen Untersuchungen.

# Abbreviations

<i>Abbreviation</i>	<i>Term</i>	<i>See page</i>
$\mathbb{1}_{\{x\}}$	Indicator function	20
$a_j^i$	Allele at SNP position $j$ in haplotype $i$	20
C	A computer programming language	27
CD/CV	Common Disease/Common Variants hypothesis	5
$D$	Haplotype frequency deviation from expectation	10
$\delta_i$	Frequency deviation for a particular haplotype	21
$D'$	Standardized $D$	10
$ D' $	Absolute value of $D'$	
$\Delta S$	Entropy difference	21
$\varepsilon$	Normalized entropy difference	21
$\varepsilon_m$	$\varepsilon$ for window size $m$	21
$\varepsilon_{all}$	$\varepsilon$ , calculated with all haplotypes	30f
$\varepsilon_{cmn}$	$\varepsilon$ , calculated with re-scaled common haplotypes	30f
$\varepsilon_{lve}$	$\varepsilon$ , calculated with common haplotypes	30f
HT	Haplotype	2
htSNP	Haplotype-tagging SNP	7
kb	Kilobases (distance measured in $10^3$ of nucleotides)	
LD	Linkage disequilibrium	2, 9, 20
$L_B$	Likelihood of observed SNP sequence	21
$L_E$	Likelihood of SNP sequence in equilibrium	21
$m$	Window size used in the calculation of $\varepsilon$	20
Mb	Megabases (distance measured in $10^6$ of nucleotides)	

<i>Abbreviation</i>	<i>Term</i>	<i>See page</i>
$N$	Number of all haplotypes in the sample	20
NED	Normalized entropy difference	21
$n_i$	Number of a particular haplotype in the sample	20
$p_i$	Probability of a particular haplotype	20
$p_{ij}$	Frequency of haplotype $(i, j)$	9
$p_{i\cdot}, p_{\cdot j}$	Single allele frequencies	9
Perl	A computer programming language	31
$q_i$	Expected haplotype equilibrium frequency	20
R	A statistical computing language	31
$r^2$	Standardized $D$	10
$S$	Entropy	19
$S_B$	Entropy of the observed SNP sequence	20
$S_E$	Entropy of the SNP sequence under equilibrium	20
SNP	Single nucleotide polymorphism	2
SNaP	A SNP haplotype data simulation program	27
$\theta$	Recombination fraction	3, 11

# Bibliography

- [1] G. R. Abecasis, E. Noguchi, A. Heinzmann, J. A. Traherne, S. Bhattacharyya, N. I. Leaves, G. G. Anderson, Y. Zhang, N. J. Lench, A. Carey, L. R. Cardon, M. F. Moffatt, and W. O. Cookson. Extent and distribution of linkage disequilibrium in three genomic regions. *Am J Hum Genet*, 68(1):191–197, Jan 2001.
- [2] H. C. Ackerman, G. Ribas, M. Jallow, R. Mott, M. Neville, F. Sisay-Joof, M. Pinder, R. D. Campbell, and D. P. Kwiatkowski. Complex haplotypic structure of the central MHC region flanking TNF in a west african population. *Genes Immun*, 4(7):476–86, Oct 2003.
- [3] Hans Ackerman, Stanley Usen, Richard Mott, Anna Richardson, Fatoumatta Sisay-Joof, Pauline Katundu, Terrie Taylor, Ryk Ward, Malcolm Molyneux, Margaret Pinder, and Dominic P. Kwiatkowski. Haplotypic analysis of the TNF locus by association efficiency and entropy. *Genome Biol*, 4(4):R24, 2003.
- [4] Alan Agresti. *An Introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics. Applied Probability and Statistics. John Wiley & Sons, New York, 1996.
- [5] Joshua M. Akey, Kun Zhang, Momiao Xiong, and Li Jin. The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol Biol Evol*, 20(2):232–42, Feb 2003.
- [6] D. B. Allison, M. Heo, N. Kaplan, and E. R. Martin. Sibling-based tests of linkage and association for quantitative traits. *Am J Hum Genet*, 64(6):1754–63, Jun 1999.
- [7] Eric C. Anderson and John Novembre. Finding haplotype block boundaries by using the minimum-description-length principle. *Am J Hum Genet*, 73(2):336–54, Aug 2003.
- [8] Kristin G. Ardlie, Leonid Kruglyak, and Mark Seielstad. Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet*, 3(4):299–309, Apr 2002.
- [9] K. L. Ayres and D. J. Balding. Measuring gametic disequilibrium from multilocus data. *Genetics*, 157(1):413–23, Jan 2001.

- [10] S. A. Bacanu, B. Devlin, and K. Roeder. The power of genomic control. *Am J Hum Genet*, 66(6):1933–44, Jun 2000.
- [11] D.J. Balding, M. Bishop, and C. Cannings, editors. *Handbook of Statistical Genetics*. Wiley, Jan 2001.
- [12] Thomas M. Balding and Joy A. Thomas. *Elements of Information Theory*. Wiley series in telecommunications. Wiley, New York, 1991.
- [13] L. Beckmann, C. Fischer, K. G. Deck, I. M. Nolte, te G. Meerman, and J. Chang-Claude. Exploring haplotype sharing methods in general and isolated populations to detect gene(s) of a complex genetic trait. *Genet Epidemiol*, 21 Suppl 1:S554–9, 2001.
- [14] J. H. Bennett. On the theory of random mating. *Ann Eugen*, 18:311–317, 1954.
- [15] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull Calcutta Math Soc*, 35:99, 1943.
- [16] H Bielka and T Börner. *Molekulare Biologie der Zelle*. Fischer, Jena, Stuttgart, 1995.
- [17] J Blangero, JT Williams, and L Almasy. Variance component methods for detecting complex trait loci. In DC Rao, editor, *Advances in Genetics*, volume 42, pages 151–81, San Diego, 2000. Academic Press.
- [18] S. Bolk, J. Higgins, J. Moore, H. Nguyen, J. Roy, S. Schaffner, E.S. Lander, M.J. Daly, and D. Altshuler. The extent and diversity of common human haplotypes. *Am J Hum Genet*, 69 (Suppl.)(4):176, Oct 2001.
- [19] L. R. Cardon and J. I. Bell. Association study designs for complex diseases. *Nat Rev Genet*, 2(2):91–9, Feb 2001.
- [20] L. R. Cardon and D. W. Fulker. The power of interval mapping of quantitative trait loci, using selected sib pairs. *Am J Hum Genet*, 55(4):825–33, Oct 1994.
- [21] Lon R. Cardon and Goncalo R. Abecasis. Using haplotype blocks to map human complex trait loci. *Trends Genet*, 19(3):135–40, Mar 2003.
- [22] A. Chakravarti. Population genetics—making sense out of sequence. *Nat Genet*, 21(1 Suppl):56–60, Jan 1999.
- [23] M. N. Chiano and D. G. Clayton. Fine genetic mapping using haplotype analysis and the missing data problem. *Ann Hum Genet*, 62 ( Pt 1):55–60, Jan 1998.
- [24] A. G. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol*, 7(2):111–22, Mar 1990.
- [25] D. Clayton. Population association. In Balding et al. [11], pages 519–540.

- [26] F. Clerget-Darpoux, C. Bonaiti-Pellie, and J. Hochez. Effects of misspecifying genetic parameters in lod score analysis. *Biometrics*, 42(2):393–9, Jun 1986.
- [27] Helen M. Colhoun, Paul M. McKeigue, and Davey George Smith. Problems of reporting genetic associations with complex outcomes. *Lancet*, 361(9360):865–72, Mar 8 2003.
- [28] A. Collins, S. Ennis, P. Taillon-Miller, P. Y. Kwok, and N. E. Morton. Allelic association with SNPs: metrics, populations, and the linkage disequilibrium map. *Hum Mutat*, 17(4):255–62, Apr 2001.
- [29] The International HapMap Consortium. The International HapMap Project. *Nature*, 426(6968):789–96, Dec 18 2003.
- [30] M. J. Daly, J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. High-resolution haplotype structure in the human genome. *Nat Genet*, 29(2):229–32, Oct 2001.
- [31] M.D. Daly, J.D. Rioux, S.F. Schaffner, T.J. Hudson, and E.S. Lander. Fine-structure haplotype map of 5q31: implications for gene-based association studies and genomic ld mapping. *Am J Hum Genet*, 69 (Suppl.)(4):181, Oct 2001.
- [32] Elisabeth Dawson, Goncalo R. Abecasis, Suzannah Bumpstead, Yuan Chen, Sarah Hunt, David M. Beare, Jagjit Pabial, Thomas Dibling, Emma Tinsley, Susan Kirby, David Carter, Marianna Papaspyridonos, Simon Livingstone, Rocky Ganske, Elin Lohmussaar, Jana Zernant, Neeme Tonisson, Maido Remm, Reedik Magi, Tarmo Puurand, Jaak Vilo, Ants Kurg, Kate Rice, Panos Deloukas, Richard Mott, Andres Metspalu, David R. Bentley, Lon R. Cardon, and Ian Dunham. A first-generation linkage disequilibrium map of human chromosome 22. *Nature*, 418(6897):544–8, Aug 1 2002.
- [33] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–22, Sep 20 1995.
- [34] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, Dec 1999.
- [35] B. Devlin, K. Roeder, and S. A. Bacanu. Unbiased methods for population-based association studies. *Genet Epidemiol*, 21(4):273–84, Dec 2001.
- [36] B. Devlin, K. Roeder, and L. Wasserman. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol*, 60(3):155–66, Nov 2001.
- [37] A. M. Dunning, F. Durocher, C. S. Healey, M. D. Teare, S. E. McBride, F. Carlungo, C. F. Xu, E. Dawson, S. Rhodes, S. Ueda, E. Lai, R. N. Luben, Van E. J. Rensburg, A. Mannermaa, V. Kataja, G. Rennart, I. Dunham, I. Purvis, D. Easton,

- and B. A. Ponder. The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am J Hum Genet*, 67(6):1544–54, Dec 2000.
- [38] AWF Edwards. The early history of statistical estimation of linkage. In IH Pawlowitzki, JH Edwards, and EA Thompson, editors, *Genetic Mapping of Disease Genes*, pages 9–14, San Diego, London, 1997. Academic Press.
- [39] R. C. Elston. Linkage and association. *Genet Epidemiol*, 15(6):565–76, 1998.
- [40] ES Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 15 2001.
- [41] JC Venter et al. The sequence of the human genome. *Science*, 291(5507):1304–51, Feb 16 2001.
- [42] L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, 12(5):921–7, Sep 1995.
- [43] RA Fisher. On the mathematical foundations of theoretical statistics. *Phil Trans R Soc, A* 222:309–68, 1922.
- [44] D. W. Fulker and L. R. Cardon. A sib-pair approach to interval mapping of quantitative trait loci. *Am J Hum Genet*, 54(6):1092–103, Jun 1994.
- [45] D. W. Fulker, S. S. Cherny, P. C. Sham, and J. K. Hewitt. Combined linkage and association sib-pair analysis for quantitative traits. *Am J Hum Genet*, 64(1):259–67, Jan 1999.
- [46] 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, Adebawale Adeyemo, Richard Cooper, Ryk Ward, Eric S. Lander, Mark J. Daly, and David Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–9, Jun 21 2002.
- [47] D. B. Goldstein. Islands of linkage disequilibrium. *Nat Genet*, 29(2):109–11, Oct 2001.
- [48] D. Gordon, S. J. Finch, M. Nothnagel, and J. Ott. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Hum Hered*, 54(1):22–33, 2002.
- [49] H. H. Göring and J. D. Terwilliger. Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet*, 66(4):1310–27, Apr 2000.

- [50] I. C. Gray, D. A. Campbell, and N. K. Spurr. Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet*, 9(16):2403–8, Oct 2000.
- [51] Helene Guillon and Bernard de Massy. An initiation site for meiotic crossing-over and gene conversion in the mouse. *Nat Genet*, 32(2):296–9, Oct 2002.
- [52] S. W. Guo. Linkage disequilibrium measures for fine-scale mapping: a comparison. *Hum Hered*, 47(6):301–14, Nov-Dec 1997.
- [53] D. Gusfield. Inference of haplotypes from samples of diploid populations: complexity and algorithms. *J Comput Biol*, 8(3):305–23, 2001.
- [54] Christopher A. Haiman, Daniel O. Stram, Malcolm C. Pike, Laurence N. Kolonel, Noel P. Burt, David Altshuler, Joel Hirschhorn, and Brian E. Henderson. A comprehensive haplotype analysis of CYP19 and breast cancer risk: the multiethnic cohort. *Hum Mol Genet*, 12(20):2679–92, Oct 15 2003.
- [55] Jochen Hampe, Stefan Schreiber, and Michael Krawczak. Entropy-based SNP selection for genetic association studies. *Hum Genet*, 114(1):36–43, Dec 2003.
- [56] Paul Hardenbol, Johan Baner, Maneesh Jain, Mats Nilsson, Eugeni A. Namsaraev, George A. Karlin-Neumann, Hossein Fakhrai-Rad, Mostafa Ronaghi, Thomas D. Willis, Ulf Landegren, and Ronald W. Davis. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nat Biotechnol*, 21(6):673–8, Jun 2003.
- [57] Daniel L. Hartl and Andrew G. Clark. *Principles of Population Genetics*. Sinauer, Sunderland, MA, 1988.
- [58] J. K. Haseman and R. C. Elston. The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet*, 2(1):3–19, Mar 1972.
- [59] M. E. Hawley and K. K. Kidd. HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J Hered*, 86(5):409–11, Sep-Oct 1995.
- [60] P. W. Hedrick. Gametic disequilibrium measures: proceed with caution. *Genetics*, 117(2):331–41, Oct 1987.
- [61] N. A. Holtzman. Putting the search for genes in perspective. *Int J Health Serv*, 31(2):445–61, 2001.
- [62] S. Horvath, X. Xu, and N. M. Laird. The family based association test method: strategies for studying general genotype–phenotype associations. *Eur J Hum Genet*, 9(4):301–6, Apr 2001.
- [63] Richard R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–8, Feb 2002.

- [64] R.R. Hudson. Linkage disequilibrium and recombination. In Balding et al. [11], pages 309–324.
- [65] Damini Jawaheer, Wentian Li, Robert R. Graham, Wei Chen, Aarti Damle, Xi-angli Xiao, Joanita Monteiro, Houman Khalili, Annette Lee, Robert Lundsten, Ann Begovich, Teodorica Bugawan, Henry Erlich, James T. Elder, Lindsey A. Criswell, Michael F. Seldin, Christopher I. Amos, Timothy W. Behrens, and Peter K. Gregersen. Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. *Am J Hum Genet*, 71(3):585–94, Sep 2002.
- [66] A. J. Jeffreys, L. Kauppi, and R. Neumann. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet*, 29(2):217–22, Oct 2001.
- [67] Alec J. Jeffreys and Rita Neumann. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet*, 31(3):267–71, Jul 2002.
- [68] S. E. Johnatty, M. Abdellatif, L. Shimmin, R. B. Clark, and E. Boerwinkle. beta(2) adrenergic receptor 5' haplotypes influence promoter activity. *Br J Pharmacol*, 137(8):1213–1216, Dec 8 2002.
- [69] G. C. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, Di G. Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S. C. Gough, D. G. Clayton, and J. A. Todd. Haplotype tagging for the identification of common disease genes. *Nat Genet*, 29(2):233–7, Oct 2001.
- [70] L. B. Jorde. Linkage disequilibrium and the search for complex disease genes. *Genome Res*, 10(10):1435–44, Oct 2000.
- [71] Richard Judson, Benjamin Salisbury, Julie Schneider, Andreas Windemuth, and J. Claiborne Stephens. How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics*, 3(3):379, May 2002.
- [72] Xiayi Ke and Lon R. Cardon. Efficient selective screening of haplotype tag SNPs. *Bioinformatics*, 19(2):287–8, Jan 2003.
- [73] Brian W. Kernighan and Dennis M. Ritchie. *The C Programming Language*. Prentice Hall Software Series, Upper Saddle River, NJ, second edition, 1988.
- [74] J. R. Kidd, A. J. Pakstis, H. Zhao, R. B. Lu, F. E. Okonofua, A. Odunsi, E. Grigorenko, B. B. Tamir, J. Friedlaender, L. O. Schulz, J. Parnas, and K. K. Kidd. Haplotypes and linkage disequilibrium at the phenylalanine hydroxylase locus, PAH, in a global representation of populations. *Am J Hum Genet*, 66(6):1882–99, Jun 2000.

- [75] Motoo Kimura. *Population Genetics, Molecular Evolution & the Neutral Theory: Selected Papers*. University of Chicago Press, 1994.
- [76] M. Knapp. The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/ disequilibrium test. *Am J Hum Genet*, 64(3):861–70, Mar 1999.
- [77] M Knapp, SA Seuchter, and MP Baur. Linkage analysis in nuclear families. 1: Optimality criteria for affected sib-pair tests. *Hum Hered*, 44(1):37–43, Jan-Feb 1995.
- [78] Michael Knapp and Tim Becker. Family-based association analysis with tightly linked markers. *Hum Hered*, 56(1-3):2–9, 2003.
- [79] Hans Knoblauch, Anja Bauerfeind, Christine Krahenbuhl, Aurelie Daury, Klaus Rohde, Stephane Bejanin, Laurent Essioux, Herbert Schuster, Friedrich C. Luft, and Jens Georg Reich. Common haplotypes in five genes influence genetic variance of LDL and HDL cholesterol in the general population. *Hum Mol Genet*, 11(12):1477–85, Jun 1 2002.
- [80] M. Krawczak, N. A. Chuzhanova, P. D. Stenson, B. N. Johansen, E. V. Ball, and D. N. Cooper. Changes in primary DNA sequence complexity influence the phenotypic consequences of mutations in human gene regulatory regions. *Hum Genet*, 107(4):362–5, Oct 2000.
- [81] L. Kruglyak. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet*, 22(2):139–44, Jun 1999.
- [82] L. Kruglyak, M. J. Daly, M. P. Reeve-Daly, and E. S. Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet*, 58(6):1347–63, Jun 1996.
- [83] J. C. Lam, K. Roeder, and B. Devlin. Haplotype fine mapping by evolutionary trees. *Am J Hum Genet*, 66(2):659–73, Feb 2000.
- [84] E. Lander and L. Kruglyak. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet*, 11(3):241–7, Nov 1995.
- [85] E. S. Lander and N. J. Schork. Genetic dissection of complex traits. *Science*, 265(5181):2037–48, Sep 30 1994.
- [86] D. F. Levinson, I. Nolte, and te G. J. Meerman. Haplotype sharing tests of linkage disequilibrium in a hutterite asthma data set. *Genet Epidemiol*, 21 Suppl 1:S308–11, 2001.
- [87] R. C. Lewontin. On measures of gametic disequilibrium. *Genetics*, 120(3):849–52, Nov 1988.

- [88] Shuying Sue Li, Najma Khalid, Christopher Carlson, and Lue Ping Zhao. Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. *Biostatistics*, 4(4):513–22, Oct 2003.
- [89] Kirk E. Lohmueller, Celeste L. Pearce, Malcolm Pike, Eric S. Lander, and Joel N. Hirschhorn. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet*, 33(2):177–82, Feb 2003.
- [90] J. C. Long, R. C. Williams, and M. Urbanek. An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am J Hum Genet*, 56(3):799–810, Mar 1995.
- [91] C. J. MacLean, R. B. Martin, P. C. Sham, H. Wang, R. E. Straub, and K. S. Kendler. The trimmed-haplotype test for linkage disequilibrium. *Am J Hum Genet*, 66(3):1062–75, Mar 2000.
- [92] H. Mannila, M. Koivisto, M. Perola, T. Varilo, W. Hennah, J. Ekelund, M. Lukk, L. Peltonen, and E. Ukkonen. Minimum description length block finder, a method to identify haplotype blocks and to compare the strength of block boundaries. *Am J Hum Genet*, 73(1):86–94, Jul 2003.
- [93] N. Martin, D. Boomsma, and G. Machin. A twin-pronged attack on complex traits. *Nat Genet*, 17(4):387–92, Dec 1997.
- [94] Celia A. May, Angela C. Shone, Luba Kalaydjieva, Antti Sajantila, and Alec J. Jeffreys. Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. *Nat Genet*, 31(3):272–5, Jul 2002.
- [95] Christian Meisel, Thomas Gerloff, Julia Kirchheiner, Przemyslaw M. Mrozkiewicz, Przemyslaw Niewinski, Jurgen Brockmoller, and Ivar Roots. Implications of pharmacogenetics for individualizing drug treatment and for study design. *J Mol Med*, 81(3):154–67, Mar 2003.
- [96] Zhaoling Meng, Dmitri V. Zaykin, Chun-Fang Xu, Michael Wagner, and Margaret G. Ehm. Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *Am J Hum Genet*, 73(1):115–30, Jul 2003.
- [97] Richard W. Morris and Norman L. Kaplan. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol*, 23(3):221–33, Oct 2002.
- [98] N. E. Morton, W. Zhang, P. Taillon-Miller, S. Ennis, P. Y. Kwok, and A. Collins. The optimal measure of allelic association. *Proc Natl Acad Sci U S A*, 98(9):5217–21, Apr 24 2001.
- [99] N.E. Morton. Sequential tests for the detection of linkage. *Am J Hum Genet*, 7(3):277–318, Sep 1955.

- [100] Tianhua Niu, Zhaohui S. Qin, Xiping Xu, and Jun S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet*, 70(1):157–69, Jan 2002.
- [101] M. Nordborg. Coalescent theory. In Balding et al. [11], pages 179–212.
- [102] M. Nothnagel. Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-control data by supervised learning methods. *Am J Hum Genet*, 71(Suppl 4)(4):A2363, Oct 2002.
- [103] M. Nothnagel, R. Fürst, and K. Rohde. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum Hered*, 54(4):186–98, 2002.
- [104] M. Nothnagel and J. Ott. Statistical gene mapping of traits in humans—hypertension as a complex trait: is it amenable to genetic analysis? *Semin Nephrol*, 22(2):105–14, Mar 2002.
- [105] D. R. Nyholt. All LODs are not created equal. *Am J Hum Genet*, 67(2):282–8, Aug 2000.
- [106] Jürg Ott. *Analysis of human genetic linkage*. Johns Hopkins University Press, Baltimore and London, third edition, 1999.
- [107] Csaba Pal and Laurence D. Hurst. Evidence for co-evolution of gene order and recombination rate. *Nat Genet*, 33(3):392–5, Mar 2003.
- [108] N. Patil, A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, D. H. Lee, C. Marjoribanks, D. P. McDonough, B. T. Nguyen, M. C. Norris, J. B. Sheehan, N. Shen, D. Stern, R. P. Stokowski, D. J. Thomas, M. O. Trulson, K. R. Vyas, K. A. Frazer, S. P. Fodor, and D. R. Cox. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science*, 294(5547):1719–23, Nov 23 2001.
- [109] M. S. Phillips, R. Lawrence, R. Sachidanandam, A. P. Morris, D. J. Balding, M. A. Donaldson, J. F. Studebaker, W. M. Ankener, S. V. Alfisi, F-S Kuo, A. L. Camisa, V. Pazorov, K. E. Scott, B. J. Carey, J. Faith, G. Katari, H. A. Bhatti, J. M. Cyr, V. Derohannessian, C. Elosua, A. M. Forman, N. M. Grecco, C. R. Hock, J. M. Kuebler, J. A. Lathrop, M. A. Mockler, E. P. Nachtman, S. L. Restine, S. A. Varde, M. J. Hozza, C. A. Gelfand, J. Broxholme, G. R. Abecasis, M. T. Boyce-Jacino, and L. R. Cardon. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet*, 33(3):382–7, Mar 2003.
- [110] David Posada and Carsten Wiuf. Simulating haplotype blocks in the human genome. *Bioinformatics*, 19(2):289–90, Jan 22 2003.
- [111] J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*, 69(1):124–37, Jul 2001.

- [112] J. K. Pritchard and P. Donnelly. Case-control studies of association in structured or admixed populations. *Theor Popul Biol*, 60(3):227–37, Nov 2001.
- [113] J. K. Pritchard and M. Przeworski. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69(1):1–14, Jul 2001.
- [114] J. K. Pritchard and N. A. Rosenberg. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet*, 65(1):220–8, Jul 1999.
- [115] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–59, Jun 2000.
- [116] J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly. Association mapping in structured populations. *Am J Hum Genet*, 67(1):170–81, Jul 2000.
- [117] Jonathan K. Pritchard and Nancy J. Cox. The allelic architecture of human disease genes: common disease-common variant. or not? *Hum Mol Genet*, 11(20):2417–23, Oct 1 2002.
- [118] D. Rabinowitz. A transmission disequilibrium test for quantitative trait loci. *Hum Hered*, 47(6):342–50, Nov-Dec 1997.
- [119] Steven J. Raynard, Leah R. Read, and Mark D. Baker. Evidence for the murine IgH mu locus acting as a hot spot for intrachromosomal homologous recombination. *J Immunol*, 168(5):2332–9, Mar 1 2002.
- [120] D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, May 10 2001.
- [121] D. E. Reich and E. S. Lander. On the allelic spectrum of human disease. *Trends Genet*, 17(9):502–10, Sep 2001.
- [122] David E. Reich, Stephen F. Schaffner, Mark J. Daly, Gil McVean, James C. Mullikin, John M. Higgins, Daniel J. Richter, Eric S. Lander, and David Altshuler. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet*, 32(1):135–42, Sep 2002.
- [123] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–7, Sep 13 1996.
- [124] K. Rohde and R. Fürst. Haplotyping and estimation of haplotype frequencies for closely linked biallelic multilocus genetic phenotypes including nuclear family information. *Hum Mutat*, 17(4):289–95, Apr 2001.
- [125] Noah A. Rosenberg and Magnus Nordborg. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nat Rev Genet*, 3(5):380–90, May 2002.

- [126] P Rubinstein, M Walker, C Carpenter, C Carrier, J Krassner, CT Falk, and F Ginsburg. Genetics of HLA disease associations. the use of the haplotype relative risk (HRR) and the 'haplo-delta" (Dh) estimates in juvenile diabetes from three racial groups. *Hum Immunol*, 3:384, 1981.
- [127] Chiara Sabatti and Neil Risch. Homozygosity and linkage disequilibrium. *Genetics*, 160(4):1707–19, Apr 2002.
- [128] Benjamin A. Salisbury, Manish Pungliya, Julie Y. Choi, Ruhong Jiang, Xiao Jenny Sun, and J. Claiborne Stephens. SNP and haplotype variation in the human genome. *Mutat Res*, 526(1-2):53–61, May 15 2003.
- [129] NJ Schork and X Xu. The use of twins in quantitative trait locus mapping. In TD Spector, H Snieder, and AJ MacGregor, editors, *Advances in Twin and Sib-pair Analysis*, pages 189–202, London, 2000. Greenwich Medical Media.
- [130] Pak Sham. *Statistics in Human Genetics*. Arnold Applications of Statistics. Arnold, London, 1998.
- [131] C. E. Shannon. A mathematical theory of communication. *Bell System Techn J*, 27:379–423,623–656, Jul,Oct 1948.
- [132] Sagiv Shifman, Jane Kuypers, Mark Kokoris, Benjamin Yakir, and Ariel Darvasi. Linkage disequilibrium patterns of the human genome across populations. *Hum Mol Genet*, 12(7):771–6, Apr 1 2003.
- [133] S. L. Slager, J. Huang, and V. J. Vieland. Effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genet Epidemiol*, 18(2):143–56, Feb 2000.
- [134] R. S. Spielman and W. J. Ewens. A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet*, 62(2):450–8, Feb 1998.
- [135] R. S. Spielman, R. E. McGinnis, and W. J. Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*, 52(3):506–16, Mar 1993.
- [136] J. C. Stephens. Single-nucleotide polymorphisms, haplotypes, and their relevance to pharmacogenetics. *Mol Diagn*, 4(4):309–17, Dec 1999.
- [137] J. C. Stephens, J. A. Schneider, D. A. Tanguay, J. Choi, T. Acharya, S. E. Stanley, R. Jiang, C. J. Messer, A. Chew, J. H. Han, J. Duan, J. L. Carr, M. S. Lee, B. Koshy, A. M. Kumar, G. Zhang, W. R. Newell, A. Windemuth, C. Xu, T. S. Kalbfleisch, S. L. Shaner, K. Arnold, V. Schulz, C. M. Drysdale, K. Nandabalan, R. S. Judson, G. Ruano, and G. F. Vovis. Haplotype variation and linkage disequilibrium in 313 human genes. *Science*, 293(5529):489–93, Jul 20 2001.

- [138] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, 68(4):978–89, Apr 2001.
- [139] Daniel O. Stram, Christopher A. Haiman, Joel N. Hirschhorn, David Altshuler, Laurence N. Kolonel, Brian E. Henderson, and Malcolm C. Pike. Choosing Haplotype-Tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the multiethnic cohort study. *Hum Hered*, 55(1):27–36, 2003.
- [140] MW Strickberger. *Genetics*. Macmillan, New York, third edition, 1985.
- [141] Chris Stringer. Modern human origins: progress and prospects. *Philos Trans R Soc Lond B Biol Sci*, 357(1420):563–79, Apr 29 2002.
- [142] Richard Strohman. Maneuvering in the complex path from genotype to phenotype. *Science*, 296(5568):701–3, Apr 26 2002.
- [143] Michael H. P Stumpf. Haplotype diversity and the block structure of linkage disequilibrium. *Trends Genet*, 18(5):226–8, May 2002.
- [144] L. Subrahmanyam, M. A. Eberle, A. G. Clark, L. Kruglyak, and D. A. Nickerson. Sequence variation and linkage disequilibrium in the human T-cell receptor beta (TCRB) locus. *Am J Hum Genet*, 69(2):381–95, Aug 2001.
- [145] P. Taillon-Miller, I. Bauer-Sardina, N. L. Saccone, J. Putzel, T. Laitinen, A. Cao, J. Kere, G. Pilia, J. P. Rice, and P. Y. Kwok. Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat Genet*, 25(3):324–8, Jul 2000.
- [146] M. D. Teare, A. M. Dunning, F. Durocher, G. Rennart, and D. F. Easton. Sampling distribution of summary linkage disequilibrium measures. *Ann Hum Genet*, 66(Pt 3):223–33, May 2002.
- [147] J. D. Terwilliger and J. Ott. A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum Hered*, 42(6):337–46, 1992.
- [148] Joseph D. Terwilliger, Fatemeh Haghighi, Tero S. Hiekkalinna, and Harald H. H Goring. A bias-ed assessment of the use of SNPs in human complex traits. *Curr Opin Genet Dev*, 12(6):726–34, Dec 2002.
- [149] G. Thomson and M. P. Baur. Third order linkage disequilibrium. *Tissue Antigens*, 24(4):250–5, Oct 1984.
- [150] Rebecca J. C Twells, Charles A. Mein, Michael S. Phillips, J. Fred Hess, Riitta Veijola, Matthew Gilbey, Matthew Bright, Michael Metzker, Benedicte A. Lie, Amanda Kingsnorth, Edward Gregory, Yusuke Nakagawa, Hywel Snook, William S. Y Wang,

- Jennifer Masters, Gillian Johnson, Iain Eaves, Joanna M. M Howson, David Clayton, Heather J. Cordell, Sarah Nutland, Helen Rance, Philippa Carr, and John A. Todd. Haplotype structure, LD blocks, and uneven recombination within the LRP5 gene. *Genome Res*, 13(5):845–55, May 2003.
- [151] Jeffrey D. Wall and Jonathan K. Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet*, 4(8):587–97, Aug 2003.
- [152] Larry Wall, Tom Christiansen, and Randal L. Schwartz. *Programmieren mit Perl*. O’Reilly, Köln, second edition, 1997.
- [153] Lusheng Wang and Ying Xu. Haplotype inference by maximum parsimony. *Bioinformatics*, 19(14):1773–80, Sep 22 2003.
- [154] Ning Wang, Joshua M. Akey, Kun Zhang, Ranajit Chakraborty, and Li Jin. Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am J Hum Genet*, 71(5):1227–34, Nov 2002.
- [155] D. E. Weeks and K. Lange. The affected-pedigree-member method of linkage analysis. *Am J Hum Genet*, 42(2):315–26, Feb 1988.
- [156] Bruce S. Weir. *Genetic Data Analysis II: Methods for discrete population genetic data*. Sinauer, Sunderland, MA, 1996.
- [157] S. S. Wilks. *Mathematical Statistics*. Wiley, New York, 1962.
- [158] A. Wille. *Sum Statistics for the Joint Detection of Multiple Disease Loci in Complex Traits*. Ph.D. thesis, The Rockefeller University, 1230 York Avenue, New York, NY 10021, U.S.A., June 2003.
- [159] M. H. Wolpoff, J. Hawks, D. W. Frayer, and K. Hunley. Modern human ancestry at the peripheries: a test of the replacement theory. *Science*, 291(5502):293–7, Jan 12 2001.
- [160] Momiao Xiong, Jinying Zhao, and Eric Boerwinkle. Haplotype block linkage disequilibrium mapping. *Front Biosci*, 8:a85–93, May 1 2003.
- [161] X. Xu, S. Weiss, X. Xu, and L. J. Wei. A unified Haseman-Elston method for testing linkage with quantitative traits. *Am J Hum Genet*, 67(4):1025–8, Oct 2000.
- [162] Cyrus P. Zabetian, Sarah G. Buxbaum, Robert C. Elston, Michael D. Kohnke, George M. Anderson, Joel Gelernter, and Joseph F. Cubells. The structure of linkage disequilibrium at the DBH locus strongly influences the magnitude of association between diallelic markers and plasma dopamine beta-hydroxylase activity. *Am J Hum Genet*, 72(6):1389–400, Jun 2003.

- [163] Kui Zhang, Peter Calabrese, Magnus Nordborg, and Fengzhu Sun. Haplotype block structure and its applications to association studies: power and study designs. *Am J Hum Genet*, 71(6):1386–94, Dec 2002.
- [164] Kui Zhang, Minghua Deng, Ting Chen, Michael S. Waterman, and Fengzhu Sun. A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci U S A*, 99(11):7335–9, May 28 2002.
- [165] Kui Zhang, Fengzhu Sun, Michael S. Waterman, and Ting Chen. Haplotype block partition with limited resources and applications to human chromosome 21 haplotype data. *Am J Hum Genet*, 73(1):63–73, Jul 2003.
- [166] Kun Zhang and Li Jin. HaploBlockFinder: haplotype block analyses. *Bioinformatics*, 19(10):1300–1, Jul 1 2003.
- [167] H. Zhao, A. J. Pakstis, J. R. Kidd, and K. K. Kidd. Assessing linkage disequilibrium in a complex genetic system. I. overall deviation from random association. *Ann Hum Genet*, 63 ( Pt 2):167–79, Mar 1999.
- [168] Xiaofeng Zhu, Denise Yan, Richard S. Cooper, Amy Luke, Morna A. Ikeda, Yen-Pei C. Chang, Alan Weder, and Aravinda Chakravarti. Linkage disequilibrium and haplotype diversity in the genes of the renin-angiotensin system: findings from the family blood pressure program. *Genome Res*, 13(2):173–81, Feb 2003.

# CURRICULUM VITAE

Michael Nothnagel

Sonntagstr. 8

D-10245 Berlin

Geburtsdatum 22. Juli 1971

Geburtsort Berlin

Nationalität deutsch

## **Max-Delbrück-Centrum für Molekulare Medizin**

Berlin

Doktorand in der Arbeitsgruppe Bioinformatik 3/2002 -  
(Prof. Dr. J.G. Reich)

NGFN: Bioinformatik am Gene Mapping Center,  
Organisation des Gene Mapping Course 2003

Thema: "The Definition of Multilocus Haplotype  
Blocks and Common Diseases"

## **Baylor College of Medicine**

Houston, Texas, U.S.A.

Predoctoral Fellow im Laboratory of Molecular 1/2003 - 4/2003  
and Human Genetics (Prof. Dr. S.M. Leal)

Statistische Analyse von SNP-Haplotyp-Daten

## **Rockefeller University**

New York, New York, U.S.A.

Visiting Student im Laboratory of 1/2001 - 11/2001  
Statistical Genetics (Prof. Dr. J. Ott)

Einarbeitung in die Theorie der Statistischen Genetik,  
Anwendung diskriminanzanalytischer Methoden  
für die Genkartierung

# CURRICULUM VITAE

**Franz-Volhard-Klinik der Charité/  
INFOGEN Medizinische Genetik GmbH**  
Berlin

Project Manager Statistics 9/2000 - 2/2002  
Berater für INFOGEN 2/1999 - 8/2000  
Freier Mitarbeiter 4/1995 - 1/1999

Arbeitsgruppe Genetik Fieldworking, später für die  
Ausgründung INFOGEN (Prof. Dr. H. Schuster)  
Datenbank-Design, -Programmierung, -Auswertung,  
Statistische Analysen, Produktentwicklung,  
Prozeß- und Projekt-Management, Dokumentation

**Humboldt-Universität zu Berlin**

Diplomand im Institut für Mathematik 4/1999 - 10/1999  
(Prof. Dr. O. Bunke)

Thema: "Verfahren der Diskriminanzanalyse -  
eine vergleichende und integrierende Übersicht"

Studium der Mathematik und Biologie; 10/1991 - 10/1999  
Spezialisierung in Statistik / Stochastik

**Städtisches Klinikum**

Berlin

Zivildienstleistender in der Rehabilitationsstation 9/1990 - 8/1991  
für körperlich behinderte Jugendliche

**Bildungsweg**

Schulbildung, Berlin 9/1978 - 8/1990

Abitur der Spezialschule mathematisch-  
naturwissenschaftlich-technischer Richtung  
"Heinrich Hertz"

Michael Nothnagel  
Sonntagstr. 8  
D-10245 Berlin, Germany

## Publikationen und Präsentationen

Nothnagel M, Fürst R, Rohde K. (2002) Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. *Hum Hered*; 54(4):186-198.

Nothnagel M, Ott J. (2002) Statistical gene mapping of traits in humans-hypertension as a complex trait: Is it amenable to genetic analysis? *Semin Nephrol*; 22(2):105-14.

Nothnagel M. (2002) Simulation of LD block-structured SNP haplotype data and its use for the analysis of case-control data by supervised learning methods. American Society of Human Genetics, October 15-19, Baltimore, MD, U.S.A., *Am J Hum Genet* 71 (Suppl)(4), A2363.

Nothnagel M. (2002) A software program to simulate SNP haplotype data with blocks of high linkage disequilibrium. NGFN Symposium: The genetic and molecular basis of human disease, November 17-19, Berlin.

Nothnagel M, Fürst R, Rohde K. (2002) Haplotype estimation and linkage disequilibrium for phase-unknown SNP genotypes. German Human Genome Project Meeting, September 29 - October 2, Leipzig, Germany.

Gordon D, Finch SJ, Nothnagel M, Ott J. (2002) Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Hum Hered*; 54(1):22-33.

Schuster H, Lamprecht A, Junghans C, Dietz B, Baron H, Nothnagel M, Müller-Myhsok B, Luft FC. (1998) Approaches to the genetics of cardiovascular disease through genetic field work. *Kidney Int*; 53(6):1449-54.

# Erklärung an Eides Statt

Hiermit erkläre ich, daß ich die vorliegende Arbeit mit dem Titel

## The Definition of Multilocus Haplotype Blocks and Common Diseases

zur Erlangung des akademischen Grades *Doctor rerum medicarum* selbständig und ohne die unzulässige Hilfe Dritter verfaßt habe. Die Arbeit stellt, auch in Teilen, keine Kopie anderer Arbeiten dar; die für die Arbeit benutzten Literaturstellen, Quellen und Hilfsmittel sind vollständig angegeben.

Michael Nothnagel  
29. Februar 2004