UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE, HEALTH & LIFE SCIENCES

School of Medicine

The Construction of Linkage Disequilibrium maps and their Application to Association mapping of disease genes

by

Tai-Yue Kuo

Thesis for the degree of Doctor of Philosophy

February 2008

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF MEDICINE, HEALTH & LIFE SCIENCES

SCHOOL OF MEDICINE

Doctor of Philosophy

THE CONSTRUCTION OF LINKAGE DISEQUILIBRIUM MAPS AND THEIR APPLICATION TO ASSOCIATION MAPPING OF DISEASE GENES

by **Tai-Yue Kuo**

Success in association mapping of disease genes depends on knowledge of Linkage Disequilibrium (LD) structure in candidate regions. An LD map characterising such structures is constructed by making use of the Malecot model which describes the decline of LD with physical distance based on pairwise measures of association between SNPs. The HapMap project provides a valuable resource that can be used to construct genome-wide LD maps. However, the millions of SNPs in the HapMap data pose a heavy computational challenge. This difficulty can be resolved by excluding the very distant SNP pairs without losing map quality. Modern computational technology with parallel processing can be used to speed up the process of map construction. A composite likelihood approach employing LD maps for association mapping has successfully localised several causal variants. An application to Rheumatoid Arthritis (RA) is described here. This approach, utilising the genome-wide LD map, is very suitable for genome-wide association studies.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DECLARATION OF AUTHORSHIP

I, **Tai-Yue Kuo**, declare that the thesis entitled, "**The Construction of LD maps and their Application to Association mapping of disease genes**", and the work presented in it are my own and has been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published in journals and book:

1. **Kuo, T-Y.**, Lau, W., Tapper, W., Cox, S., Collins, A. (2007) Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. *Bioinformatics*: 23; 517-519.
2. **Kuo, T-Y.**, Lau W., Collins, A. (2007) *LDMAP*: The construction of high-resolution linkage disequilibrium maps of the human genome. In 'Linkage disequilibrium and association mapping: analysis and

ACKNOWLEDGEMENTS

I would like to express my deepest respect and most sincere gratitude to my supervisor, Prof. Andrew R Collins, for his patient guidance, encouragement, and advice he has provided in the development of this thesis throughout my PhD study. I am also grateful to Prof. Newton Morton for his inspiration and encouragement from the initial conception to the end of this research.

I would like to thank Dr. Nikolas Maniatis, Dr. Sarah Ennis and Dr. Weihua Zhang for their assistance, constructive criticism and comments for the thesis. I would also like to thank Dr. William Tapper, Dr. Winston Lau and Jane Gibson, who have helped me in many ways throughout.

A sincere appreciation goes to Dr. John Holloway and Dr. Dawn Teare for willingly accepting to evaluate the thesis and sit in the examination committee. I also acknowledge Dawn Teare for the provision of the rheumatoid arthritis data in the thesis. In addition, I would also like to acknowledge Ministry of Education in Taiwan for giving me the scholarship which enabled me to undertake this study.

Finally, thanks to my beloved parents and my wife, Yu-Hua Lee, for their constant support and understanding. I am also greatly indebted to my daughter Yen-Chen Kuo for enduring my absence for so long.

LIST OF ABBREVIATIONS

CEPH	Centre d'Etude du Polymorphisme Humain (CEU)
CEU	Utah Residents with Northern and Western European Ancestry
CHB	Han Chinese in Beijing, China
cM	CentiMorgan
CDCV	Common disease/common variant
95%CI	95% confidence interval
DNA	Deoxyribonucleic acid
df	Degrees of Freedom
EM	Expectation-Maximization algorithm
FDR	False discovery rate
GWA	Genome-wide association
GAW	Genetic Analysis Workshop
HapMap	Haplotype Map of the human genome
HWE	Hardy-Weinberg equilibrium
JPT	Japanese in Tokyo, Japan
LD	Linkage Disequilibrium
LDUs	Linkage Disequilibrium Units
LDDB	Linkage Disequilibrium Database
lnL	ln (natural logarithm) Likelihood
MCMC	Markov chain Monte Carlo
MALD	Mapping by admixture linkage disequilibrium
MAF	Minor allele frequencies
MHC	Major Histocompatibility complex
Mb	Megabases
max_intv	The maximum number of intervals between any pair of SNPs
max_dist	The maximum distance in kb between any pair of SNPs
NCBI	National Center for Biotechnology Information
NARAC	The North American Rheumatoid Arthritis Consortium
QC	Quality control
RA	Rheumatoid Arthritis
SNP	Single nucleotide polymorphism
TDT	Transmission disequilibrium test
UCSC	University of California, Santa Cruz
YRI	Yoruba in Ibadan, Nigeria

Chapter 1 Literature review

1.1 Introduction

Human Genetics is a study of DNA, genes, gene expression, and their applications to human health. It is particularly concerned with human diseases that are caused by genetic variants. DNA is comprised of four nucleotide bases: Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). The order of these nucleotide bases along a DNA strand, which is known as DNA sequence, encodes the genetic information in a precise order of base pairs. Genes are the DNA sequences that contain the genetic information necessary for building proteins. The information that is used to make proteins has to pass through a two-stage process known as transcription and translation. This process is also called gene expression.

Any change in DNA sequence is called mutation. Mutations may be large or small scale. A large scale mutation includes gain or loss of a region of a chromosome and a small scale mutation may be only a small change in a nucleotide base, such as a substitution, deletion, or insertion. In evolutionary terms, mutation provides genetic diversity but, in human health, mutation may affect the expression of genes, resulting in different types of diseases. If a mutation is present at relatively high frequency (>1%) in a population, it is called a polymorphism. The most common polymorphisms in the human genome are single nucleotide polymorphisms (SNPs). However, SNPs with known locations in the genome can be used as genetic markers to localise disease genes. Approximately 10 million SNPs existing in the human genome can be used for disease mapping.

Mendelian diseases, such as Cystic Fibrosis (CF), are single gene disorders, which are rare but with large phenotypic effects. Complex diseases, such as diabetes, heart diseases, and rheumatoid arthritis, do not follow Mendelian inheritance patterns but also exhibit familial aggregation. This may be due to sharing the same genes or environment. Such diseases are common but complex in nature because they are influenced by multiple genes and environmental risk factors. Therefore, each gene has only a modest effect. This is the main reason that few genes involved in common complex diseases have been identified to date.

Association mapping is a strategy that identifies the location of disease genes from the human genome. It has an advantage of requiring no prior knowledge about disease mechanism. The process of localising a disease gene is from several megabase regions previously identified by linkage to eventually identifying the location of the disease gene.

For mapping disease genes in the human genome, the analysis of genetic recombination is an essential method. Recombination is the process of exchanging genetic material between maternal and paternal chromosomes by crossing over during meiosis (the process of cell division to form gametes). This is an important process as it is the basis of genetic diversity. Analysing the recombination frequency between two loci allows the estimation of the genetic distance between them. This is the basic principal of linkage analyses and identification of disease genes. Genetic distance between two particular loci on a chromosome is measured by the number of recombination events divided by the total number of meioses. If the distance between two loci is very small, recombination is rare and the loci are tightly linked.

Linkage analyses can narrow down the candidate regions from the entire human genome to regions of several megabases (Mb). These regions, however, are still too large for fine mapping and hence for the localisation of most disease genes. Another approach is so called linkage disequilibrium (LD) analysis, also known as association mapping, which can further refine the candidate region. This approach employs LD, also called allelic association (described in the next section), which has much higher resolution than linkage analysis, because it exploits the information from historical recombination events over many generations. These recombination events break up large shared regions into smaller segments (See Figure 1.1); therefore, this approach can further narrow down the candidate region.

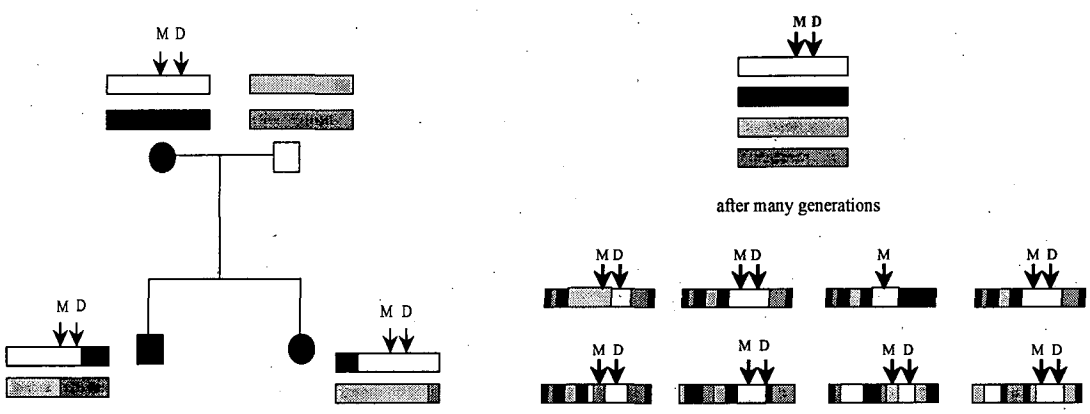


Figure 1.1 Linkage mapping (left) and association mapping (right)

M is a marker allele and D is a disease allele. This figure shows that after many generations with many historical recombination events, the region with both marker and disease alleles has been narrowed (right figure), compared to only one generation (left figure)

This chapter describes how LD can be used to localise disease genes. It also includes recent findings about LD patterns in the genome and in different populations. These findings motivated the international HapMap project and the development of a LD map for the entire genome. The last section introduces different methods using LD for mapping disease genes and the challenges of dealing with common complex diseases.

1.2 Linkage Disequilibrium

1.2.1 Introduction

Linkage disequilibrium (LD) is the non-random association of alleles at adjacent loci. It is present when two alleles at adjacent loci are found together more often than would be expected under random segregation. That is to say, the strong association between two alleles at small distance is retained after many generations. This is because recombination events occur infrequently at small distances. However LD is not only influenced by recombination; other historical events such as population admixture, genetic drift, natural selection and mutation may obscure the relationship between LD and distance between two alleles. This chapter reviews the literature on LD measures and approaches to model LD patterns in the genome.

1.2.2 The measures of LD

A variety of measures of LD have been proposed (Table 1.1), differing in the use of marginal allele frequencies (Devlin and Risch 1995). However, only three measures have been commonly used by the scientific community.

The **D' measure** is one of the commonly-used measures of LD. it is derived from the covariance D, which is calculated as $D = (P_{11}P_{22} - P_{12}P_{21})$, where P_{11}, P_{22}, P_{12},

P_{21} are the frequencies of four haplotypes respectively in a 2X2 table (see Table 1.2). A standardization method is applied that divides D by the minimum value of [QR, (1-Q)(1-R)] when D is negative or by the minimum value of [Q(1-R),R(1-Q)] when D is positive (Lewontin 1964). This method to normalise D is less dependent upon allele frequencies (Hedrick 1987), although some dependency remains.

Another common measure is the **r^2 measure**, which is presented as $r^2 = \frac{D^2}{Q(1-Q)R(1-R)}$ (Hill and Robertson 1968). It can be used to test the statistical significance of LD with the total number of haplotypes. At equilibrium, D equals 0; thus D' and r^2 equal 0 too. However, in some cases, these two measures may not be consistent with each other (Pritchard and Przeworski 2001).

The **ρ measure**, proposed by Collins and Morton 1998 , is based on population genetics theory. It is calculated as $\rho = \frac{D}{Q(1-R)}$, where D is the covariance; Q is the frequency of the putative youngest allele and (1-R) is the frequency of one of the marker alleles at a particular locus (See Table 1.2). An interchange process of the frequencies of four haplotypes is performed in order to ensure that Q < (1-Q), R, (1-R) and that D>0. The measure ρ is equivalent to the absolute maximum value of D' in a random sample, but accommodates case enrichment in case/control samples. When modelling the decline of LD with distance, ρ yields the smallest error variance compared to other metrics (Morton et al. 2001). This model will be described in the next section.

Table 1.1 A variety of measures of LD

Definition	Symbol	Estimate $\hat{\psi}=D/C$
Covariance	**D**	$D=\lvert\pi_{11}\pi_{22}-\pi_{12}\pi_{21}\rvert$
Association	**ρ**	$D/Q(1-R)$
Correlation	**r**	$D/\sqrt{Q(1-Q)R(1-R)}$
Regression	**b**	$D/R(1-R)$
Frequency difference	**f**	$D/Q(1-Q)$
Delta	**δ**	$D/Q(1-R-Q+RQ+D)$
Yule	**y**	$D/[2Q(1-Q)R(1-R)+D(1-2Q)(1-2R)+2D^2]$

Table 1.2 Haplotype frequencies for a 2X2 table

		Locus A		
		1	2	
Locus B 1	Observed	P_{11}	P_{12}	$Q=P_{11}+P_{12}$
	Expected	QR+D	Q(1-R)-D	
2	Observed	P_{21}	P_{22}	$1-Q=P_{21}+P_{22}$
	Expected	(1-Q)R-D	(1-Q)(1-R)+D	
		R	1-R	N
		$=P_{11}+P_{21}$	$=P_{12}+P_{22}$	=1

The actual haplotype frequencies and the expected haplotype frequencies for two alleles at each of two loci. The expected haplotype frequencies are given at equilibrium (D=0). The marginal frequencies Q, 1-Q, R, and 1-R, represent the allele frequencies.

1.2.3 Modelling LD

The covariance D can be modelled by $D_t = D_0(1-\theta)^t$ (Falconer and Mackay 1960), where D_0 is the disequilibrium in generation 0, θ is the recombination rate per generation and t is the number of generations since a mutation took place at t = 0. When θ is small and t is large, the equation can be simplified as $D_t = D_0 e^{-\theta t}$, which describes the exponential decline of LD with recombination and generations. In addition, the equation assumes the constant recombination rate and constant population size in every generation (Jorde 2000). In fact, most populations have undergone rapid population growth.

The expected value of r^2 can be written as $E(r^2) = \frac{1}{1+4N_e\theta}$ (Ota and Kimura 1971; Pritchard and Przeworski 2001), where θ is the recombination rate per generation and N_e is the effective population size. The equation considers the population size during each generation, which is proportional to the time since a mutation occurred (Hill and Robertson 1968; Kaplan et al. 1995; Jorde 2000). N_e is the harmonic mean of the population size of each generation (Gillespie 1998), so a dramatic decrease in the population size over one generation would have much impact on the extent of LD. This is a "population bottleneck" (Wright 1969). This formula has also been used frequently with coalescence theory to estimate the population recombination rate (Fearnhead and Donnelly 2001; Li and Stephens 2003; McVean et al. 2004).

The Malecot model (Malecot 1948) was first applied by Collins and Morton 1998 to describe the relationship between LD and distance (See Figure 1.2), which is written as $\rho = (1-L)Me^{-cd} + L$. In the equation, d is the distance

between two loci; L is the residual LD at large distance, referring to the bias; M is the association at 0 distance. M is 1 if the youngest allele is monophyletic and less than 1 if it is polyphyletic. ε is proportional to the product of recombination and time. The parameter M is affected by the population size and mutation rate (Morton et al. 2001). This formula estimates $\varepsilon d \approx \theta t$ which is more appropriate for modelling LD (Collins and Morton 1998).

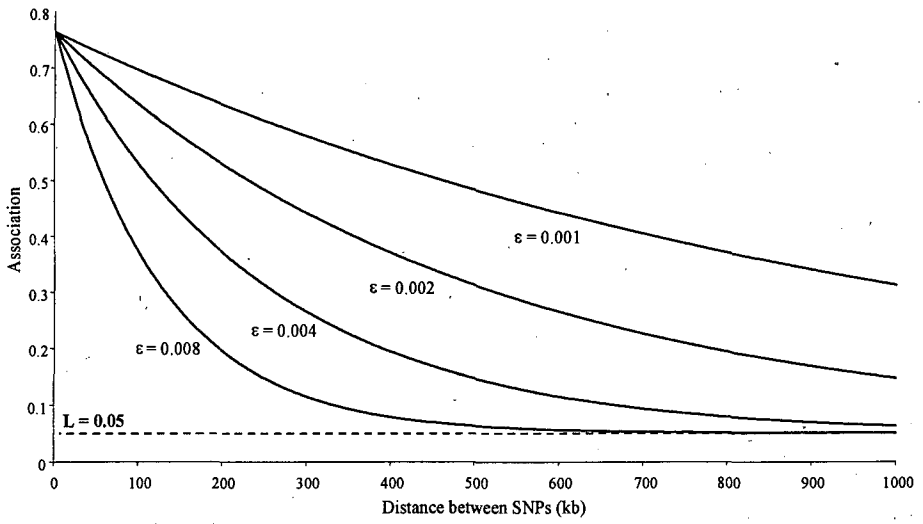


Figure 1.2 An example of the Malecot model where M=0.75, L=0.05, and a range of values for ε

1.2.4 Linkage Disequilibrium maps (LD maps)

The term of "LD map" is commonly used to describe LD patterns for a particular region or a whole chromosome. The most frequent approach is the use of D' or r^2, which plots average values in a sliding window against the corresponding physical locations in kilobase (kb) (Dawson et al. 2002; Taillon-Miller et al. 2004; Miretti et al. 2005). However this approach does not provide the relative location for each locus and smooths the LD patterns. The construction of LD maps has

been proposed by Maniatis et al. 2002 . Such LD maps have additive distances and locations in linkage disequilibrium units (LDUs) for all markers, which make LD maps unique compared to other alternative maps (Maniatis et al. 2002; Zhang et al. 2002a).

This method estimates the parameter ε of the Malecot model (See 1.2.2) for each interval by fitting the model to all marker-by-marker measures informative for that interval. The length of the i^{th} interval is computed as $\varepsilon_i d_i$ in LDUs, where ε_i is the Malecot parameter and d_i is the length of the interval on the physical map in kb. The total map length for a region is $\sum \varepsilon_i d_i$, which is the sum of the length of all intervals in this region (See Figure 1.3). An LD map (See Figure 1.4) exhibits block-step structures, in which blocks (i.e., $\varepsilon_i=0$) represent the regions of high LD and steps (i.e., $\varepsilon_i>0$) represent the regions of low LD (Maniatis et al. 2002; Zhang et al. 2002a). A value of $\varepsilon_i d_i>2.5$ indicates "a hole" in the map. The mean of ε for a region is computed as $\frac{\sum \varepsilon_i d_i}{\sum d_i}$. The swept radius is defined as 1/ε, reflecting the extent of useful LD.

An LD map is a very useful tool for association studies. It can also be used to determine suitable marker densities, compare populations and detect selective sweeps and other phenomena of evolutionary interest (Ennis et al. 2001).

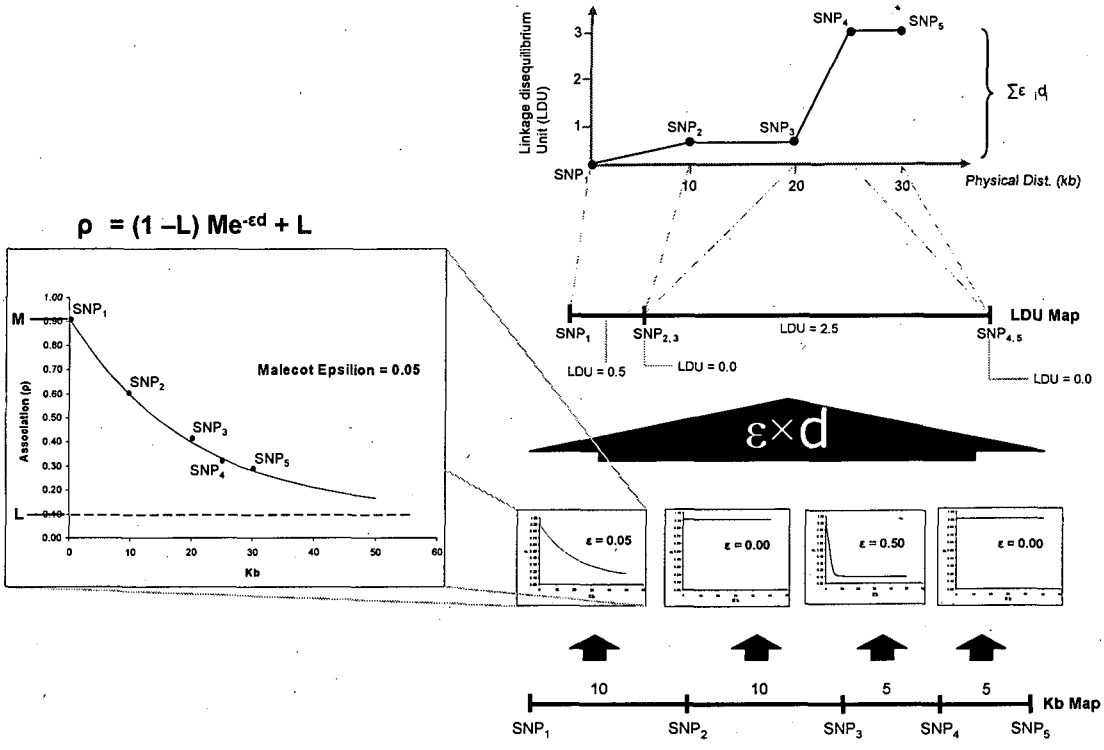


Figure 1.3 The construction of LD maps

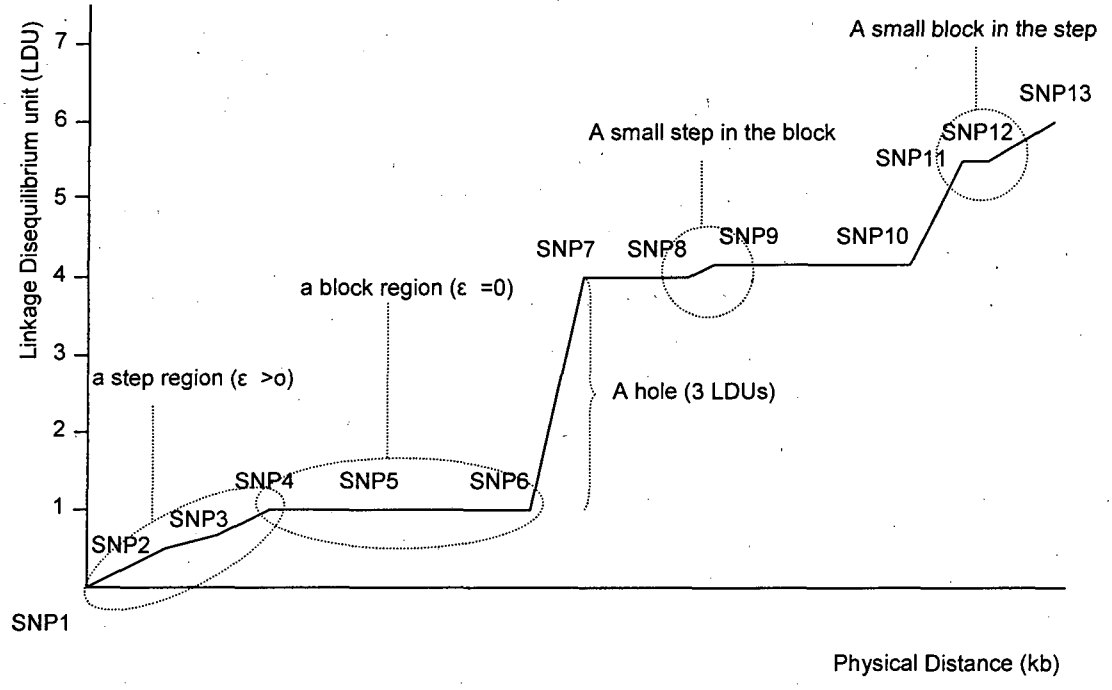


Figure 1.4 An illustration of an LD map

1.3 LD patterns in the human genome

1.3.1 Introduction

The success of association studies for disease gene mapping depends on knowledge of the LD structure. However, the extent of LD varies across the genome and in different populations. The international HapMap project (Consortium 2003) that genotyped more than five million SNPs in four different populations has provided useful data to understand haplotype, recombination hotspots and LD between different individuals and different populations. These data are also very suitable for the construction of a genome-wide LD map.

1.3.2 The patterns of LD in the genome

A simulation study by Kruglyak (1999) suggested that the extent of "useful" LD is less than 3 kb. However, this study did not take into account the effects of natural selection and demographic history in populations (Thompson and Neel 1997; Collins et al. 1999). Several empirical studies have found genomic regions of long-range LD in many populations (Collins et al. 1999; Huttley et al. 1999; Reich et al. 2001; Abecasis et al. 2001b). A block-like LD structure with limited haplotype diversity was first described on chromosome 5q31 (Daly et al. 2001). A study of chromosome 21 also found few haplotypes in each LD block (Patil et al. 2001). One study of the Major Histocompatibility Complex (MHC) on 6p21.3, using sperm typing techniques, showed that the areas of LD breakdown correspond precisely to recombination hotspots (Jeffreys et al. 2001). This suggests that recombination plays an important role in determining LD patterns. Dawson et al. 2002 studying chromosome 22, also reported a correlation between the intensity of LD and recombination. Maps constructed from the same data in the two published papers (Daly et al. 2001; Jeffreys et al. 2001), illustrate

block-step structures that match perfectly with their results (Zhang et al. 2002) (See Figure 1.5).

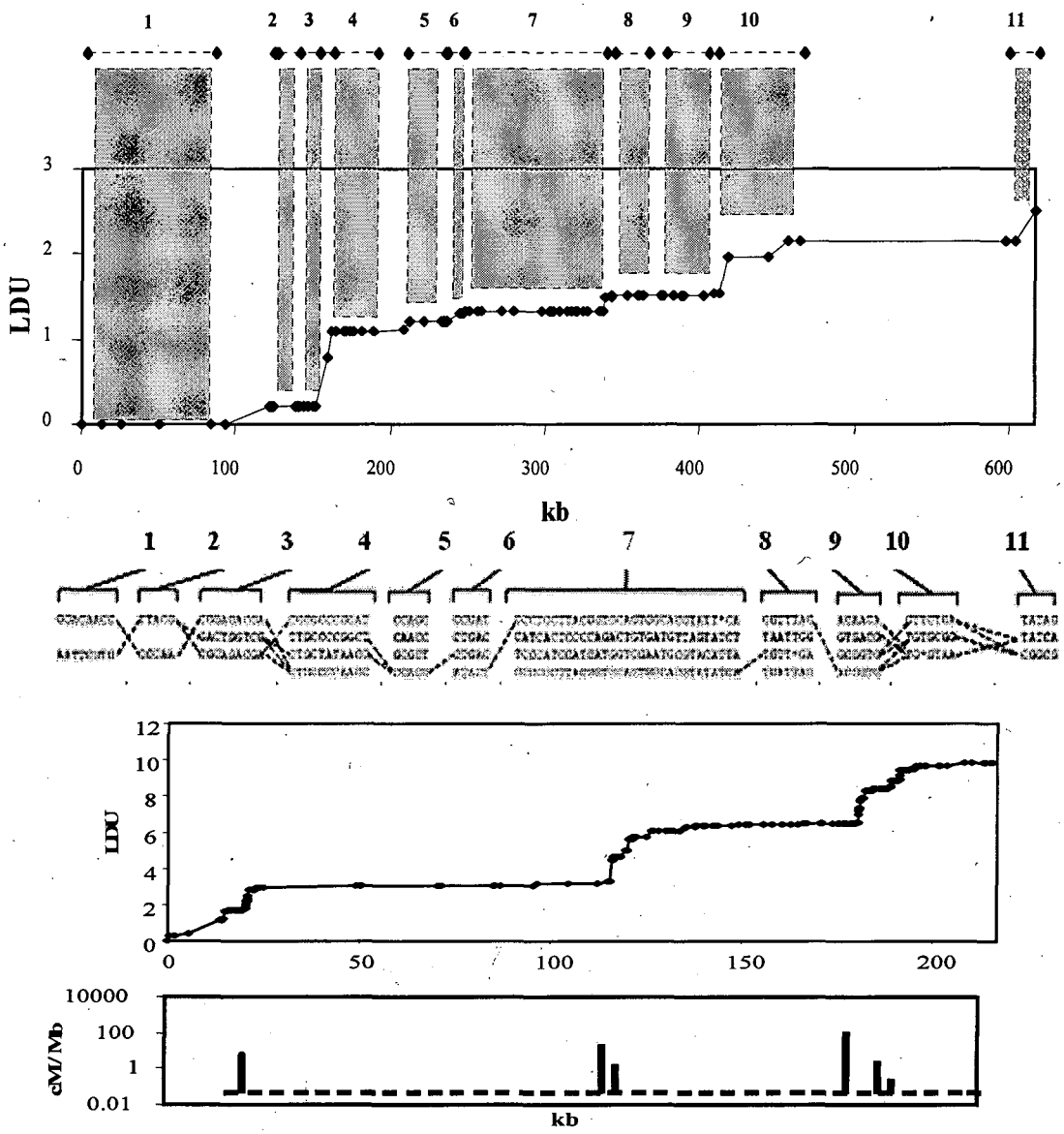


Figure 1.5 Remarkable agreement between LD maps and other results

The blocks in the LD map of 5q31 agree remarkably well with the 11 haplotype blocks inferred by Daly et al. 2001 (the upper figure). The positions of the steps in the LD map of 6p21.3 correspond to the sites of the recombination hotspots reported by Jeffreys et al. 2001 (the lower figure). The source is Zhang et al. 2002a

1.3.3 The patterns of LD in different populations

To understand more about LD patterns, researchers have investigated more regions and different populations. Several studies have found that the extent of LD is greater in non-African populations than in African populations (Gabriel et al. 2002; Altshuler et al. 2005; De La Vega et al. 2005). A study of the Finnish population found that the extensive LD blocks in the young sub-isolates are much longer than in the general Finnish population (Varilo and Peltonen 2004). This study confirmed the previous finding of population isolates exhibiting more extensive LD (Service et al. 2001). An explanation is that non-African populations and population isolates have experienced more intense population bottlenecks, through processes such as migration, which reduced their population size dramatically in the past (Lonjou et al. 2003). Other environmental and demographic changes such as famine, war and epidemic diseases can also generate new population bottlenecks (Slatkin and Veuille 2002; Morton 2005). Despite the variations in the LD patterns between different populations, there is a remarkable agreement in the locations of the common recombination hotspots in different populations (De La Vega et al. 2005). Although the same recombination hotspots exist in most populations, Kauppi et al. 2003 have found that haplotype composition in the same blocks can be different between populations. This result has been confirmed by several studies (Crawford et al. 2004; Liu et al. 2004). In addition, long-range haplotypes may not always break at recombination hotspots (Altshuler et al. 2005).

1.3.4 The HapMap project

Initially, the patterns of LD were studied in small regions of the genome or in a single population using low marker densities. These studies provided an important contribution to our initial understanding of the structure of LD. Most importantly, they motivated the international collaboration of the HapMap project (Consortium 2003), which aimed to develop a map describing common haplotypes in the human genome. The entire human genome contains approximately 10 million common SNPs that constitute 90% of the variation in populations (Kruglyak and Nickerson 2001; Reich et al. 2003). The Phase I data in the HapMap Project contains at least one million SNPs (one SNP per 5 kb) across the whole genome. The latest released Phase II data includes an additional 4.6 million SNPs, giving a density of one SNP per 1 kb. These SNPs are genotyped in the 269 DNA samples: 30 trios (two parents and a child) from a US Utah population with Northern and Western European ancestry; 30 trios from Yoruba people in Ibadan, Nigeria; 44 unrelated Japanese in Tokyo, Japan; and 45 unrelated Han Chinese in Beijing, China. These four populations are abbreviated as CEU, YRI, JPT and CHB respectively. The HapMap data is very valuable resource that will enable understanding of the genetic variation, LD structure and recombination hotspots across the human genome and in different populations. These data can also be used to construct genome-wide or population-specific LD maps.

1.4 Association studies for identifying causal variants

1.4.1 Introduction

The principle of association studies is to detect genetic markers that are associated with disease phenotype. It compares the difference in allele frequencies of genetic markers between affected individuals (cases) and healthy individuals (controls). Therefore, a case-control study design is commonly used for association studies. If a marker exhibits a significant difference in allele frequency between cases and controls, this marker may be close to a causal allele. However, there may be a spurious association caused by genotyping or sampling errors. This chapter introduces several common approaches for mapping disease genes, including single SNP tests, haplotype analyses, and composite likelihood methods. Their advantages and challenges are also described. All of these approaches have been successful in localising several major genes. However, the effectiveness of these approaches is still unknown when applied to common diseases.

1.4.2 Single SNP tests

A chi-squared test between affection status and every SNP in the data is the simplest and the most common method used in association studies. SNPs are often chosen from the coding regions under the assumption that any change in sequence of amino acid would lead to a change in protein function, which is likely to cause diseases (Cargill et al. 1999; Botstein and Risch 2003). However, several studies have shown that some SNPs in non-coding regions may also be associated with disease (Duan et al. 2003; Lin et al. 2003; Tokuhiro et al. 2003). The use of a single SNP test has several disadvantages. The main drawback is that it does not take into account the LD between SNPs. Marker SNPs that are close together, are correlated with one another and therefore, it is difficult to determine which SNP

has an effect on the disease phenotype. A false positive association can also arise from population stratification, improper case-control matching, or chance due to multiple testing (Zondervan et al. 2002; Cardon and Palmer 2003). It is generally believed that analysing multiple SNPs simultaneously is more efficient and appropriate than a single-SNP test.

1.4.3 Haplotype analyses

Haplotype analyses have received a great deal of attention. A review of the literature by Salem et al. 2005 has reported more than 40 haplotype methods for association mapping between cases and controls. A haplotype can be estimated either molecularly or probabilistically (Yan et al. 2000; Douglas et al. 2001; Niu 2002). However, molecular methods are expensive and labour-intensive. Probabilistic methods, statistical inference, such as Bayesian methods (Stephens and Donnelly 2003) and Expectation-Maximisation (EM) algorithm methods (Hawley and Kidd 1995) have been suggested but using pedigree analyses can obtain haplotypes with greater accuracy than random SNPs (Tishkoff et al. 2000; Zhang et al. 2001; Schaid 2002; Thomas et al. 2004).

Since SNPs in the same LD block are highly correlated, many have redundant information and can be eliminated. However, the highest power in Haplotype analyses is achieved when the disease SNP itself is typed. In addition, most studies infer block structure and boundaries by their own definitions. Some studies have used pairwise measures to determine blocks, whereby all pairwise coefficients exceed a predefined threshold (Daly et al. 2001; Reich et al. 2001; Gabriel et al. 2002). Other studies have defined blocks by using a small number of haplotypes that account for a high proportion of observations (75~90%) (Johnson

et al. 2001; Patil et al. 2001; Zhang et al. 2002b). However, block definitions vary depending on the threshold used, and hence are subjective and arbitrary (Cardon and Abecasis 2003; Tapper et al. 2003).

1.4.4 Composite likelihood methods

An alternative approach that uses a composite likelihood approach and the Malecot model under different hypotheses has also been proposed (Maniatis et al. 2004; Maniatis et al. 2005). This approach utilises LD information from an LD map and estimates a maximum-likelihood location of a causal polymorphism. This method was firstly applied in the CYP2D6 region which is associated with the poor drug metabolizing activity. It was shown that an LD map is more powerful compared to a physical map, which yields an error of only 15 kb away from the real causal variant (Maniatis et al. 2005).

1.4.5 Alternative approaches

There are other alternative approaches for association mapping such as meta-analysis and admixture mapping. Meta-analysis (Hirschhorn et al. 2002; Lohmueller et al. 2003; Hirschhorn 2005) is a common method that utilises the results from published studies in order to validate findings and significance. This method requires detailed information on the sample and methodology that are used for each study (Craddock et al. 2001). However, the main drawback of this approach is that the sample size and statistic metrics vary substantially among studies and investigators often fail to report the negative results. Another approach is admixture mapping, which is also known as mapping by admixture linkage disequilibrium (MALD) (Patterson et al. 2004). This approach localises disease-causing variants that are different in the frequency between two

historically separated populations. It is expected that affected populations derived from the recent mixture of two or more ethnic populations should have higher frequency of the alleles near the disease gene, which are co-inherited with the disease genes from the ancestral population that carries more disease-susceptibility alleles. The advantage of this approach is that it greatly reduces the number of markers required for genome-wide scans. However, a dense map that identifies SNPs with significant difference in allele frequency between two populations is required (Smith et al. 2004).

1.5 Genome-wide association studies for Common diseases

1.5.1 Introduction

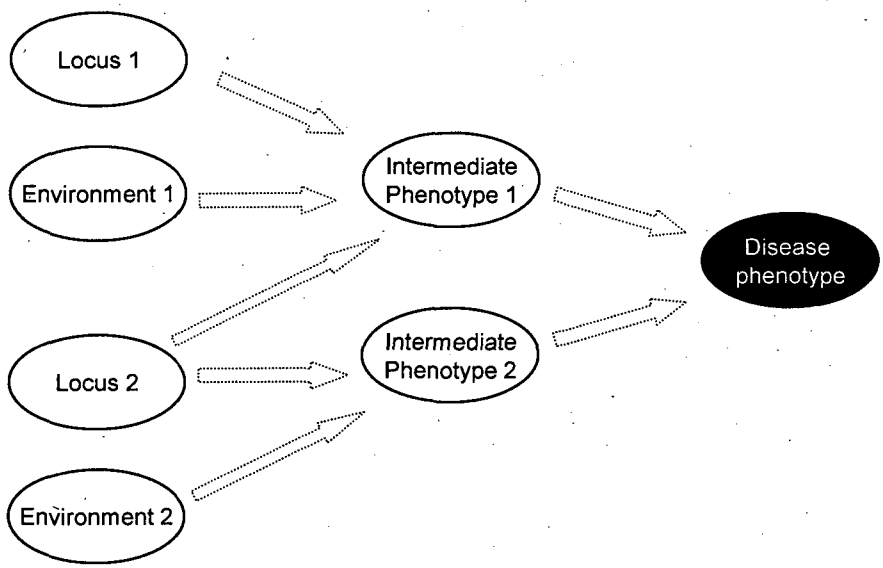
Linkage studies for single gene Mendelian disorders have been very successful but mapping genes for common diseases is extremely challenging (Altmuller et al. 2001). Recent advances in high-throughput genotyping techniques (Syvanen 2005) and the abundance of SNP resources, such as dbSNP. have made genome-wide association (GWA) studies feasible. The advantage of GWA is that investigators do not need to determine possible candidate regions ahead of the genome-wide screen. Such studies examine thousands of SNPs across the whole genome in order to identify short regions that harbour susceptibility loci for common diseases. GWA scans are potentially powerful so the development of analytical tools is necessary in order to ensure success in disease gene mapping.

1.5.2 Common diseases

Unlike single gene disorders showing Mendelian inheritance patterns, common diseases are more complex. Such diseases are influenced by a mixture of multiple genetic variants and environmental risk factors (Figure 1.6); therefore, the contribution of each genetic variant is relatively small. For example, more than 150 rare high-risk alleles have been identified for Alzheimer's disease, but all of these alleles contribute to less than 5% of the disease cases; the remaining 95% of the disease cases arise from complex interactions between environmental and genetic factors of each individual (Rocchi et al. 2003). Recent studies have suggested that common genetic variants account for a proportion of common diseases, which is the common disease/common variant (CD/CV) hypothesis (Reich and Lander 2001). It is still debatable whether most of complex disease is caused by variants that are common or rare (Risch and Merikangas 1996; Pritchard 2001; Pritchard and Cox 2002; Smith and Lusis 2002). However, recent studies(Consortium 2007) have revealed a number of common causal variants.

Several studies have suggested multivariant approaches, such as logistic regression (Hosmer and Lemeshow 2000) and multi-factorial methods (Ritchie et al. 2001) can be applied for the identification of gene × gene and gene × environment interactions. These approaches have been used in several studies of common diseases such as hypertension (Clark et al. 2000) and breast cancer (Ritchie et al. 2001). However a large sample sizes are still needed when there are many independent variables (Moore and Williams 2002). Furthermore, the environmental variance may be minimised by matching cases and controls. The selection of extreme phenotypes has been suggested in order to reduce the confounding effects with environmental risk factors (Long and Langley 1999).

Nevertheless, the study of gene × environment interaction can only be meaningful when the genes of the phenotype in question are well established (Figure 1.6).



Souce: Carlson et al. 2004

Figure 1.6 The complex interplay of genetic and environmental factors.

1.5.3 Genome-wide association studies

Studies of common diseases using a small number of markers genotyped in few candidate regions have reported several significant associations with diseases. Although some results could explain a proportion of the effects of disease phenotype, many of them have been difficult to replicate (Hirschhorn et al. 2002; Page et al. 2003). This is perhaps due to the existence of other common variants with modest phenotypic effects that lie outside these candidate regions (Lohmueller et al. 2003); therefore, analysis of the entire genome could provide robust results. Therefore GWA studies are potentially powerful.

GWA involves multiple tests across the genome and hence the inflation of the number of false positives is inevitable. Investigators need to adjust the significance thresholds to control the false positive rate. The most popular method for p value correction is the Bonferroni correction. However, this method is very conservative because of the very large number of SNPs that are involved (e.g. at least 500,000 SNPs across the genome)

The false discovery rate (FDR) has been proposed to control for multiple testing (Morton 1955; Benjamini et al. 2001). The FDR is described as the proportion of false positives in all significant results (See Figure 1.7). It has been frequently applied to microarray analyses. The success of employing FDR depends mainly on knowing the distribution of true significant results among all tests, but this is usually unknown. FDR methods operate under the assumption that nominal p values under the null hypothesis are uniformly distributed (Storey and Tibshirani 2003). However, a uniform distribution is not achieved in most cases due to stochastic variation; therefore, failure to take this into account will inflate the nominal significance (Yang 2004). Several programs, such as Q value (Storey and Tibshirani 2003), BUM (Pounds and Morris 2003), SPLOSH (Pounds and Cheng 2004) and LBE (Dalmasso et al. 2005), have been proposed to estimate FDR. They are different in their ways of modelling the distribution of p values. For example, Q value assumes a uniform distribution of p values; BUM uses a beta-uniform function for the distribution of p values; and LBE is based on the expectation of the transformed p value.

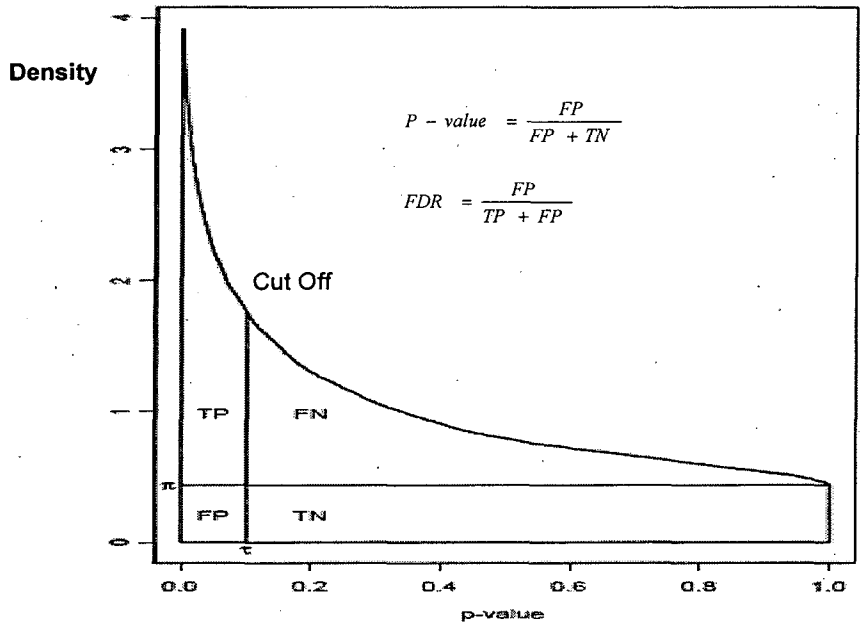


Figure 1.7 The Estimation of False Positive Rate (P value) and False Discovery Rate (FDR).

The areas of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) give the estimation of the false positive rates (p value) and the false discovery rate (FDR). τ is the threshold for determining significant results. π is the proportion of true null results among all tests.

Localising causal genetic variants for common diseases is very challenging. There are several methods that can be employed for association mapping. In order to ensure the success in disease gene mapping, the LD pattern needs to be taken into account. A variety of methods using LD for mapping common disease have been proposed, but the effectiveness of these methods needs to be examined in a large data set in GWA scan. GWA studies with the use of genome-wide LD maps offer the greatest prospect to unravel the cause of common diseases.

Chapter 2 Strategies to construct a whole genome LD map

2.1 Introduction

A map describing the patterns of LD in the human genome is a powerful tool for LD mapping and population genetics. LD maps identify regions with recombination hotspots where higher SNP density may be required for localizing causal polymorphisms. Differences in the map lengths between populations reflect different population histories, in which populations experienced one or multiple different population bottlenecks. Most importantly, an LD map plays an equivalent role to a linkage map; a linkage map provides the genetic location for each SNP in centimorgans (cM) whereas an LD map provides that for each SNP in LDUs. The genetic location, in either cM or LDUs for each SNP, can be used to predict possible locations of causal polymorphisms by linkage and association respectively, but the location on the LDU scale has much higher resolution than that on the cM scale.

The data for LD map construction can be phase-known haplotype data or phase-unknown genotype data from a sample of unrelated individuals. In the HapMap project (Consortium 2003), the phase I data contains at least 1 million SNPs and the phase II data contains an additional 4.6 million SNPs. These SNPs were genotyped in 269 individuals from four different populations from Utah (CEU), Japanese (JPT), Chinese (CHB), and Yoruban (YRI) residents (See Chapter 1). The abundant SNPs and population-specific data sets make the HapMap data very suitable for constructing an LD map for the whole genome and for different populations.

An LD map is constructed from multiple pairwise data from any pair of SNPs. A gap between any two adjacent SNPs is defined as "an interval". If there are n SNPs in a region, there are n-1 intervals. The total number of possible pairs between any two SNPs in the region is $\frac{n(n-1)}{2}$. For estimating the LDU length for an interval, any pairs of SNPs that span and include this interval contain part of the LD information for this interval, but the information declines with increasing distance. Pairs within a certain distance are defined as "the informative pairs" related to the interval (See Figure 2.1).

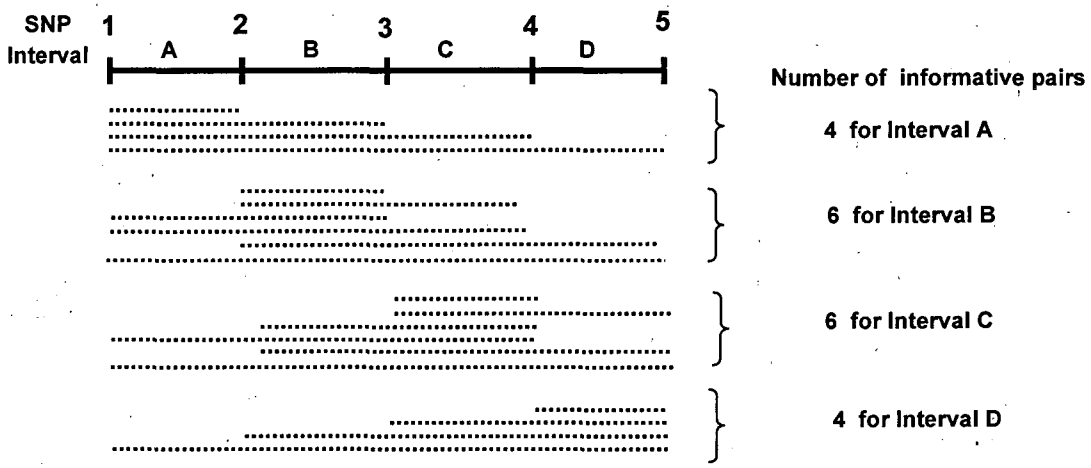


Figure 2.1 The number of Informative Pairs in different intervals

A challenge for the construction of LD maps is the management of the computational load posed by the volume of pairwise data, which leads to a poor computer performance and an insufficient memory failure. Therefore it is necessary to optimise the numbers of SNP pairs used in analyses of a large data set.

There are three methods to remove redundant SNP pairs during the preparation of a data set.

1) **Separating a large data set into smaller sub-sets**

   A large data set with a large number of SNPs could be separated into several sub-sets each containing fewer SNP pairs.

2) **Reducing the SNP density**

   The SNP density is calculated by the number of SNPs over the physical distance (kb) within specific genomic region. When the SNP density is reduced, the total possible pairs are reduced sharply.

3) **Excluding uninformative pairs**

   In practice, all possible pairs could be used in a data set. However, uninformative SNP pairs may be removed in order to reduce the computational load. At a very large distance (kb), SNP pairs may contain negligible information about the LD structure of a given interval and are thus uninformative, and can be removed from the data set (See Figure 2.2).

These three methods can be used separately or together to reduce the number of pairs within a data set. However reducing a number of pairs runs the risk of reducing the quality of an LD map. Even though we use the same raw data, LD maps will not be identical if the data sets are made under different methods and

limitations, and differences might indicate significant losses of information.

In this study, I have evaluated the impact of these three methods on LD map construction. I also have compared the impact on two regions with very different LD patterns. Some useful criteria to evaluate the quality of an LD map are described and an optimal strategy to construct an LD map is suggested in the chapter.

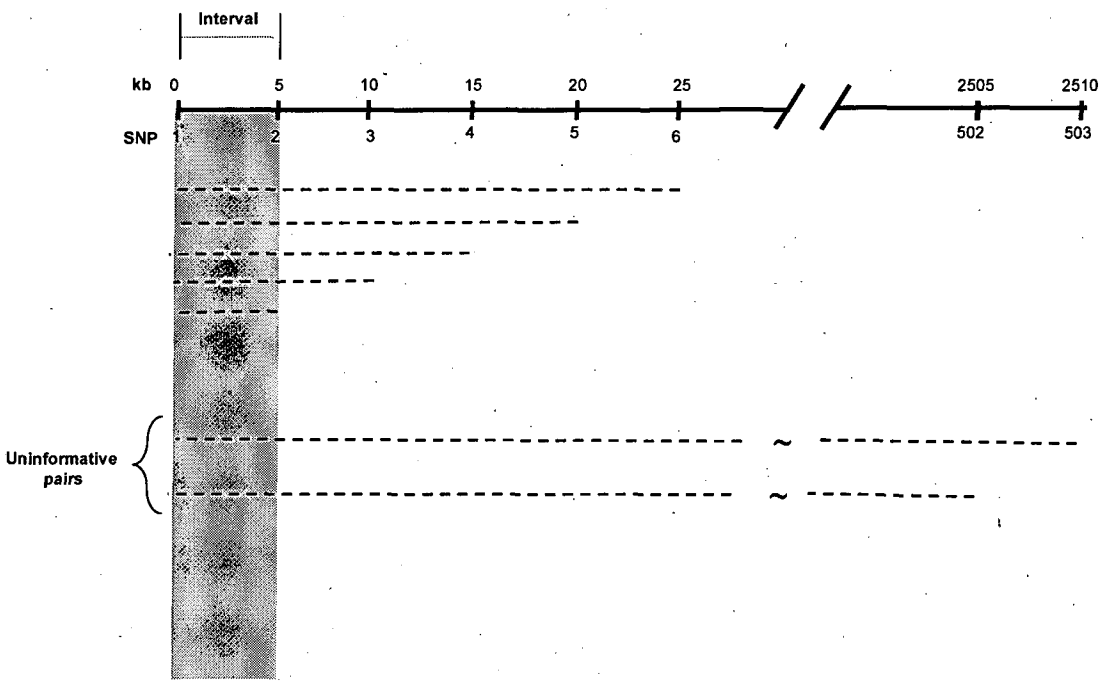


Figure 2.2 Uninformative pairs for the interval

Any pairs that are at very large distance beyond the extent of LD are defined as uninformative pairs. For instance, there are two uninformative pairs at the bottom because the physical distance between two SNPs in kb is very large (> 2500 kb). The average extent of LD in the human genome is ~ 50 kb.

2.2 Materials and Methods

2.2.1 The LDMAP program

I used the LDMAP program (http://cedar.genetics.soton.ac.uk/pub/ Program /LDMAP ; Maniatis et al. 2002) to construct LD maps based on pairwise SNP data. The LDMAP program estimates the parameter ε from each interval by fitting the Malecot model, $\rho = (1-L)Me^{-\varepsilon d} + L$, to pairwise measures that are informative for each interval (Fig 2.2). The length of the i^{th} interval is computed as $\varepsilon_i d_i$ in LDUs and the total map length for a region is $\sum \varepsilon_i d_i$ (See 1.2.4).

2.2.2 The study samples

The study chromosome

Firstly, I constructed an LD map using the genotype data for chromosome 22 from the CEU samples in the HapMap data (Phase II), which was released in October 2005. This data set included 30 trios (parents and a child), but only the 60 unrelated parental samples were used. SNPs with minor allele frequencies (MAF) less than 5% and any with significant deviations from Hardy-Weinberg equilibrium (HWE) ($X^2 > 10$), were removed from the data (Gomes et al. 1999). A total of 27060 SNPs were genotyped in the sample. The physical length of chromosome 22 is ~35 Mb (34,924 kb). The total map length of the LD map for this chromosome is 1137 LDUs. The LDU/Mb is approximately 32.

The study regions

The LD map of chromosome 22 was used to indicate the regions of interest for this study. I selected two regions with the same physical length (5 Mb) with very different magnitudes of LD (See Figure 2.3 and Table 2.1).In this chapter, "the low LD region" and "the high LD region" refer to these two regions. Only the genotype

data in these two regions were used on the samples in this study.

1) The low LD region has a step-like structure with 244 LDUs. It is located between 21.5 Mb and 26.5 Mb. This region was genotyped with 4499 SNPs at an average density of one SNP every 1.1 kb. The LDU/Mb is approximately 48.8.

2) In contrast, the high LD region is a block-like structure with only 76 LDUs. It is located from 26.5 Mb to 31.5 Mb almost adjacent to the low LD region. This region was genotyped with 3124 SNPs at an average density of one SNP every 1.6 kb. The LDU/Mb is approximately 15.2.

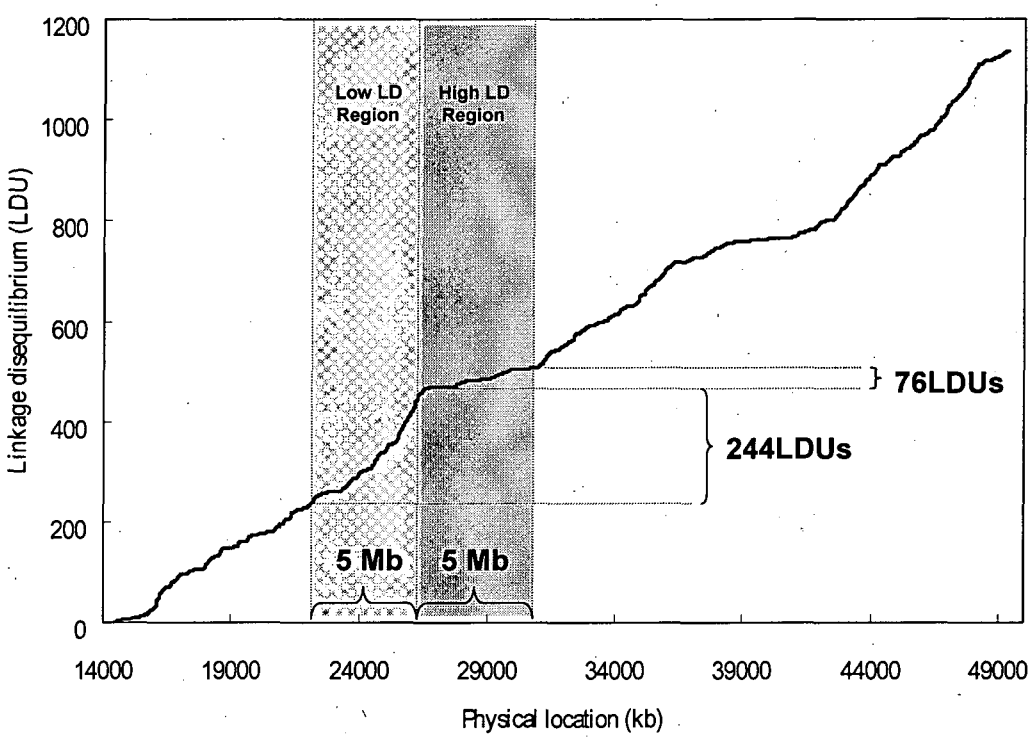


Figure 2.3 The study regions chosen from an LD map of chromosome 22

A low LD region with 244 LDU and a high LD region with 76 LDU were selected for the study regions. These two regions have the same physical length of 5 Mb.

Table 2.1 The descriptions for the study regions

Description	Low LD region	High LD region
Physical range	5 Mb	5 Mb
Map Length	244 LDU	76 LDU
Number of SNPs	4499	3124
ε (kb)	0.03	0.003
Swept radius (1/ε)	32.8 kb	335.8 kb

Although two regions are close together with the same physical length, they exhibit very different patterns of LD, which is reflected in their LDU map length, the parameter ε and the swept radius. The swept radius, 1/ε, is the distance at which LD declines to e^{-1}~0.37 of its original value. It is usually described as the average extent of useful LD.

2.2.3 LD maps based on different data sets

Making Different data sets

By taking the two contrasting 5 Mb regions shown in Figure 2.3, the properties of LD maps using their respective data sets under alternative approaches can be examined. Here are the detailed descriptions about how these data sets were made.

1) The SNP density:

Data sets for the low and high LD regions were modified by gradually reducing their SNP densities from the sample data. In all cases, the first and the last SNP were chosen so that each data set maintained constant length in physical distance (kb) for the region. Other SNPs were then chosen to satisfy alternative SNP density requirements. For example, to achieve a 1 SNP per 2 kb density, the second SNP was chosen precisely at a location 2 kb away from the first SNP. If there was no SNP at precisely 2 kb distal to the first SNP, the SNP which is the closest to the location was chosen. The new chosen SNP was

then used to select another SNP 2 kb away from the chosen SNP. This process was repeated along the region until the last SNP was selected. The data sets were made from the two samples using the SNP density at 2kb, 3kb, 4kb, 6kb, 8kb, 10kb and 12kb per SNP.

2) Limiting the informative pairs

There are two constraints in the LDMAP program for removing uninformative pairs from all possible pairs. The first one is the maximum distance in kb between any pair of SNPs (max_dist) and the other is the maximum number of intervals between any pair of SNPs (max_intv). The default values are 500 kb for the max_dist and 100 for the max_intv. Using the max_dist at 500 kb means that if the distance between a pair of SNPs is over 500 kb, this pair will be removed from the data set. Using the max_intv at 100 means that if a pair of SNPs is separated by more than 100 intervals, this pair will be removed. If both constraints are applied, any pairs that contravene either two will be removed. For simplicity, I only used the max_intv to constrain the informative pairs in the study. The data sets were made from the two samples using the max_intv at 25, 50, 75, 100, 125, and 150. The max_dist remained at 500 kb in all data sets.

3) The number of segments

If an original data set contains a large number of SNPs, it can be separated into several sub-data sets with approximately the same number of SNPs. The LD maps based on different sub-data sets can be constructed separately and then connected together to form an integrated LD map. For example, if there are 1000 SNPs in a region, each segment contains 500 SNPs by dividing two segments, but 250 SNPs by dividing four segments. In this study, my two selected regions contain 4499 SNPs and 3124 SNPs respectively. The data sets

were made from each of the two samples using 2, 3, 4, 6, 8, 10, and 12 segments. The number of SNPs per segment in these data is 375 ~ 2250 and 260~1562 for the low LD and high LD regions respectively. In addition, each segment includes additional SNPs from its two neighboring segments. The region where additional SNPs are from is defined as an overlapping region. Each segment contains two overlapping regions except the first and the last segment which has only one overlapping region. The overlapping region is used for the connection of two segmental LD maps of two neighbouring segments. The number of additional SNPs in the overlapping region is set to a default of 25 SNPs. In this study, I used the same default overlap value for all the data sets that were made using the assembly method. When connecting all overlapping sections of segmental LD maps to form an integrated LD map, the length in LDU of each interval within the overlapping regions were replaced by the mean of the length of that interval from the two neighbouring segments (See Figure 2.4).

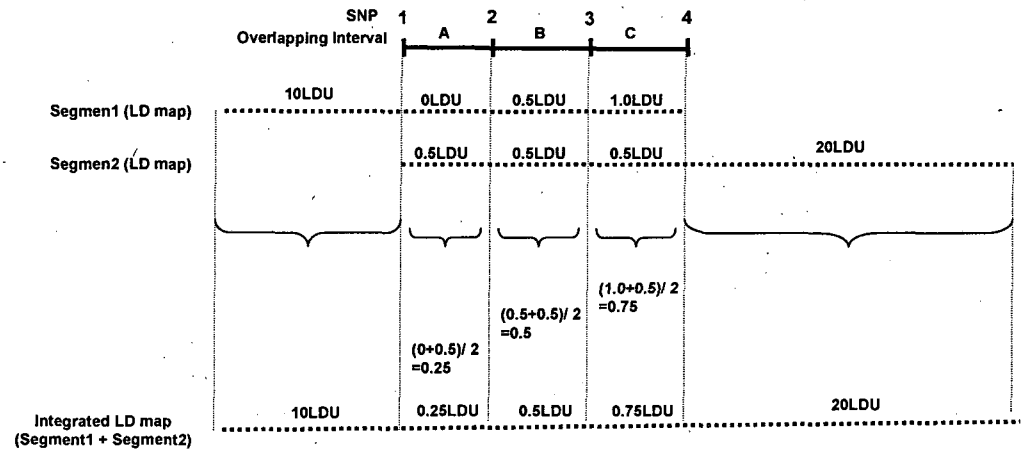


Figure 2.4 The mean of the LDU length of each interval within overlapping sections from two neighbouring segments.

Constructing LD maps from these data sets

22 LD maps for each of the two selected regions (high and low LD regions) were constructed using the LDMAP program. The data sets for the 22 LD maps were created under the criteria for reducing number of pairs, including 8 by reducing the SNP density, 6 by limiting the informative pairs and 8 by separating segments. For a given region, all LD maps had the same number of SNPs except the LD maps that were constructed using reduced SNP density. Therefore, inserting into the LD maps the SNPs removed when making a data set was necessary for the comparison of these LD maps. Linear interpolation was applied to give a relative LDU location from an LD map for these SNPs according to their kb location (See Figure 2.5). For example, if a given SNP being inserted is between SNP_1 and SNP_2, given two locations, kb_1, LDU_1 and kb_2, LDU_2 respectively for each SNP, the LDU location for this SNP, given kb_i for its kb location, is calculated as the equation,

$$LDU_i = LDU_1 + \left(\frac{LDU_2 - LDU_1}{kb_2 - kb_1}\right)(kb_i - kb_1)$$

By this interpolation method, all SNP removed during map construction were positioned back into the maps, but the length and the shape of the LD maps were not changed.

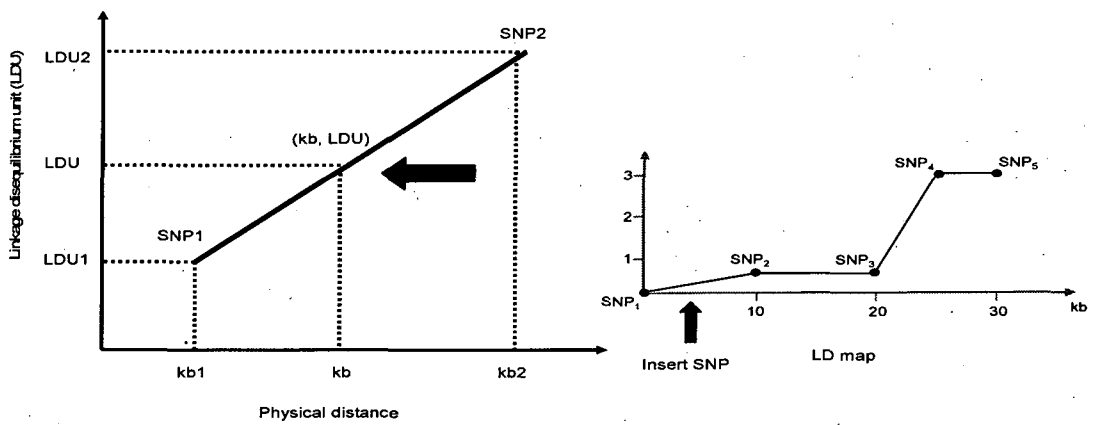


Figure 2.5 Linear interpolation method provides SNPs LDU locations based on their kb location

2.2.4 Comparisons between LD maps

Standard data sets and a default LD map

The quality of an LD map is considered to depend on how well it fits the pairwise data in a data set (Maniatis et al. 2002). If a particular LD map has a reduced error variance relative to other LD maps, this LD map is taken to have higher accuracy for the data set. To measure how well an LD map fits a data set, we can use the residual error variance, $V = \frac{-2\ln lk}{(n-m)}$, where $-2\ln lk$ is the composite log likelihood computed as $-2\ln lk = \sum k_\rho (\hat{\rho} - \rho)^2$; n is the number of pairwise data points in the data set; and m is the degrees of freedom referring to the number of parameters estimated. In the $-2\ln lk$, $\hat{\rho}$ is the association probability estimated from the 2×2 haplotype table; ρ is the predicted association probability given from the LD map using the Malecot model; and k_ρ is the information about $\hat{\rho}$, computed as $\frac{NQ(1-R)}{(1-Q)R}$ (See Table 1.2).

To compare alternative LD maps, its is important that all are evaluated against the same pairwise data set. In this study, five standard datasets were made; each of them including all SNPs but with different max_intv at 100, 200, 300, 400, and 500 respectively, yielding five residual error variances for different standard data sets

The LD map constructed from the data set using the default value of 100 max_intv was defined as "the default LD map". For simplicity, all LD maps were compared with the default LD map. The map length, the residual error variances and the block proportions were used in the comparison between these LD maps.

These three elements related to the quality of an LD map are described in detail in the following section.

The criteria for the comparison

Here I define three criteria in order to compare the difference between these LD maps.

1) The relative length:

   The map length of all LD maps was compared individually with that of the default LD map. The relative length is defined as the map length of each LD map divided by the map length of the default map.

2) The relative efficiency:

   This criterion is used for comparing the residual error variance of each LD map individually with that of the default LD map. The ratio (V_D/V_E) between the residual error variance of the default LD map (V_D) and each LD map (V_E) is defined as the relative efficiency. Five different standard data sets were used so that each LD map has five different values of the relative efficiency.

3) The block ratio:

   An interval in which $\varepsilon_i = 0$ is defined as a block here. The proportion of LD blocks is defined as the sum of all intervals in kb where $\varepsilon_i = 0$ divided by the entire length of the region. The block ratio is defined as the proportion of LD blocks in each of the alternative LD maps divided by the proportion of LD blocks of the default LD map.

2.3 Results

2.3.1 The impact on the relative map length

1) SNP density

For the low LD region, while reducing the SNP density from 1 SNP per 2 kb to 10 kb, the relative map length is between 0.91~1.03, but drops to 0.78 when the density is reduced to 1 SNP per 12 kb. However, for the high LD region, the relative map length tends to reduce gradually as the SNP density is decreased. The relative map length reduces to 0.82 when the density is reduced to 1 SNP per 12 kb (Figure 2.6 a). For the high LD region, using higher SNP density will break up large blocks into several smaller blocks. However, for the low LD region, the map length is limited by the fewer intervals in the region due to low SNP density.

2) Maximum interval between pairs of SNPs

When the max_intv is reduced, the relative map length tends to increase for the both regions, although this is more apparent for the low LD region. For the high LD region, the relative map length is 1.13 and 1.11 at the max_intv of 25 and 50 respectively, whereas it is 1.26 and 1.15 respectively for the low LD region. In other words, there is at least a 10% increase in the map length compared to the default map when using the max_intv as ≤50. The relative map length reduces gradually while the max_intv increases for the low LD region, but remains more stable while the max_intv is over 75 for the high LD region (Figure 2.6 b). A possible reason is that the estimation of the parameter ε may not be accurate if the max_intv is not large enough to cover the mean extent of LD.

3) Number of segments

The assembly method using different numbers of segments to construct the LD maps increases marginally the map length for the both regions, but this is less than 4%, compared to the default LD map (Figure 2.6 c). The reason that causes the map length slightly longer perhaps may be the same as using an insufficient max_intv. When we divide a larger segment into several smaller segments, the number of pairs used to estimate the parameter ε for the intervals at the end parts of each smaller segment is fewer than the requirement of the max_intv due to the truncated side of those intervals. Therefore, we may expect the map length to be longer when more segments used.

2.3.2 The impact on the relative efficiency

1) SNP density

The relative efficiency decreases gradually for the both regions when the SNP density is reduced, but the decline is more rapid for the low LD region. When the density is reduced to 1SNP per 12 kb, the relative efficiency for the low LD and high LD regions is 0.73 and 0.79 respectively (Figure 2.7 a). This trend of declining relative efficiency with the reducing SNP density is not different when using different standard data sets. Once again, reducing SNP density over the range examined decreases the relative efficiency by up to 20%.

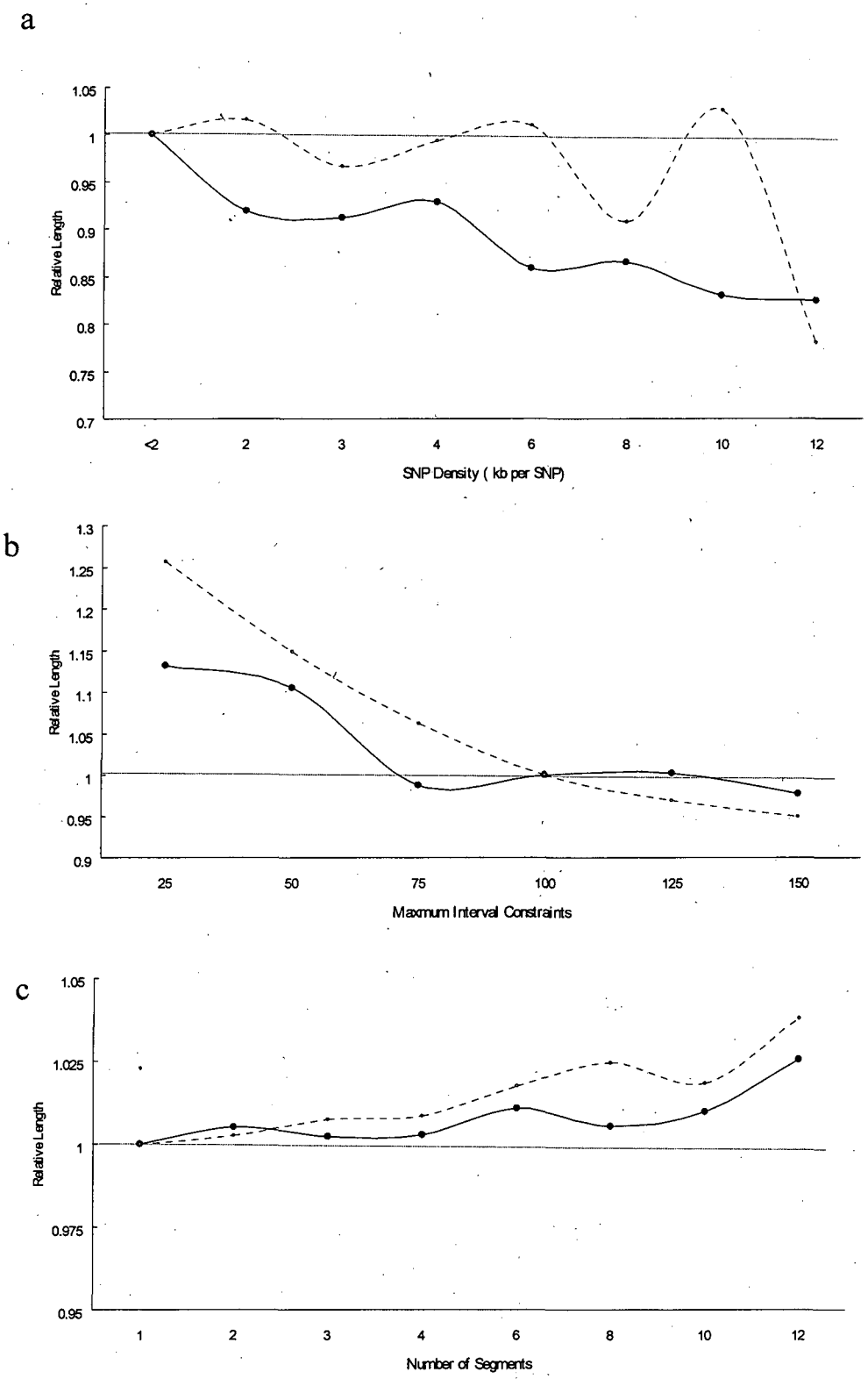


Figure 2.6 The impact on the relative map length

LD maps constructed based on alternative methods to reduce the pairwise data from the high LD region (solid line) and the low LD region (dashed line). The results are shown for a) SNP density; b) Maximum interval constraints; c) Number of segments.

2) Maximum interval between pair of SNPs

Fitting alternative maps to the standard data set using the 100 max_intv, the relative efficiency increases while the max_intv increases from 25 to 50, and is maximal at 75 intervals and declines slightly when the max_intv is over 100. Similar results are found for both regions (Figure 2.7 b). When fitting them to other standard data sets, for the high LD region, the relative efficiency increases dramatically between 25 and 75 intervals, but improves only slightly when the max_intv is over 100 (Figure 2.8 a). However, for the low LD region, the highest relative efficiency is evident at the 75 max_intv but it decreases slightly while the max_intv increases above 75 (Figure 2.8 b). The results also shows that using fewer intervals (the max_intv =25), the relative efficiency falls more rapidly for the high LD region than for the low LD region. The decline in relative efficiency when using large number of pairs may reflect the dependency between pairs which is reduced by using a smaller sub-set.

3) Number of segments

The assembly method has a very little impact on the relative efficiency for the both regions when divided by up to 6 segments, which is about 500~750 SNP per segment. However, only when the region is divided into 8 or more segments, there is an evident reduction in the relative efficiency and this is more apparent for the high LD region than the low LD region. The relative efficiency remains between 0.95~1 for all LD maps constructed (Figure 2.7 c).

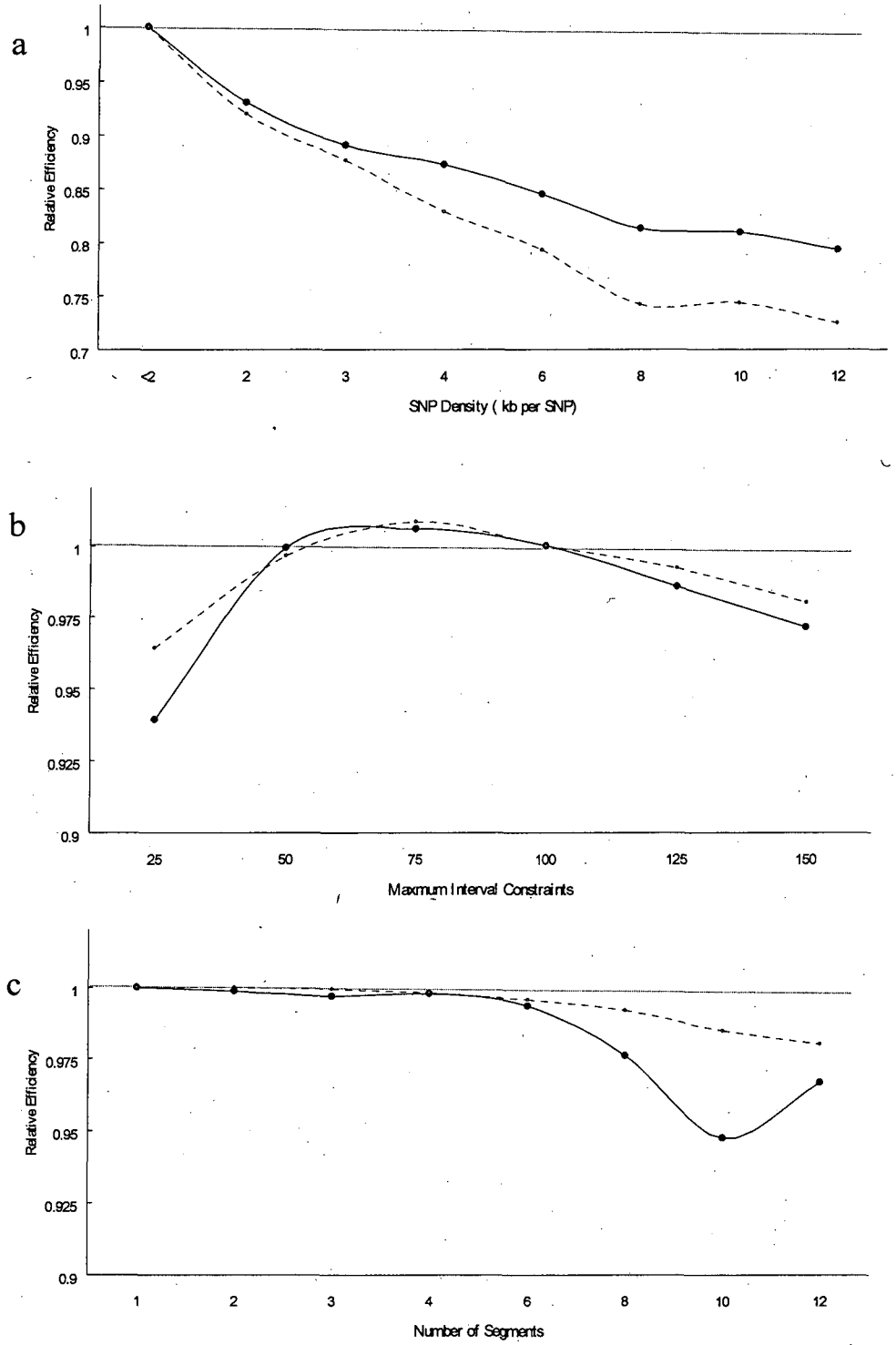


Figure 2.7 The impact on the relative efficiency

LD maps constructed based on alternative methods to reduce the pairwise data from the high LD region (solid line) and the low LD region (dashed line). The results are shown for a) SNP density; b) Maximum interval constraints; c) Number of segments. These LD maps fitted the standard data set which includes all SNPs with the max_intv,100.

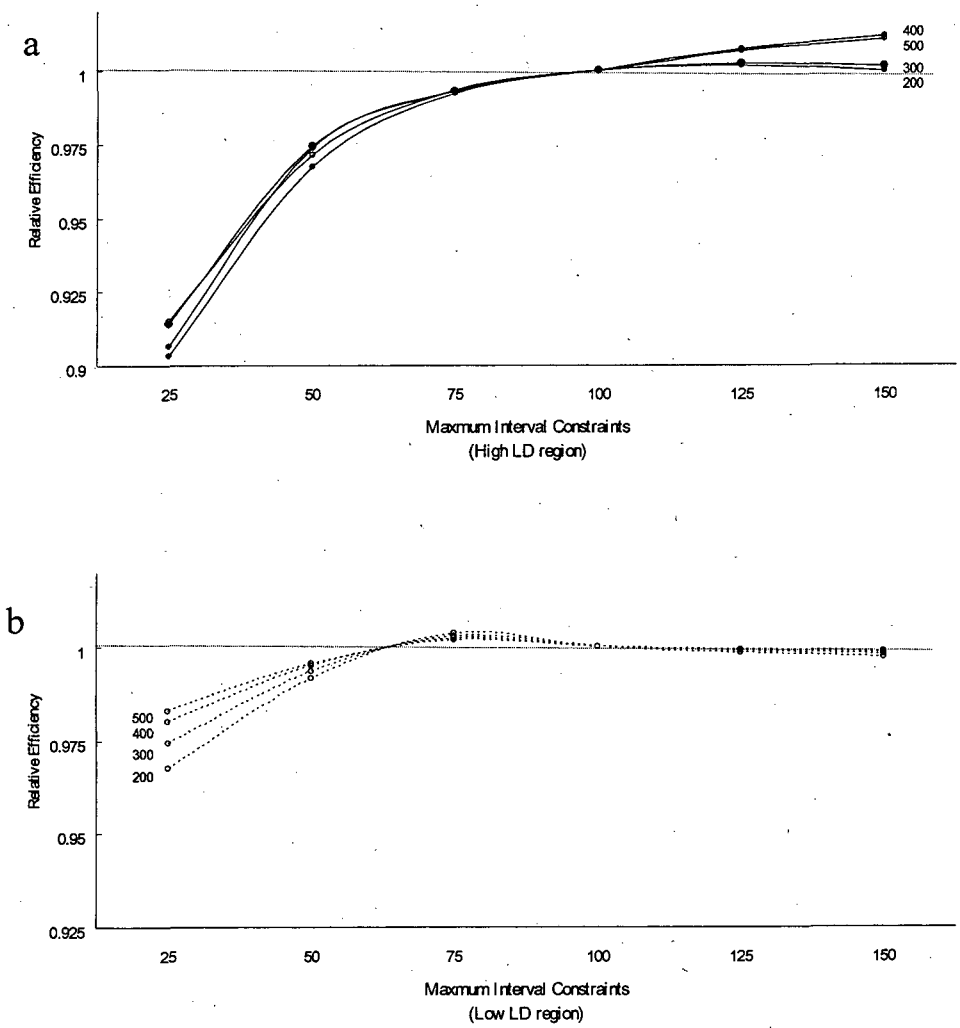


Figure 2.8 The impact on the relative efficiency

Each LD map for the high LD region (the upper figure) and the low LD region (the lower figure) fitted to another standard datasets which includes all SNPs with the max_intv at 200, 300, 400, and 500 respectively.

2.3.3 The impact on block ratio

1) SNP density

The LD block proportion in the default map for the low and high LD region is approximately 69% and 81% respectively. Results show that while reducing the SNP density to 1 SNP per 12 kb, the block ratio for the low LD region decreases very rapidly to 0.59 whereas it only decreases to 0.74 for the high LD region (Figure 2.9 a). This shows that using lower SNP density has less ability to delimit LD blocks for both regions.

2) Maximum interval between pair of SNPs

As seen in Figure 2.9 b, when the max_intv is 25, the block ratio is 0.95 and 0.92 for the high and low LD regions respectively. This increases with the increasing the max_intv, but tends to stabilise when the max_intv is over 100.

3) Number of segments

The block ratio slightly decreases while the number of segments increases for the high LD region. It reduces only 5% even dividing by 12 segments in the assembly method. However, the trend of the block ratio is more unpredictable for the low LD region ranging between 0.9~1 for the various number of segments used (Figure 2.9 c).

These results above agree with the results in Figure 2.6, in which the longer map length is correlated with a smaller proportion of LD blocks. The results here give an evidence for the presumption that the longer map length reflects a poor characterisation of LD blocks.

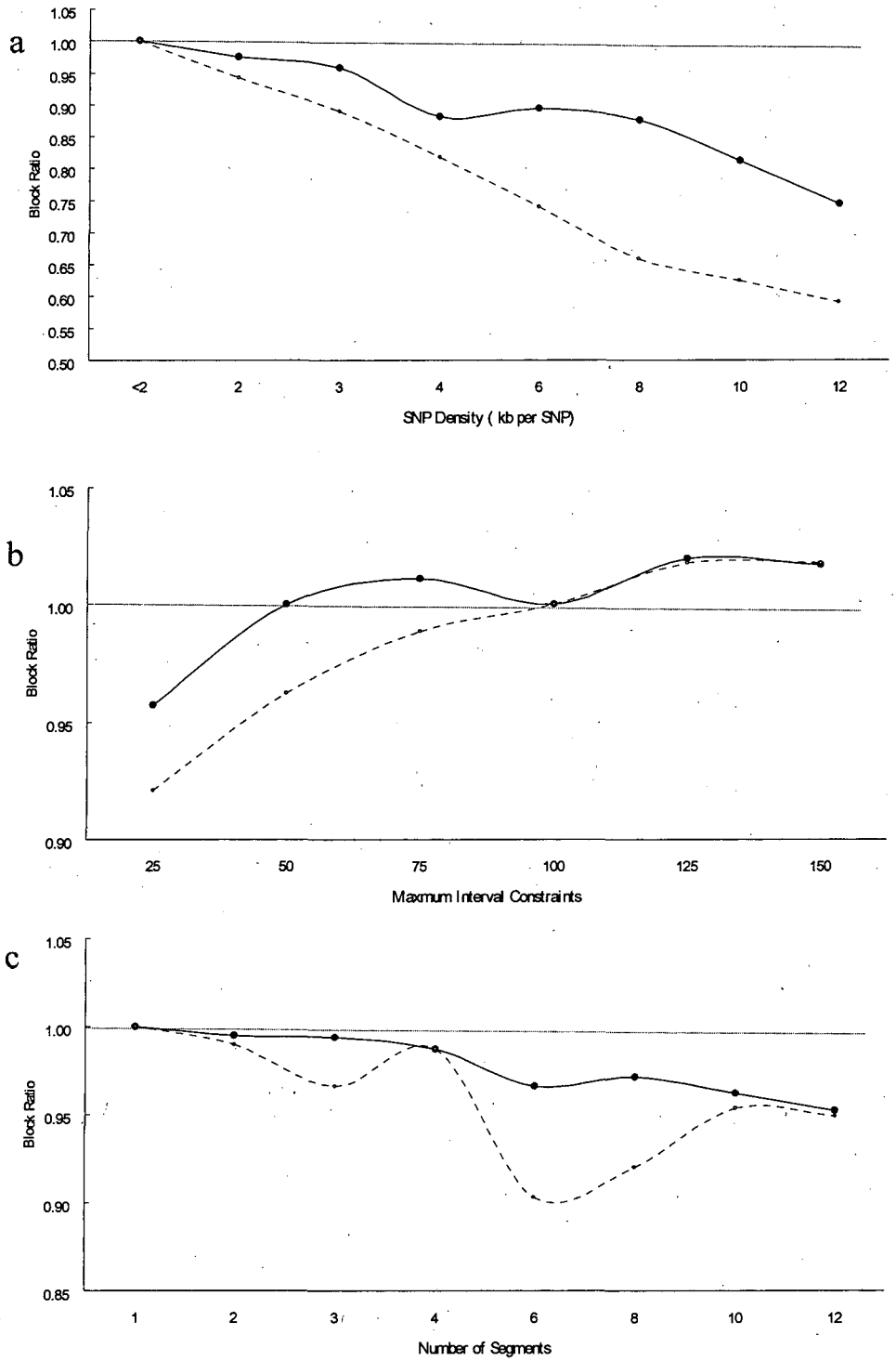


Figure 2.9 The impact on the block ratio

LD maps constructed based on alternative methods to reduce the pairwise data from the high LD region (solid line) and the low LD region (dashed line). The results are shown for a) SNP density; b) Maximum interval constraints; c) Number of segments.

2.3.4 Processing time

In this study, I attempted to estimate the time it took when different LD maps were constructed. The actual time was difficult to estimate because it was influenced not only by the procedures used but also by how busy the server was during the map construction. Generally speaking, it took about 3~6 hrs to complete an LD map when a data set included 1000 SNPs with the use of the default value (100 max_intv and 500 max_dist), but only took 30 mins to 1 hr to complete it when a data set included only 500 SNPs with only 50 max_intv.

2.4 Discussion

Using a high SNP density for a data set enables the construction of an LD map with a very high resolution. A large number of pairs generated from the data set leads to a heavy computational challenge. This difficulty does not exist for a small region or a region with a very low SNP density. However, investigators have to deal with the computational burden imposed by large data sets when studies target whole chromosomes or the entire human genome. The problem has had an impact on investigations into the LD pattern for the whole genome, and will increase when more genome-wide association studies are conducted. In this study, I proposed three approaches to reduce the volume of the pairwise data: reducing the SNP density, constraining the max_intv for pairs, and dividing a large chromosome into smaller segments. It is encouraging to know that LD maps are very robust to the approaches that are suggested to reduce the max_intv and to use segments, but are not robust to reductions in SNP density. Results show that there is a large impact on the map length, the relative efficiency and the block ratio when the SNP density is reduced. A great deal of information is lost even

when the SNP density is reduced only from 1 SNP per kb to 1 per 2 kb. Therefore, using all the SNPs available in a data set is suggested to achieve high resolution LD maps. Using the max_intv at 100 is sufficient to obtain informative pairs for the high and low LD regions. Although increasing the max_intv for the high LD region may continue to increase the relative efficiency, this is modest compared to the map using the max_intv at 100 only. However, using a large max_intv would increase the processing time dramatically and perhaps would decrease the relative efficiency for the low LD region. Therefore, these should be included in the consideration for cost-benefits when constructing an LD map. The results also show that using segments for constructing an LD map has an even smaller impact on the quality of an LD map. However, it is uncertain that the tiny effect is from many overlapping regions or from insufficient SNPs used in a segment. These two factors are correlated, because there are fewer SNPs in a segment as more segments are used. Remarkably, the relative efficiency remains at more than 70%, even for the worst LD map that was constructed at 1 SNP per 12 kb. This relative efficiency is much higher than the 41% on average in the kb map that was constructed from a data set including all SNPs.

In this study, I have measured the map length, the residual error variance, and the block proportion for each LD map to evaluate the quality of these maps. According to the Figures 2.6 and 2.9, it is apparent that the increase in the map length is correlated with the decline of the block proportion, except for those LD maps that were constructed using different SNP densities. These maps have different numbers of intervals, making the map length variable. If the number of informative pairs is constrained stringently, the block structure may not be characterised properly. For example, the map length is much longer but the block

ratio is much less for the low and high LD regions when the max_intv is at 25. This may also explain why the map length increases with the increase in the number of segments. The reason is that the informative pairs needed to estimate the length of each interval at the end regions on two sides of any segment are limited. Although extending 25 SNPs in an overlapping region was applied, it was still not enough for these intervals according to the results. The inaccurate estimation for the map length in these intervals can be improved by extending the overlapping regions to 100 SNPs. Therefore, each interval in any segment will use the same number of pairwise data for the estimation of its length. This tiny revision in the process ensures using enough information pairs for each interval and it would further improve the quality of an LD map.

In addition to reducing the number of pairs in a data set, using the segment method to construct LD maps has another advantage. It allows several LD maps for these segments to be constructed simultaneously because each segmental map construction can be considered as an independent job. It would decrease the processing time markedly, because many jobs can be processed in parallel. For example, given 100 computers, if a long chromosome is divided into 100 segments and they are processed on these 100 computers simultaneously, it only takes the time which is required for constructing one segmental LD map. GRID computing technology (Rowe et al. 2003; Sulakhe et al. 2005) coordinates and shares network resources by utilizing many servers and is perfectly suitable for the assembly method. We will implement this technology with the LDMAP program when constructing LD maps for the human genome.

In summary, the computational difficulty for constructing a genome-wide LD map

can be resolved by limiting the size of a data set without losing the quality of the map. New technologies and powerful computers can also accelerate the process. This study guides the choice of optimal strategies in the LDMAP program when constructing an LD map for a large data set and also reveals the reasons for differences between LD maps.

Chapter 3 The Construction and analysis of whole Genome LD Maps from the HapMap data.

3.1 Introduction

LD maps are useful tools that describe the structure and magnitude of LD in genomic regions. They are applicable to different fields in genetics, most importantly to association mapping. LD maps guide the design and analysis of association studies, and also identify regions that may have been subject to natural selection during human history. In the past few years, the construction of LD maps was limited to a small number of genomic regions. The first LD map of a whole human chromosome was constructed in 2003 (Tapper et al. 2003) using a dataset of chromosome 22 from a published paper (Dawson et al. 2002). This LD map consists of approximately one thousand SNPs across the entire chromosome. However, constructing a genome-wide LD map for the whole human genome seemed inconceivable until the HapMap project was launched (Consortium 2003). With advancement in high-throughput genotyping techniques and reduced cost, the International HapMap Project combined effort to provide a public database of common variation in the form of numerous SNPs across the human genome. These data provide the best source for genome-wide LD map construction at present.

Our research group made the first whole genome LD map from the HapMap data release #16 in 2005 (Tapper et al. 2005). The dataset included 0.7 million SNPs genotyped in each of 269 DNA sample at the density of one SNP per 5 kb in four different populations. The first construction of the entire genome LD map gave

experience in dealing with the computational difficulty resulting from such large datasets. The construction of an LD map is time-consuming and computationally intensive. The problem of handling such large datasets can be addressed by several strategies, for example, excluding pairs at large distance which have reduced information and creating LD maps in segments which are then rejoined to make a complete LD map (See chapter 2). The first genome-wide LD map was made by these strategies (50 max_intv, 500 max_dist, 1000 SNPs per segment and 25- SNP overlap), which increased the speed in map construction with little loss of information.

In January 2006 the HapMap project provided a new release #20, which increased SNP density from 1 SNP per 5 kb to 1 SNP per 1 kb. This release required more computational load and processing time to construct the whole genome LD map. It would be a huge task to construct the map using the same strategies with the same criteria as the previous work. Therefore, new strategies were developed with the latest computational technologies to aid map construction.

In this chapter, I describe how the genome-wide LD map was constructed efficiently using parallel processing in a GRID-based computational system. Furthermore, I compare the differences in LD maps between chromosomes and between populations. I also compare the difference between the release #20 and the release #16 LD maps.

3.2 Materials and Methods

3.2.1 Source of genotype data

The genotype datasets for constructing the whole genome LD maps were downloaded from the HapMap public release #20 (January 2006) at http://www.hapmap.org/genotypes/latest_ncbi_build35/non-redundant/. This release contains a remapping of the previous release #19 on NCBI Build 35 coordinates and has excluded SNPs inconsistent in mapping between Builds 34 and 35. These datasets are classified by chromosome and population (Chromosome: 1 to 22, X and Y; Population: CEU, CHB, JPT and YRI). Approximately 3.7 million SNPs were genotyped across the whole genome of the four population samples that comprises 90 CEU individuals (30 parent-offspring trios), 90 YRI individuals (30 trios), 45 CHB and 44 JPT unrelated individuals. However, only parental DNA samples and unrelated individuals were used in map construction (60 CEU, 60 YRI, 45 CHB and 44 JPT). Genotype data from all chromosomes, except chromosome Y, for each of the four population samples were then used to construct the whole genome population-specific LD maps. To construct an LD map of chromosome X, only female DNA samples were used. (30 CEU, 30 YRI, 23 JPT and 23 CHB female individuals).

3.2.2 SNP screen procedure

For quality control (QC) in genotype data, each SNP had to pass a screening procedure that discarded SNPs showing strong deviation from Hardy-Weinberg equilibrium with χ^2>10 (Gomes et al. 1999) and rare SNPs with minor allele frequency (MAF) less than 0.05 including monomorphic SNPs.

3.2.3 Strategies with specific criteria for the map construction

The previous whole genome LD maps (Tapper et al. 2005) had been constructed in segments with the removal of uninformative SNP pairs. The criteria used in the map construction were 50 max_intv, 500 kb max_dist, 1000 SNPs per segment with 25-SNP overlap, and overlap distance being averaged. In the new dataset of the release #20, the SNP density was approximately 1 SNP per 1-1.5 kb in different population samples. Therefore, I increased the max_intv to 100 in order to sufficiently cover the average physical distance of useful LD (the swept radius), approximately 50 kb in the genome. Other criteria in map construction were 500 kb max_dist, 2000 SNPs per segment with 100-SNP overlap and overlap distance not being averaged. Increasing the size of overlap for every segment ensures that the LDU length for each interval was estimated from sufficient informative pairs, including the first and the last intervals. The number of SNPs per segment increased from 1000 to 2000 SNPs, determined according to computational performance (Lau et al. 2007). Any segment with less than 1000 SNPs was combined with the preceding segment to ensure sufficient SNPs in every segment. Overlapping regions were only used to enable better estimations for the intervals at the two distal regions of every segment. This is a difference from the previous construction where averaging of the overlap region was used. After segmental LD maps had been constructed, overlap regions were removed entirely. For some large chromosomes, with the number of SNPs greater than 70,000, and to avoid memory problems in computing, datasets were divided into 2 or 3 smaller sub-datasets each one containing 1000- SNP overlap.

3.2.4 The software program: LDMAP-Cluster

In the previous map construction, the whole genome LD map was assembled from segmental LD maps constructed independently by the LDMAP program. This program only constructed one map a time in a sequential process, so the entire process was extremely slow, especially when there were many segments. In contrast, the new version named LDMAP-Cluster (Lau et al. 2007) is able to perform map construction in a parallel process (See Figure3.1). This program is based on the original LDMAP program but manages the submission of multiple datasets to a computing cluster of numerous dual-processor servers under a Linux environment. Therefore, each dataset is processed independently to make a segmental LD map. This feature greatly speeds up map construction for the whole genome. LDMAP-Cluster also provides a useful function to merge segmental maps into a complete map. Further information can be found on the website (http://www.som.soton.ac.uk/research/geneticsdiv/epidemiology/LDMAP/default.htm).

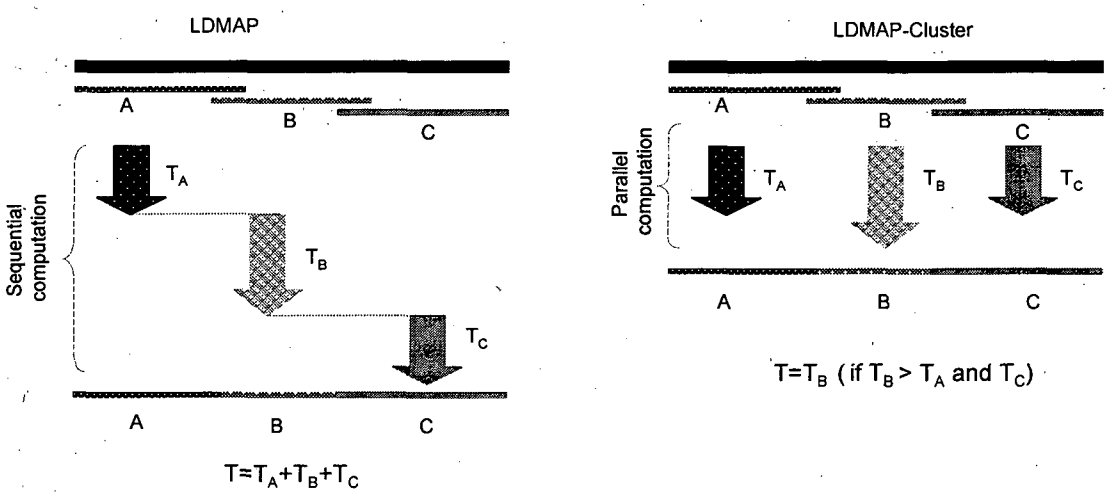


Figure 3.1 Sequential and parallel computation for map construction

T_A, T_B and T_C are required time for processing segments A, B and C respectively. T is the total required time.

3.2.5 Special terms and their descriptions

Several special terms are used in the study for characterisation of the LD patterns in these LD maps and also for comparison between populations. These terms and their descriptions are listed in this section.

1. **LDU/Mb ratio:** It is a measure for the intensity of LD in a particular region, calculated as the total LDU length divided by the physical length in megabases.

2. **A block:** A region with consecutive intervals which have an LDU length of zero.

3. **A hole:** An interval with LDU length greater than 2.5.

4. **Mean block size:** Calculated as the total kb length of intervals in which LDU length equals zero divided by the total number of blocks.

5. **Block coverage:** Calculated as the total kb length of intervals in which LDU length equals zero divided by the total kb length of all intervals.

6. **Specific block proportion:** Calculated as the number of specific blocks divided by the total number of blocks.

7. **Hole coverage:** Calculated as the total kb length of intervals in which LDU length greater than 2.5 LDUs divided by the total kb length of all intervals.

8. **Hole contribution:** Calculated as the total LDU length of intervals in which LDU length greater than 2.5 LDUs divided by the entire LDU length.

3.3 Results

3.3.1 The removal of SNPs

Initially, each downloaded dataset of the four population samples contained approximately 3.7 million SNPs genotyped across the whole genome. The YRI dataset had slightly fewer SNPs. After the SNP screening procedure, 23%-40% of these SNPs were monomorphic and excluded from the datasets. Other potentially problematic SNPs (less than 1% with HWE χ^2 >10 and roughly 10 % with MAF<0.05) were also excluded. More than a half of the SNPs (1.9-2.3 million) remaining in the post-screened datasets were then used for the LD map construction (See Table 3.1). Although the YRI dataset had the smallest number of SNPs initially, it contained the highest number of SNPs after screening. The CHB and JPT post-screened datasets both have very similar numbers of SNPs (1 SNP per 1.55 kb) but hundreds of thousands fewer SNPs compared to the YRI and CEU samples (1 SNP per 1.26-1.39 kb). Approximately 1.3 million common SNPs (55%-68%) are shared in four population post-screened datasets, but the proportion (55%) in the YRI dataset is much less due to more population specific SNPs in its total. More than 81% of intervals between two adjacent SNPs are less than 2 kb and over 93% are less than 4 kb.

During the screen procedure, many problematic SNPs in the download datasets were found. Some of them have two different reference IDs in UCSC genome browser databases. For instance, one SNP at 51,575,414 base pair (bp) of chromosome 2 has two IDs, rs17868116 and rs10184263; another SNP at 118,641,217 bp of chromosome 4 also has two IDs, rs17861176 and rs11729803. A total of 228 SNPs with this problem were identified. Another type of problematic

SNPs were those SNPs that are monomorphic in all parental chromosomes but not in children's. This type of problem indicates genotyping error. The total of 12,730 SNPs in the CEU dataset and 11,539 SNPs in the YRI dataset with this problem were identified and all problematic SNPs were removed from the datasets.

Table 3-1 The SNPs removed in the datasets of the four population samples after the SNP screen procedure

population	Download SNPs	Monomorphic SNPs	*Rare SNPs	HWE χ^2 >10 SNPs	Post-Dataset SNPs
CEU	3,720,803	1,224,673	376,657	20,122	2,110,581
	100%	32.92%	10.13%	0.54%	56.74%
CHB	3,715,927	1,447,862	365,703	19,533	1,894,783
	100%	38.96%	9.84%	0.53%	50.99%
JPT	3,715,927	1,489,758	337,557	19,704	1,880,578
	100%	40.09%	9.08%	0.53%	50.61%
YRI	3,641,870	851,075	441,914	29,091	2,336,706
	100%	23.37%	12.13%	0.80%	64.16%

* Rare means SNPs with minor allele frequencies (MAF) less than 5% but greater than 0 in the sample. Some rare SNPs with MAF less than 5 % could also have significant deviation from HWE, so these SNPs were counted in both columns.

3.3.2 The completion of the whole genome LD maps

The human genome comprises of 23 chromosomes, 1-22 and X or Y covering 2933 Mb of the euchromatin. In the map construction, the whole genome was analysed in approximately 1000 segments with 2000 SNPs each. In general, map construction in a segment of 2000 SNPs required 5 – 10 hrs of computation time. It could have taken at least 5000 hours equivalent to approximately 200 days to construct the whole map on a sequential process. However, it only took approximately 20 days in a parallel process with at least 10 servers available to us.

Currently, all information on these LD maps were stored in a collection of flat files arranged by populations and chromosomes. Each flat file includes SNPs with their rs-ID, kb locations and LDU locations. Our research group has been developing an online Linkage Disequilibrium Database (LDDB) which integrates these LD maps with useful information from other genetic maps. This web-based database is available at http://cedar.genetics.soton.ac.uk/ public_html/ .

3.3.3 Comparison between populations

Table 3.2 shows that the YRI LD map has the longest map among four population-specific genome-wide LD maps, resulting in an LDU/Mb ratio much greater than the other population's. The map lengths in the other three sample LD maps are more similar, but the CHB LD map is slightly longer.

The total block coverage reflecting high LD regions, accounts for up to 67.74%-71.26% of the entire genome sequence with a mean block size ranging from 6.2-9.1 kb in the four population samples (See Table 3.3). The majority of blocks are less than 30 kb long and very few blocks (less than 1%) are over 100 kb. The YRI LD map contains the highest number of blocks and the shortest block size among the four population samples reflecting accumulated recombination events over the long history of this population. Although the YRI LD map has 100,000 blocks more than the JPT and CHB maps, the difference in block coverage is very small (<2%). The CEU LD map has slightly higher block coverage, but only 3.52% higher than the YRI map. In other words, the large difference in the map length among populations is not influenced by the composition of blocks, but, mainly by the intensity of recombination in inter-block regions.

Table 3.2 The general information of the whole genome LD maps for the four population samples

population	Number of SNP	Physical Length (kb)	SNP Density (per kb)	LDU length	Ratio (LDU/Mb)
CEU	2,110,581	2,932,892	1.3896	57,820	19.71
CHB	1,894,783	2,932,921	1.5479	64,931	22.14
JPT	1,880,578	2,932,911	1.5596	58,731	20.02
YRI	2,336,706	2,932,878	1.2551	81,346	27.74

Table 3.3 The block information of the genome-wide LD maps for the four population samples

chromosome	Number of blocks	Mean block sizes (kb)	Block coverage	Specific block proportion				
				<2 kb	<5 kb	<10 kb	<30 kb	<100 kb
CEU	223,918	8.55	71.26%	32.37%	55.18%	73.58%	93.65%	99.47%
CHB	207,158	8.82	68.84%	31.67%	54.30%	72.76%	93.17%	99.36%
JPT	201,343	9.07	69.55%	31.14%	53.47%	71.91%	92.74%	99.28%
YRI	303,018	6.20	67.74%	38.58%	63.85%	81.85%	96.97%	99.76%

Holes are defined here as intervals that exceed 2.5 LDUs. They are likely to reflect both uneven marker coverage and particularly recombination intense regions. They account for less than 1% of the genome sequence but contribute to 4.41-17.54% of map length among populations. In general, the number of holes can be reduced by increasing the SNP density at the regions with extremely low LD (Tapper et al. 2003). For this reason, the YRI LD map with the highest SNP density has fewer holes than the other LD maps. By contrast, the CHB and JPT LD maps have many more holes. Because the maximum LDU value for a hole is constrained to 3, adding more SNPs into a hole may contribute to increase or decrease in map length. Therefore, map length is less reliable in a region with many holes. However, locations of holes tend to be different between populations. The result (Table 3.5) shows that very few of them (1-2%) are shared by all populations and many more (8-23%) are shared in at least one populations.

Table 3.4 The hole information for the genome-wide LD maps for the four populations

population	Number of Hole	*Hole coverage	*Hole contribution
CEU	2,033	0.53%	10.38%
CHB	3,838	0.94%	17.54%
JPT	2,900	0.81%	14.64%
YRI	1,216	0.37%	4.41%

Table 3.5 The proportion of holes shared between populations

	The proportion of population-specific holes shared					
	CEU	CHB	JPT	YRI	non-Africa populations	All populations
CEU	1.00	0.23	0.18	0.12	0.07	0.02
CHB	0.13	1.00	0.18	0.08	0.04	0.01
JPT	0.13	0.23	1.00	0.09	0.05	0.01
YRI	0.16	0.18	0.17	1.00	0.02	0.02

To compare the local variations in patterns of LD between populations, each map was divided into non-overlapping segments and the number of LDU per megabase for each segment calculated (LD intensity). The correlation coefficient between LD intensities of any two maps was calculated. This was repeated by using 1000, 500, 100, 50, and 10 kb respectively for each segment. The results show that local patterns of LD between any two of populations are highly correlated (see Table 3.6). The correlation coefficient decreases with the length of segment, reflecting local variation in patterns of LD between populations. However, the coefficient remains very high even when 10 kb per segment is used (0.55-0.64).

Table 3.6 The correlation coefficients of LD intensities between any two populations

	1000 kb	500 kb	100 kb	50 kb	10 kb
CEU-YRI	0.923	0.894	0.777	0.717	0.584
CEU-CHB	0.926	0.890	0.771	0.719	0.609
CEU-JPT	0.919	0.882	0.762	0.706	0.597
CHB-JPT	0.929	0.892	0.777	0.730	0.637
CHB-YRI	0.922	0.885	0.761	0.698	0.559
YRI-JPT	0.913	0.877	0.756	0.695	0.554

3.3.4 Comparison between chromosomes

Although the SNP density in the release #20 dataset has reached approximately one SNP per 1 kb, these SNPs are not distributed evenly across the whole genome (Figure 3.2). Some genomic regions have a SNP density of less than 1 SNP per 0.5 kb, but others with intervals greater than 100 kb between two adjacent SNPs. The size and the number of regions with extremely low SNP density varies between chromosomes.

Figure 3.3 shows the map length of each chromosome in four population samples. Obviously, physical length in kb is strongly correlated with map length in LDU. All chromosomes in the YRI sample always present the longest map length compared to the other three population samples. It also shows that the CEU and the JPT LD maps reveal high similarity in the map length of all chromosomes, but they are much shorter than the YRI map. The CHB map length is intermediate, and slightly longer than the JPT and CEU LD maps.

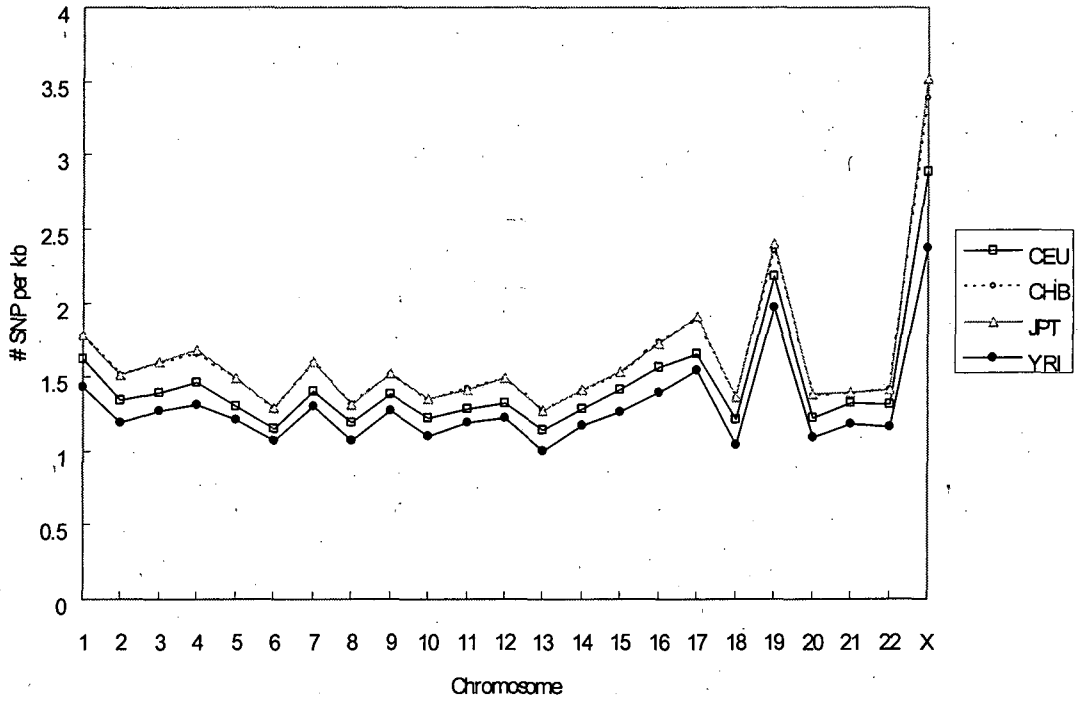


Figure 3.2 The SNP densities in the datasets of all chromosomes among the four population samples

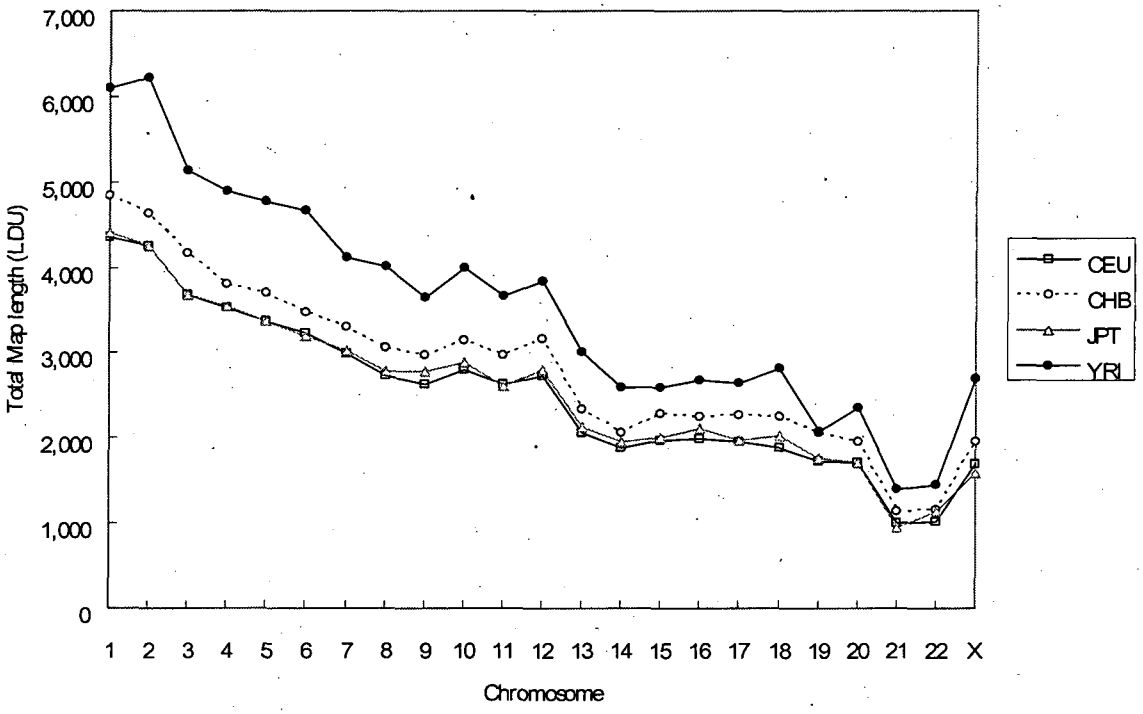


Figure 3.3 The total map length of all chromosomes among the four population samples

Figure 3.4 presents the LDU/Mb ratio in all chromosomes among the four population samples. Every chromosome in the YRI sample has the highest LDU/Mb ratio compared to the other populations. The JPT and CEU samples have very similar LDU/Mb ratios in their corresponding chromosomes but fewer than both the CHB and the YRI samples. The average difference in LDU/Mb ratio between the YRI and the CHB for corresponding chromosomes is 5.61. The YRI sample has an unusually low LDU/Mb ratio on chromosome 19, reflecting more extensive LD in this chromosome than on average. In addition, shorter chromosomes, such as chromosome 17-22, have slightly higher LDU/Mb ratio than other large chromosomes. This is because the small chromosomes have higher recombination rates (Kaback et al. 1992). Not surprisingly, chromosome X has extraordinarily low LDU/Mb ratio, reflecting extremely high LD in this chromosome, because of the peculiar recombination pattern and effects of selection (Tapper et al. 2005)

Figure 3.5 shows that the block coverage in the majority of chromosomes among the four population is between 65%-75% with the exception of chromosomes 1, 9 and 16 which have extraordinarily low values. This is because these three chromosomes contain regions of heterochromatin with extremely low SNP density, resulting in poor characterisation of block structures. However, the corresponding chromosomes among these population samples have shown very consistent values in their block coverage, implying that the same chromosomes have very similar block distributions.

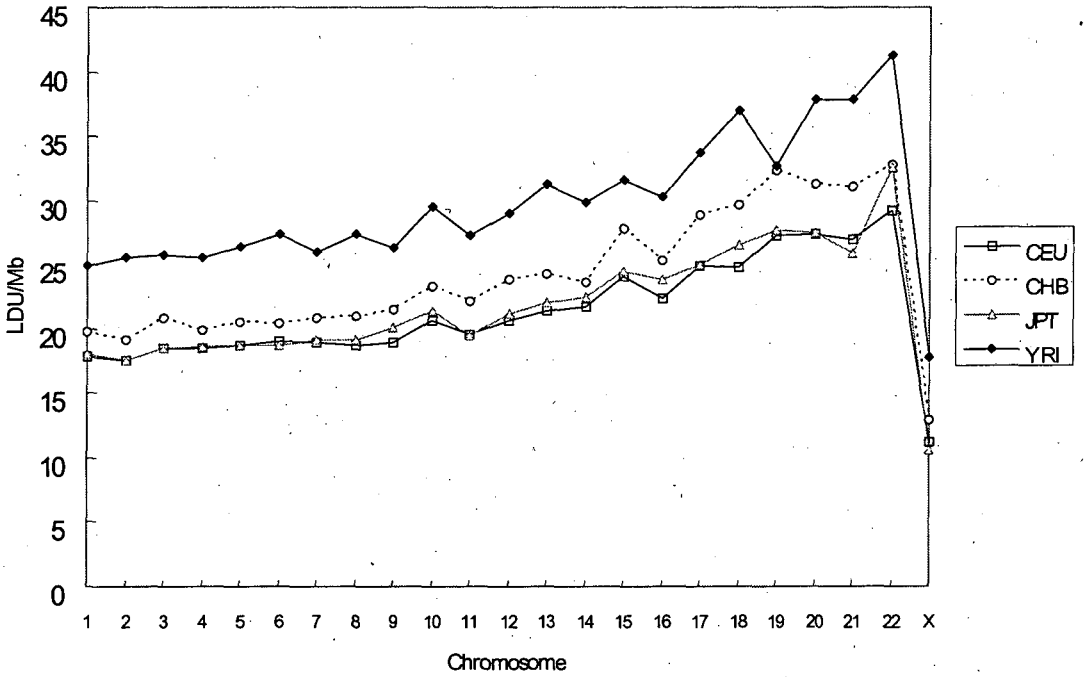


Figure 3.4 The LDU/Mb ratio of all chromosomes among the four population samples

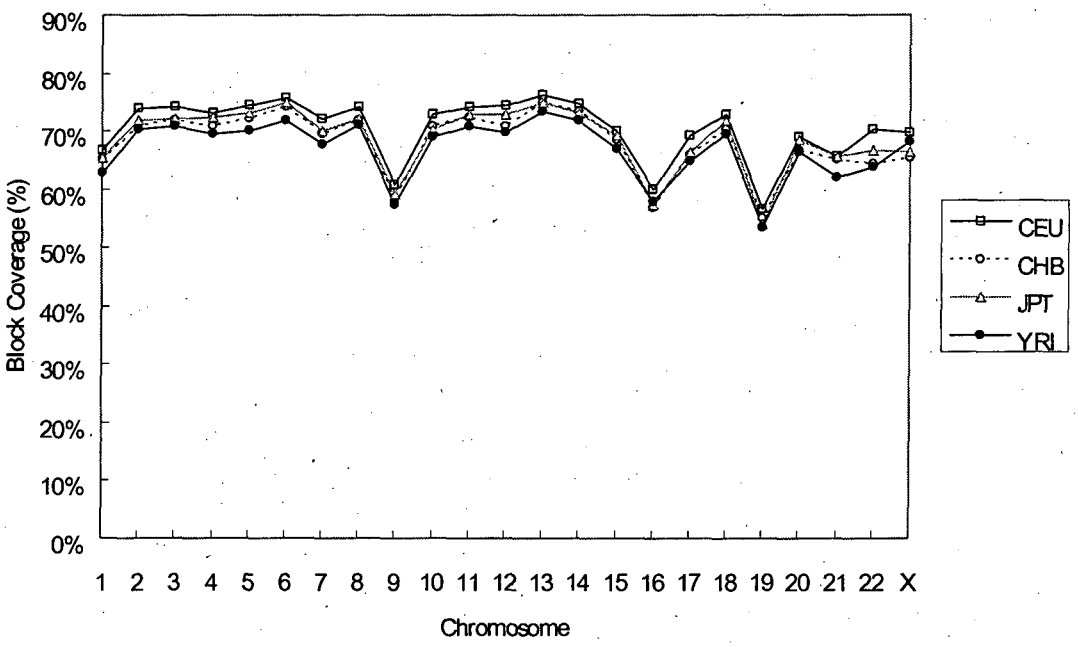


Figure 3.5 The block coverage of all chromosomes among the four population samples

3.3.5 Comparison between release #16 and #20 LD maps

Release #20 has 2.8 times more SNPs than the release #16 dataset. The physical location of these SNPs in the release #20 were remapped on NCBI 35 coordinates and inconsistent SNPs between Build 34 and 35 were removed (See http://genome.ucsc.edu). This revision made the whole genome sequences 2-3 megabases shorter in the release #20 than in the release #16. Although they might have some impacts on the total map length, the results (See Table 3.7 and Figure 3.6) show very little difference in the map length between these two releases (2.3-3.7% only).

Despite the consistent map length between the release #16 and #20 LD maps, the increase in the SNP density in the release #20 dataset improved the LD map with clearer resolution of blocks and steps. In the maps created using more densely typed SNPs, large block regions have been separated into many smaller discrete blocks. The number of blocks has doubled in the new maps, but the block coverage has increased only 11%. Table 3.8 shows that the proportion of smaller blocks in the genome has increased in the new maps. For instance, in the CEU map, the proportion of blocks that are less than 2 kb has increased from 11.94% to 32.37%, and those that are less than 5 kb has increased from 31.80 % to 55.18 %. Furthermore, high SNP density filled many of the holes resulting from large gaps or recombination hot spots in the previous map. Table 3.7 shows that the number of holes in the new maps has reduced in all four population samples.

Table 3.7 The comparison between the release #16 and #20 LD maps

Populations	CEU	CHB	JPT	YRI	mean
Number of SNPs					
Release #16	761,968	673,232	667,370	783,366	721,484
Release #20	2,110,581	1,894,783	1,880,578	2,336,706	2,055,662
Physical map (kb)					
Release #16	2,935,830	2,935,112	2,935,075	2,935,396	2,935,353
Release #20	2,932,892	2,932,921	2,932,911	2,932,878	2,932,901
LD map (LD Units)					
Release #16	56,250	62,686	56,656	79,499	63,773
Release #20	57,820	64,931	58,731	81,346	65,707
*Block coverage (%)					
Release #16	62%	58%	59%	57%	59%
Release #20	71%	69%	70%	68%	70%
Number of Blocks					
Release #16	119,300	107,298	104,212	141,714	118,131
Release #20	223,918	207,158	201,343	303,018	233,859
Number of Holes					
Release #16	2,911	4,879	3,731	2,979	3,625
Release #20	2,033	3,838	2,900	1,216	2,497

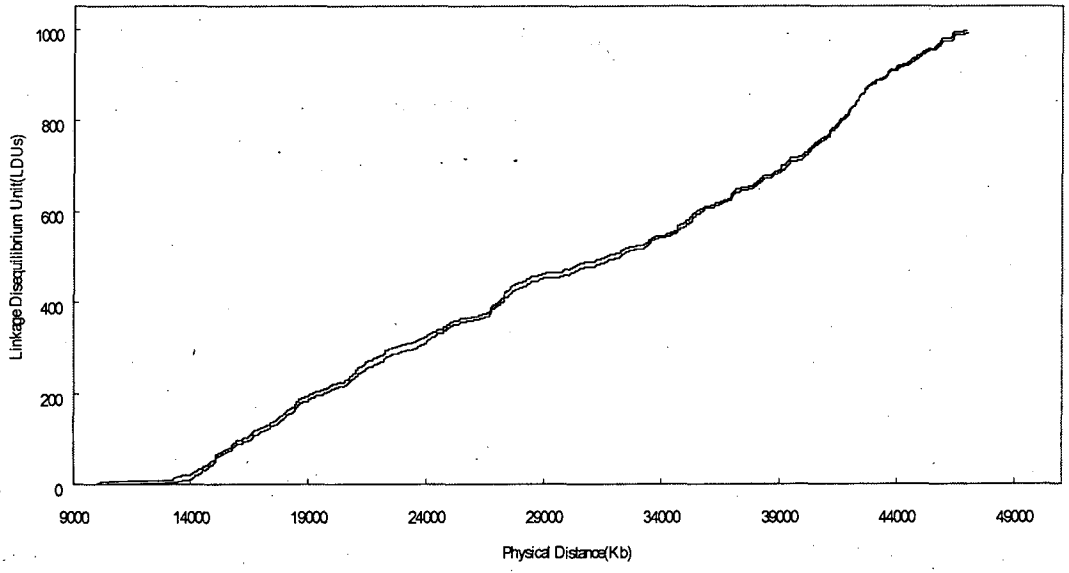


Figure 3.6 The LD maps of chromosome 21 for the CEU sample constructed from the releases #16 and #20 datasets.

The two lines on the figure are almost overlapped and difficult to distinguish, indicating that the two LD maps have very similar LDU length and patterns.

Table 3.8 The comparison of the block structure between the release #16 and #20 LD maps

Dense SNP coverage enhances resolution by increasing the number of small blocks in the map. For example, the block proportion less than 2 kb has increased from 11.94% (release **#16)** to 32.37%(release **#20)** in the CEU LD map.

	Specific block proportion					
release #16	<2 kb	<5 kb	<10 kb	<30 kb	<50 kb	<100 kb
CEU	11.94%	31.80%	55.23%	87.49%	94.93%	98.91%
CHB	11.61%	30.80%	53.94%	86.50%	94.39%	98.74%
JPT	11.05%	29.68%	52.53%	85.67%	93.91%	98.55%
YRI	15.44%	38.45%	63.64%	92.29%	97.21%	99.43%
release #20	<2 kb	<5 kb	<10 kb	<30 kb	<50 kb	<100 kb
CEU	32.37%	55.18%	73.58%	93.65%	97.52%	99.47%
CHB	31.67%	54.30%	72.76%	93.17%	97.23%	99.36%
JPT	31.14%	53.47%	71.91%	92.74%	97.05%	99.28%
YRI	38.58%	63.85%	81.85%	96.97%	98.92%	99.76%

3.4 Discussion

To construct higher resolution LD maps, more informative pairs, more SNPs per segment, and much longer overlapping regions were used in the new map construction. This generated more pairwise data points in the datasets and consequently increased the time required for computation. The previous study (chapter 2) has shown that using pairs with insufficient numbers of flanking intervals (max_intv <50) increases map length but using too many pairs at very large distance increases the error variance. The appropriate value of the max_intv depends on the SNP density in a dataset. In general, the limitation should not be less than the averaged swept radius of approximately 50 kb in the Human genome. The segmental method enabled efficient map construction by processing several segments simultaneously and overlapping regions were used to manage discrete segments. In the previous map construction, the LDU ratio in overlapping regions were calculated by averaging the LDU value in each interval within overlapping regions from two adjacent segments. This method was simple but the averaged values might not represent reliably the real values in these intervals. Instead of using averaged values, the length of overlapping regions was extended in the new map construction, and these regions were only used to assist the estimation of the LDU values in the main segments. I tested this method in several smaller chromosomes and found that this resulted in a slightly smaller error variance. Although a great number of pairwise data were generated by using these criteria in the new map construction, the computational load was no longer an issue because using the parallel process was considerably faster than the sequential process.

The segmental method not only provides an efficient way to construct an LD map but also has the advantage of permitting efficient update. If new SNPs are added in a segment, this segment can be updated independently and inserted back into the current LD map. Furthermore, according to the comparison between the release #16 and #20 LD maps that shows the map length and LD patterns are highly consistent even though they have different SNP density, this implies that these genome-wide LD maps are highly robust and do not require frequent reconstruction unless there is another dataset with much higher SNP density than the release #20 dataset.

The ratio of LDU/Mb is a good indicator to measure the magnitude of LD in a region. A high LDU/Mb ratio indicates that LD erodes more rapidly in that region. The present study has shown that this value is quite different between the YRI population and the other populations. The YRI population always has the highest LDU/Mb ratio in its LD maps compared to CEU, JPT and CHB populations. These latter populations are "out-of-Africa" populations, and are likely to have experienced several population bottlenecks in their histories. The most intense bottleneck was the migration of ancestors from Africa, which took place roughly 100,000 years ago (Lonjou et al. 2003). Other subsequent bottlenecks such as famine, wars and pandemic diseases, contribute to different effective bottleneck times among these populations (Zhang et al. 2004a; Morton 2005). On the other hand, the average LDU/Mb ratio is more consistent among chromosomes, except in some shorter chromosomes. The reason for higher values in shorter chromosomes, such as chromosome 21 and 22, is due to the higher recombination rate on smaller chromosomes (Kaback et al. 1992). Although many chromosomes in the same population have very similar LDU/Mb ratio, some of them display

remarkably extensive block structures in particular genomic regions. These regions could be caused by extremely low SNP density, low recombination rates and natural selection. For example, centromeric regions always have extensive LD, which can extend several megabases across the centromeres with very few SNPs. In addition, chromosome X in all populations has an extraordinary low LDU/Mb ratio resulting from multiple regions with very high LD. Such high LD regions are believed to be the results of several influences. Firstly, unlike a female with a pair of X chromosomes, a male has one X and one Y, so recombination only occurs in 2/3 of the X chromosomes every generation. Second, there is evidence for more intense selection against deleterious mutations when X chromosome is monosomic in males (Giannelli and Green 2000).

Although LD patterns of the same chromosomes are very similar between populations, local variations are found in different genomic regions. Recombination events dominate LD patterns, accounting for 95% of the variation (Tapper et al. 2005). Other factors specific in one or few particular populations, such as demographic history and nature selection, would generate diversity and divergence in LD patterns between populations. However, demographic history affects the entire genome whereas nature selection affects specific genomic regions causing local variations (Akey et al. 2004; Stajich and Hahn 2005). Therefore, selection, either being beneficial or deleterious, results in local reduction of variation in genomic regions, which can reduce haplotype diversity and hence increase local intensity of LD (Kim and Nielsen 2004; Nielsen et al. 2005). The identification of the signals of excess LD attributable to selection is very important, because it indicates functional importance of DNA sequences. In fact, it is challenging to identify such signals, because many regions where

selection takes place may not be identified by comparing two populations. For signals to be detectable, differences in genetic and environment backgrounds between populations are required, and signals should be strong enough to withstand the disruption of recurrent recombination. In addition, excess LD may also caused by other unpredictable factors, such as random genetic drift, density of SNPs and genotyping error.

The construction of the whole genome LD maps with extremely high resolution for the four human populations has been completed. These maps with unique LDU locations have great value in the study of genetic epidemiology and human evolution. Each population-specific LD map described recurrent recombination, selection and demographic evolution in its history. In order to identify selection, a large-scale comparison between these LD maps can be performed to search for substantial difference in the strength and distribution of LD between populations, which could be the signal of local selection taking place over history. For example, if there is a substantial difference in the map length of a corresponding region between any two populations, this region might have biological interpretation or evolution interests in one of their population histories.

The new genome-wide LD maps have an extremely high SNP density that characterises block and step structures more clearly. The number of LD blocks in these maps has increased and the averaged block size is shorter because large blocks have been broken up into several smaller blocks. The increase in the proportion of small blocks is advantageous to disease mapping. It means that candidate regions can be further refined. However, in some genomic regions with very low SNP density, the LD structure is difficult to characterise and the overall

block coverage could be underestimated. On the other hand, the majority of steps is limited to small inter-block regions as the proportion of blocks increase. Steps with indeterminable LDUs, known as holes, might be the regions of intense recombination. Such regions could be limited to only 0.5-5 kb width (Jeffreys et al. 2001). However, there are other factors which may cause holes, such as insufficient SNP density, the criteria to declare a hole and errors in estimating LDU values. In the present study, I only looked at the intervals which are over 2.5 LDUs and ignored the regions with many small steps that could be recombination hotspots as well. So, in order to identify recombination hotspots across the whole genome, using more flexible declaration for holes or recombination hotspots is necessary for further studies .

The present study provides a general view of the whole genome LD maps. The comparison between these LD maps is only at chromosome level. Further studies focusing on particular genomic regions are underway. Despite individual variations in particularly local regions, these population-specific LD maps reveal very similar LD patterns in corresponding chromosomes. Therefore, to extend the applications of LD maps to other populations, a standard cosmopolitan LD map can be made by averaging the LDU length from these four LD maps (Gibson et al. 2005). This map would be a convenient and useful tool for any population.

Chapter 4 Association Mapping for Rheumatoid Arthritis in the MHC candidate region

4.1 Introduction

Rheumatoid Arthritis (RA) is an autoimmune disease that affects people of all ages, although women are more frequently affected than men. The prevalence rate of RA remains relatively constant at 0.5~1.0% in many populations (Alamanos and Drosos 2005). This disease not only results in pain, swelling, and loss of function in the joints, but also attacks other organs including lung, heart and kidney (Rodevand et al. 1999). The causes of RA are related to multiple genetic and environmental factors, but genetic factors account for approximately 60% of the variation in the disease (MacGregor et al. 2000). So far, researchers have found several regions that appear to be significantly associated with RA on different chromosomes including 1p, 6, 8p, 12, 16 and 18q by genome-wide linkage studies (Jawaheer et al. 2003; Yamamoto and Yamada 2005; Choi et al. 2006). In this and the next chapters, I investigated two of these RA candidate regions respectively with the application of genome-wide LD maps for localisation of causal variants.

The major Histocompatibility complex (MHC) region located on chromosome 6p21.3 (See figure 4.1) has been investigated frequently because it involves many diseases including RA (Deighton et al. 1989; Ioannidis et al. 2002; Gorman et al. 2004). This region has strong LD (Jeffreys et al. 2001; Kauppi et al. 2005; Miretti et al. 2005) and contains more than 280 genes (Consortium 1999), making it extremely difficult to precisely localise disease susceptibility genes. The DRB1

gene in the MHC region has been found to be strongly associated with RA (Gregersen et al. 1987; Dizier et al. 1993). However, recent studies have suggested there might be an additional causal variant which is independent of the DRB1 gene (Brintnell et al. 2004; Kochi et al. 2004). A study using a transmission disequilibrium test (TDT) suggested that this causal variant is likely to be located near the junction of the MHC class I and class III region (Kilding et al. 2004). A susceptibility locus near the tumor necrosis factor (TNF) gene at the telomeric end of the class III region has been reported in different studies (Hajeer et al. 2000; Martinez et al. 2000; Castro et al. 2001; Ota et al. 2001). Therefore, using association approaches with SNP markers to refine the candidate region and identify another causal variant is necessary.

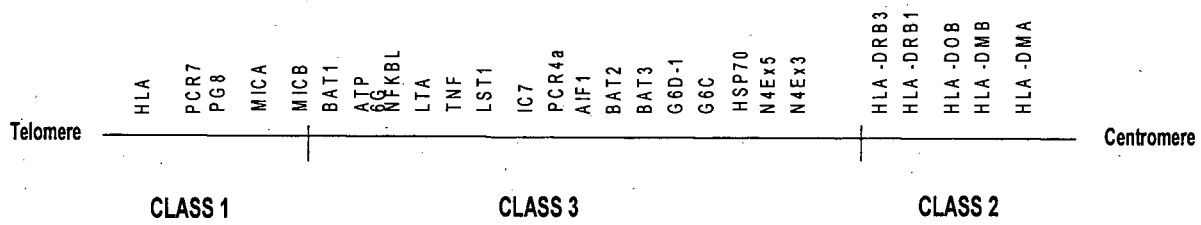


Figure 4.1 The major Histocompatibility complex (MHC) region on 6p21.3

Single SNP testing and haplotype analysis are two common association approaches for mapping susceptibility variants of common diseases in a candidate region. Single SNP testing applies a χ^2 statistic test for each SNP individually. It relies on a Bonferroni correction to reduce false positive rate when number of SNPs is large. On the other hand, haplotype analysis identifies haplotypes which are significantly over or under-represented in patients in comparison to healthy individuals. The separation of phase-unknown genotypes into haplotypes in

population-based studies is labor-intensive and very expensive in lab, It is possible to use statistical methods for haplotype inference from genotype data, but this relies on correct haplotype estimation. The present study used a composite likelihood method avoiding these two restrictions. This method considers all SNPs in a candidate region simultaneously and applies the Malecot model which estimates the location of a causal variant (Maniatis et al. 2004).

Under the composite likelihood method, each SNP must have a relative location reflecting the correlation between this SNP and other SNPs. This location can be provided from a physical map, a linkage map or an LD map. However, using an LD map as a reference map is more appropriate because it represents allelic association estimated using observed pairwise SNP data from a real population. An LD map can be constructed from a study sample or obtained from the genome-wide LD maps constructed using the HapMap data (See Chapter 3). It is preferable to use the latter map because the HapMap data usually has much higher SNP density than any other study samples at present. In addition, the genome-wide LD maps can also be used if an LD map cannot be constructed from a study sample.

This study used two case/control samples by genotyping for a number of SNPs in the MHC candidate region to identify the causal variant associated with RA. The first sample was from a British Caucasian population and the other was from a Japanese population. The two samples have different sample size and SNP coverage. These two factors are important for study design in association approaches because smaller sample size and insufficient SNP coverage cause unreliable results. In this study, I compared the results between the single SNP

test and the composite likelihood method. Furthermore, I investigated the difference in the results when using different maps for SNP locations in the composite likelihood analyses.

4.2 Materials and Methods

4.2.1 Case/Control samples

Two case/control samples from population data of unrelated individuals were used to search for RA causal variant at the MHC region in the present study. The first sample was from British Caucasian population and the other was from a Japanese population.

The British Caucasian sample

This sample consists of 316 RA patients and 210 healthy Individuals from a British Caucasian population. All of the patients were recruited from the Arthritis Research Campaign national repository. 20 SNPs located in an 1850 kb wide candidate region covering the class I and the class III of the MHC region were genotyped in both patients and healthy controls. The physical locations on the kb scale for these SNPs were obtained by matching the DNA forward and reverse primer sequences (See Table 4.1) with the Human Genome sequence assembly (NCBI build 35, UCSC May 2004) using the BLAST program (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/informa tion3.html).

Table 4.1 The forward and reverse primer sequences for each SNP in the British Caucasian sample

Name of SNP	Forward Primer	Reverse Primer
PCR7	TGCTCAAAGGACTGCAGGAA	GAACTTGGGCTGCAAATACA
PG82271	GCTGTTTGTCAAGGAGACAACCT	CTCCAACTGTCAGCTGCTTA
PG8436	TTGGTGCAGCCTCTGAACCT	CCTGCGTGCTGCTTTGG
MICA2	GAAGACAACAGCACCAGGAGCT	CTGACGTTCATGGCCAAGGT
MICA1	GAGCTCCCAGCATTTCTACTACGA	GGCATCTTCCTTCAAGAAATTCCT
BAT1991	GCCCTCCGCAAATACCAA	TTCCAATGGGTTCTTCTCATA
NFKBL	AACGCCCCTCACAGTTCACTT	TCCAGGCTGGAGGAAATGG
TNFbeta	CAGTCTCATTGTCTCTGTCACACATT	ATCGACAGAGAAGGGGACAAGAT
TNF308	GGCCACTGACTGATTTGTGTGT	CAAAAGAAATGGAGGCAATAGGTT
LST1	AGTCATGAGCTGCATACA	TAATGTTATCGCGGAATGATG
IC7	GGCCTCCTAGAGACCCTGACAT	CAGGGACCTCGAGCATCAAA
PCR4A	CCTCCTCAGCCTCCCAAAGT	GTGCAGCAGCGACAGAAAAGT
AIF1	TCTCCTCCACCTAGCAGTTGGT	TCCATTAAGGTCAAACTCCATGTATTT
BAT3	CCTGTGGTGGTGCATGGA	ACCGGCGCCCTGCT
G6D1	CCTCACTGCCCCAGAAGGA	ATCTGCAAGGGCTGCAGATG
G6C2	CCCCAAAGACCTGGTTTGC	GTCATAGGGAAGCCTGGTCTTG
G6C1	GCATGCTGGTGGAAATTGG	GGCATCACAGAAGCCATCAGT
HSP70	CTTGGTAGAGTTTTGTGATG	TCGTGGCTGGAGGTCAA
N4E$_x$5	AGCCCATCCTGGCAAGTG	TGTGAGGTGAATCCAGACAA
N4E$_x$3	ACCCAGCTTCTTGTGCACTTG	CGGCCCCTTTTGGAACA

The Japanese sample

The second sample came from a report published by Okamoto et al. 2003 . It consists of 116 RA patients and 100 unaffected controls from a Japanese population. The RA patients were diagnosed according to the American Rheumatism Association's criteria (Arnett et al. 1988). All individuals were genotyped with 35 SNPs in a 44 kb region including TNF, ATP6G and BAT1 genes. The physical locations for these SNPs were obtained through a Genome browser according to their rs-identifier, which were based on the same Genome sequence assembly as used for the British Caucasian sample. The report provides the allele frequencies of those SNPs in the case and control groups without detailed genotype information for each individual.

4.2.2 Obtaining LD maps for the candidate region

An LD map for the British Caucasian sample was constructed by the LDMAP program using the healthy control data in the sample. For quality control, each SNP had been tested for Hardy-Weinberg equilibrium (HWE) using a likelihood ratio test before the map construction. The same procedure could not be applied for the Japanese sample because the limited information from the report (only allele frequencies for each SNP were presented) was not enough to construct an LD map. Therefore, I also used the two HapMap LD maps (CEU and JPT) for the LDU locations of those SNPs in the British Caucasian and the Japanese samples respectively. If SNPs were not included in the HapMap LD maps, the LDU locations for them were linearly interpolated (See chapter 2, Figure 2.5).

4.2.3 Statistic analysis

Single SNP test using Pearson's χ^2

Pearson's χ^2 from a two by two contingency table between affection status and a diallelic SNP (see Table 4.2) were calculated to test non-random allelic association between cases and controls. The values of a, b, c and d represent the allele counts of a SNP in the case and control groups. The χ^2 value for each SNP was calculated as $\chi_1^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$ with one degree of freedom. The allele counts in the British Caucasian sample were obtained by pooling SNP genotype data together from individuals. In the Japanese sample, they were calculated from their allele frequencies provided in the paper (Okamoto et al. 2003).

Table 4.2 Four counts, a, b, c and d in a 2×2 table between disease status and a diallelic marker

	Marker		
Affection status	+	-	
Case	a	b	a+b
Control	c	d	c+d
	a+c	b+d	n=a+b+c+d

Multiple SNPs test using the Composite likelihood method

First, the observed $\hat{Z}_i$ and expected Z_i associations between a disease and any SNP i are estimated respectively. The observed $\hat{Z}_i$ is estimated as $\hat{Z}_i = \frac{(ad-bc)}{(a+b)(b+d)}$ where a, b, c and d are allele counts in case/control groups (See Table 4.2) and the expected association Z_i is obtained from the Malecot equation $Z_i = (1-L)Me^{-\varepsilon d_i} + L$. Kz_i is the corresponding information for $\hat{Z}_i$, calculated as $Kz_i = \frac{n(a+b)(b+d)}{(a+c)(c+d)}$. In this study, a composite likelihood method was used to test whether there was significant evidence in the candidate region and to estimate the location of the causal variant. The composite likelihood method is based on the Malecot equation, but the distance d_i is replaced with $(S_i - S)$, where S_i is the location of the i^{th} SNP in either kb or LDU, and S is the location of the causal polymorphism. The composite log likelihood is calculated as $-2\ln lk = \sum Kz_i(\hat{Z}_i - Z_i)^2$.

To test significance of a region, two 'sub-hypotheses (models A and B) are contrasted. Model A is the null hypothesis H_0 of no association between the disease phenotype and SNPs in this region. It assumes the parameter M is 0 and L is fixed to the predicted L_p (Morton et al. 2001). However, model B replaces the predicted L_p with the estimated L. The χ^2 for the A-B contrast is calculated as $\chi^2_{df=1} = \frac{[(-2\ln lk)_A - (-2\ln lk)_B]}{V_B}$, where V_B is the residual error variance of model B, calculated as $V_B = \frac{(-2\ln lk)_B}{(m-k)}$ (m is the number of SNPs and k is the degree of freedom. If the A-B contrast shows nominal significance, there is significant evidence for causal polymorphisms within the region. If a

region is significantly associated with disease, additional contrasts (the A-C , A-D, A-C' and A-D' contrasts) are used to test for a causal polymorphism at location S depending on the number of Malecot parameters estimated. Model C estimates both parameters M and S but uses the predicted L_p. The χ^2 value for the A-C contrast is calculated as $\chi^2_{df=2} = \frac{[(-2\ln lk)_A - (-2\ln lk)_C]}{V_C}$ with 2 degrees of freedom, where V_C is the residual error variance of model C. Model D further replaces L_p with the estimated L, giving 3 degrees of freedom in the A-D contrast and $\chi^2_{df=3} = \frac{[(-2\ln lk)_A - (-2\ln lk)_D]}{V_D}$. If the A-C and A-D contrasts show nominal significance, the parameter S could be the best-predicted location of the causal polymorphism. For convenience and simplicity, the parameter ε is usually fixed to 1 for an LD map, because LD maps are constructed such that ε~1 (Maniatis et al. 2002). However, for physical maps, the ε is obtained by fitting observed pairwise data to the model. If the parameter ε is also estimated, one more degree of freedom is added for the χ^2 and C' and D' are used to distinguish them from the former models. The contrasts between null and alternative hypotheses with the number of degree of freedom are shown in Figure 4.2.

Furthermore, for region with small number of SNPs, an F-test is more reliable than a χ^2 test (Maniatis et al. 2006). An F-value is estimated as the mean variance between models divided by the mean variance within model. The latter is the residual error variance. For instance, the F-value for the A-C contrast is $F_{2,m-2} = \frac{\frac{[(-2\ln lk)_A - (-2\ln lk)_C]}{2}}{V_C}$. For simplicity, any p value from either χ^2-test or F-test can be converted into χ^2 with one degree of freedom by using the

Hastings approximation (Abramowitz and Stegun 1965). The variance (V_S) for the causal location ($\hat{S}$) is the inverse of the information that is estimated in an information matrix with simultaneous estimates of M, S and L. The standard error for $\hat{S}$ is $Se = \sqrt{V_S}$, and 95% confidence interval (95% CI) is $\hat{S} \pm 1.96Se$.

$$Z_i = (1 - L)Me^{-\varepsilon(S_i - S)} + L$$

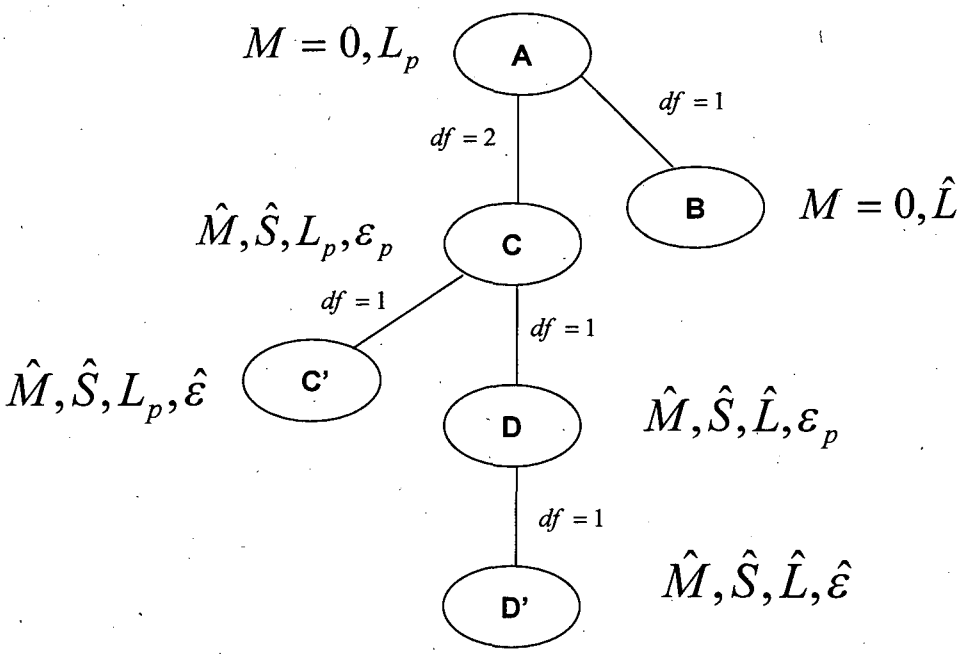


Figure 4.2 Sub-hypothesis under the Malecot model

Different hypotheses use different estimated parameters (with a circumflex) and predicted parameters (with a small p). The number of degrees of freedom are shown between models, indicating the difference in the number of the estimated parameters between models. For example, there are 2 degrees of freedom in the A-C contrast (M and S are estimated) and 3 degrees in the A-D contrast (M,S and L are estimated).

4.2.4 The LOCATE program

The LOCATE program implements the algorithms described in the last section. This program generates two output files of results. The intermediate output contains the association and its corresponding information between affection status and each SNP, which are necessary for the composite likelihood method. The results of the single SNP tests are also shown in this output. The final output shows the results of the composite likelihood method including the optimal estimations of the Malecot parameters for each Model. Significant tests for the contrasts between null hypothesis and each of the alternative hypotheses are shown in the final output. This program using association approaches is useful for refining a candidate region.

4.3 Results

4.3.1 LD maps for the RA candidate region

The results of the HWE tests for the 20 SNPs in the British Caucasian sample are shown in Table 4.3. Only N4Ex3 shows significant deviation from HWE (p value = 0.013). However, after correction for multiple tests, this is not significant (p value: $0.013 \times 20 = 0.26$, p value >0.05). Therefore, all SNPs were used to construct an LD map termed as the sample LD map (Figure 4.3a).

Figure 4.3b shows the HapMap LD maps of the candidate region for the two samples, which were obtained from the genome-wide LD maps described in Chapter 3. The two HapMap LD maps are much longer than the sample LD map. This must reflect the much dense SNP coverage and higher resolution and suggested that the sample map is poorly characterised. Despite different

lengths in those maps, the three maps have very similar block-step structures. In the Caucasian sample, 20 SNPs are located in the near 2 Megabase (Mb) region, but the majority of SNPs are clustered in the TNF gene. Several large gaps with the distance between two adjacent SNPs greater than 200 kb or 2.5 LDUs on the map due to insufficient SNP density are not ideal for association mapping. On the other hand, In the Japanese sample, 35 SNPs are clustered in a 44 kb wide region with only 0.224 LDUs at the TNF gene.

Table 4.3 The HWE tests for the 20 SNPs in the British Caucasian sample

SNP	Physical location(kb)	MAF	HWE χ^2 test	Uncorrected Pvalue
PCR7	30,455.298	0.155	0.000	0.987
Pg82271	31,218.370	0.336	1.288	0.256
Pg8436	31,230.307	0.474	0.100	0.752
MICA-2	31,486.936	0.250	0.612	0.434
MICA-1	31,486.954	0.136	1.208	0.272
BAT1991	31,617.417	0.281	0.039	0.844
NFKBL	31,623.321	0.326	0.042	0.837
TNFbeta	31,648.318	0.338	0.379	0.538
TNF-308	31,651.018	0.189	2.359	0.125
LST1	31,663.111	0.405	0.209	0.648
IC7	31,668.682	0.152	0.361	0.548
PCR4a	31,680.864	0.355	2.820	0.093
AIF1	31,691.823	0.350	1.682	0.195
BAT3	31,719.754	0.171	2.076	0.150
G6D-1	31,783.708	0.160	0.479	0.489
G6C-2	31,797.299	0.162	3.160	0.075
G6C-1	31,797.974	0.195	0.874	0.350
HSP70	31,885.940	0.308	0.010	0.920
N4Ex5	32,296.583	0.288	0.668	0.414
N4Ex3	32,298.372	0.381	6.152	0.013*

* <0.05

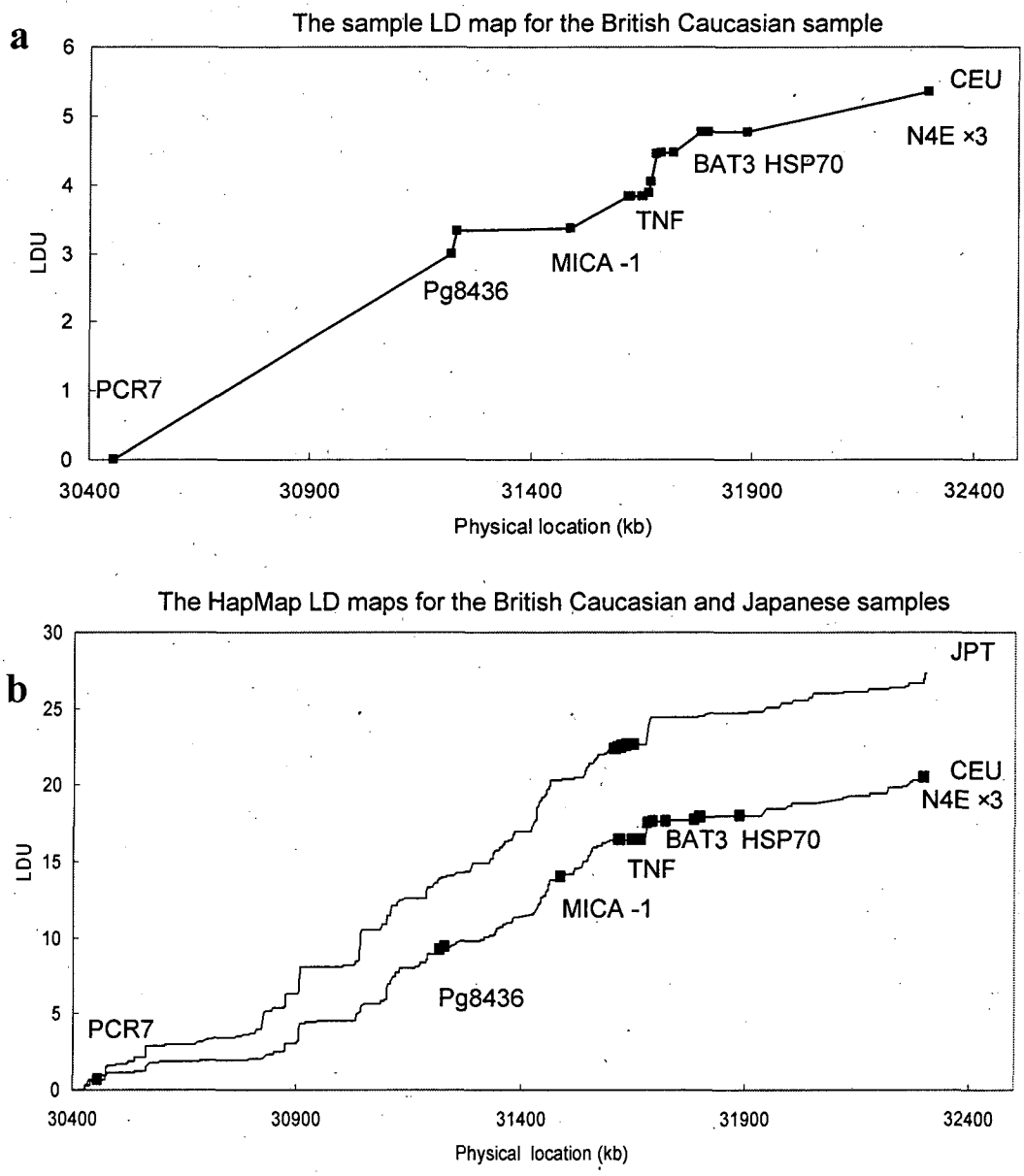


Figure 4.3 The HapMap and the control LD maps of the RA candidate region for the two samples

The dots indicate the locations of the SNPs used in both samples.

4.3.2 Results from the single SNP test

For the Caucasian sample

Table 4.4 shows the results from the single SNP test for the 20 SNPs in the Caucasian sample. A total of 7 SNPs show significant differences in allelic frequency between the cases and controls (p value <0.05) but only 3 SNPs (NFKBL, TNF-β and HSP70) remain significant after Bonferroni correction. The most significant SNP is HSP70 (p value = 0.00055). The other two significant SNPs are close to the TNF gene.

For the Japanese sample

Table 4.5 shows the results for the 35 SNPs in the Japanese sample. Before Bonferroni correction, 7 SNPs show significant associations with the RA (p value<0.05). The most significant one is rs1799724 with χ^2=7.853 (p value= 0.005). However, after Bonferroni correction, none show significant association.

Table 4.4 Single SNP tests for the 20 SNPs in the British Caucasian sample

SNP	Kb Map	χ^2	Uncorrected p value	[1]Corrected p value
PCR7	30455.298	1.173	0.278840	
Pg82271	31218.370	0.450	0.502191	
Pg8436	31230.307	0.014	0.905186	
MICA2	31486.936	1.122	0.289592	
MICA1	31486.954	4.615	0.031696	0.633927
BAT1991	31617.417	1.111	0.291928	
NFKBL	31623.321	9.513	0.002040	0.040798
TNFbeta	31648.318	10.685	0.001080	0.021598
TNF308	31651.018	0.625	0.429053	
LST1	31663.111	5.317	0.021120	0.422409
IC7	31668.682	0.033	0.855052	
PCR4a	31680.864	1.893	0.168920	
AIF1	31691.823	5.001	0.025340	0.506798
BAT3	31719.754	4.966	0.025851	0.517013
G6D1	31783.708	0.530	0.466702	
G6C2	31797.299	0.012	0.913933	
G6C1	31797.974	0.805	0.369521	
HSP70	31885.940	17.582	0.000028	0.000550
N4Ex5	32296.583	0.419	0.517367	
N4Ex3	32298.372	0.010	0.921962	

p values <0.05 are underlined.

[1]Bonferroni correction for the p values.

Table 4.5 Single SNP tests for the 35 SNPs in the Japanese sample

SNP	Kb Map	χ^2	Uncorrected p value	[1]Corrected p value
rs3219189	31606.025	0.370	0.5432	
rs2516478	31606.717	0.542	0.4617	
rs929138	31611.678	6.493	0.0108	0.3791
rs1129640	31614.604	0.592	0.4417	
rs933208	31614.627	0.293	0.5883	
rs2071596	31614.670	1.311	0.2522	
rs2516393	31614.723	0.070	0.7907	
rs2523512	31614.779	0.367	0.5448	
rs2523511	31614.832	0.875	0.3496	
rs2071595	31615.041	0.000	0.9883	
rs2239527	31617.758	3.683	0.0550	
rs2523506	31617.945	0.558	0.4549	
rs2239528	31618.084	0.001	0.9695	
rs2071594	31620.699	4.409	0.0357	1.2511
rs2071593	31620.777	0.077	0.7810	
rs2239705	31621.381	6.823	0.0090	0.3151
rs2523503	31621.537	0.542	0.4617	
rs2523502	31621.844	0.215	0.6425	
rs3219186	31622.961	1.202	0.2728	
rs3219185	31623.057	3.684	0.0549	
rs3219184	31623.119	1.470	0.2254	
rs2071592	31623.318	7.480	0.0062	0.2184
rs2239708	31623.742	0.060	0.8065	
rs2071591	31623.777	3.971	0.0463	1.6198
rs3219183	31624.342	0.060	0.8065	
rs3219182	31625.094	0.265	0.6067	
rs3219180	31625.152	0.000	0.9961	
rs2857605	31632.830	0.011	0.9178	
rs2857604	31633.084	5.856	0.0155	0.5433
rs3093949	31633.162	3.161	0.0754	
rs2239707	31633.299	0.407	0.5236	
rs2230365	31633.428	0.703	0.4016	
rs1799964	31650.287	0.592	0.4415	
rs1800630	31650.455	0.542	0.4617	
rs1799724	31650.461	7.853	0.0051	0.1775

p values <0.05 are underlined.

[1]Bonferroni correction for the p values.

4.3.3 Results from the composite likelihood method

Tables 4.6 and 4.7 list the SNP information in the two samples necessary to the composite likelihood method, including the association Z and the corresponding information K_Z from each SNP data, and the SNP locations provided from the kb, the sample LD and the HapMap LD maps.

Table 4.6 The association information of the 20 SNPs in the British Caucasian sample

SNP	Kb Map	Sample	HapMap	Z	K_Z	χ^2
PCR7	30455.298	0.000	0.000	0.030	1295.432	1.173
Pg82271	31218.370	3.000	8.601	0.059	129.661	0.450
Pg8436	31230.307	3.329	8.796	0.007	282.858	0.014
MICA2	31486.936	3.365	13.351	0.113	88.591	1.122
MICA1	31486.954	3.365	13.351	0.311	47.863	4.615
BAT1991	31617.417	3.837	15.718	0.104	103.233	1.111
NFKBL	31623.321	3.837	15.718	0.139	490.544	9.513
TNFbeta	31648.318	3.837	15.775	0.152	464.747	10.685
TNF308	31651.018	3.837	15.775	0.025	1017.063	0.625
LST1	31663.111	3.875	15.775	0.171	180.818	5.317
IC7	31668.682	4.044	15.775	0.005	1400.543	0.033
PCR4a	31680.864	4.440	16.842	0.065	451.627	1.893
AIF1	31691.823	4.465	16.970	0.105	452.524	5.001
BAT3	31719.754	4.465	16.970	0.288	60.008	4.966
G6D1	31783.708	4.775	17.087	0.020	1279.638	0.530
G6C2	31797.299	4.775	17.230	0.016	44.909	0.012
G6C1	31797.974	4.775	17.230	0.112	63.845	0.805
HSP70	31885.940	4.775	17.324	0.186	506.262	17.582
N4Ex5	32296.583	5.357	19.803	0.063	105.887	0.419
N4Ex3	32298.372	5.357	19.816	0.008	154.391	0.010

Table 4.7 The association information of the 35 SNPs in the Japanese sample

SNP	Kb Map	Sample LD Map*	HapMap LD Map	Z	K_Z	χ^2
rs3219189	31606.025	-	0.000	0.069	77.714	0.370
rs2516478	31606.717	-	0.000	0.069	113.891	0.542
rs929138	31611.678	-	0.024	0.218	136.709	6.493
rs1129640	31614.604	-	0.065	0.151	25.811	0.592
rs933208	31614.627	-	0.065	0.048	128.194	0.293
rs2071596	31614.670	-	0.065	0.076	227.782	1.311
rs2516393	31614.723	-	0.065	0.047	31.814	0.070
rs2523512	31614.779	-	0.065	0.057	112.149	0.367
rs2523511	31614.832	-	0.065	0.181	26.799	0.875
rs2071595	31615.041	-	0.074	0.001	95.254	0.000
rs2239527	31617.758	-	0.102	0.120	255.817	3.683
rs2523506	31617.945	-	0.102	0.069	117.405	0.558
rs2239528	31618.084	-	0.102	0.003	148.982	0.001
rs2071594	31620.699	-	0.102	0.135	240.263	4.409
rs2071593	31620.777	-	0.102	0.028	95.254	0.077
rs2239705	31621.381	-	0.102	0.239	119.582	6.823
rs2523503	31621.537	-	0.102	0.069	113.891	0.542
rs2523502	31621.844	-	0.102	0.084	30.801	0.215
rs3219186	31622.961	-	0.162	0.208	27.792	1.202
rs3219185	31623.057	-	0.167	0.379	25.605	3.684
rs3219184	31623.119	-	0.167	0.158	58.907	1.470
rs2071592	31623.318	-	0.167	0.171	255.817	7.480
rs2239708	31623.742	-	0.167	0.026	88.754	0.060
rs2071591	31623.777	-	0.167	0.129	237.733	3.971
rs3219183	31624.342	-	0.167	0.026	88.754	0.060
rs3219182	31625.094	-	0.167	0.055	88.754	0.265
rs3219180	31625.152	-	0.167	0.001	68.361	0.000
rs2857605	31632.830	-	0.167	0.010	106.404	0.011
rs2857604	31633.084	-	0.198	0.215	127.222	5.856
rs3093949	31633.162	-	0.207	0.117	232.724	3.161
rs2239707	31633.299	-	0.224	0.036	318.895	0.407
rs2230365	31633.428	-	0.224	0.055	230.244	0.703
rs1799964	31650.287	-	0.224	0.069	124.556	0.592
rs1800630	31650.455	-	0.224	0.069	113.891	0.542
rs1799724	31650.461	-	0.224	0.253	122.610	7.853

* The sample LD map cannot be constructed from the study sample due to lack of genotype information for each individual.

For the British Caucasian sample

Table 4.8 shows the results from the composite likelihood method for the British Caucasian samples. Models A and B do not take SNP locations into account, so the values of the Malecot parameters and the likelihood in the two models are not affected by the choice of the three maps. Because of this reason, the χ^2 values for the A-B contrast in all analyses with different maps are the same. In the British Caucasian sample, the χ^2 value for the A-B contrast is 4.522 (p value=0.034) implying that this candidate region is significantly associated with RA. However, models C, D, C' and D' estimate S and other parameters, resulting in different estimations if the reference map is changed. The same location $\hat{S}$ at 31675-31676 kb was estimated in these models whether the sample LD map or the HapMap LD map was used. However, the latter map results in higher χ^2 value with smaller 95% CI. Differently, the $\hat{S}$ is at 31886 kb when the kb map was used. The 95% CI in the analysis using the kb map is much wider than that using LD maps.

Table 4.8 The analysis of the British Caucasian sample by the composite likelihood method

SNP locations based on the kb map

model	df	$-2\ln lk$	V	L	M	ε	S
A	20	35.169	1.759	0.0371			
B	19	27.552	1.450	0.0670			
C	18	21.585	1.199	0.0371	0.0732	0.0022	31886
D	17	18.566	1.092	0.0000	0.1315	0.0022	31886
C'	17	18.092	1.064	0.0371	0.1362	0.0057	31886
D'	16	17.662	1.104	0.0155	0.1490	0.0039	31886

Contrast	χ_1^2	p value	Se	95%CI in kb	
A – B	4.522	0.034*			
A – C	6.261	0.012*	167.9	31533-32239	(706)
A – D	6.481	0.010**	95.1	31685-32087	(402)
A – C'	6.853	0.009**	106.7	31661-32111	(450)
A – D'	5.400	0.020*	54.2	31771-32001	(230)

* <0.05 **<0.01

SNP locations based on the sample LD map

model	df	$-2\ln lk$	V	L	M	ε	S
A	20	35.169	1.759	0.0371			
B	19	27.552	1.450	0.0670			
C	18	22.475	1.249	0.0371	0.0784	1.256	31675
D	17	20.759	1.221	0.0009	0.1406	1.256	31675
C'	17	21.520	1.266	0.0371	0.1168	2.220	31675
D'	16	20.745	1.297	0.0054	0.1399	1.377	31675

Contrast	χ_1^2	p value	Se	95%CI in kb	
A – B	4.522	0.034*			
A – C	5.620	0.018*	77.8	31590-31895	(305)
A – D	4.915	0.027*	16.7	31666-31731	(65)
A – C'	4.427	0.035*	17.5	31665-31734	(69)
A – D'	3.465	0.063	16.7	31666-31731	(65)

* <0.05

SNP locations based on the HapMap LD map

model	df	$-2\ln lk$	V	L	M	ε	S
A	20	35.169	1.759	0.0371			
B	19	27.552	1.450	0.0670			
C	18	19.621	1.090	0.0371	0.1121	1	31676
D	17	19.405	1.142	0.0292	0.1272	1	31676
C'	17	19.484	1.146	0.0371	0.0952	0.743	31676
D'	16	18.814	1.176	0.0171	0.1074	0.480	31676

Contrast	χ_1^2	p value	Se	95%CI in kb	
A – B	4.522	0.034*			
A – C	7.798	0.005**	15.6	31670-31731	(61)
A – D	5.854	0.016*	4.6	31670-31688	(18)
A – C'	5.796	0.016*	46.0	31615-31795	(180)
A – D'	4.618	0.032*	69.4	31606-31878	(272)

* <0.05 **<0.01

For the Japanese sample

The A-B contrast for the Japanese sample also shows significant association (χ^2 =5.66, p value=0.017) in this region (See Table 4.9). When the SNP locations are based on the kb map , the $\hat{S}$ is 31650 kb in models C and D whereas 31623 kb in models C' and D'. Models C' and D' that estimate an additional ε may be less reliable than models C and D. This is because the parameter ε is greatly over-estimated and has a very large error in models C' and D' when a region is in strong LD (Maniatis et al. 2004). In this study, the parameter ε in C' and D' models are 1.58 and 2.12 for the kb map and 6.43 and 171.86 for the LD map respectively.

Although the A-B contrast indicates the association with RA in this candidate region, further contrasts of A-C and A-D do not support this association. The two contrasts both indicate the same location at 31633 kb, but the χ^2 values for them do not indicate such significant association. It is possible that the causal variant is near but not within the region. It is also possible that the sample size in the Japanese sample is too small to replicate this association.

Table 4.9 The analysis of the Japanese sample by the composite likelihood method

SNP locations based on the kb map

model	df	$-2\ln lk$	V	L	M	ε	S
A	35	26.232	0.750	0.0676			
B	34	22.156	0.652	0.0980			
C	33	22.127	0.671	0.0676	0.0346	0.0022	31650
D	32	22.079	0.690	0.0000	0.1037	0.0022	31650
C'	32	19.409	0.607	0.0676	0.3556	1.5816	31623
D'	31	18.526	0.598	0.0846	0.4085	2.1163	31623

Contrast	χ_1^2	p value	Se	95%CI in kb	
A – B	5.660	0.017*			
A – C	3.530	0.060	818.9	31606-31650	(44)
A – D	2.260	0.133	830.6	31606-31650	(44)
A – C'	5.372	0.020*	0.23	31622-31623	(1)
A – D'	5.005	0.025*	0.27	31622-31623	(1)

* <0.05

SNP locations based on the HapMap LD map

model	df	$-2\ln lk$	V	L	M	ε	S
A	35	26.232	0.750	0.0676			
B	34	22.156	0.652	0.0980			
C	33	22.100	0.670	0.0676	0.0351	1	31633
D	32	22.034	0.689	0.0000	0.1041	1	31632
C'	32	21.979	0.689	0.0676	0.0465	6.43	31633
D'	31	20.247	0.653	0.0922	0.2728	171.86	31633

Contrast	χ_1^2	p value	Se	95%CI in kb	
A – B	5.660	0.017*			
A – C	3.564	0.059	11.3	31606-31650	(44)
A – D	2.305	0.129	11.3	31606-31650	(44)
A – C'	2.359	0.125	9.5	31613-31650	(37)
A – D'	3.027	0.082	11.3	31606-31650	(44)

*, <0.05

4.4 Discussion

In addition to the DRB1 gene playing an important role in RA (Gregersen et al. 1987; Dizier et al. 1993), many studies indicated a range of possible locations for additional variants in the MHC candidate region (Singal et al. 1999; Ota et al. 2001; Newton et al. 2003; Kilding et al. 2004). The study of the British Caucasian sample has confirmed the evidence of association in this region and suggested a possible location near the TNF gene. The study of the Japanese sample also shows a weak evidence of association in this region and supports the suspected location within NFKBL1 gene. The function of the NFKBL1 has not been determined, but it produces NF-kappaB like protein. Asahara et al. 1995 found high activity of NF-kappa B in the chronic inflammation of the joint in RA patients. Bondeson et al. 1999 reported that blocking NF-kappaB reduces the inflammatory response in the rheumatoid joint.

Failure to replicate the previous findings could be due to low SNP density and small sample size in a study design (Zondervan and Cardon 2004). In the British Caucasian sample, the total of 20 SNPs were genotyped in a 2 Mb wide candidate region (near 20 LDUs on the HapMap LD map), which is equivalent to the density of approximately 1 SNP per 100 kb or per 1 LDU. However, those SNPs were not equally distributed across the region in which 4 large intervals between two adjacent SNPs are greater than 200 kb or 2.5 LDUs. Such large gaps are not ideal for association mapping. It is important to use optimal SNP density based on the LDU scale to ensure coverage of a region. Several SNPs per LDU spanning a range of frequencies would provide better localisation of causal variants (Tapper et al. 2003). By contrast, the study region in the Japanese sample is only 44 kb (0.224 LDUs), but the region was genotyped for 35 SNPs, which is equivalent to

156 SNPs per LDU. Genotyping many SNPs in a small region with high LD is not efficient in disease mapping unless causality in a region is strongly suspected. In this case, large sample sizes might be more useful rather than genotyping more SNPs.

The composite likelihood method for association mapping requires a reference map to provide genetic locations for all SNPs in a candidate region. This study shows that the choice of maps influences the estimates of the Malecot parameters even though the association and the corresponding information between the affection status and SNPs remain the same. In general, the estimated $\hat{S}$ is robust whether the sample or the HapMap LD maps is used, but using the latter map usually has smaller 95% CI with higher χ^2 value in comparison with other map. The best reference map at present is the genome-wide LD map constructed from the HapMap data because of the high SNP density and resolution. Therefore, it is unnecessary for researchers to construct an LD map based on a low resolution SNP sample unless its density is higher than the HapMap data.

Results from the composite likelihood method and the single SNP test are generally consistent. In general, the possible location of a causal variant estimated from the composite likelihood method is highly correlated with a cluster of significant SNPs, but may not be the SNP with the highest χ^2 value. For instance, the analysis of the British Caucasian sample indicates the possible location at 31676 kb, surrounded by 5 SNPs with χ^2 value, between 4.9 and 10.7. However, the SNP with the highest χ^2 is 200 kb away from the estimated location. The composite likelihood method considers association between affection status and all SNPs in the sample. Therefore, the localisation of causal variants

would improve with increases in SNP density. By contrast, higher SNP density may create many false positive results, which is a major problem for single SNP tests.

Chapter 5 Association mapping for Rheumatoid Arthritis at chromosome 18q

5.1 Introduction

Genome-wide association (GWA) studies have recently become feasible in the field of association mapping since recent advances in DNA technologies with high throughput and low cost (Klein et al. 2005; Maraganore et al. 2005; Syvanen 2005). Such studies using a large number of SNPs that are genotyped across the whole human genome give better resolution for disease mapping. However, the development of appropriate tools for analysing those SNP genotype data lags behind the development of GWA studies.

It is challenging to analyse a large number of SNP genotypes from GWA studies. Such studies may involve thousands of SNP tests and thus false positive results will inevitably occur by chance. For haplotype analysis, reconstruction of haplotypes is difficult and unreliable when a region involves many SNPs. Therefore, it is important to apply a two-stage design for disease gene mapping (Zhang et al. 2004a). The first stage is to perform a rapid screen in order to identify candidate regions from the whole genome. The second stage is to further localise any putative causal polymorphisms in these candidate regions.

The composite likelihood method (Maniatis et al. 2004) that considers all SNPs simultaneously and estimates a possible location of a causal variant in a region can be used in GWA studies. However, for this method, the issue is not the number of SNPs but the size of a candidate region. Locations estimated by this method

may be less reliable if a region is very large and other causal variants in the region may be missed. One feasible solution for the composite likelihood method is to analyse a large region using segments (non-overlapping windows). A large region can be seen as an assembly of many separated segments and each segment can be studied independently.

A program called CHROMSCAN has been developed for genome-wide association studies of complex disease (Morton et al. 2007). This program is a development of the LOCATE program that uses a composite likelihood under the Malecot model for disease mapping (See Chapter 4). The application of this approach has been extended to manage multiple segments from a large region rather than a single region. I used here a 10 Megabase (Mb) wide Rheumatoid Arthritis (RA) candidate region on chromosome 18q that has shown strong evidence of linkage in the US genome-wide linkage studies (Jawaheer et al. 2003). This region was genotyped for 2300 SNPs in 460 cases and 460 controls. The analysis of this sample can be used to evaluate the performance of the CHROMSCAN program.

5.2 Materials and Methods

5.2.1 Study sample and SNPs

The study sample, provided by the Genetic Analysis Workshop (GAW) in 2006, consists of 460 RA patients and 460 unaffected controls. All patients were provided by the North American Rheumatoid Arthritis Consortium (NARAC) and the controls were recruited from a New York City population. All individuals were genotyped for 2300 SNPs across an approximately 10 Mb candidate region

(48,896-58415 kb on the physical map of UCSC May 2004) of chromosome 18q. This region has shown strong evidence of linkage in the US genome-wide linkage studies (Jawaheer et al. 2003).

5.2.2 LD maps for the candidate region

Two LD maps for the candidate region were used to assign LDU locations to each of these SNPs (See Figure 5.1). The first map, termed the GAW LD map, was constructed from the study sample of the unaffected controls after the removal of seven SNPs showing significant departure from Hardy-Weinberg equilibrium ($\chi^2 \geq 10$) and 81 SNPs with minor allele frequencies (MAF) less than 5 percent. This LD map contains the remaining 2212 SNPs and generated 151 LDUs. The second LD map was extracted directly from the CEU genome-wide LD map that was constructed using the HapMap data (See chapter 3), and termed the HapMap LD map. This map contains 8086 SNPs within the same 10 Mb region and generated 202 LDUs. Despite its higher SNP density, 185 out of the 2300 study SNPs were missing and therefore their LDU locations were linearly interpolated (See chapter 2). Physical locations for these SNP were based on build 35 (UCSC May 2004) of the human genome sequence.

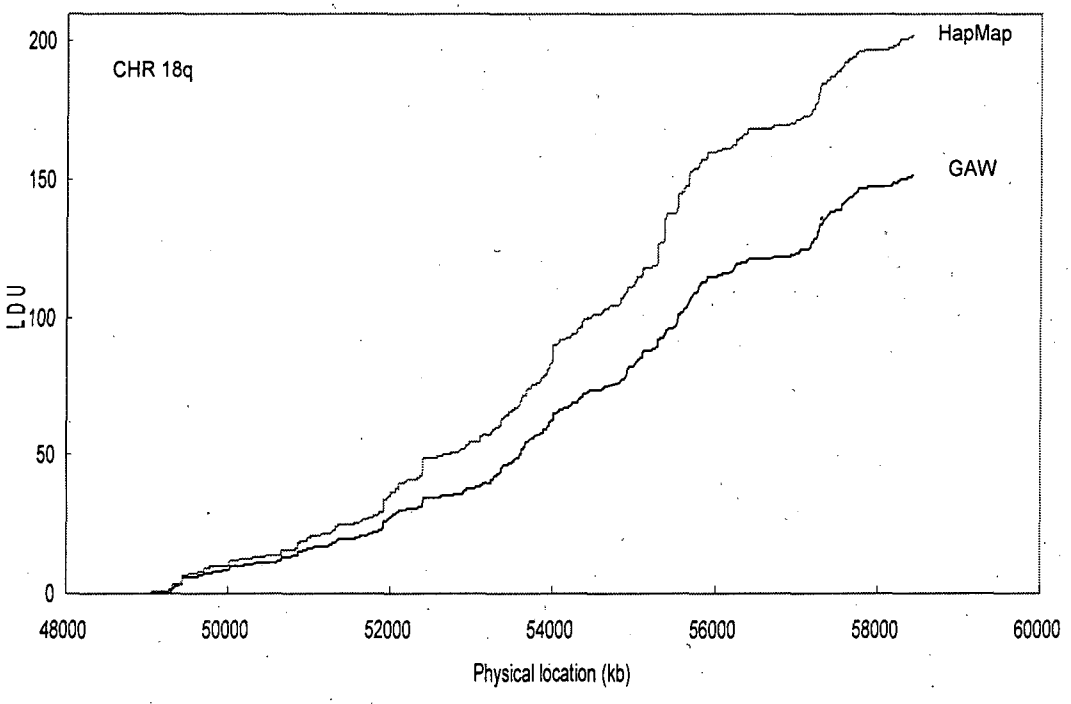


Figure 5.1 The GAW and the HapMap LD maps for the candidate region of chromosome 18q

The two maps have very similar LD patterns but different map lengths. This is due to different SNP densities in the datasets from which they were constructed. The higher density of markers in the HapMap resolved some of the poorly characterised regions in the GAW sample, particularly " holes" where an upper limit on LDUs is applied.

5.2.3 Subdivision of the candidate region

The entire 10 Mb region was divided into contiguous but non-overlapping segments, each with a minimum of 10 LDUs without breaking a block and no less than 30 SNPs. The size of segment is constrained to 10 LDUs and the restriction of 30 SNPs ensures sufficient SNPs in each segment for better estimation. In addition, I also evaluated the effect of using 5 LDUs per segment, while other restrictions remained the same in comparison to the former analyses. Table 5.1 describes the four analyses, each of which uses 5 or 10 LDUs on the scale of one of

the two LD maps respectively for each segment. The analyses 1 and 2 are based on the GAW LD map whereas the analyses 3 and 4 are based on the HapMap LD map. Despite a fixed minimal length on LDU scale for each segment, more segments results in fewer SNPs per segment with shorter physical length. More segments were expected in the HapMap LD map than the GAW LD map because the LDU length in the HapMap LD map is longer. Each segment in these analyses was considered as an independent study of a small candidate region. Tables 5.2-5.5 show the detailed information for each of segments in the four analyses including the number of SNPs, the physical and the LDU distances.

Table 5.1 The general description of the four analyses in this study.

# analyses	LD map used	Minimal Length	Number of segments	Mean of SNP number	Mean of kb length	Mean of LDU length
1	GAW	10 LDUs	14	164	673 Kb	10.35 LDUs
2	GAW	5 LDUs	27	85	349 Kb	5.54 LDUs
3	HapMap	10 LDUs	18	127	523 Kb	11.02 LDUs
4	HapMap	5 LDUs	31	74	302 Kb	6.37 LDUs

Table 5.2 The detailed description of each segment in analyses #1

Analysis #1 (10 LDUs minimal and the GAW LD map)

#Segment	Number of SNPs	Physical Length (kb)	LDU Length (LDUs)
1	342	48896 - 50159 (1262)	0.00 - 10.04 (10.04)
2	262	50159 - 51575 (1417)	10.04 - 20.06 (10.02)
3	146	51575 - 52100 (525)	20.06 - 30.09 (10.03)
4	312	52102 - 53241 (1139)	30.21 - 40.34 (10.13)
5	124	53242 - 53613 (371)	40.91 - 51.14 (10.23)
6	92	53615 - 53966 (352)	52.07 - 62.08 (10.01)
7	117	53967 - 54410 (443)	62.08 - 72.40 (10.32)
8	53	54436 - 55002 (566)	72.97 - 83.45 (10.48)
9	128	55017 - 55381 (364)	84.35 - 94.60 (10.25)
10	92	55382 - 55678 (296)	95.25 - 107.24 (11.99)
11	94	55721 - 56245 (524)	108.65 - 119.06 (10.41)
12	275	56246 - 57248 (1001)	119.24 - 130.06 (10.82)
13	59	57253 - 57555 (302)	130.90 - 141.05 (10.15)
14	197	57556 - 58415 (859)	141.05 - 151.11 (10.06)

Table 5.3 The detailed description of each segment in analyses #2

Analysis #2 (5 LDUs minimal and the GAW LD map)

#Segment	Number of SNPs	Physical Length (kb)	LDU Length (LDUs)
1	99	48896 - 49443 (547)	0.00 - 5.25 (5.25)
2	274	49446 - 50231 (785)	5.25 - 10.54 (5.29)
3	117	50233 - 50981 (748)	10.54 - 16.07 (5.53)
4	137	50984 - 51687 (703)	16.07 - 21.12 (5.05)
5	73	51713 - 51947 (234)	21.12 - 26.23 (5.11)
6	109	51947 - 52341 (393)	26.23 - 31.53 (5.30)
7	159	52351 - 52912 (560)	31.56 - 36.92 (5.36)
8	101	52915 - 53271 (357)	36.92 - 42.17 (5.25)
9	92	53274 - 53496 (222)	42.41 - 47.79 (5.38)
10	46	53500 - 53668 (168)	47.79 - 54.50 (6.71)
11	47	53669 - 53919 (250)	54.50 - 59.52 (5.02)
12	40	53920 - 54006 (86)	59.53 - 64.92 (5.39)
13	84	54008 - 54312 (304)	65.00 - 70.48 (5.48)
14	34	54315 - 54799 (484)	70.48 - 76.15 (5.67)
15	30	54799 - 54968 (169)	76.24 - 82.04 (5.80)
16	30	54970 - 55102 (132)	82.04 - 87.72 (5.68)
17	90	55114 - 55355 (241)	87.72 - 92.84 (5.12)
18	66	55356 - 55520 (164)	92.84 - 98.37 (5.53)
19	35	55521 - 55620 (100)	98.37 - 103.78 (5.41)
20	30	55623 - 55837 (214)	103.78 - 112.98 (9.20)
21	69	55838 - 56245 (408)	112.98 - 119.06 (6.08)
22	222	56246 - 57034 (788)	119.24 - 124.50 (5.26)
23	53	57035 - 57248 (213)	124.50 - 130.06 (5.56)
24	30	57253 - 57410 (157)	130.90 - 138.30 (7.40)
25	34	57417 - 57636 (219)	138.30 - 143.36 (5.06)
26	138	57636 - 58178 (542)	143.36 - 148.67 (5.31)
27	54	58183 - 58415 (233)	148.74 - 151.11 (2.37)

Table 5.4 The detailed description of each segment in analyses #3

Analysis #3 (10 LDUs minimal and the HapMap LD map)

#Segment	Number of SNPs	Physical Length (kb)	LDU Length (LDUs)
1	255	48896 - 49928 (1032)	0.00 - 10.01 (10.01)
2	238	49935 - 51009 (1073)	10.02 - 20.29 (10.27)
3	198	51010 - 51919 (908)	20.30 - 33.67 (13.37)
4	134	51926 - 52398 (472)	33.67 - 48.58 (14.91)
5	241	52399 - 53262 (863)	48.79 - 59.20 (10.41)
6	113	53262 - 53596 (333)	59.25 - 69.56 (10.31)
7	75	53599 - 53919 (319)	69.56 - 79.64 (10.08)
8	40	53920 - 54006 (86)	79.64 - 89.92 (10.28)
9	103	54008 - 54448 (440)	89.95 - 99.98 (10.03)
10	49	54496 - 54974 (478)	101.27 - 111.40 (10.13)
11	70	54977 - 55285 (308)	111.40 - 125.97 (14.57)
12	60	55286 - 55381 (95)	126.33 - 136.43 (10.10)
13	89	55382 - 55634 (252)	137.27 - 147.28 (10.01)
14	30	55641 - 55838 (196)	147.35 - 157.38 (10.03)
15	134	55839 - 56399 (560)	157.38 - 168.39 (11.01)
16	208	56408 - 57248 (840)	168.39 - 179.99 (11.60)
17	55	57253 - 57545 (293)	180.67 - 190.67 (10.00)
18	201	57548 - 58415 (867)	190.67 - 201.94 (11.27)

Table 5.5 The detailed description of each segment in analyses #4

Analysis #4 (5 LDUs minimal and the HapMap LD map)

#Segment	Number of SNPs	Physical Length (kb)	LDU Length (LDUs)
1	99	48896 - 49443 (547)	0.00 - 6.37 (6.37)
2	172	49446 - 50028 (582)	6.37 - 11.47 (5.10)
3	199	50029 - 50869 (839)	11.61 - 18.43 (6.82)
4	99	50885 - 51334 (449)	18.43 - 24.02 (5.59)
5	104	51337 - 51884 (547)	24.29 - 29.31 (5.02)
6	34	51885 - 51962 (77)	29.31 - 34.69 (5.38)
7	43	51963 - 52100 (137)	34.74 - 39.91 (5.17)
8	75	52102 - 52398 (296)	39.97 - 48.58 (8.61)
9	159	52399 - 52984 (586)	48.79 - 54.49 (5.70)
10	86	52993 - 53274 (281)	54.49 - 59.55 (5.06)
11	82	53274 - 53456 (183)	59.55 - 65.57 (6.02)
12	35	53468 - 53615 (146)	65.65 - 71.25 (5.60)
13	47	53618 - 53807 (190)	71.25 - 76.34 (5.09)
14	36	53818 - 53938 (120)	76.36 - 81.49 (5.13)
15	30	53938 - 54017 (79)	81.51 - 89.95 (8.44)
16	74	54020 - 54300 (281)	89.95 - 95.27 (5.32)
17	30	54301 - 54606 (305)	95.27 - 102.37 (7.10)
18	30	54607 - 54947 (340)	102.37 - 111.06 (8.69)
19	31	54949 - 55095 (145)	111.06 - 116.52 (5.46)
20	51	55095 - 55285 (190)	116.83 - 125.97 (9.14)
21	51	55286 - 55370 (84)	126.33 - 135.68 (9.35)
22	65	55371 - 55528 (157)	135.68 - 143.68 (8.00)
23	36	55530 - 55678 (148)	144.90 - 152.74 (7.84)
24	44	55721 - 55880 (159)	153.62 - 159.51 (5.89)
25	69	55884 - 56287 (403)	159.74 - 164.74 (5.00)
26	147	56288 - 56895 (607)	164.74 - 169.76 (5.02)
27	102	56903 - 57177 (274)	169.87 - 175.06 (5.19)
28	30	57197 - 57397 (200)	175.28 - 186.67 (11.39)
29	41	57397 - 57636 (239)	186.67 - 193.57 (6.90)
30	140	57636 - 58206 (570)	193.57 - 198.59 (5.02)
31	52	58210 - 58415 (205)	198.76 - 201.94 (3.18)

5.2.4 The composite likelihood method

The previous chapter has described the composite likelihood method that estimates the Malecot parameters (M, S, L and ε) for association mapping of causal variants. Significance tests for this method are based on contrasts between null and alternative models. The null hypothesis of model A assumes no association between SNPs and disease status, which does not estimate any Malecot parameters. On the other hand, the alternative hypothesis assumes association with disease and estimates partial or all Malecot parameters, depending on which model is used (See chapter 4). In this study, I used model D because its absolute deviation of estimated S from the true location is relatively small in the simulation test (Morton et al. 2007). This model estimates parameters M, S and L and takes the parameter ε as 1. The parameter ε is taken as 1 because it is always ~1 when an LD map is used to indicate SNP locations. χ^2 with 3 degree of freedom for the A-D contrast is estimated as X/V, where $X=[(-2\ln lk)_A-(-2\ln lk)_D]$ and V is error variance. However, when SNP density is very high, the estimation is distorted by autocorrelation due to non-independent SNPs in high LD, resulting in an inaccurate error variance (V). This problem can be solved by a permutation method that randomly shuffles case/controls status under the assumption of no association without any changes in SNP genotype to create many replicates (i.e. 1,000-10,000). Here I used 1000 replicates and each replicate j was estimated for X_j. All X_j of 1000 replicates were then ranked according to their values. p value (p_j) for each replicate was determined by the fraction of its rank in the 1000 replicates. For each corresponding replicate, p_j was converted into χ_j^2 with 3 degrees of freedom

(Abramowitz and Stegun 1965) and its error variance V_j for this replicate was estimated as X_j / χ_j^2. To estimate the error variance V from the real data (H_1), a regression: $\ln V_j = a + b \ln X_j$ was applied by fitting 20 replicates (V_j and X_j) on both sides centered on X to calculate a and b. If X is an outlier, the 20 closest replicates are taken. Therefore, V is calculated as $\exp(a + b \ln X)$. By estimating V from this method, the autocorrelation effect is avoided.

All segments in the four analyses were analysed in the same way as described above. A segment with nominally significant association with RA was identified where p value $\leq$ 0.05. For each segment, a location ($\hat{S}$) and its corresponding information (K) for $\hat{S}$ were estimated. The information K is estimated as $Kss/(V/3)$, where Kss is an information matrix with simultaneous estimates of M, S and L. The standard error (Se) was calculated as $\sqrt{1/K}$ and the 95% confidence interval (CI) was calculated as $\hat{S} \pm 1.96Se$. The 95% confidence interval (95% CI) on LDU scale were then converted to more standardised scale in kb. The model is implemented in the CHROMSCAN program.

5.2.5 Haplotype analysis for significant segments

Haplotype analyses were performed on segments if they were nominally significant in the analyses using the composite likelihood method. Common haplotypes (>1%) and their frequencies were estimated for cases and controls. The analyses were performed by the PHASE program (version 2), which implements Gibbs sampling, a form of Markov chain Monte Carlo (MCMC) algorithm, for reconstructing haplotypes from population data (Stephens et al. 2001; Stephens

and Donnelly 2003). This program also assigns a pair of the maximum likelihood haplotypes for each individual.

A simple χ^2 test was performed to identify significant haplotypes between cases and controls by testing each haplotype in turn against the rest of others. The χ^2 value for each suspected haplotype was calculated as $\chi_1^2 = \frac{N(p-q)^2}{(p+q)(2-p-q)}$ with one degree of freedom, where p and q are the haplotype frequencies in case and control groups respectively and N is the total number of haplotypes in the sample. This study collected genotype data in 460 cases and 460 controls, so N is 1840 (920×2=1840).

5.2.6 Evaluation for the performance in the CHROMSCAN program

The CHROMSCAN program is a development from the LOCATE program for genome-wide association studies. It manages multiple segments of a large region and performs a permutation test with many replicates under null hypothesis for estimation of the error variance in a significance test. However, size of segment, breakpoints of segment and number of replicates are set as optional in this program. It remains unclear how robust the findings are to varying these limits. Therefore, this study performs a test to evaluate the effects of these variables. Three point estimates were chosen, including a high significant, a moderately significant and a non-significant locus in the candidate region. Then, I performed the test by changing these variables in which three loci were located in order to evaluate their influences on the results.

5.3 Results

5.3.1 The significant segments indicated by the composite likelihood method

Tables 5.6-5.9 shows all of the results from the composite likelihood method performed by the CHROMSCAN program, including χ^2 values of the significance test, point estimates ($\hat{S}$) and 95% CI for all segments. Two segments showing significant association with RA were identified. The point estimate for the first and the most significant segment (S_1) is at 53306 or 53308 kb, indicated by all tables. The second one (S_2) at 51584 or 51585 kb is less significant and only detectable in the analyses using 5 LDUs per segment (Tables 5.7 and 5.9). The results show that the point estimates are highly consistent in the four analyses. However, χ^2 values are higher and 95% CI are smaller in the analyses using 5 LDUs per segment. In addition, using the HapMap LD map as the reference map in the analyses seems to show smaller 95% CI in kb.

Table 5.6 Results of Analysis #1 from the composite likelihood method

Analysis #1 (10 LDUs minimal and the GAW map)

#Segment	X_{A-D}	V	χ^2_{A-D} (df=3)	p value	Se	S (LDU)	95%CI(LDU)	S(kb)	95%CI(kb)
1	6.30	12.95	0.49	0.9220	0.70	8.97	7.59 - 10.35	50020	49807 - 50158
2	12.97	7.95	1.63	0.6521	11.42	20.04	-2.35 – 42.42	51574	50159 - 51575
3	20.31	3.18	6.38	0.0943	0.49	20.06	19.09 - 21.03	51575	51577 - 51680
4	10.78	12.25	0.88	0.8303	0.60	35.22	34.05 – 36.40	52683	52398 - 52898
5*	85.36	6.43	13.27	0.0041	0.18	42.70	42.35 - 43.05	53306	53273 - 53342
6	18.40	3.25	5.66	0.1294	0.29	56.50	55.93 - 57.06	53752	53732 - 53781
7	25.95	4.80	5.41	0.1440	0.24	68.62	68.15 - 69.09	54230	54215 - 54277
8	3.20	1.52	2.10	0.5510	0.40	75.65	74.87 - 76.43	54742	54636 - 54804
9	2.95	4.28	0.69	0.8755	0.69	85.21	83.85 - 86.56	55071	55027 - 55095
10	5.05	3.24	1.56	0.6687	0.55	96.50	95.43 - 97.57	55491	55384 - 55517
11	16.62	3.91	4.25	0.2361	0.90	110.45	108.68 - 112.22	55782	55723 - 55805
12	3.89	41.11	0.09	0.9925	2.67	125.64	120.41 - 130.86	57158	56371 - 57248
13	9.44	2.23	4.24	0.2366	0.22	139.18	138.75 - 139.61	57498	57466 - 57518
14	21.86	4.99	4.38	0.2235	0.53	150.90	149.86 - 151.95	58401	58264 - 58415

A segment showing significant association with RA is marked in grey colour.

Table 5.7 Results of Analysis #2 from the composite likelihood method

Analysis #2 (5 LDUs minimal and the GAW map)

#Segment	X_{A-D}	V	χ^2_{A-D} (df=3)	p value	Se	S (LDU)	95%CI(LDU)	S(kb)	95%CI(kb)
1	4.75	3.21	1.48	0.6874	0.50	2.80	1.82 - 3.78	49369	49341 - 49439
2	8.93	8.68	1.03	0.7942	0.37	8.64	7.91 - 9.37	50018	49918 - 50023
3	9.71	3.76	2.58	0.4607	0.71	10.54	9.14 - 11.93	50233	50235 - 50648
4*	44.66	5.28	8.46	0.0373	0.18	20.06	19.71 - 20.41	51585	51539 - 51628
5	9.71	1.88	5.18	0.1592	0.31	24.07	23.47 - 24.67	51915	51913 - 51916
6	8.53	3.94	2.17	0.5389	0.42	27.77	26.95 - 28.6	52005	51975 - 52040
7	3.13	6.32	0.50	0.9198	0.44	35.22	34.36 - 36.08	52682	52418 - 52864
8	5.98	3.94	1.52	0.6780	0.21	38.97	38.57 - 39.38	53113	53108 - 53125
9*	86.94	5.98	14.55	0.0022	0.13	42.70	42.44 - 42.96	53306	53295 - 53330
10	1.00	1.72	0.58	0.9015	0.00	54.50	54.5 - 54.5	53668	53668 - 53668
11	20.88	3.09	6.77	0.0797	0.43	56.52	55.68 - 57.35	53753	53722 - 53845
12	6.82	2.12	3.22	0.3590	0.94	59.53	57.69 - 61.38	53920	53922 - 53960
13	21.67	4.36	4.97	0.1739	0.22	68.59	68.16 - 69.01	54229	54215 - 54265
14	1.63	0.92	1.77	0.6208	1.69	76.15	72.85 - 79.45	54799	54430 - 54799
15	2.30	1.13	2.04	0.5644	0.00	77.27	77.27 - 77.27	54827	54827 - 54827
16	4.71	1.10	4.29	0.2315	0.17	85.24	84.9 - 85.58	55086	55054 - 55094
17	2.34	2.37	0.98	0.8052	0.62	92.66	91.44 - 93.87	55351	55285 - 55355
18	4.22	2.73	1.54	0.6721	0.33	96.45	95.79 - 97.1	55483	55454 - 55503
19	2.87	1.42	2.02	0.5680	0.89	103.61	101.86 - 105.36	55615	55583 - 55620
20	1.50	1.05	1.43	0.6991	1.48	108.06	105.17 - 110.96	55703	55657 - 55786
21	15.02	3.55	4.23	0.2375	0.61	114.03	112.82 - 115.23	55875	55838 - 56064
22	3.98	13.24	0.30	0.9599	0.86	122.72	121.04 - 124.4	56966	56395 - 57034
23	0.14	3.34	0.04	0.9976	1.49	125.75	122.84 - 128.66	57159	57035 - 57241
24	0.32	2.65	0.12	0.9895	2.12	138.06	133.89 - 142.22	57398	57305 - 57410
25	10.25	1.76	5.81	0.1213	0.15	139.18	138.89 - 139.47	57498	57466 - 57514
26	14.72	4.05	3.64	0.3034	8.62	145.24	128.35 - 162.14	57734	57644 - 58178
27	5.50	1.82	3.03	0.3871	0.29	150.88	150.31 - 151.46	58400	58344 - 58415

*Segments showing significant association with RA are marked in grey colour.

Table 5.8 Results of Analysis #3 from the composite likelihood method

Analysis #3 (10 LDUs minimal and the HapMap map)

#Segment	X_{A-D}	V	χ^2_{A-D} (df=3)	p value	Se	S (LDU)	95%CI(LDU)	S(kb)	95%CI(kb)
1	5.27	20.75	0.25	0.9684	1.26	3.41	0.94 – 5.88	49391	49313 - 49442
2	4.68	9.53	0.49	0.9209	4.41	10.02	1.37 - 18.67	49935	49936 - 50901
3	37.32	5.33	7.00	0.0719	0.19	25.45	25.07 - 25.82	51584	51535 - 51614
4	6.01	4.95	1.21	0.7498	0.73	35.97	34.55 - 37.39	52005	51961 - 52057
5	2.59	17.74	0.15	0.9858	1.29	50.22	47.69 - 52.75	52683	52401 - 52916
6*	88.10	7.04	12.51	0.0058	0.18	59.92	59.57 - 60.26	53308	53296 - 53332
7	20.04	3.00	6.68	0.0827	0.31	75.64	75.03 - 76.25	53777	53734 - 53803
8	7.06	2.13	3.32	0.3451	0.49	79.91	78.95 - 80.87	53932	53922 - 53934
9	19.77	3.88	5.10	0.1649	0.27	93.74	93.21 - 94.26	54232	54215 - 54279
10	4.40	1.64	2.69	0.4423	0.40	103.98	103.19 - 104.77	54703	54653 - 54807
11	3.94	1.79	2.21	0.5302	0.87	115.14	113.43 - 116.85	55088	55012 - 55095
12	0.51	3.60	0.14	0.9865	0.94	126.94	125.09 - 128.78	55336	55288 - 55352
13	4.60	2.67	1.73	0.6313	0.00	137.90	137.9 - 137.9	55488	55488 - 55488
14	1.33	1.13	1.18	0.7580	1.09	150.25	148.11 - 152.39	55664	55651 - 55676
15	22.73	5.10	4.46	0.2162	0.31	165.07	164.45 - 165.68	56306	56254 - 56320
16	5.04	10.89	0.46	0.9269	0.66	170.59	169.29 - 171.88	56969	56725 - 57032
17	10.10	2.10	4.82	0.1855	0.29	189.68	189.11 - 190.24	57537	57519 - 57542
18	23.16	4.97	4.66	0.1983	0.35	201.70	201.01 - 202.39	58387	58313 - 58415

*A segment showing significant association with RA is marked in grey colour.

Table 5.9 Results of Analysis #4 from the composite likelihood method

Analysis #4 (5 LDUs minimal and the HapMap map)

#Segment	X_{A-D}	V	χ^2_{A-D} (df=3)	p value	Se	S (LDU)	95%CI(LDU)	S(kb)	95%CI(kb)
1	4.61	3.22	1.43	0.6976	1.30	4.81	2.26 - 7.35	49440	49330 - 49443
2	3.99	7.42	0.54	0.9104	1.89	10.02	6.31 - 13.72	49954	49446 - 50028
3	6.00	7.24	0.83	0.8424	0.29	12.70	12.13 - 13.27	50299	50157 - 50425
4	5.76	3.01	1.91	0.5906	0.76	19.16	17.67 - 20.64	50937	50886 - 51042
5*	41.22	3.95	10.42	0.0153	0.13	25.44	25.19 - 25.7	51584	51564 - 51604
6	8.50	1.24	6.83	0.0777	4.30	30.63	22.2 - 39.07	51912	51886 - 51962
7	1.23	2.44	0.50	0.9180	0.35	36.00	35.31 - 36.69	52006	51981 - 52048
8	6.78	3.33	2.03	0.5654	0.61	45.65	44.47 - 46.84	52396	52396 - 52397
9	7.38	5.13	1.44	0.6966	0.29	50.18	49.61 - 50.75	52660	52637 - 52765
10	3.63	4.03	0.90	0.8254	2.04	59.55	55.56 - 63.55	53274	53109 - 53274
11*	86.24	6.22	13.86	0.0031	0.14	59.92	59.64 - 60.2	53308	53296 - 53330
12	0.52	1.46	0.36	0.9491	1.34	67.51	64.88 - 70.14	53569	53472 - 53613
13	16.92	2.72	6.22	0.1013	0.60	75.51	74.33 - 76.69	53771	53704 - 53807
14	4.91	1.85	2.66	0.4473	0.32	76.96	76.35 - 77.58	53837	53818 - 53848
15	5.16	1.68	3.08	0.3796	1.82	81.51	77.94 - 85.08	53938	53943 - 53996
16	22.39	3.88	5.77	0.1234	0.22	93.68	93.25 - 94.1	54230	54216 - 54269
17	0.95	1.56	0.61	0.8947	1.03	99.98	97.96 - 102	54448	54365 - 54606
18	1.02	1.30	0.79	0.8529	0.55	104.04	102.97 - 105.11	54706	54624 - 54811
19	5.27	1.27	4.14	0.2466	0.42	115.14	114.31 - 115.98	55088	55024 - 55094
20	2.08	1.51	1.38	0.7107	1.64	118.83	115.61 - 122.05	55269	55097 - 55282
21	0.47	2.86	0.16	0.9833	0.35	126.91	126.23 - 127.6	55311	55288 - 55347
22	2.79	2.64	1.06	0.7871	0.00	137.90	137.9 - 137.9	55488	55488 - 55488
23	3.00	1.42	2.12	0.5485	0.49	146.88	145.92 - 147.84	55613	55591 - 55650
24	6.41	1.79	3.58	0.3111	0.43	155.18	154.34 - 156.03	55785	55772 - 55802
25	11.48	3.22	3.57	0.3122	1.63	164.69	161.49 - 167.88	56274	56166 - 56287
26	5.89	7.96	0.74	0.8637	0.81	164.92	163.33 - 166.52	56297	56289 - 56378
27	0.57	10.07	0.06	0.9965	3.57	169.89	162.9 - 176.89	56915	56908 - 57177
28	0.39	4.38	0.09	0.9931	3.86	177.34	169.77 - 184.91	57220	57200 - 57329
29	8.44	1.83	4.63	0.2013	0.22	188.63	188.2 - 189.06	57475	57469 - 57503
30	15.46	4.21	3.67	0.2994	1.79	195.58	192.08 - 199.08	57730	57644 - 58206
31	4.97	1.79	2.78	0.4275	0.23	201.71	201.25 - 202.17	58387	58363 - 58415

*Segments showing significant association with RA are marked in grey colour.

5.3.2 Haplotype analyses for the significant segments

Figure 5.2 presents the HapMap LD maps for the two significant segments including the point estimates, S_1 and S_2, and their 95% confidence intervals. The black dots on the maps are those SNPs showing significant association with RA in the single SNP test (p value < 0.05).

Haplotype analyses were performed on each of the two significant segments, focusing on the area within a small LDU distance that contains the putative caused locus and the majority of significant SNPs (See table 5.10). The first area (S_1) is between 53297 and 53312 kb with 0.043 LDUs. This area contains 16 SNPs in which 14 SNPs are highly significant. Another area (S_2) is between 51555 and 51616 kb with 0.368 LDUs. This area contains 21 SNPs in which 12 SNPs show significant association.

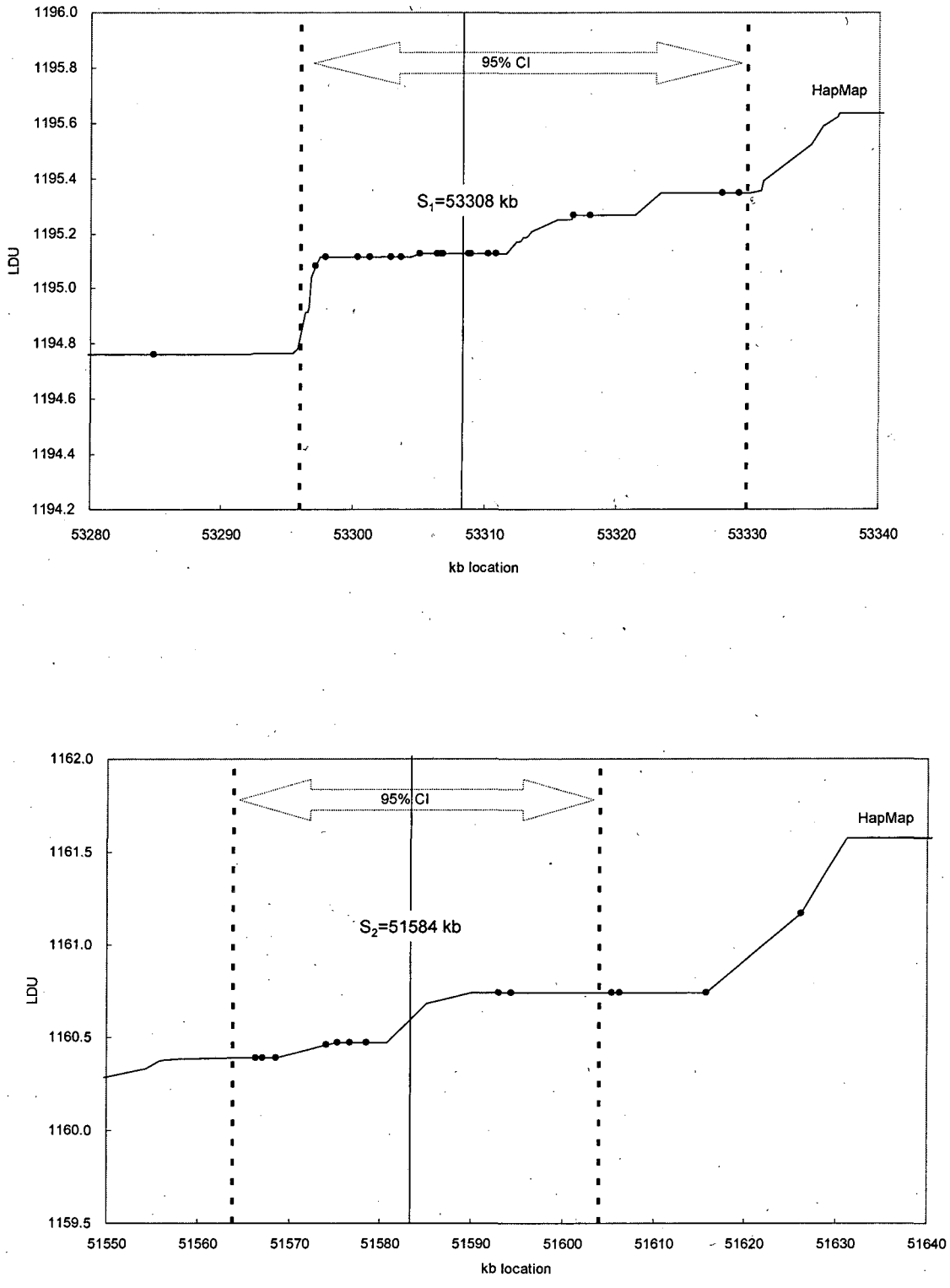


Figure 5.2 LD maps for the significant regions S_1 and S_2

The black dots on the maps are those SNPs showing significant association with RA in the single SNP test (p value < 0.05).

Table 5.10 Selected SNPs from the two significant segments with their locations for haplotype analyses

The selected SNPs in the S_1 area

No.	rs-number	kb location	LDU distance	χ_1^2
1	rs660936	53297.068	0	5.22234
2	rs674849	53297.884	0.032	9.97169
3	rs615030	53300.352	0.032	9.79587
4	rs629737	53301.269	0.032	11.61718
5	rs519596	53302.863	0.032	11.95221
6	rs660626	53303.627	0.032	10.54578
7	rs3745070	53303.942	0.032	1.54977
8	rs3745064	53305.064	0.043	12.24676
9	rs3848516	53306.378	0.043	10.65911
10	rs608017	53306.709	0.043	10.69989
11	rs608823	53306.868	0.043	10.69989
12	rs552396	53308.810	0.043	11.0443
13	rs2279096	53308.956	0.043	10.65911
14	rs1217583	53310.272	0.043	8.07478
15	rs3899444	53310.926	0.043	11.00203
16	rs4940796	53311.675	0.043	0.29391

The selected SNPs in the S_2 area

No.	rs-number	kb location	LDU location	χ_1^2
1	rs813043	51555.692	0	2.51912
2	rs784254	51556.380	0.008	2.51912
3	rs711745	51558.095	0.011	2.82461
4	rs784251	51563.901	0.018	2.66863
5	rs4800995	51566.375	0.018	8.06634
6	rs784237	51567.112	0.018	8.89857
7	rs796743	51568.637	0.018	8.67925
8	rs784235	51574.142	0.09	8.50657
9	rs784233	51575.244	0.098	0.05456
10	rs4800996	51575.356	0.099	9.16822
11	rs3745044	51576.668	0.099	10.24285
12	rs784232	51578.479	0.099	10.51579
13	rs1642295	51580.697	0.099	0.33626
14	rs784240	51585.120	0.309	0.33626
15	rs1362781	51590.034	0.368	3.56218
16	rs2306163	51593.047	0.368	6.49247
17	rs931040	51594.427	0.368	9.42302
18	rs4996482	51605.367	0.368	6.26716
19	rs899101	51606.229	0.368	5.9369
20	rs899102	51606.327	0.368	2.7005
21	rs1031830	51615.765	0.368	10.82539

There are 22 haplotypes in the S_1 region, but four common haplotypes (H_1, H_2, H_3 and H_4) represent 95% of individuals in the sample, shown in table 5.11. Haplotypes H_1 and H_3, appear to have very significant difference in haplotype frequencies between cases and controls. The S_2 area contains more SNPs and is not in a well-characterised block. Therefore, the total number of haplotypes is 58. It requires at least 9 haplotypes to represent 95% of individuals in the sample (Table 5.12). The results show that only haplotype H_5 is significantly associated with the disease. Haplotype H_1 in the S_1 area and haplotype H_5 in the S_2 area have similar properties (protective effect against the disease), but there is no association between the two haplotypes in the control group, examined by a simple χ^2 test (χ^2=0.396).

Table 5.11 Haplotypes and haplotype frequencies in the significant area of S_1

Haplotype number	Haplotype	Frequency				
		Total	Case	Control	χ^2	p value*
	++++++-++++++++-					
H_1	1001010001001000	0.661	0.628	0.695	9.22	0.0024 (0.0096)
H_2	0110100110110010	0.169	0.180	0.158	1.59	0.2073
H_3	0110100110110110	0.101	0.120	0.082	7.32	0.0068 (0.0272)
H_4	0001010001001000	0.019	0.013	0.025	3.55	0.0595

+: SNP being significant; -: SNP being non-significant

*: p values <0.05 are underlined; p values with brackets are corrected by Bonferroni correction.

Table 5.12 Haplotypes and haplotype frequencies in the significant area of S_2

Haplotype number	Haplotype ----++++-+++---++++-+	Frequency				
		Total	Case	Control	χ^2	p value*
H_1	01011000011011011111	0.339	0.343	0.336	0.10	0.7518
H_2	10101000011011101001	0.256	0.274	0.238	3.13	0.0769
H_3	10101000011011011100	0.100	0.089	0.110	2.26	0.1328
H_4	10100111000111010011	0.079	0.074	0.084	0.63	0.4274
H_5	01010111000111010011	0.075	0.056	0.094	9.57	<u>0.0020</u> <u>(0.018)</u>
H_6	10101000011011001100	0.048	0.056	0.039	2.94	0.0864
H_7	01010000100000011011	0.040	0.036	0.043	0.59	0.4424
H_8	10100000011011011100	0.012	0.014	0.011	0.34	0.5598
H_9	01011111000111010011	0.012	0.010	0.014	0.62	0.431

+: SNP being significant; -: SNP being non-significant

*: p values <0.05 are underlined; p values with brackets are corrected by Bonferroni correction.

5.3.3 Effects on the performance in the program

To evaluate effects from the options used in the program, a test was performed and three point estimates were chosen, including the most significant S_1 locus at 53,308 kb, the moderately significant S_2 locus at 51,584 kb and a non-significant locus at 54,231 kb. Table 5.13 shows that point estimates and 95 % CI are highly robust to size of segment; only a slight increase in 95% CI as size of segment increases. Although enlarged length of segment would include many irrelevant SNPs and therefore decrease χ^2 value, the association of the S_1 locus is still detectable when the length increases to 40 LDUs (near 2 Mb of the physical

distance). If a segment includes both the S_1 and the S_2 loci, only the S_1 locus with stronger association is indicated. Table 5.14 shows the effects of number of replicates on the results for the S_1 locus. The χ^2 value remains very stable as the number of replicates is more than 250, but less reliable as it below 100. In addition, if the S_1 locus is on the edge of a segment, the effect is very little even though the S_1 locus is not in the segment (See table 5.15).

Table 5.13 The effects of size of segment on the results for three loci with different intensities of association

segment	Segment Size(LDUs)	Physical Size(kb)	SNP Number	$\chi^2_{A\text{-}D}$ (df=3)	p value	Point estimate	95%CI
highly	2	102	57	12.782	0.0051	53,308	30
significant (S_1)	5	139	72	15.218	0.0016	53,308	32
	10	351	146	9.782	0.0205	53,308	60
	20	971	303	9.755	0.0208	53,308	60
	40	1,834	517	7.989	0.0462	53,308	62
	80	3,463	906	6.955	0.0733	53,308	67
moderately	2	297	51	10.901	0.0123	51,586	28
significant (S_2)	5	503	89	9.473	0.0236	51,583	48
	10	934	200	6.305	0.0977	51,584	89
	20	1,318	262	6.506	0.0894	51,584	92
	40	2,959	728	2.481	0.4787	51,584	218
	80	4,560	1,152	5.086	0.1656	53,308*	71
Non-significant	2	98	33	5.758	0.1240	54,236	63
	5	251	60	5.075	0.1664	54,235	90
	10	416	103	5.303	0.1509	54,231	59
	20	715	118	4.874	0.1813	54,233	69
	40	1,359	244	3.768	0.2876	53,777	70
	80	2,405	585	9.795	0.0204	53,308*	62

The point estimates with * indicate the S_1 locus because the size of those segments is so large that the S_1 locus is included.

Table 5.14 The effects of number of replicates on the results for the most significant locus

Number of replicates	Segment Size(LDUs)	SNP Number	$\chi^2_{A\text{-}D}$ (df=3)	p value	Point estimate	95%CI
50	10	146	8.474	0.0372	53,308	61
100	10	146	12.670	0.0054	53,308	36
250	10	146	10.334	0.0159	53,308	59
500	10	146	10.280	0.0163	53,308	59
1000	10	146	9.782	0.0205	53,308	60
2500	10	146	10.524	0.0146	53,308	59

Table 5.15 The effects of breakpoints in segment on the results for the most significant locus

LDUs between S_1 and the nearby breakpoint*	Segment Size(LDUs)	SNP Number	$X^2_{A\text{-}D}$ (df=3)	p value	Point estimate	95%CI
-0.3	10	91	14.874	0.0019	53,305	40
0	10	96	14.204	0.0026	53,301	35
0.5	10	230	8.783	0.0323	53,308	56
1	10	115	12.502	0.0058	53,308	35
2	10	118	12.625	0.0055	53,308	35
2.5	10	192	10.286	0.0163	53,308	37

A value of zero indicates that the S_1 locus (53308 kb) is on one of the breakpoints in a segment. If a value is minus, it means that the S_1 locus is not in the segment.

5.4 Discussion

Since many genome-wide association studies for different common diseases are being carried out, developing powerful analytical approaches for identification of genetic susceptibility variants from a large region is increasingly important. It is a challenge for almost all association approaches to analyse a large dataset including thousands of SNPs. An effective solution, at the initial stage of disease mapping, is to screen a large region in segments with fewer SNPs, in order to identify segments that are significantly associated with disease, followed by more detailed analyses of significant segments and replication along with meta-analysis (Morton et al. 2007). Since the majority of segments are not expected to be associated with disease, excluding non-significant segments after the initial screen could save lots of time and resource.

The composite likelihood method, which evaluates whether a segment is associated with disease of interest by considering association information from all SNPs in a segment simultaneously, is capable of screening a large region for signals of association. This study is a good illustration of this powerful method. In this study, I identified two significant segments that contain a group of SNPs and one or two particular haplotypes with significant differences in frequencies between cases and controls. This suggests that this method can facilitate single SNP testing and haplotype analysis for characterisation of significant regions. An important haplotype in the significant segment at 53297-53312 kb is strongly associated with RA. This haplotype exists in the majority of individuals, accounting for 66.1% of DNA samples, and has protective effect against RA. Currently, no gene has been reported in this area, but it contains 4 human

mRNAs (CR590917, AK021717, AK124558 and BC013134) and highly conserved DNA sequences in different species, implying some functional importance. In the other significant segment, a gene named AK127787 is located at 51595-51600 kb. The knowledge of this gene is very limited. DNA sequences in this area also show potential importance of functional mechanism. Further analyses using functional tests or DNA sequencing are necessary to confirm these findings.

The composite likelihood method provides point estimates and 95% CI on the LDU scale, which are variable between populations. The locations can be transformed by interpolation into locations on the kb scale. This study shows that point estimates and 95% CI are very robust to size of segment and choice of reference map. There are some limitations to this approach. If a segment includes multiple causal variants with strong effects at different loci, the point estimate for this segment could be distorted by the interference of these variants and p value increased. This will happen more frequently if a segment is very large. Secondly, if a causal variant is very close to one of the breakpoints in a segment, parts of SNPs surrounding the causal variant will be truncated in analyses, resulting in loss of information. This could happen in any segmentation of a large region, but it is more likely to happen as smaller segments are examined. However, this study also shows that the effect is very small from these two cases if association is very strong. In addition, using smaller segments in analyses tends to generate higher χ^2 values and smaller confidence intervals, but this does not support using fewer than 5 LDUs per segment, at the initial screen, due to more tests and fewer SNPs in a segment. On the other hand, enlarged length of segment decreases the χ^2 value, which could be due to noise from distant and irrelevant SNPs, and other sites with independent and confounding effects on the trait.

This study with a moderate-size region provides a good opportunity to evaluate the performance of the CHROMSCAN program. The required computing time to analyse one segment depends on the number of replicates and the number of SNPs in the segment. More replicates and more SNPs per segment require more time. To use the program efficiently for a large data, it is suggested to run 100-250 replicates for each segment at the beginning to identify significant regions. Then, further investigation on significant regions uses 1,000 or more replicates to increase accuracy of χ^2 and 95% CI. In fact, this study shows that the results are highly stable even when the number of replicates is only 250.

The increase in false positive rates by chance in multiple testing on SNPs, haplotypes and segments is a common problem in genome-wide disease mapping. A simple way to reduce the error rate is Bonferroni correction using strict statistical criteria for significance level. However, as the number of tests increases, many putatively positive results do not satisfy the criteria after the correction. In the case of this study, the nominal p value for the most significant SNP among 2293 SNPs is 4.66×10^{-4}, which is not significant after correction as it would need to be 2.18×10^{-5} (0.05/2293) using Bonferroni. Although the composite likelihood method based on a multi-markers approach reduces 2293 tests to 18 tests, the p value for the most significant S_1 locus, 5.8×10^{-3}, is close to the corrected significance level of 2.8×10^{-3} (0.05/18). Because the Bonferroni criterion is thought to be too conservative, many significant results are omitted in multiple-hypothesis tests including the true ones. Controlling the false positive rate without missing causal polymorphisms is essential. One possible route to the management of this problem is through the development of the false discovery rate (FDR). Another feasible solution is to stratify the same samples into different groups according to

some common variables, such as age and sex, and re-analyse, test a second or another samples and combine evidence across samples using meta-analysis (Morton 2007). Confirmation of evidence is best achieved by replication studies for any putatively significant region using additional samples if budget and time are feasible.

Chapter 6 Summary

Linkage Disequilibrium (LD) is a measure of the degree of association between alleles in a population. Previous studies have shown that the pattern of LD is highly variable in different chromosome regions and different populations. Therefore, it is useful to construct a genome-wide LD map at high resolution throughout the human genome. Association between pairs of SNPs can be modeled using the Malecot equation that describes the decline in LD with distance under the composite likelihood method. The magnitude of the ε parameter of the Malecot equation indicates the region of the genome with extensive and less extensive LD.

The construction of a genome-wide LD map encounters computational difficulties induced by the volume of SNP pairwise data generated from a large genotype sample. This can be solved by separating a large dataset into smaller sub-datasets and excluding the uninformative distant pairs. The estimation of a distance in LDU between any of two adjacent SNPs is highly robust at sufficient marker densities. Powerful computers with parallel computing technology can also facilitate the map construction more efficiently.

In the thesis, these strategies were used to construct genome-wide LD maps for four major populations (Caucasian, Chinese, Japanese and African) from the phase II data of the HapMap project. The LDMAP-Cluster program exploiting parallel computation process was used for rapid map construction. A comparison of patterns of LD across the four populations are also presented. The results show

"out of African" populations exhibit more extensive LD than African for all chromosomes, highlighting the importance of population demography in shaping the pattern of LD. Despite those differences, the general view of LD patterns is similar across the populations, indicating recombination dominates these patterns.

The application of LD to map genes of complex disease using high density maps of SNPs in candidate region is currently an active research area in human genetics. I describe an association approach that utilises LD maps for reliable localisation of disease-causing variants. This method uses composite likelihood estimate of location for a causal variant by combining association information from all SNPs. I also examine the performance of this mapping approach using three case/control studies of Rheumatoid Arthritis (RA). The results of these studies demonstrate the great potential of the genome-wide LD maps for high-resolution mapping of disease genes, and practical implications for appropriate design and selection of SNPs for disease association studies.

Genome-wide association studies (GWAS) involving hundreds of thousands of SNPs in cases and controls are getting common today. Recently, a joint GWA study of several common diseases using 500,000 SNPs has identified association signals at many loci across the genome (Consortium 2007). A challenge for these studies is the analyses of a huge number of SNP, contributing to many false positive results. A two-stage design for disease gene mapping is therefore suggested. The first stage identifies significant regions associated with disease of interest from the whole genome followed by the second stage that further localise the causal polymorphisms in these regions. The CHROMSCAN program analysis by

segments is very suitable for genome-wide association studies. The results of this study demonstrate the efficiency and robustness in this approach for the localisation of variants which contribute to human diseases.

Several novel methods and extensions of existing methods, using multiple SNP analysis, are proposed for candidate regions and genome-wide association studies. These methods are believed to have greater power than single SNP analysis because they combine more information from multiple SNPs. Despite different strategies, methodologies, algorithms and statistical measures in these methods, they all require heavy computation (permutation, simulation and iteration) and suitable software. Unlike CHROMSCAN which combines information across several SNPs simultaneously and takes into account LD information, many other methods derive haplotypes or "optimal" sets of markers based on unique algorithms and then perform chi-square tests. These methods include the localised haplotype cluster algorithm (Browning and Browning 2007), sequential haplotype scan (Yu and Schaid 2007b), pattern-based data mining (Li et al. 2007) and backward search algorithm (Lo and Zheng 2002). Although methods for inferring haplotypes or optimal sets of SNPs for analyses are not the same, and results are often inconsistent, one common feature of these methods is that they can detect association through a combination of SNPs that is not detectable for individual SNPs. In the case of rheumatoid arthritis association studies of GAW 15, the sequential haplotype scan approach (Yu and Schaid 2007a) indicated a set of SNPs between 53,716-53,747 kb containing only 1 significant SNP (uncorrected P=0.015) amongst five SNPs. The localised haplotype cluster algorithm (Browning and Thomas 2007) indicated that the most significant haplotype contains two SNPs at 555,158 and 555,159 kb respectively, but none of those and their

neighboring SNPs show significant association in single SNP analysis. Among these methods, only the pattern-based mining strategy (Li et al. 2007) indicated the significant region (51566-51594 kb) agreeing with the CHROMSCAN results.

Thomas and Camp (2004) developed a graphic model to describe allelic association between SNPs. It can be used to detect allele-phenotype association by fitting the model. Another approach that takes into account the shared ancestry of sampled chromosomes is based on a coalescent model for fine mapping (Morris et al. 2002). However, the two approaches are computationally intensive and may not be suitable for large-scale genome-wide association studies. Instead, clustering haplotypes through similarity, without an explicit model, could be a much faster approach (Molitor et al. 2003). An efficient analysis using the CHROMSCAN program has been performed on several large-scale genome-wide association datasets using a segmental method and parallel processing. There are other advantages of CHROMSCAN for association mapping. Large samples containing thousands of individuals do not pose a computational issue for CHROMSCAN, but severely limit methods that require haplotype reconstruction. Although CHROMSCAN employs EM algorithm for haplotype frequencies from unphased genotype data, this error is likely to be minimal in 2-SNP haplotypes. However, estimating haplotypes over large distances with many SNPs and individuals may increase error. Therefore, investigators should be always aware of potential biases in multiple SNP analysis.

Appendix A: General information for population-specific genome-wide LD maps

A-1: The CEU genome-wide LD map

Chromosome	Number of SNP	Physical Length (kb)	LDU length	Block coverage	Number of Hole
1	150,782	244,659	4,354	66.66%	138
2	181,137	242,758	4,245	73.93%	127
3	143,316	199,300	3,689	74.09%	142
4	130,618	191,392	3,527	73.19%	125
5	138,633	180,570	3,366	74.40%	92
6	149,019	170,739	3,240	75.78%	110
7	112,786	158,494	2,997	72.08%	99
8	122,449	146,116	2,723	74.09%	82
9	100,395	138,349	2,619	60.58%	92
10	111,087	135,330	2,794	72.83%	81
11	104,696	134,262	2,630	74.28%	81
12	100,054	132,330	2,731	74.33%	95
13	84,148	96,202	2,052	76.29%	62
14	68,323	87,141	1,890	74.82%	55
15	58,353	82,048	1,971	70.09%	94
16	56,920	88,667	1,984	59.95%	65
17	47,488	78,587	1,955	69.17%	80
18	62,838	76,115	1,891	72.99%	75
19	29,248	63,584	1,732	56.66%	93
20	50,997	62,376	1,707	69.05%	63
21	27,925	36,996	997	65.52%	49
22	26,721	35,081	1,023	70.41%	28
X	52,648	151,794	1,702	69.76%	105
Total	2,110,581	2,932,892	57,820	71.26%	2,033

A-2: The CHB genome-wide LD map

Chromosome	Number of SNP	Physical Length (kb)	LDU length	Block coverage	Number of Hole
1	138,265	244,823	4,832	65.06%	259
2	161,035	242,749	4,633	70.91%	236
3	125,336	199,300	4,159	71.74%	244
4	115,535	191,357	3,808	70.91%	209
5	122,041	180,570	3,694	72.08%	195
6	133,588	170,731	3,477	74.08%	173
7	99,565	158,489	3,304	69.39%	171
8	111,829	146,115	3,066	71.77%	154
9	91,408	138,347	2,969	58.12%	160
10	100,721	135,304	3,154	70.74%	173
11	94,756	134,261	2,970	72.32%	180
12	89,331	132,330	3,160	70.80%	192
13	76,026	96,206	2,339	74.98%	136
14	62,366	87,141	2,058	73.42%	104
15	54,387	82,014	2,283	68.56%	164
16	51,250	88,667	2,251	56.78%	131
17	41,641	78,583	2,267	65.88%	146
18	56,551	76,113	2,255	70.36%	164
19	27,067	63,584	2,056	54.88%	195
20	45,551	62,376	1,947	66.75%	125
21	26,825	36,996	1,147	64.99%	87
22	24,854	35,070	1,150	64.42%	61
X	44,855	151,794	1,955	65.44%	179
Total	1,894,783	2,932,920	64,931	68.84%	3,838

A-3: The JPT genome-wide LD map

Chromosome	Number of SNP	Physical Length (kb)	LDU length	Block coverage	Number of Hole
1	137,432	244,815	4,412	65.24%	211
2	160,220	242,749	4,257	71.83%	185
3	124,031	199,300	3,688	71.98%	164
4	113,996	191,357	3,547	72.39%	166
5	121,285	180,570	3,379	73.21%	145
6	132,778	170,731	3,195	74.93%	130
7	98,849	158,489	3,023	70.12%	143
8	111,336	146,115	2,791	72.18%	121
9	91,166	138,347	2,779	59.06%	142
10	100,032	135,304	2,891	70.33%	116
11	95,208	134,261	2,617	72.89%	103
12	88,627	132,330	2,801	72.90%	137
13	75,508	96,206	2,118	74.98%	107
14	61,653	87,141	1,953	73.11%	99
15	53,661	82,014	2,008	69.16%	114
16	51,419	88,667	2,115	57.39%	121
17	41,216	78,583	1,970	66.52%	107
18	55,796	76,113	2,023	71.94%	113
19	26,469	63,584	1,763	54.04%	132
20	45,216	62,376	1,713	68.58%	93
21	26,700	36,996	956	65.65%	46
22	24,824	35,070	1,140	66.55%	81
X	43,156	151,794	1,593	66.30%	124
Total	1,880,578	2,932,912	58,731	69.55%	2,900

A-4: The YRI genome-wide LD map

Chromosome	Number of SNP	Physical Length (kb)	LDU length	Block coverage	Number of Hole
1	171,661	244,820	6,101	62.74%	74
2	203,378	242,801	6,214	70.36%	79
3	156,422	199,327	5,139	70.85%	73
4	145,604	191,379	4,895	69.60%	62
5	149,100	180,573	4,768	69.96%	60
6	159,954	170,736	4,670	71.94%	50
7	121,524	158,489	4,117	67.82%	52
8	136,719	146,116	4,004	70.99%	36
9	108,871	138,349	3,641	57.42%	71
10	123,472	135,261	3,989	69.04%	51
11	112,579	134,261	3,667	70.86%	46
12	108,658	132,370	3,835	69.71%	58
13	96,127	96,203	3,009	73.50%	36
14	74,677	87,095	2,599	71.81%	38
15	65,087	81,906	2,589	67.02%	28
16	63,607	88,667	2,685	57.74%	47
17	50,978	78,583	2,648	64.88%	59
18	73,149	76,115	2,813	69.43%	51
19	32,306	63,580	2,075	53.32%	53
20	57,125	62,376	2,356	66.27%	43
21	31,421	36,996	1,399	61.94%	19
22	30,230	35,081	1,445	63.83%	27
X	64,057	151,794	2,689	68.28%	103
Total	2,336,706	2,932,878	81,346	67.74%	1,216

Appendix B: Block structure information for population-specific genome-wide LD maps

B-1: The CEU genome-wide LD map

Chromosome	Block number	Mean size (kb)	<2 kb	<5 kb	<10 kb	<30 kb	<100 kb
1	16,590	8.89	30.80%	54.11%	72.39%	93.16%	99.39%
2	18,518	8.88	30.91%	53.64%	72.28%	93.34%	99.40%
3	14,835	9.12	30.54%	52.89%	71.39%	92.96%	99.45%
4	13,764	9.65	29.10%	51.33%	69.59%	92.33%	99.62%
5	14,329	8.69	32.33%	54.59%	73.22%	93.45%	99.54%
6	14,491	8.41	32.65%	55.42%	74.21%	93.93%	99.62%
7	11,871	8.90	31.77%	53.72%	71.86%	93.08%	99.47%
8	12,396	8.18	34.23%	56.32%	74.33%	94.43%	99.61%
9	10,435	7.44	34.89%	59.17%	77.29%	95.09%	99.59%
10	11,530	7.86	34.35%	57.25%	75.59%	94.82%	99.55%
11	10,936	8.36	32.97%	55.72%	73.77%	94.05%	99.46%
12	10,708	8.44	32.66%	56.04%	74.26%	93.48%	99.51%
13	8,569	8.18	34.45%	56.61%	74.56%	94.46%	99.71%
14	7,264	8.27	33.12%	56.10%	74.23%	94.23%	99.53%
15	6,583	7.79	34.23%	58.09%	75.86%	94.74%	99.41%
16	6,857	7.07	37.38%	61.92%	79.41%	95.26%	99.55%
17	5,809	8.17	32.14%	55.86%	75.50%	94.22%	99.28%
18	6,763	7.82	33.65%	57.08%	75.81%	94.81%	99.69%
19	4,030	8.43	29.01%	52.68%	73.52%	94.84%	99.55%
20	5,853	7.08	36.53%	60.11%	78.97%	95.80%	99.76%
21	3,257	7.28	33.65%	58.40%	77.71%	95.98%	99.85%
22	3,190	6.84	39.40%	62.88%	80.44%	95.30%	99.47%
X	5,340	15.10	19.76%	37.60%	56.18%	81.80%	96.99%
Total	223,918	8.55	32.37%	55.18%	73.58%	93.65%	99.47%

B-2: The CHB genome-wide LD map

Chromosome	Block number	Mean size (kb)	<2 kb	<5 kb	<10 kb	<30 kb	<100 kb
1	15,256	9.26	30.08%	52.40%	71.29%	92.48%	99.19%
2	16,712	9.44	29.97%	52.08%	70.57%	92.44%	99.42%
3	13,650	9.30	29.37%	52.10%	70.91%	92.44%	99.17%
4	12,788	9.95	28.08%	50.09%	68.83%	92.04%	99.48%
5	13,029	9.11	31.58%	53.29%	71.71%	92.91%	99.39%
6	13,502	8.67	32.57%	55.44%	73.42%	93.32%	99.49%
7	11,065	9.12	30.97%	53.50%	71.77%	92.78%	99.39%
8	11,523	8.41	33.29%	55.85%	73.94%	93.67%	99.52%
9	9,733	7.81	34.76%	58.25%	76.48%	94.68%	99.69%
10	10,833	8.15	34.27%	56.76%	74.75%	94.32%	99.51%
11	9,939	8.73	32.27%	54.76%	72.77%	93.20%	99.28%
12	9,889	8.73	31.05%	53.82%	73.00%	93.38%	99.44%
13	7,931	8.41	32.08%	54.92%	73.61%	93.83%	99.51%
14	6,969	8.55	33.25%	55.72%	73.38%	93.49%	99.61%
15	6,392	7.67	33.50%	57.13%	76.25%	94.87%	99.30%
16	6,318	7.24	36.10%	60.19%	78.30%	95.25%	99.51%
17	5,329	8.38	32.11%	55.23%	74.89%	93.51%	99.16%
18	6,429	7.91	33.60%	56.53%	75.61%	94.71%	99.70%
19	3,752	8.60	28.84%	53.76%	73.77%	94.19%	99.41%
20	5,289	7.39	35.53%	59.41%	77.75%	95.58%	99.62%
21	3,177	7.37	33.55%	58.33%	76.52%	96.07%	99.84%
22	2,974	6.91	38.16%	62.11%	79.69%	95.36%	99.43%
X	4,679	15.27	20.13%	37.79%	54.91%	80.83%	96.35%
Total	207,158	8.82	31.67%	54.30%	72.76%	93.17%	99.36%

B-3: The JPT genome-wide LD map

Chromosome	Block number	Mean size (kb)	<2 kb	<5 kb	<10 kb	<30 kb	<100 kb
1	15,221	9.27	30.12%	52.31%	71.24%	92.46%	99.13%
2	16,246	9.72	29.79%	51.61%	69.61%	91.89%	99.30%
3	13,363	9.55	30.26%	51.96%	70.51%	91.94%	99.18%
4	12,299	10.52	27.95%	48.68%	66.99%	90.80%	99.44%
5	12,632	9.45	30.29%	51.65%	70.27%	92.48%	99.27%
6	13,027	9.02	30.86%	52.69%	71.88%	92.88%	99.39%
7	10,517	9.52	30.69%	52.33%	70.46%	91.95%	99.26%
8	11,105	8.75	33.02%	55.54%	72.99%	93.10%	99.51%
9	9,644	7.80	33.30%	57.28%	76.00%	94.97%	99.55%
10	10,594	8.22	33.82%	56.84%	74.70%	94.01%	99.48%
11	9,852	9.01	31.63%	54.02%	72.20%	92.85%	99.46%
12	9,426	9.22	30.08%	52.41%	71.47%	93.02%	99.32%
13	7,927	8.42	31.68%	54.55%	73.47%	93.76%	99.48%
14	6,784	8.62	32.36%	55.13%	73.14%	93.59%	99.48%
15	6,125	8.02	33.70%	57.32%	74.96%	93.93%	99.20%
16	6,237	7.24	36.09%	60.11%	78.07%	95.17%	99.42%
17	5,255	8.68	31.76%	54.96%	74.37%	93.00%	99.18%
18	6,014	8.44	31.69%	54.59%	73.83%	94.05%	99.55%
19	3,604	8.74	29.50%	53.66%	73.20%	93.65%	99.33%
20	5,153	7.68	33.94%	58.22%	76.46%	94.97%	99.53%
21	3,076	7.53	32.51%	56.11%	75.88%	95.74%	99.68%
22	2,865	7.32	35.95%	59.65%	78.33%	94.83%	99.37%
X	4,377	15.40	20.04%	37.45%	54.19%	79.92%	95.59%
Total	201,343	9.07	31.14%	53.47%	71.91%	92.74%	99.28%

B-4: The YRI genome-wide LD map

Chromosome	Block number	Mean size (kb)	<2 kb	<5 kb	<10 kb	<30 kb	<100 kb
1	22,807	6.33	37.83%	63.04%	81.12%	96.84%	99.72%
2	25,598	6.33	37.57%	62.72%	80.91%	96.91%	99.76%
3	19,879	6.71	35.95%	61.32%	79.90%	96.49%	99.74%
4	18,891	6.77	35.41%	59.88%	78.90%	96.79%	99.78%
5	18,992	6.37	38.20%	63.11%	81.07%	96.80%	99.80%
6	19,536	6.10	38.71%	63.83%	82.24%	97.16%	99.84%
7	15,854	6.39	37.37%	62.86%	80.91%	96.90%	99.76%
8	16,973	5.88	40.54%	65.17%	82.77%	97.42%	99.85%
9	13,799	5.49	40.97%	67.14%	84.59%	97.86%	99.83%
10	15,668	5.58	40.84%	66.39%	83.94%	97.58%	99.76%
11	14,285	6.28	38.42%	63.82%	81.39%	96.89%	99.76%
12	14,151	6.23	38.44%	63.64%	81.97%	96.83%	99.77%
13	12,081	5.64	40.19%	65.93%	83.90%	97.77%	99.86%
14	9,695	6.14	38.49%	64.35%	82.07%	97.08%	99.78%
15	8,862	5.77	41.13%	66.28%	84.09%	97.17%	99.72%
16	8,852	5.18	44.22%	69.66%	86.00%	97.80%	99.71%
17	7,508	6.19	38.11%	63.95%	82.78%	96.67%	99.60%
18	9,658	5.32	42.18%	68.05%	85.20%	97.95%	99.89%
19	4,953	6.61	35.11%	61.20%	81.00%	96.73%	99.84%
20	7,822	5.10	43.81%	69.75%	86.62%	97.92%	99.86%
21	4,462	4.97	41.04%	68.62%	86.69%	98.57%	99.87%
22	4,361	4.87	46.89%	73.19%	87.92%	97.78%	99.79%
X	8,331	10.75	25.20%	47.07%	66.34%	90.61%	98.91%
Total	303,018	6.20	38.58%	63.85%	81.85%	96.97%	99.76%

LIST OF REFERENCES

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

Abramowitz M, Stegun A (1965) Handbook of mathematical functions with formulas, graphs and mathematical tables. Dover publications, Inc New York

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLoS Biol 2:e286.

Alamanos Y, Drosos AA (2005) Epidemiology of adult rheumatoid arthritis. Autoimmun Rev 4:130-136.

Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 69:936-950.

Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P (2005) A haplotype map of the human genome. Nature 437:1299-1320.

Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, Healey LA, Kaplan SR, Liang MH, Luthra HS, et al. (1988) The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 31:315-324.

Asahara H, Asanuma M, Ogawa N, Nishibayashi S, Inoue H (1995) High DNA-binding activity of transcription factor NF-kappa B in synovial membranes of patients with rheumatoid arthritis. Biochem Mol Biol Int 37:827-832.

Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125:279-284.

Bondeson J, Foxwell B, Brennan F, Feldmann M (1999) Defining therapeutic targets by using adenovirus: blocking NF-kappaB inhibits both inflammatory and destructive mechanisms in rheumatoid synovium but spares anti-inflammatory mediators. Proc Natl Acad Sci U S A 96:5668-5673.

Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat

Genet 33:228-237.

Brintnell W, Zeggini E, Barton A, Thomson W, Eyre S, Hinks A, Silman AJ, Worthington J (2004) Evidence for a novel rheumatoid arthritis susceptibility locus on chromosome 6p. Arthritis Rheum 50:3823-3830.

Browning S, Thomas J (2007) Multilocus analysis of GAW15 NARAC chromosome 18 case-control data. BMC proceedings 1(Suppl 1):S11

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084-1097.

Cardon LR, Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. Trends Genet 19:135-140.

Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361:598-604.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22:231-238.

Carlson CS, Eberle MA, Kruglyak L, Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. Nature 429:446-452.

Castro F, Acevedo E, Ciusani E, Angulo JA, Wollheim FA, Sandberg-Wollheim M (2001) Tumour necrosis factor microsatellites and HLA-DRB1*, HLA-DQA1*, and HLA-DQB1* alleles in Peruvian patients with rheumatoid arthritis. Ann Rheum Dis 60:791-795.

Choi SJ, Rho YH, Ji JD, Song GG, Lee YH (2006) Genome scan meta-analysis of rheumatoid arthritis. Rheumatology (Oxford) 45:166-170.

Clark CJ, Davies E, Anderson NH, Farmer R, Friel EC, Fraser R, Connell JM (2000) alpha-adducin and angiotensin I-converting enzyme polymorphisms in essential hypertension. Hypertension 36:990-994.

Collins A, Lonjou C, Morton NE (1999) Genetic epidemiology of single-nucleotide polymorphisms. Proc Natl Acad Sci U S A 96:15173-15177.

Collins A, Morton NE (1998) Mapping a disease locus by allelic association. Proc Natl Acad Sci U S A 95:1741-1745.

Consortium IH (2003) The International HapMap Project. Nature 426:789-796.

Consortium M (1999) Complete sequence and gene map of a human major histocompatibility complex. The MHC sequencing consortium. Nature 401:921-923.

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

Craddock N, Dave S, Greening J (2001) Association studies of bipolar disorder. Bipolar Disord 3:284-298.

Crawford DC, Carlson CS, Rieder MJ, Carrington DP, Yi Q, Smith JD, Eberle MA, Kruglyak L, Nickerson DA (2004) Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. Am J Hum Genet 74:610-622.

Dalmasso C, Broet P, Moreau T (2005) A simple procedure for estimating the false discovery rate. Bioinformatics 21:660-668.

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229-232.

Dawson E, Abecasis GR, Bumpstead S, Chen Y, Hunt S, Beare DM, Pabial J, et al. (2002) A first-generation linkage disequilibrium map of human chromosome 22. Nature 418:544-548.

De La Vega FM, Isaac H, Collins A, Scafe CR, Halldorsson BV, Su X, Lippert RA, et al. (2005) The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. Genome Res 15:454-462.

Deighton CM, Walker DJ, Griffiths ID, Roberts DF (1989) The contribution of HLA to rheumatoid arthritis. Clin Genet 36:178-182.

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

Dizier MH, Eliaou JF, Babron MC, Combe B, Sany J, Clot J, Clerget-Darpoux F (1993) Investigation of the HLA component involved in rheumatoid arthritis (RA) by using the marker association-segregation chi-square (MASC) method: rejection of the unifying-shared-epitope hypothesis. Am J Hum Genet 53:715-721.

Douglas JA, Boehnke M, Gillanders E, Trent JM, Gruber SB (2001)

Experimentally-derived haplotypes substantially increase the efficiency of linkage disequilibrium studies. Nat Genet 28:361-364.

Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, Gejman PV (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum Mol Genet 12:205-216.

Ennis S, Maniatis N, Collins A (2001) Allelic association and disease mapping. Brief Bioinform 2:375-387.

Falconer DS, Mackay FC (1960) Introduction to Quantitative genetics.18-19

Fearnhead P, Donnelly P (2001) Estimating recombination rates from population genetic data. Genetics 159:1299-1318.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D (2002) The structure of haplotype blocks in the human genome. Science 296:2225-2229.

Giannelli F, Green PM (2000) The X chromosome and the rate of deleterious mutations in humans. Am J Hum Genet 67:515-517.

Gibson J, Tapper W, Zhang W, Morton N, Collins A (2005) Cosmopolitan linkage disequilibrium maps.20-27.

Gillespie JH (1998) Population Genetics A Concise Guide. The Johns Hopkins University Press Baltimore and London:36-37

Gomes I, Collins A, Lonjou C, Thomas NS, Wilkinson J, Watson M, Morton N (1999) Hardy-Weinberg quality control. Ann Hum Genet 63:535-538.

Gorman JD, Lum RF, Chen JJ, Suarez-Almazor ME, Thomson G, Criswell LA (2004) Impact of shared epitope genotype and ethnicity on erosive disease: a meta-analysis of 3,240 rheumatoid arthritis patients. Arthritis Rheum 50:400-412.

Gregersen PK, Silver J, Winchester RJ (1987) The shared epitope hypothesis. An approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. Arthritis Rheum 30:1205-1213.

Hajeer AH, Dababneh A, Makki RF, Thomson W, Poulton K, Gonzalez-Gay MA, Garcia-Porrua C, Mattey DL, Ollier WE (2000) Different gene loci within the HLA-DR and TNF regions are independently associated with susceptibility and severity in Spanish rheumatoid arthritis patients. Tissue Antigens 55:319-325.

Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Hered 86:409-411.

Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. Genetics 117:331-341.

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. TheorAppl Genet 38:226-231

Hirschhorn JN (2005) Genetic approaches to studying common diseases and complex traits. Pediatr Res 57:74R-77R.

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

Hosmer D, Lemeshow S (2000) Applied Logistic Regression. New York: John Wiley & Sons Inc:1-175

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

Ioannidis JP, Tarassi K, Papadopoulos IA, Voulgari PV, Boki KA, Papasteriades CA, Drosos AA (2002) Shared epitopes and rheumatoid arthritis: disease associations in Greece and meta-analysis of Mediterranean European populations. Semin Arthritis Rheum 31:361-370.

Jawaheer D, Seldin MF, Amos CI, Chen WV, Shigeta R, Etzel C, Damle A, et al. (2003) Screening the genome for rheumatoid arthritis susceptibility genes: a replication study and combined analysis of 512 multicase families. Arthritis Rheum 48:906-916.

Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217-222.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA (2001) Haplotype tagging for the identification of common disease genes. Nat Genet 29:233-237.

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

Kaback DB, Guacci V, Barber D, Mahon JW (1992) Chromosome size-dependent control of

meiotic recombination. Science 256:228-232.

Kaplan NL, Hill WG, Weir BS (1995) Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 56:18-32.

Kauppi L, Sajantila A, Jeffreys AJ (2003) Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. Hum Mol Genet 12:33-40.

Kauppi L, Stumpf MP, Jeffreys AJ (2005) Localized breakdown in linkage disequilibrium does not always predict sperm crossover hot spots in the human MHC class II region. Genomics 86:13-24.

Kilding R, Iles MM, Timms JM, Worthington J, Wilson AG (2004) Additional genetic susceptibility for rheumatoid arthritis telomeric of the DRB1 locus. Arthritis Rheum 50:763-769.

Kim Y, Nielsen R (2004) Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513-1524.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385-389.

Kochi Y, Yamada R, Kobayashi K, Takahashi A, Suzuki A, Sekine A, Mabuchi A, Akiyama F, Tsunoda T, Nakamura Y, Yamamoto K (2004) Analysis of single-nucleotide polymorphisms in Japanese rheumatoid arthritis patients shows additional susceptibility markers besides the classic shared epitope susceptibility sequences. Arthritis Rheum 50:63-71.

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139-144.

Kruglyak L, Nickerson DA (2001) Variation is the spice of life. Nat Genet 27:234-236.

Lau W, Kuo TY, Tapper W, Cox S, Collins A (2007) Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. Bioinformatics 23:517-519.

Lewontin RC (1964) THE INTERACTION OF SELECTION AND LINKAGE. II. OPTIMUM MODELS. Genetics 50:757-782.

Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination

hotspots using single-nucleotide polymorphism data. Genetics 165:2213-2233.

Li Z, Zheng T, Califano A, Floratos A (2007) Pattern-based mining strategy to detect multi-locus association and gene-environment interaction. BMC proceedings 1 (Suppl 1):S16

Lin MT, Storer B, Martin PJ, Tseng LH, Gooley T, Chen PJ, Hansen JA (2003) Relation of an interleukin-10 promoter polymorphism to graft-versus-host disease and survival after hematopoietic-cell transplantation. N Engl J Med 349:2201-2210.

Liu N, Sawyer SL, Mukherjee N, Pakstis AJ, Kidd JR, Kidd KK, Brookes AJ, Zhao H (2004) Haplotype block structures show significant variation among populations. Genet Epidemiol 27:385-400.

Lo SH, Zheng T (2002) Backward Haplotype Transmission Association (BHTA) algorithm - a fast multiple-marker screening method. Hum Hered 53:197-215.

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat Genet 33:177-182.

Long AD, Langley CH (1999) The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res 9:720-731.

Lonjou C, Zhang W, Collins A, Tapper WJ, Elahi E, Maniatis N, Morton NE (2003) Linkage disequilibrium in human populations. Proc Natl Acad Sci U S A 100:6069-6074.

MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K, Silman AJ (2000) Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. Arthritis Rheum 43:30-37.

Malecot G (1948) Les Mathematiques de l'Heredite. Maison et Cie, Paris

Maniatis N, Collins A, Gibson J, Zhang W, Tapper W, Morton NE (2004) Positional cloning by linkage disequilibrium. Am J Hum Genet 74:846-855.

Maniatis N, Collins A, Xu CF, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke X, Morton NE (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. Proc Natl Acad Sci U S A 99:2228-2233.

Maniatis N, Morton NE, Collins A (2005) The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. Hum Mol Genet 14:145-153.

Maniatis N, Morton NE, Collins A (2006) Effects of single SNPs, haplotypes, and whole Genome LD Maps on accuracy of association mapping. Genetic Epidemiology

Maraganore DM, de Andrade M, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PV, Frazer KA, Cox DR, Ballinger DG (2005) High-resolution whole-genome association study of Parkinson disease. Am J Hum Genet 77:685-693.

Martinez A, Fernandez-Arquero M, Pascual-Salcedo D, Conejero L, Alves H, Balsa A, de la Concha EG (2000) Primary association of tumor necrosis factor-region genetic markers with susceptibility to rheumatoid arthritis. Arthritis Rheum 43:1366-1370.

McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304:581-584.

Miretti MM, Walsh EC, Ke X, Delgado M, Griffiths M, Hunt S, Morrison J, Whittaker P, Lander ES, Cardon LR, Bentley DR, Rioux JD, Beck S, Deloukas P (2005) A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. Am J Hum Genet 76:634-646.

Molitor J, Marjoram P, Thomas D (2003) Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. Genet Epidemiol 25:95-105.

Moore JH, Williams SM (2002) New strategies for identifying gene-gene interactions in hypertension. Ann Med 34:88-95.

Morris AP, Whittaker JC, Balding DJ (2002) Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am J Hum Genet 70:686-707.

Morton N, Maniatis N, Zhang W, Ennis S, Collins A (2007) Genome scanning by composite likelihood. Am J Hum Genet 80:19-28.

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277-318.

Morton NE (2005) Linkage disequilibrium maps and association mapping. J Clin Invest 115:1425-1430.

Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok PY, Collins A (2001) The optimal measure of allelic association. Proc Natl Acad Sci U S A 98:5217-5221.

Newton J, Brown MA, Milicic A, Ackerman H, Darke C, Wilson JN, Wordsworth BP, Kwiatkowski D (2003) The effect of HLA-DR on susceptibility to rheumatoid arthritis is influenced by the associated lymphotoxin alpha-tumor necrosis factor haplotype. Arthritis Rheum 48:90-96.

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C (2005) Genomic scans for selective sweeps using SNP data. Genome Res 15:1566-1575.

Niu (2002) Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms. Am J Hum Genet 78:174.

Okamoto K, Makino S, Yoshikawa Y, Takaki A, Nagatsuka Y, Ota M, Tamiya G, Kimura A, Bahram S, Inoko H (2003) Identification of I kappa BL as the second major histocompatibility complex-linked susceptibility locus for rheumatoid arthritis. Am J Hum Genet 72:303-312.

Ota M, Katsuyama Y, Kimura A, Tsuchiya K, Kondo M, Naruse T, Mizuki N, Itoh K, Sasazuki T, Inoko H (2001) A second susceptibility gene for developing rheumatoid arthritis in the human MHC is localized within a 70-kb interval telomeric of the TNF genes in the HLA class III region. Genomics 71:263-270.

Ota T, Kimura M (1971) Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics 68:571-580.

Page GP, George V, Go RC, Page PZ, Allison DB (2003) "Are we there yet?": Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. Am J Hum Genet 73:711-719.

Patil N, Berno AJ, Hinds DA, Barrett WA, Doshi JM, Hacker CR, Kautzer CR, Lee DH, Marjoribanks C, McDonough DP, Nguyen BT, Norris MC, Sheehan JB, Shen N, Stern D, Stokowski RP, Thomas DJ, Trulson MO, Vyas KR, Frazer KA, Fodor SP, Cox DR (2001) Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719-1723.

Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly MJ, Reich D (2004) Methods for high-density admixture mapping of disease genes. Am J Hum Genet 74:979-1000.

Pounds S, Cheng C (2004) Improving false discovery rate estimation. Bioinformatics 20:1737-1745.

Pounds S, Morris SW (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical

distribution of p-values. Bioinformatics 19:1236-1242.

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124-137.

Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not? Hum Mol Genet 11:2417-2423.

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. Am J Hum Genet 69:1-14.

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001) Linkage disequilibrium in the human genome. Nature 411:199-204.

Reich DE, Gabriel SB, Altshuler D (2003) Quality and completeness of SNP databases. Nat Genet 33:457-458.

Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17:502-510.

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516-1517.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138-147.

Rocchi A, Pellegrini S, Siciliano G, Murri L (2003) Causative and susceptibility genes for Alzheimer's disease: a review. Brain Res Bull 61:1-24.

Rodevand E, Bathen J, Ostensen M (1999) [Rheumatoid arthritis and heart disease]. Tidsskr Nor Laegeforen 119:223-225.

Rowe A, Kalaitzopoulos D, Osmond M, Ghanem M, Guo Y (2003) The discovery net system for high throughput bioinformatics. Bioinformatics 19:i225-231.

Salem RM, Wessel J, Schork NJ (2005) A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. Hum Genomics 2:39-66.

Schaid DJ (2002) Relative efficiency of ambiguous vs. directly measured haplotype frequencies. Genet Epidemiol 23:426-443.

Service SK, Ophoff RA, Freimer NB (2001) The genome-wide distribution of background

linkage disequilibrium in a population isolate. Hum Mol Genet 10:545-551.

Singal DP, Li J, Lei K (1999) Genetics of rheumatoid arthritis (RA): two separate regions in the major histocompatibility complex contribute to susceptibility to RA. Immunol Lett 69:301-306.

Slatkin M, Veuille M (2002) Modern developments in theoretical population genetics. The legacy of Gustave Malecot. Oxford University Press Oxford, United Kingdom/New York, New York, USA:280pp.

Smith DJ, Lusis AJ (2002) The allelic structure of common disease. Hum Mol Genet 11:2455-2461.

Smith MW, Patterson N, Lautenberger JA, Truelove AL, McDonald GJ, Waliszewska A, Kessing BD, et al. (2004) A high-density admixture map for disease gene discovery in african americans. Am J Hum Genet 74:1001-1013.

Stajich JE, Hahn MW (2005) Disentangling the effects of demography and selection in human history. Mol Biol Evol 22:63-73.

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162-1169.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978-989.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100:9440-9445.

Sulakhe D, Rodriguez A, D'Souza M, Wilde M, Nefedova V, Foster I, Maltsev N (2005) GNARE: automated system for high-throughput genome analysis with grid computational backend. J Clin Monit Comput 19:361-369.

Syvanen AC (2005) Toward genome-wide SNP genotyping. Nat Genet 37:S5-10.

Taillon-Miller P, Saccone SF, Saccone NL, Duan S, Kloss EF, Lovins EG, Donaldson R, Phong A, Ha C, Flagstad L, Miller S, Drendel A, Lind D, Miller RD, Rice JP, Kwok PY (2004) Linkage disequilibrium maps constructed with common SNPs are useful for first-pass disease association screens. Genomics 84:899-912.

Tapper W, Collins A, Gibson J, Maniatis N, Ennis S, Morton NE (2005) A map of the human genome in linkage disequilibrium units. Proc Natl Acad Sci U S A 102:11835-11839.

Tapper WJ, Maniatis N, Morton NE, Collins A (2003) A metric linkage disequilibrium map

of a human chromosome. Ann Hum Genet 67:487-494.

Thomas A, Camp NJ (2004) Graphical modeling of the joint distribution of alleles at associated loci. Am J Hum Genet 74:1088-1101.

Thomas S, Porteous D, Visscher PM (2004) Power of direct vs. indirect haplotyping in association studies. Genet Epidemiol 26:116-124.

Thompson EA, Neel JV (1997) Allelic disequilibrium and allele frequency distribution as a function of social and demographic history. Am J Hum Genet 60:197-204.

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK (2000) The accuracy of statistical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67:518-522.

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

Varilo T, Peltonen L (2004) Isolates and their potential use in complex gene mapping efforts. Curr Opin Genet Dev 14:316-323.

Wright S (1969) *Evolution and the Genetic of Populations. University of Chicago Press*, Chicago 2

Yamamoto K, Yamada R (2005) Genome-wide single nucleotide polymorphism analyses of rheumatoid arthritis. J Autoimmun 25:12-15.

Yan H, Papadopoulos N, Marra G, Perrera C, Jiricny J, Boland CR, Lynch HT, Chadwick RB, de la Chapelle A, Berg K, Eshleman JR, Yuan W, Markowitz S, Laken SJ, Lengauer C, Kinzler KW, Vogelstein B (2000) Conversion of diploidy to haploidy. Nature 403:723-724.

Yang X (2004) Qvalue Method May Not Always Control False Discovery Rate in Genomics Applications. IEEE Computer Society Technical Journals:556-557

Yu Z, Schaid D (2007a) Application of sequential haplotype scan methods to case-control data. BMC proceedings 1(Suppl 1):S21

Yu Z, Schaid DJ (2007b) Sequential haplotype scan methods for association analysis. Genet Epidemiol 31:553-564.

Zhang K, Deng M, Chen T, Waterman MS, Sun F (2002b) A dynamic programming

algorithm for haplotype block partitioning. Proc Natl Acad Sci U S A 99:7335-7339.

Zhang S, Pakstis AJ, Kidd KK, Zhao H (2001) Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. Am J Hum Genet 69:906-914.

Zhang W, Collins A, Gibson J, Tapper WJ, Hunt S, Deloukas P, Bentley DR, Morton NE (2004a) Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. Proc Natl Acad Sci U S A 101:18075-18080.

Zhang W, Collins A, Maniatis N, Tapper W, Morton NE (2002) Properties of linkage disequilibrium (LD) maps. Proc Natl Acad Sci U S A 99:17004-17007.

Zondervan KT, Cardon LR (2004) The complex interplay among factors that influence allelic association. Nat Rev Genet 5:89-100.

Zondervan KT, Cardon LR, Kennedy SH (2002) What makes a good case-control study? Design issues for complex traits such as endometriosis. Hum Reprod 17:1415-1423.

LIST OF PUBLICATIONS

1. **Kuo, T-Y.**, Lau, W., Tapper, W., Cox, S., Collins, A. (2007) Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. *Bioinformatics*: 23; 517-519.

2. **Kuo, T-Y.**, Lau W., Collins, A. (2007) *LDMAP*: The construction of high-resolution linkage disequilibrium maps of the human genome. In 'Linkage disequilibrium and association mapping: analysis and applications', Ed. Andrew Collins, Humana Press, New Jersey.

3. Zhang, W, Lau, WWS, Hu, C, **Kuo, T-Y.** (2007) Impact of marker density on the accuracy of association mapping. *BMC Proceedings* 1(Suppl 1):S166.

4. **Kuo, T-Y**, Lau W, Hu, C, Zhang, W. (2007) Association mapping of susceptibility loci for rheumatoid arthritis. *BMC Proceedings* 1(Suppl 1):S15.

BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 4 2007, pages 517–519
doi:10.1093/bioinformatics/btl615

Genetics and population analysis

Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome

Winston Lau†, Tai-Yue Kuo†, William Tapper, Simon Cox[1] and Andrew Collins*

Human Genetics Division, Duthie Building (Mailpoint 808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK and [1]Southampton Regional e-Science Centre, School of Engineering Sciences, University of Southampton, Southampton SO17 1BJ, UK

Received on September 21, 2006; revised on November 14, 2006; accepted on November 28, 2006
Advance Access publication December 1, 2006
Associate Editor: Keith A Crandall

ABSTRACT

Summary: Linkage disequilibrium (LD) maps increase power and precision in association mapping, define optimal marker spacing and identify recombination hot-spots and regions influenced by natural selection. Phase II of HapMap provides ~2.8-fold more single nucleotide polymorphisms (SNPs) than phase I for constructing higher resolution maps. *LDMAP-cluster*, is a parallel program for rapid map construction in a Linux environment used here to construct genome-wide LD maps with >8.2 million SNPs from the phase II data.

Availability: The LD maps, *LDMAP-cluster* and documentation are available from: http://www.som.soton.ac.uk/research/geneticsdiv/epidemiology/LDMAP

Contact: arc@soton.ac.uk

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Linkage disequilibrium (LD) describes the tendency of alleles at markers in close proximity to be inherited together more frequently than expected under random segregation. Precise characterization of LD structure underpins efficient mapping of disease genes by association. Maniatis *et al.* (2002) developed an analogue to linkage maps in centimorgans with maps expressed in LD units (LDUs), which have ~1500-fold higher resolution (Tapper *et al.*, 2005), and lengths reflecting the number of generations since an 'effective' bottleneck (Zhang *et al.*, 2004). Improved localization and substantial increases in power are found when disease mapping with LDU maps (Maniatis *et al.*, 2005).

The *LDMAP* program constructs LD maps from single nucleotide polymorphism (SNP) data in population samples using the 'interval' algorithm (Maniatis *et al.*, 2002). The program constructs LD maps from either phase unknown (genotypic) data or phase-known (haplotypic) data. Further details of the core methodology are given in Supplementary material. Map construction is computationally intensive employing composite likelihood to estimate a parameter, epsilon (ε), describing the decline of association in each interval between adjacent SNPs.

Phase II of HapMap (International HapMap Consortium, 2005), provides ~2.8-fold more SNPs than phase I. The huge volume of data imposes a considerable computational burden addressed here through the implementation of a parallel algorithm, in the program *LDMAP-cluster*, deployed on a Linux Beowulf cluster. We have used this program to construct genome-wide LDU maps from phase II data for the four HapMap populations. A detailed description of the data are given in the Supplementary materials.

2 IMPLEMENTATION

LDMAP-cluster is written in C, as a wrapper program that encapsulates *LDMAP*. We deployed the program on a Linux Beowulf cluster of over 900 processors. The batch queuing and job management is administrated by Open-PBS (Portable Batch System), http://www.openpbs.org/.

The segment-based parallel approach is illustrated in Figure 1. We established that assembly of maps in segments of ~2000 SNPs loses minimal information and provides substantial reductions in computing time (Supplementary Figure 1). We also examined the effect on map quality of varying the number of pairwise observations used to estimate epsilon in each map interval. An optimum 'interval window' of informative SNP pairs separated by no more than ~100 intervals was identified (Supplementary Figure 2). Map segments are submitted and constructed as individual jobs on the cluster. The parallel processing is accomplished by the concurrent submission of all segments.

LDMAP-cluster is a 64 bit program, enabling access to more memory than conventional 32 bit platforms. The program features synchronous processing supporting multiple SNP dataset submissions. To efficiently utilize dual-processor machines in the cluster, segments are assigned as two jobs per submission. In addition to job monitoring commands (i.e. '*showq*' and '*qstat*') supplied by Open-PBS, a custom-made program, '*checkSeg*', tracks the status of the submitted jobs grouped by SNP dataset.

A segment of 2000 SNPs requires 5–10 h of computation (AMD Opteron 2 GHz with 2 GB RAM), corresponding to the minimum time for construction of the whole map given complete parallelization.

LDMAP-cluster is compatible with a Linux Beowulf cluster with Open-PBS installed as the batch scheduler. Recompilation of the program is essential for linking to the platform specific libraries. Minor modification of the code responsible for job submission is required for porting onto a Linux cluster with a different batch scheduler. Compatibility across all platforms is difficult to

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

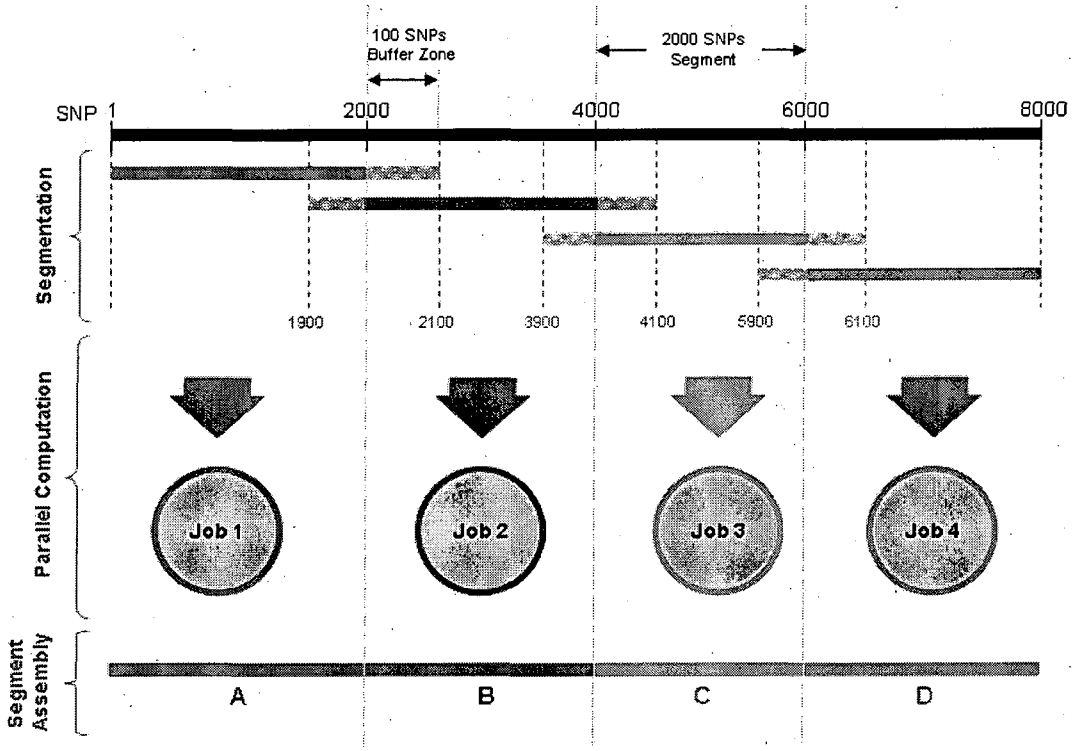


Fig. 1. A chromosome is divided into segments of ~2000 SNPs. A 'buffer zone' of 100 SNPs extends from the ends of each segment to minimize loss of information. Buffer zones are eliminated in map assembly and segments are connected end to end to form the complete map.

guarantee given different hardware (e.g. 32 or 64 bit), software (e.g. PBS or Condor) and administrative environments (e.g. versions of glibc and Tcl/Tk libraries), but modification for local systems should be straightforward as the software is written in standard C. Further technical issues are discussed in detail in the Supplementary materials and supporting website.

3 RESULTS

Tapper *et al.* (2005) describe a genome-wide LD map constructed from ~490 k SNPs (post-screening) from HapMap phase I public release #16 for the CEU population. We describe here maps from all four HapMap populations with 1.9–2.3 million SNPs per population. These data were analyzed in 4195 segments of ~2000 SNPs. Approximately 8.2 million SNPs were processed in ~25 170 computing hours achieved over about one month real-time. The phase II LD maps resolve ~31% of the 'holes' (intervals constrained to the upper limit of three LDUs, Service *et al.*, 2006) in the phase I maps where the LD structure is not fully characterized. Such regions are more frequent in large outbred populations, such as those represented in HapMap, where recombination events have accumulated in narrow regions over many generations creating locally high-haplotype diversity. Considering the hugely increased marker density the relatively small proportion of resolved holes suggests that many holes correspond to particularly intense recombination hot-spots. Disease gene mapping by association is expected to be particularly difficult in these areas (Service *et al.*, 2006).

Although the broad pattern of LD is consistent between the two HapMap phases (Fig. 2), the fine scale structure of steps and blocks differs in many regions. Increasing SNP density recovers structural details from regions with lower marker coverage in phase I but differences also reflect changes in the sequence build and the resolution of some holes, (which may locally increase or decrease map length).

Overall the phase II maps are 3.1% longer (Table 1), a modest increase consistent with the essentially additive property of the LDU map distances noted previously (Ke *et al.*, 2004).

4 DISCUSSION

Genome-wide LDU maps constructed using *LDMAP-cluster* have substantially higher marker density than maps published for the CEU population (Tapper *et al.*, 2005). The maps should guide marker selection, empower genome-wide association studies and facilitate other genomic studies. The LD pattern at fine scale is described by these maps, and applications to disease association mapping are expected to increase power and precision for localization of disease genes, consistent with existing evidence (Maniatis *et al.*, 2005). The LD pattern is highly consistent between the high-resolution (HapMap release #20) and low-resolution (release #16) maps, despite small differences in overall map length attributable to changes in the sequence and the better characterized LD structure.

Efforts are now underway to generate large case-control and other phenotype samples for association studies with many thousands of

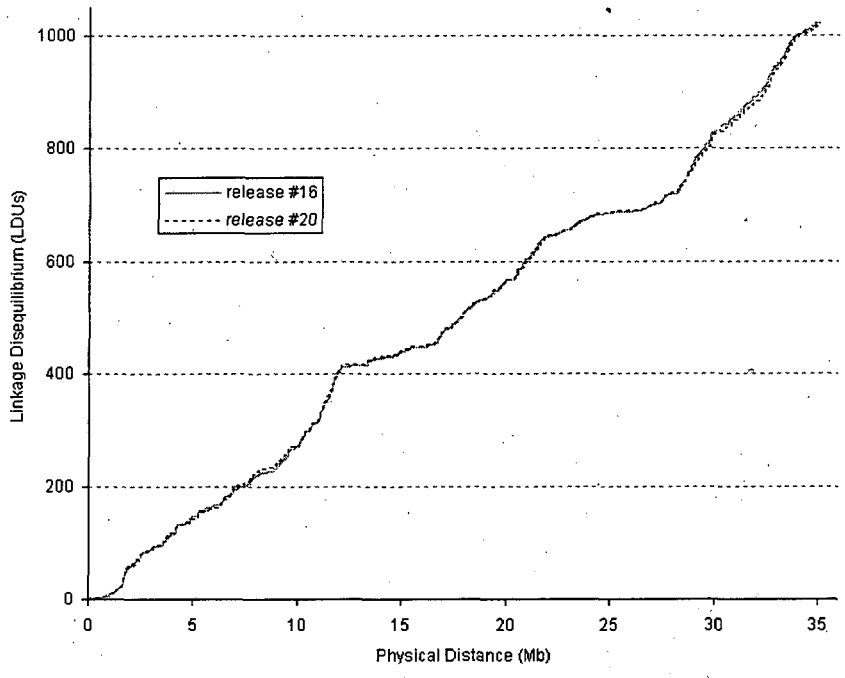


Fig. 2. LD maps of chromosome 22 (CEU) constructed from HapMap #16 (13 959 SNPs) and #20 (26 721 SNPs). The LD pattern is highly consistent between the two HapMap phases.

Table 1. Characteristics of the LDU maps

Populations	CEU	CHB	JPT	YRI	Σ
No. of holes in LD map					
Phase I release #16	2911	4879	3731	2979	14 500
Phase II release #20	2033	3838	2900	1216	9987
Diff.	−30%	−21%	−22%	−59%	(avg.) −31%
Overall LD map length (in LDUs)					
Phase I release #16	56 250	62 686	56 655	79 499	255 091
Phase II release #20	57 819	64 930	58 730	81 345	262 826
Diff.	+2.8%	+3.6%	+3.7%	+2.3%	(avg.) +3.1%

SNPs. The complexities of processing and analyzing such huge bodies of data are an area of rapid research. We anticipate that the genome-wide LDU maps and software tools developed will facilitate association mapping in these samples and contribute to studies of recombination, selection and population history. Applications to data from other organisms, including a recent application to the Bovine genome (Khatkar *et al.*, 2006), demonstrate the wide-applicability and utility of this form of genetic map for describing and analyzing LD structure with high-resolution.

ACKNOWLEDGEMENTS

This research is supported by a University of Southampton e-Science centre Postgraduate Research grant.

Conflict of Interest: none declared.

REFERENCES

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

Ke,X. *et al.* (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.*, **13**, 577–588.

Khatkar,M.S. *et al.* (2006) A first generation metric linkage disequilibrium map of bovine chromosome 6. *Genetics*, **174**, 79–85.

Maniatis,N. *et al.* (2002) The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl Acad. Sci. USA*, **99**, 2228–2233.

Maniatis,N. *et al.* (2005) The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. *Hum. Mol. Genet.*, **14**, 145–153.

Service,S. *et al.* (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat. Genet.*, **38**, 556–560.

Tapper,W. *et al.* (2005) A map of the human genome in linkage disequilibrium units. *Proc. Natl Acad. Sci. USA*, **102**, 11835–11839.

Zhang,W. *et al.* (2004) Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc. Natl Acad. Sci. USA*, **101**, 18075–18080.

BMC Proceedings

BioMed Central

Proceedings

Open Access

Association mapping of susceptibility loci for rheumatoid arthritis

Tai-Yue Kuo[1,2], Winston Lau[1], Cheng Hu[3] and Weihua Zhang*[4,5]

Address: [1]Human Genetics Division, Duthie Building (Mailpoint 808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK, [2]National Cheng Kung University Hospital, No. 138, Shengli Road, Tainan City, Taiwan, [3]Shanghai Diabetes Institute, Shanghai Jiaotong University, 600 Yishan Road, Shanghai 200233, People's Republic of China, [4]Section of Cancer Genetics, The Institute of Cancer Research, 15 Cotswold Road, Belmont, Sutton Surrey SM2 5NG, UK and [5]Department of Cardiology, Ealing Hospital NHS Trust, Uxbridge Road, Southall, Middlesex, UB1 3HW, UK

Email: Tai-Yue Kuo - kuotaiyu@soton.ac.uk; Winston Lau - wwsl@soton.ac.uk; Cheng Hu - alfredhc@sjtu.edu.cn; Weihua Zhang* - weihua.zhang@eht.nhs.uk

* Corresponding author

from Genetic Analysis Workshop 15
St. Pete Beach, Florida, USA. 11–15 November 2006

Published: 18 December 2007

BMC Proceedings 2007, 1(Suppl 1):S15

This article is available from: http://www.biomedcentral.com/1753-6561/1/S1/S15

Abstract

We analyzed a case-control data set for chromosome 18q from the Genetic Analysis Workshop 15 to detect susceptibility loci for rheumatoid arthritis (RA). A total number of 460 cases and 460 unaffected controls were genotyped on 2300 single-nucleotide polymorphisms (SNPs) by the North American Rheumatoid Arthritis Consortium. Using a multimarker approach for association mapping under the framework of the Malecot model and composite likelihood, we identified a region showing significant association with RA ($p < 0.002$) and the predicted disease locus was at a genomic location of 53,306 kb with a 95% confidence interval (CI) of 53,295–53,331 kb. A common haplotype in this region was protective against RA ($p = 0.002$). In another region showing nominal significant association (51,585 kb, 95% CI: 51,541–51,628 kb, $p = 0.037$), a haplotype was also protective ($p = 0.002$). We further demonstrated that reducing SNP density decreased power and accuracy of association mapping. SNP selection based on equal linkage disequilibrium (LD) distance generally produced higher accuracy than that based on equal kilobase distance or tagging.

Background

Rheumatoid arthritis (RA) is a common chronic disease, with a moderately strong genetic component. Chromosome 18q has shown evidence for linkage in the U.S. and French linkage scans [1]. The North American Rheumatoid Arthritis Consortium (NARAC) performed fine mapping on a 10-Mb region on 18q with a dense single-nucleotide polymorphism (SNP) map and the data were collected by the Genetic Association Workshop (GAW) 15 for Problem 2. Here we applied a novel association mapping approach based on the Malecot model and composite likelihood to identify disease associated regions and predict the locations of possible disease loci [2]. Haplotype analysis on the candidate regions was performed. We also studied the effect of region length and SNP density on the accuracy of association mapping in comparison with our analysis of the simulated data in Problem 3 of GAW15 [3].

Methods

Data

A total of 2300 SNPs in a 9,519.224 kb region of 18q were genotyped by NARAC in 460 cases of RA and 460 controls. Controls were recruited from a New York City population. Seven SNPs showing significant departure from Hardy-Weinberg equilibrium (HWE) in the control samples using a likelihood ratio chi-square test ($\chi^2 \geq 10$) were removed, resulting in a total of 2293 SNPs [4]. Further removal of 81 SNPs with a minor allele frequency (MAF) of less than 5% resulted in a total of 2212 SNPs for our main data analysis.

LD map

Physical locations of these SNPs were determined from build 35 (UCSC May 2004) of the human genome sequence. An LD map expressed in linkage disequilibrium (LD) units (LDUs) was created using the control samples with the LDMAP-cluster program, a parallel version of LDMAP program that rapidly constructed the map http:www.som.soton.ac.uk/research/geneticsdiepidemiology/ldmap/[5]. LDU is determined by the product of the ε and distances in kilobases for an interval of two adjacent SNPs and is additive, where ε represents the exponential decline of LD with distance for that interval. The LD map length was 151.115 LDUs, which is essentially the same as the 2293 SNPs containing rare SNPs.

We also used the LD map built from the CEU samples of the HapMap Phase II data http://www.som.soton.ac.uk/research/geneticsdiv/epidemiology/LDMAP/map2.htm[5]. The same region on the CEU LD map contains 8086 SNPs with a length of 202 LDUs. Despite its higher SNP density, 185 SNPs were missing and therefore their LDU locations were linearly interpolated. However, alternative LD maps did not seem to exert a significant effect on the results (data not shown), and we hereby only report the results using the LD map constructed from the GAW15 data.

Association mapping

The 10-Mb segment of 18q was divided into 14 non-overlapping consecutive regions. Each region had a minimum of 10 LDUs and 30 SNPs by default without breaking LD blocks and was analyzed individually. We also used a 5-LDU region length, resulting in a total of 26 regions for association analysis. In the Malecot model, association is a function of several parameters, the most important of which is S, the predicted location of the disease variant [2]. Composite likelihood combines information of all pairwise marker-disease associations in each region. S and its 95% confidence interval (CI) are estimated by fitting the model to the data and maximizing the composite likelihood. Significance tests are carried out by contrasting two hierarchical models. Model A assumes no association and no parameters are estimated. Model D assumes an association and S and two other parameters are estimated with ε specified. The difference in the -2 natural log composite likelihood (denoted as Λ) between the two models (denoted as $\Lambda_A - \Lambda_D$) is a statistic monotonic to a chi-square with 3 degrees of freedom (χ^2_3). A permutation test was performed for each region with hundreds of replicates under the null hypothesis of no association by shuffling case-control status to obtain an empirical p-value [2]. The specified value for ε of 1.0543 was obtained by fitting the LD map to the genotype data for the control samples. However, similar values for ε did not seem to have an appreciable effect on the results (data not shown).

This approach has been implemented in the CHROMSCAN program and a parallel version, CHROMSCAN-cluster, based on cluster computing was used for permutations with 1000 replicates http://www.som.soton.ac.uk/research/geneticsdiv/epidemiology/chromscan/. Pearson's χ^2s were obtained for allelic associations between single SNPs and RA.

Haplotype analysis for candidate regions

Haplotype analysis using the PHASE program version 2 [6] was performed for candidate regions showing nominal significant association. The five most common haplotypes and their frequencies were compared between cases and controls. A chi-square test was applied to identify significant associations by testing each haplotype in turn against all others, including rare ones.

SNP density and region length

To generate different SNP density, we used Tagger in Haploview software to select tagging SNPs based on pairwise LD (r^2) using the control samples with rare SNPs included [7]. For comparison, we selected the same number of SNPs as Tagger but by equidistance in LDU or kilobases. To do this, SNPs with the same LDU were reassigned LDU locations by linear interpolation (tilting) so that every SNP would have a unique location. By centering at the "disease locus", we also studied the effect of LDU region length on the results.

Results and discussion

Association mapping of disease locus

We found a nominally significant association between region 5 and RA. The estimated location of the disease locus S_1 was at 53,306 kb near a SNP of global maximal chi-square (rs3745064, $\chi^2 = 12.25$, $p = 0.00047$) using the LD map with a 10-LDU region length ($p = 0.002$). After Bonferroni correction for 14 regions the results were still statistically significant ($p_c = 0.02$). Removing this SNP did not change the results. However, inclusion of SNPs with MAF < 5% resulted in a wider 95% CI (53,274–53,342 kb, point location at 53,307 kb). Figure 1 shows that the estimated location for the disease variant was in a 10-kb LD

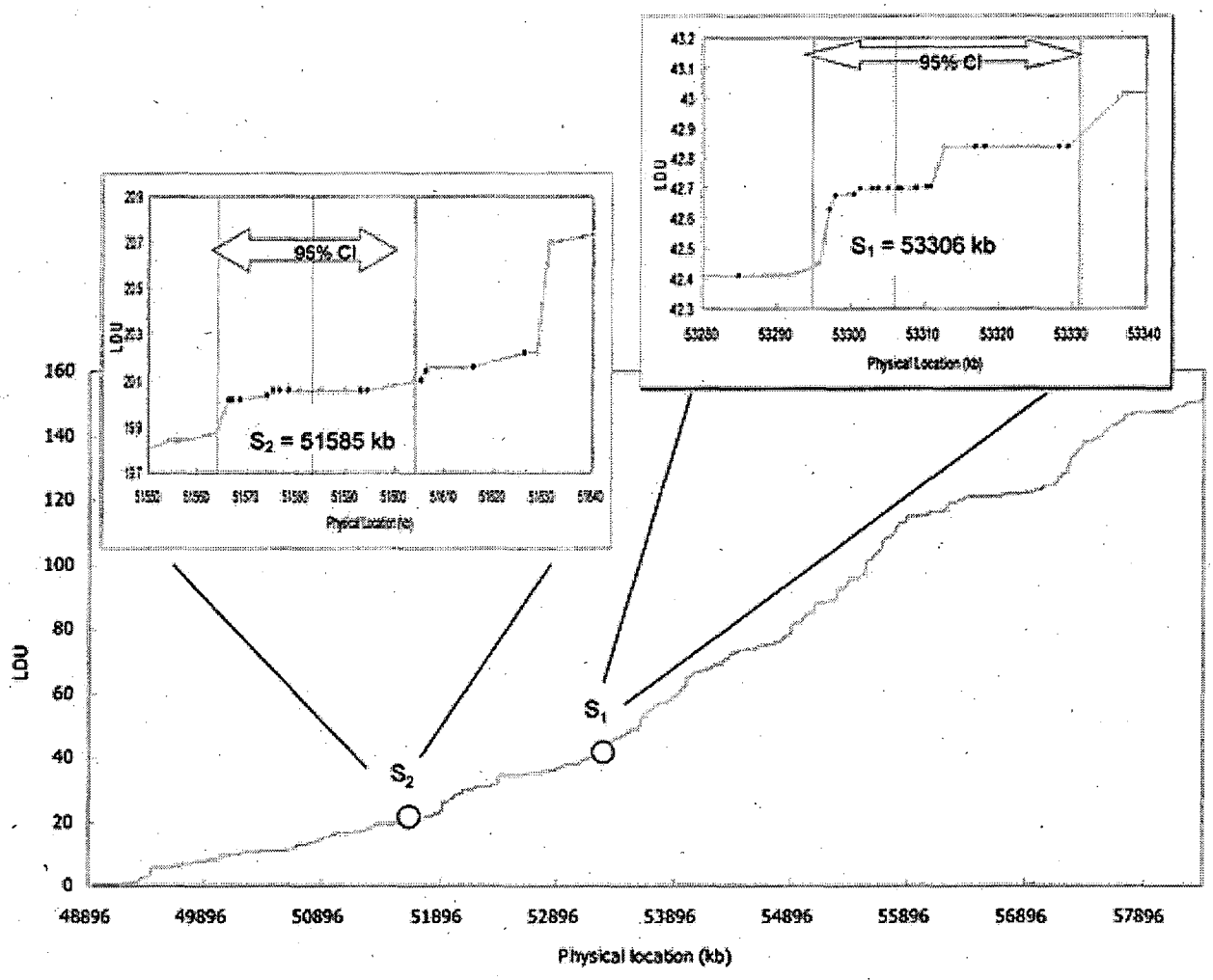


Figure 1

An LD map in relation to the putative disease loci S_1 and S_2. The details of the SNPs and LD patterns for the regions around S_1 and S_2 are enlarged in the upper diagrams. The vertical black solid line indicates the location of the point estimate within the 95% CI. The black dots on the map represent SNPs showing nominal significant association with RA ($p < 0.05$) and the gray dots represent SNPs showing no association.

Table 1: Association mapping of RA susceptibility loci

Loci	LDU length	No. of regions	Region (no. of SNP)	Length (kb)	Location S (kb)	95% CI (kb)	χ^2_3	p
S_1	10.2	14	5th(112)	371	53,306	53,295–53,331	15	0.0017
S_2	5.1	26	4th(132)	703	51,585	51,541–51,628	8	0.0370

block where a cluster of SNPs showed modest association with RA. At such a significance level, none of the SNPs were statistically significant after correction for multiple comparisons. This is in contrast to multimarker approaches, in which one cluster of SNPs is considered at a time in light of LD among nearby SNPs, markedly reducing the number of tests.

At 5-LDU region length with a total of 26 regions, we found a locus (S_2) at 51,585 kb showing nominal significant association ($p = 0.04$, Table 1 and Figure 1). Similar results were obtained using the data with rare SNPs (MAF < 5%) included. In this region there was also a cluster of SNPs associated with RA. However, consideration of multiple testing this region would not be statistically significant.

Haplotype analysis for candidate regions

Haplotype analysis was performed on sub-regions containing S_1 and S_2 with the majority of nominally associated SNPs included. S_1 sub-region (53,297–53,312 kb, 0.043 LDUs) contained 16 SNPs (Table 2) and S_2 sub-region (51,556–51,616 kb, 0.368 LDUs) contained 21 SNPs (Table 3). Haplotypes H_1 and H_3 at S_1 and H_5 at S_2 appeared to be significantly associated with RA, with the first two haplotypes being almost complementary. Both H_1 at S_1 and H_5 at S_2 showed protective effects against RA.

Further analyses of S_1 categorized all individuals into three groups, H_1/H_1, H_1/H- and H-/H-, where H- is a haplotype other than H_1. There was a significant association between haplotype pairs and disease status ($\chi^2_2 = 10.3$, $p = 0.006$). An individual carrying H_1/H_1 had a lower risk of RA than those carrying H_1/H- or H-/H-. The odds ratios (ORs) were 0.58 (95% CI: 0.37–0.92) and 0.85 (0.54–1.34) for an individual carrying H_1/H_1 or H_1/H- compared to H-/H- haplotypes, respectively. We performed analyses conditional on whether an individual carried H_1/H_1. Interestingly, H_5 appeared to be significant only in H_1/H- or H-/H- carriers ($\chi^2 = 7.647$, $p < 0.05$), but not in H_1/H_1 carriers ($\chi^2 = 1.509$), indicating a possible interaction between the two haplotypes.

Genes and mRNA at the candidate regions

The UCSC genome browser (May 2004) was used to find genes and mRNAs within the 95% CI of loci S_1 and S_2. No known genes have been found nearby S_1, but the area of the 95% CI for locus S_1 contains four human mRNA (CR590917, AK021717, AK124558, and BC013134), two of which span the point estimate of S_1. Therefore, this region might contain genes not yet identified. In addition, this area is highly conserved across species, implying functional importance of the genomic sequence. A known

Table 2: Common haplotype analysis of the S_1 candidate region

Code	S_1 Haplotype[a]	Frequency			χ^2	p
	+ + + + + + - + + + + + + + + -	Total	Case	Control		
H_1	2 1 1 2 1 2 1 1 1 2 1 1 2 1 1 1	0.661	0.628	0.695	9.22	0.002
H_2	1 2 2 1 2 1 1 2 2 1 2 2 1 1 2 1	0.169	0.180	0.158	1.59	0.2
H_3	1 2 2 1 2 1 1 2 2 1 2 2 1 2 2 1	0.101	0.120	0.082	7.32	0.007
H_4	1 1 1 2 1 2 1 1 1 2 1 1 2 1 1 1	0.019	0.013	0.025	3.55	0.06
H_5	1 1 2 1 2 2 2 2 2 1 2 2 1 2 2 2	0.017	0.017	0.017	0.00	1

[a]SNPs from left to right: rs660936, rs674849, rs615030, rs629737, rs519596, rs660626, rs3745070, rs3745064, rs3848516, rs608017, rs608823, rs552396, rs2279096, rs1217583, rs3899444, rs4940796, '+', '-' denote a SNP with nominal association (+) or no association (-) with RA. '1', '2' denote the alleles of a SNP.

Table 3: Common haplotype analysis of the S_2 candidate region

Code	S_2Haplotype[a]	Frequency			χ^2	*p*
	----++++-+++---++++-+	Total	Case	Control		
H_1	1 2 1 2 2 1 1 1 1 2 2 1 2 2 1 2 2 1 2 2 2	0.339	0.343	0.336	0.10	0.8
H_2	2 1 2 1 2 1 1 1 1 2 2 1 2 2 2 1 2 2 1 1 2	0.256	0.274	0.238	3.13	0.08
H_3	2 1 2 1 2 1 1 1 1 2 2 1 2 2 1 2 2 1 2 1 1	0.100	0.089	0.110	2.26	0.1
H_4	2 1 2 1 1 2 2 2 1 1 1 2 2 2 1 2 1 1 2 2 1	0.079	0.074	0.084	0.63	0.4
H_5	1 2 1 2 1 2 2 2 1 1 1 2 2 2 1 2 1 1 2 2 1	0.075	0.056	0.094	9.57	0.002

[a]SNPs from left to right: rs813043, rs784254, rs711745, rs784251, rs4800995, rs784237, rs796743, rs784235, rs784233, rs4800996, rs3745044, rs784232, rs1642295, rs784240, rs1362781, rs2306163, rs931040, rs4996482, rs899101, rs899102, rs1031830, '+', '-' denote a SNP with nominal association (+) or no association (-) with RA. '1', '2' denote the alleles of a SNP.

gene (*AK127787*) is within the 95% CI of S_2, but is 10 kb away from its point estimate.

Region length and SNP density

Point estimates of *S* were identical for all region lengths centred at S_1. The 95% CI was also relatively stable, with a slow increase with region length (Table 4). Enlarged region length compromised the significance levels, perhaps due to noise from distant SNPs, given that the informative SNPs were clustered in a rather small region. Computing time prolonged with increasing number of SNPs. Small region lengths, however, resulted in a heavy penalty for multiple testing. Four LDUs provided the most significant result for S_1($P_c = 30 \times 0.0008 = 0.02$, Table 4).

Our analysis of simulated data indicated that reduced SNP density decreased mapping accuracy, and SNP selection based on equal LD distance produced smaller location errors than that based on equal kilobase distance or tagging [3]. Interestingly, among the three selection approaches for S_1 region, Tagger selected the most number of SNPs while equal kilobase distance, the least number. SNPs selected by equal LDU distance generally provided the highest location accuracy (Table 5). Power was reduced with decreasing density, as indicated by the values of $\Lambda_A - \Lambda_D$ (Table 5). Using the kilobase map resulted in higher location errors in most cases and lower $\Lambda_A - \Lambda_D$ values, indicating reduced power in all circumstances compared with using the LD map (data not shown).

Conclusion

We reported a significant association between a region of 18q and RA. The estimated genomic location of the disease variant was at 53,306 kb. The Malecot model and composite likelihood approach has narrowed the possible disease locus to a 36-kb candidate region. A haplotype significantly associated with reduced risk of RA was identified in this region. DNA sequences between 53,295–53,331 kb of this region are highly conserved in vertebrates. A haplotype around 51,585 kb was also identified as reducing the risk of RA. Further sequencing or functional studies may be helpful to identify the disease variants. Reducing SNP density decreases power and location accuracy. We also conclude that SNP selection based on

Table 4: The impact of region length on association mapping

Region length		No. of regions	No. of SNPs in the region		χ^2_3	*p*	Length of 95% CI (kb)
LDU[a]	kb		All	$p < 0.05$ (%)			
1	75	58	49	22 (45)	14.02	0.0029	34
4	122	30	66	27 (41)	16.83	0.0008	40
10	578	14	191	34 (18)	10.44	0.0152	73
20	1262	7	382	37 (10)	6.75	0.0802	76
60	3732	2	946	84 (9)	7.58	0.0554	76

[a]LDUs centering S_1 at 53,306 kb to which all *S* estimates were equal.

Table 5: SNP density and accuracy – selection by tagging or equidistance

r^2/LDU/kb (kb/SNP)	No. of SNPs[a]	Location error[b]			$\Lambda_A - \Lambda_D$		
		Tagger	E_LD	E_kb	Tagger	E_LD	E_kb
Full (4)	189/189/189	0	0	0	79	79	79
1.0/0.002/1 (5)	160/163/150	1	0	-2	62	69	68
0.8/0.036/5 (11)	78/74/64	-14	4	-5	21	18	20
0.6/0.071/9 (15)	59/55/48	-26	-37	-3	10	11	10
0.4/0.134/14 (21)	43/39/33	-27	25	-5	12	11	8
0.2/0.300/26 (33)	28/21/20	-35	33	-180	7	9	1

[a]For the studied region for Tagger, Equal LD (E_LD) and kb (E_kb) distance, respectively.
[b]Assuming S_1 is the "disease locus," the region was fixed at 10 LDUs centering at S_1 and the location error was calculated as S-53,306 kb.

equal LD distance can maximally retain the prediction accuracy of the disease loci than that based on equal physical distance or SNP tagging.

Competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

T-YK, WL, and CH were supported by the Ph.D. studentships funded by the Taiwan Ministry of Education, University of Southampton, and Shanghai Jiaotong University, respectively. WZ was supported by the Institute of Cancer Research, Sutton, Surrey, UK.

This article has been published as part of *BMC Proceedings* Volume 1 Supplement 1, 2007: Genetic Analysis Workshop 15: Gene Expression Analysis and Approaches to Detecting Multiple Functional Loci. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/1?issue=S1.

References

1. Choi SJ, Rho YH, Ji JD, Song GG, Lee YH: **Genome scan meta-analysis of rheumatoid arthritis.** *Rheumatology (Oxford)* 2006, **45**:166-170.
2. Morton NE, Maniatis N, Zhang W, Ennis S, Collins A: **Genome scanning by composite likelihood.** *Am J Hum Genet* 2007, **80**:19-28.
3. Zhang W, Lau W, Hu C, Kuo T-Y: **Impact of marker density on the accuracy of association mapping.** *BMC Proc* **1(Suppl 1)**:S166.
4. Gomes I, Collins A, Lonjou C, Thomas NS, Wilkinson J, Watson M, Morton N: **Hardy-Weinberg quality control.** *Ann Hum Genet* 1999, **3**:535-538.
5. Lau W, Kuo TY, Tapper W, Cox S, Collins A: **Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome.** *Bioinformatics* 2007, **23**:517-519.
6. Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data.** *Am J Hum Genet* 2001, **68**:978-989.
7. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: **Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.** *Am J Hum Genet* 2004, **74**:106-120.

BMC Proceedings

BioMed Central

Proceedings

Open Access

Impact of marker density on the accuracy of association mapping

Weihua Zhang*†[1,4], Winston Lau†[2], Cheng Hu[3] and Tai-Yue Kuo[2]

Address: [1]Section of Cancer Genetics, The Institute of Cancer Research, 15 Cotswold Road, Belmont, Sutton, Surrey SM2 5NG, UK, [2]Human Genetics Division, Duthie Building (Mailpoint 808), Southampton General Hospital, University of Southampton, School of Medicine, Tremona Road, Southampton, SO16 6YD, UK, [3]Shanghai Diabetes Institute, Shanghai Jiaotong University, 600 Yishan Road, Shanghai 200233, People's Republic of China and [4]Department of Cardiology, Ealing Hospital NHS Trust, Uxbridge Road, Southall, Middlesex, UB1 3HW, UK

Email: Weihua Zhang* - weihua.zhang@eht.nhs.uk; Winston Lau - wwsl@soton.ac.uk; Cheng Hu - alfredhc@sjtu.edu.cn; Tai-Yue Kuo - kuotaiyu@soton.ac.uk

* Corresponding author †Equal contributors

from Genetic Analysis Workshop 15
St. Pete Beach, Florida, USA. 11–15 November 2006

Published: 18 December 2007

BMC Proceedings 2007, **1**(Suppl 1):S166

This article is available from: http://www.biomedcentral.com/1753-6561/1/S1/S166

Abstract

We studied the impact of marker density on the accuracy of association mapping using Genetic Analysis Workshop 15 simulated dense single-nucleotide polymorphism (SNP) data on chromosome 6. A total of 1500 cases and 2000 unaffected controls genotyped for 17,820 SNPs were analyzed. We applied the approach that combines information from multiple SNPs under the framework of the Malecot model and composite likelihood to non-overlapping regions of the chromosome. We successfully detected the associations with disease Loci C and D and predicted their locations as small as zero distance to Locus C when it was "typed" and 112 kb from the untyped rare Locus D. Reducing marker density decreased the accuracy of location estimates. However, the predicted locations were robust to variations in the number of SNPs. Generally, the linkage disequilibrium (LD) map reflecting distances between markers in relation to LD produced higher accuracy than the physical map. We also demonstrated that SNP selection based on equal LD distance outperforms that based on equal physical distance or SNP tagging. Furthermore, ignoring rare SNPs diminished the ability to detect rare causal variants.

Background

As the cost of genotyping decreases, genome-wide association (GWA) mapping of the predisposition genes for complex diseases is becoming a common study design in genetic epidemiology. As the huge number of single-nucleotide polymorphisms (SNPs) in the human genome is still prohibitive for exhaustive investigation, subsets of SNPs have often been selected for large scale studies. Morton et al. developed a novel GWA mapping approach based on the Malecot model and composite likelihood combining multiple marker information from non-overlapping genomic regions to predict the locations of disease variants [1]. We applied this approach to the Genetic Analysis Workshop (GAW) 15 Problem 3 simulated dense chromosome 6 data with the knowledge of the answers and we studied the effect of SNP density on the accuracy of association mapping.

Methods

Data

The simulated data set contained 1500 families with a sib pair affected with rheumatoid arthritis (RA) and a random sample of 2000 unrelated and unaffected individuals. To form a case-control study, we selected the first sibling per family as a case. A total of 1500 cases and 2000 controls from Replicate 1 were analyzed. There are three simulated disease loci. HLA-DR is at the same location of 32484.648 kb as Locus C, where a SNP denseSNP6_3437 lies, so we considered this SNP the disease variant C. Locus D is at 37233.784 kb, in very weak linkage disequilibrium (LD) with Locus C. The minor allele frequency (MAF) for the C allele was 0.4055 in control samples. The D allele has a population frequency of 0.0083, but the variant was not typed.

Genotype data were composed of 17,820 SNPs on chromosome 6, mimicking a 300 K GWA scan with no missing values. Fifty-eight SNPs showing departure from Hardy-Weinberg equilibrium (HWE) in control samples ($\chi^2_1 \geq 10$ for either Pearson's or likelihood ratio chi-square tests) were discarded [2]. Following convention, 2061 rare SNPs with MAF < 5% were further removed except when otherwise indicated. The main data set (1) was thus composed of a total number of 15,701 SNPs. In another experiment we retained all SNPs but removed 26 SNPs showing departure from HWE by the likelihood-ratio test and this generated 17,794 SNPs (data set 2).

LD map

The physical map length was 170,813 kb. LD maps expressed in LD units (LDUs) were constructed based on pair-wise LD for multiple markers in control samples [3]. LDU is the product of ε and kb distance for an interval of two adjacent SNPs and is additive, where ε represents the exponential decline of LD with distance for that interval. We used the LDMAP-cluster, a parallel version of LDMAP program that rapidly constructs the maps of equally divided chromosome segments http:www.som.soton.ac.uk/research/geneticsdiv/epidemiol ogy/ldmap/[3]. For each segment, an overall ε value was also estimated. The LD map length was 1311.225 LDUs for the main data set and 1237.923 LDUs for data set 2. SNPs can have the same LDU if they are in an LD block. Therefore, we also made tilted LD maps by reassigning LDU locations for the SNPs with the same LDU by linear interpolation.

Association mapping

A chromosome is divided into non-overlapping consecutive regions of a minimum number of 30 SNPs and a minimum length of 10 LDUs by default without breaking LD blocks. Each genomic region was then analyzed separately. Association between SNP alleles and disease status in the Malecot model is a function of several parameters. Composite likelihood combines information of all marker-disease association in a genomic region. The parameters were estimated through fitting the model to the data with a map in LDU or kilobases and by minimizing -2 natural log composite likelihood (denoted as Λ) [1]. The estimated location *S* of the disease locus is converted to a kilobase scale. The significance test is performed by contrasting two hierarchical models. Model A assumes no association with the disease, therefore *S* is not estimated. Model D assumes an association with the disease and *S* and two other parameters are estimated and ε is specified. The difference in Λ between models A and D ($\Lambda_A - \Lambda_D$) is monotonic to the magnitude of chi-square with three degrees of freedom (χ^2_3). Permutation by shuffling case-control status for each region was performed to obtain empirical *p*-values [1]. The algorithms were implemented in the CHROMSCAN program. A parallel version, CHROMSCAN-cluster, deployed on a local Beowulf cluster http://www.som.soton.ac.uk/research/geneticsdiv/epidemiology/chromscan/ was used for computing 1000 replicates.

The values of ε were obtained by averaging over eight segments in LD map construction, which were 1.14472 and 0.00568 for LD and kilobase maps, respectively, for the main data set, and 1.14386 and 0.00544 over nine segments for data set 2. Theoretically, a more accurate ε may be obtained by fitting the maps to the whole chromosome data, but the extensive computing power required for the task is impractical to implement and beyond the current computing resource. Also, slightly altered ε values did not appear to have an appreciable effect (data not shown).

For comparison, a single SNP χ^2_1 was obtained by the 2 × 2 allelic count table and the most significant SNP (msSNP) showing maximal χ^2_1 in each region was identi-

fied. Location error (in kilobases) was defined as the difference between S or the location of the msSNP and the true location of disease variant. Accuracy refers to the precision of the predicted location S. The smaller the error, the higher the accuracy.

SNP density

To generate different SNP density, we selected every i^{th} SNP (i = 2, 3, ..., 20, 25, 30) in the order of their physical locations from the full data set, representing $1/i$ the number of SNPs in the original set. For a candidate region spanning Loci C and D with rare SNPs included, we used Tagger implemented in the Haploview software to select tagging SNPs that optimally capture allelic variation among SNPs at a given r^2 threshold based on pairwise LD in control samples [4]. For comparison, we selected the same number of SNPs as Tagger but in equal LDU or kilobase distance. To do this we used the tilted LD map in which every SNP had a unique LDU location. We also studied the impact of region length and sample size.

Results and discussion

Association mapping of disease loci in full data set

Fourteen out of 126 regions showed nominal significant association with RA ($p < 0.05$), among which eight consecutive regions spanned Loci C and D (Table 1). Five regions remained significant after Bonferroni correction, among which four surrounded or spanned Locus C, and one covered Locus D (Table 1). Locus C was inside the most significant region 29. Therefore, the three regions surrounding Locus C with less significance levels must be the result of LD between variant C and other SNPs. The discontinuity of significance surrounding region 32 indicated that this region harbored another disease locus and indeed, this was where Locus D lies. Therefore, we successfully detected Loci C and D in the initial analysis. The lowest p for the rest of the regions was 0.0064. Given that there were no other disease loci, the approach had a right type I error rate (6/118 = 0.05). A lesson learned was that when there was long-range LD, consecutive regions showing association may reflect one instead of several disease loci. As an alternative to merging regions, we studied the impact of region length on accuracy (see below).

S for Locus C was reasonably accurate (55 kb apart from true location using LD map). However, the location error was 542 kb for Locus D and the 95% confidence interval did not include Locus D. Removing 10 SNPs showing significant LD with variant C did not change the results. We then divided region 32 into two or three sub-regions. Again, we did not detect significant association in the middle part where Locus D lies, although we detected the associations in the first and third sub-regions where two clusters of highly significant SNPs lay. Because Locus D is rare, the removal of rare SNPs may have had an effect. We then added rare SNPs and used the corresponding LD map and ε values, and the location accuracy was markedly improved for Locus D (Table 2). Among the added rare SNPs, three were highly associated with the disease: denseSNP6_3931, _3933, and SNP6_162 (χ^2_1 = 118, 116, and 116, respectively). It is therefore a mistake to remove rare SNPs (MAF < 0.05) in association analysis. This was in contrast to the HapMap project in which the focus was on common SNPs. However, inclusion of rare SNPs resulted in higher location error for common disease Locus C (Table 2).

Occasionally or under high marker density, the kilobase map performed better than the LD map, presumably because every SNP has a unique physical location, whereas several SNPs could have the same LDU location in LD blocks. The tilted LD map improved the location accuracy for Locus C, although not for Locus D (Table 2).

In practice, the phenomenon in this simulated data set may be too extreme. On the other hand, it is possible that several disease loci can be closely located. To distinguish such loci is a challenge to genetic epidemiologists. Under this circumstance, single SNP association plus a gene functional study may be useful.

Table 1: Association mapping of disease Loci C and D on chromosome 6

Region[a]	No. SNPs	S (kb)	$\Lambda_A - \Lambda_D$	χ^2_3	P	P_c
26	128	24882	68	14	0.002471	0.311346
27	239	26121	793	40	$<10^{-7}$	0.000001
28	348	31299	5176	128	$<10^{-12}$	<0.000001
29	176	32540	33058	322	$<10^{-12}$	<0.000001
30	153	33962	295	25	0.000017	0.002129
31	134	35638	103	15	0.001929	0.243096
32	127	37776	141	27	0.000007	0.000926
33	147	39432	50	13	0.005554	0.699754

[a]A segment of consecutive regions of 10 LDUs showing nominal significant association with RA ($p < 0.05$). See Methods for the meaning of other symbols. Loci C and D were in regions 29 and 32 at locations of 32485 and 37234 kb, respectively.
[b]P_c is Bonferroni corrected p-value for multiple tests of 126 regions ($p \times 126$).

Table 2: Candidate regions of disease Loci C and D with rare SNPs included

Locus	Map	S (kb)	$\Lambda_A - \Lambda_D$	χ^2_3	*p*	Location error with rare SNPs	
						Included	Removed
C	LD	32557	30693	360	$<10^{-12}$	72	55
	LD, tilt	32518	30799	197	$<10^{-12}$	34	21
	kb	32506	28632	496	$<10^{-12}$	22	14
D	LD	37358	154	22	0.000017	124	542
	LD, tilt	37368	156	28	0.000003	130	546
	kb	37368	148	17	0.000666	130	954

SNP density based on the order

As density decreases, location error increases whether using single or multi-SNP approaches when the disease variant was not "typed" (Table 3). There was an improvement in accuracy when the disease variant was included. In most cases, using the LD map resulted in greater accuracy than using the kilobase map, especially when the marker density was low. We also selected SNPs on the scale of one to the hundredth or even the thousandth. As long as there was one SNP highly associated with the disease (e.g., $\chi^2_1 = 27$), the association was detectable, but much compromised by precision as a result of low SNP density. These data are unusual in that the association of Locus C is extremely significant and probably would not be observed in the real data.

Although mapping accuracy decreases with marker density, even with 1/30 the number of SNPs, corresponding to a 10 K GWA scan, we could still detect Locus C (Table 3). Single SNP tests depend heavily on whether the disease variant is typed. It has less predictive value for accuracy because the SNP with maximal χ^2 is not necessarily the closest SNP to the disease variant. In contrast, methods that combine information from multiple markers predict the location of the disease variant better than single SNP tests because the location is less influenced by any single SNP effects. A multi-marker approach may therefore be more robust to genotyping errors.

We expect that the mapping accuracy will be improved further in maps with higher marker density than that assessed in this paper, such as the commercially available 500 K or more genotyping platforms for GWA studies.

SNP density based on tagging or equidistance

For the 15,805.710 kb candidate region spanning both Loci C and D, we compared location accuracy using SNPs selected with Tagger or by equidistance of LDU or kilobases (Table 4). SNPs based on equal LDU provided higher location accuracy than those based on equal kilobase distance. Equidistance generally provided higher accuracy than tagging SNP selection. Again, reducing SNP density decreases the prediction accuracy of disease Loci C and D, but this was minimally affected by selection based on equal LD distance (Table 4).

Table 3: Density and accuracy for Locus C – SNP selection by order

SNP density (kb/SNP)	No. SNPs	No. regions	msSNP		Location error by the composite likelihood approach			
					Causal SNP out		Causal SNP in	
			χ^2_1	Location error	LD	kb	LD	kb
Full (11)	15701	126	2324	153	57	13	55	14
1/2 (22)	7850	125	1762	-2	5	-19	5	-20
1/3 (33)	5233	118	2324	153	153	40	6	40
1/4 (44)	3925	106	1762	-2	-65	-56	-57	-53
1/5 (54)	3140	94	2274	42	-24	-35	-15	-36
1/6 (65)	2616	82	1285	20	20	20	10	15
1/8 (87)	1962	64	1601	65	-58	-64	-47	-58
1/10 (109)	1570	52	726	-106	-55	-59	-24	-46
1/15 (163)	1046	34	486	-887	294	26	0	3
1/20 (217)	785	26	726	-106	-97	-160	-26	-43
1/25 (272)	628	20	348	-9	69	-120	60	-79
1/30 (326)	523	17	229	188	362	-25	0	-25

Disease variant C ($\chi^2_1 = 1916$) was not present except in the full data set or specified.

Table 4: Density and accuracy – SNP selection by tagging or equidistance[a]

		Locus C			Locus D		
R^2 (LDU, kb)	No. SNPs (kb/SNP)	Tagger	E_LD	E_kb	Tagger	E_LD	E_kb
Full	1658 (10)	20	20	20	130	130	130
0.8 (0.013,7)	1080[b] (15)	35	20	30	130	124	130
0.6 (0.025,10)	874 (18)	56	-16	27	545	124	551
0.4 (0.047,15)	657 (24)	106	42	63	124	123	117
0.2 (0.099,27)	421 (38)	125	14	-23	537	112	231

[a]Location error for SNPs selected by Tagger or equal LDU (E_LD) or kb (E_kb) distance in a candidate region of 15805.710 kb with rare SNPs included. Regions were fixed at 10 LDUs for Loci C (30997–33398 kb) and D (36784–37792 kb). Disease variant C was not present except in the full data set. Tilted LD map.
[b]1079 for E_LD.

Sample size and region length

We analyzed different sample sizes based on the combination of 500, 1000, 1500, and 2000 cases or controls. Despite variations in location errors for Locus C, there was no clear trend to draw any meaningful conclusion. For Locus D, however, a high degree of accuracy appeared to be maintained when the data sets had over 1000 cases and 1500 controls. Therefore, large samples are needed for detecting rare disease loci.

With Locus C being centred, we studied region lengths from 0.2 up to 30 LDUs, with the latter starting in region 27 and ending in region 30. The location error was relatively stable but extremely small or large LDU lengths resulted in increased error. The region lengths in LDUs (location errors in kilobases) were 0.2 (107), 1 (5), 2 (82), 4 (5), 6 (5), 8 (5), 10 (5), 12 (-10), 14 (-10), 16 (-13), 18 (-14), 20 (-14), and 30 (-68). We therefore recommend 10-LDU for the maximal length while maintaining minimal error. Increasing the number of SNPs also linearly increases the computing load [3].

Fixing region length had no appreciable impact on location accuracy at high density, but the errors were greater than let-the-program-decide regions at low density (data not shown).

Conclusion

We successfully detected disease Loci C and D in the simulated dense chromosome 6 data using the Malecot model and composite likelihood approach. Decreasing SNP density compromises accuracy of association mapping. This multi-marker approach has many advantages. Firstly, it markedly decreases the number of tests in GWA studies, avoiding heavy penalty for multiple testing. Secondly, it predicts the disease loci more accurately than single SNP association tests. We also demonstrated that SNP selection by equal LD distance outperforms that by tagging or equal kilobase distance in the accuracy of association mapping. Finally, we conclude that excluding rare SNPs significantly decreases the power and accuracy in mapping rare disease loci.

Competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

WZ was supported by the Institute of Cancer Research, Sutton, Surrey, UK. WL, CH, and T-YK were supported by Ph.D. studentships funded by the University of Southampton, Shanghai Jiaotong University, and the Taiwan Ministry of Education, respectively.

This article has been published as part of *BMC Proceedings* Volume 1 Supplement 1, 2007: Genetic Analysis Workshop 15: Gene Expression Analysis and Approaches to Detecting Multiple Functional Loci. The full contents of the supplement are available online at http://www.biomedcentral.com/1753-6561/1?issue=S1.

References

1. Morton NE, Maniatis N, Zhang W, Ennis S, Collins A: **Genome scanning by composite likelihood.** *Am J Hum Genet* 2007, **80**:19-28.
2. Gomes I, Collins A, Lonjou C, Thomas NS, Wilkinson J, Watson M, Morton N: **Hardy-Weinberg quality control.** *Ann Hum Genet* 1999, **3**:535-538.
3. Lau W, Kuo TY, Tapper W, Cox S, Collins A: **Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome.** *Bioinformatics* 2007, **23**:517-519.
4. Carlson CS, Eberle MA, Rieder MJ, Yi Q, Kruglyak L, Nickerson DA: **Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium.** *Am J Hum Genet* 2004, **74**:106-120.

LDMAP: The construction of high-resolution linkage disequilibrium maps of the human genome.

Kuo T-Y, Gibson J, Lau W, Morton NE, Collins A*.

Human Genetics

School of Medicine

University of Southampton

Southampton SO16 6YD

UK

Tel: 44 (0)23 80 796939

Fax: 44 (0)23 80 794264

Email: arc@soton.ac.uk

*To whom correspondence should be addressed.

Keywords: linkage disequilibrium maps, human genome, computational load, relative efficiency, segmental assembly

Summary
The precise characterisation of the linkage disequilibrium (LD) landscape from high density single nucleotide polymorphism data underpins the association mapping of diseases and other studies. We describe the algorithm and implementation of a powerful approach for constructing LD genetic maps with meaningful map distances. The computational problems posed by the enormous number of SNPs typed in the HapMap data are addressed by developing segmental map construction with the potential for parallelization which we are developing. There is remarkably little loss of information (1-2%) through this approach but the computation times are dramatically reduced (more than 4 fold). These developments bring the construction of very high density LD maps using the 3 million SNP HapMap sample within reach. We anticipate that a whole-genome LD map will have substantial impact on disease gene mapping, genomic research and population genetics.

Introduction.
Linkage disequilibrium (LD, or allelic association), describes the statistical association between polymorphisms, such as single nucleotide polymorphisms (SNPs), and between markers and genes contributing to disease. The existence of LD reflects transmission over many generations of short segments of ancestral haplotypes comprising closely linked markers. Allelic association is evident because haplotype frequencies are not simply the products of the appropriate allele frequencies, hence 'disequilibrium'. LD is present because recombination, which destroys LD, is infrequent over small distances while other processes, such as genetic drift and population bottlenecks, act to create LD over a number of generations. A thorough understanding of the extent and structure of LD is essential for association mapping of the polymorphisms that contribute to human diseases. Given the availability of substantial bodies of high resolution SNP data (for example from the International HapMap project, http://www.hapmap.org/ , International HapMap Consortium, 2003) it is now possible to characterise LD patterns genome-wide. Once the structure is characterised there are likely to be substantial payoffs from increased resolution and power for localisation of disease genes (Maniatis et al, 2005), and for identifying genomic regions subject to selection (Sabeti et al, 2002).

It is known that LD extends for tens of kilobases, on average, in the human genome. This is true even for large heterogeneous human populations and not just isolates (Lonjou et al, 2003), suggesting that the genome might be screened with reduced numbers of SNPs because close association implies some redundancy. This is the main motivation behind the HapMap project, which aims to identify 'tag' SNPs to represent a particular haplotype with little loss of power, a strategy relying on recognition that some parts of the genome contain regions (blocks) of low haplotype diversity (Daly et al, 2001). However, much of the genome is more complex, reflecting the combined effects of intense recombination hot spots, more randomly distributed recombination events and other phenomena. Furthermore the definition of block boundaries and the instability of blocks defined with different marker densities poses difficulties (Tapper et al, 2003, Ke et al, 2004). It is also evident that a 'haplotype map' (Dawson et al, 2002), while providing annotation, is not a genetic map with meaningful distances which describe LD structure. It is also unclear how the annotation of haplotypes is directly useful for disease mapping.

A successful alternative strategy is to represent LD patterns in the form of a metric map with additive 'linkage disequilibrium unit' (LDU) distances (Maniatis et al, 2002). The low resolution features of LD maps resemble the linkage map in pattern but there are important differences which reflect population history. A whole chromosome LD map of chromosome 22 (Tapper et al, 2003) shows a close correspondence between areas of extensive LD with low recombination and areas of low LD with intense recombination. LD maps have already been used for multi-locus disease gene mapping using locations on the LDU scale as the association mapping analogue of the linkage map for localising major genes (Maniatis et al, 2004, 2005). LD units are analogous to centimorgans (cM) in that locations increase monotonically with physical distance but, whilst linkage map length is related to recombination in one generation, the LDU map length reflects accumulated recombination over many generations. The ratio of the LDU map length to the linkage map in Morgans estimates the effective number of generations over which recombination has occurred (the 'effective bottleneck time', Zhang et al, 2004), with some distortion in the LD map due to selection and because of systematic errors in estimating interference in the linkage map.

Algorithms to construct LD maps have been developed and evaluated by Maniatis et al (2002), Zhang et al (2002) and Lonjou et al (2003). The LDMAP program (http://cedar.genetics.soton.ac.uk/public_html/) described here implements and extends these algorithms. We describe here an approach for the construction of a genome-wide LD map at very high density by addressing the particular computational difficulties posed by the analysis of huge numbers of markers.

Overview of the basic algorithm.

The population genetics theory behind LD map construction is described by Morton et al (2001). The decline of LD, modelled as association ρ as a function of distance d, in Kb, is $\rho = (1\text{-}L)Me^{-\varepsilon d} + L$, in which the L parameter reflects residual association at large distance not due to linkage, M is the intercept, the association at zero distance, and ε is the exponential decline of LD as the product of recombination θ and number of generations t. The model has the same form as that developed by Malecot (1948) to describe genetic isolation by distance but has different parameters.

LD map construction estimates ε in each map interval between adjacent SNPs. For any pair of SNPs the association probability ρ and the information K_ρ form the data for LD map construction. Pairs that span a given interval contain information about association in that interval, but pairs at large distance are uninformative. The estimation of the ε vector requires the iterative substitution of distance d in the Malecot equation with distances in linkage disequilibrium units (LDUs). These are defined, for the i^{th} interval between adjacent SNPs, as $\varepsilon_i d_i$ with locations by summation over preceding intervals (Maniatis et al , 2002). The LDU locations, when plotted against Kb, typically show a pattern of steps where LD is breaking down and plateaus or blocks of high LD.

Model implementation and methods.

The raw data comprise SNP genotypes (diplotypes) from unrelated individuals with alleles coded 0 (missing), 1 and 2. Alternatively, where known with a high degree of

reliability, SNP haplotypes are used. The physical location, in kilobases from an origin closest to the p telomere for each SNP is obtained from the latest human genome sequence release.

The genotypic data are reduced to pairwise association and the corresponding information (Collins and Morton,1998, Collins et al 1999). Informative SNP pairs are selected subject to two constraints, of which the minimal set is used in the analysis. The first is the maximum distance in kilobases between any pair of SNPs, defaulted to 500 Kb. This eliminates pairs separated by a distance which greatly exceeds the range of LD in most human populations, although for isolated populations, certain genomic regions and for building LDU maps of other organisms this constraint may not be appropriate. For sub-Saharan African populations, and genomic regions with a high recombination rate, the 500 Kb distance is excessive but inclusion of these pairs only impacts on computation time. However, at the SNP densities available in the HapMap data this constraint is much less important than the second constraint which restricts the number of map intervals between any pair of SNPs. To compute ε for a given interval between adjacent SNPs, a pair that spans that interval is potentially informative but the information approaches zero if the number of intervals between the pair is large. To reduce the computational load the default maximum number of intervals, s, between a pair of SNPs informative for a given interval is 100. Therefore, for the computation of ε, there is a sliding window which encompasses all the informative pairs that span the interval. When the maximum number of intervals constraint is operating (and no pairs are eliminated by the maximum distance constraint) the total number of pairs used (N) in a map of n SNPs is:

$$N = \frac{n(n-1)}{2} - \frac{(n-s-1)(n-s-2)}{2}$$

To compute ρ for SNP pairs from diplotype data we apply the E.M. algorithm of Hill (1974) which iteratively reduces a 3x3 table of genotypic counts to four haplotype frequencies. These are converted to counts and a file which specifies the SNP pair and the sequence locations in kilobases, together with the four counts, is produced. Because no re-arrangement of the 2x2 table has taken place at this point the four counts correspond to the 11, 12, 21 and 22 haplotypes from the marker pair. This file can be concatenated with corresponding files from other populations and counts summed for shared marker pairs, assuming alleles are labelled consistently. The summed counts have been used to compute ρ for construction of 'cosmopolitan' maps (Lonjou et al, 2003, Gibson et al, 2005).

Rare SNPs with minor allele frequencies less than 0.05 are eliminated, as are any that show strong deviation from Hardy-Weinberg equilibrium (Gomes et al, 1999). The association probability ρ is obtained by re-arranging the 2x2 table (Table 1) to ensure that Q is the minimal allele frequency (Q < R, 1-R and 1-Q) and that products of haplotype frequencies give ad > bc. Conforming to this re-arrangement requires the re-labelling of SNPs (SNP_1 becoming SNP_2 and vice versa) and/or re-labelling of the SNP alleles. To achieve Q < R, markers are interchanged by switching b and c, which has the effect of exchanging Q with R and 1-Q with 1-R; for Q < 1-R markers are interchanged by switching a and d, which has the effect of exchanging Q with 1-R and 1-Q with R; for Q<1-Q alleles are interchanged (a with c and b with d) which

switches Q with 1-Q. Finally, to conform to ad > bc, alleles are interchanged, a with b and c with d which switches R with 1-R. Columns are also interchanged in the special case that disequilibrium D is zero, where b > a. The 'intermediate' file used by the program specifies the SNP pair, sequence locations (Kb), ρ, K_ρ, χ^2, sample size m, Q, R, D and the pair selection criteria (maximum number of intervals, maximum window size in Kb).

Fitting data to the kilobase map

From the intermediate file the fit of the pairwise data to the Kb map under the Malecot model is established. Pairwise data enter composite log likelihood as:

$\ln lk = -\sum K_\rho(\hat{\rho} - \rho)^2/2$, where the summation is over informative pairs (i = 1, N), ρ is the observed association between the i^{th} pair (Table 1) and $\hat{\rho}$ are the fitted values. Function minimisation is achieved using the variable metric method implemented in the subroutine *dfpmin* (Press et al, 1994, page 428). Parameter estimation for ε, L and M is controlled through a script (a 'job' file), which allows testing of hypotheses such as deviations from L=0 or M=1. In general two models (A and B) can be compared as $\chi^2_n = (-2\ln lk_A - -2\ln lk_B)/ V_B$, where model B has one or more additional parameters estimated than the simpler model A. V_B is the error variance of model B defined as $V_B = -2\ln lk_B / (N-g)$, where N is the number of pairs and g is the number of parameters estimated.

Morton et al (2001) defined a predicted value for the L parameter (Lp) which is equal to the K_ρ-weighted mean of $\sqrt{2/\pi K\rho}$ where K_ρ, the information about ρ per marker pair, is proportional to sample size. Lp depends only on the mean value of ρ for markers at large distances such that the expected value of disequilibrium D is zero.

Construction of an LD map.

The Malecot parameters from the kilobase map provide starting values for construction of the LD map. The iterative process implemented in LDMAP estimates ε, for intervals between adjacent SNPs, following the 'interval' method described by Maniatis et al (2002). Briefly, let $S_{hk} = \Sigma\varepsilon_i d_i$ where i is an interval between adjacent SNPs and summation is over all intervals contained between SNPs h and k and $\rho_{hk} = (1-L)Me^{-S_{hk}} + L$, using trial values for M, L and ε_i as described above. The estimate of ε_i, at iteration t, is given by:

$\varepsilon_i^{(t)} = \varepsilon_i^{(t-1)} + (U_i/K_i)^{(t-1)}$, where $U_i = \sum\left(\frac{\partial \ln lk}{\partial \rho_{hk}}\right)\left(\frac{\partial \rho_{hk}}{\partial \varepsilon_i}\right)$ *and* $K_i = \sum K_{\rho hk}\left(\frac{\partial \rho_{hk}}{\partial \varepsilon_i}\right)^2$.

At convergence each revised estimate ε_i contributes towards a 'global' iteration which is a complete update of the ε vector and the computation of the global composite likelihood, which is maximized iteratively. The M parameter is assumed constant for all intervals and is updated periodically at global iterations 25, 50, 100, 200, 400, 800, 1600 and so on. At these points the composite log likelihood for the LD map (-2lnlk) is obtained. This updating procedure accelerates convergence. The L parameter is optionally updated at the same points, but usually the predicted value (Lp) is used. Experience with LD map construction has shown that the estimated L may exceed Lp in small samples. This might be attributed to the local effect of block structure which can distort L (Lonjou et al, 2003). Compared to Lp, the estimation of

L typically creates more intervals where $\varepsilon_i d_i$ exceeds 3, termed 'holes' (Tapper et al, 2003). In high density maps most holes are associated with a locally high recombination rate (Tapper et al, 2003) and segments requiring local increases in marker density can thus be identified (Gibson et al, 2005).

When an estimate ε_i bounds at zero (consistent with 'complete' LD), that estimate is fixed at zero and no further iteration takes place, with a consequent reduction in computation time. We have found that removing these intervals from further iteration has very little effect on the final map, suggesting that most estimates remain at the zero limit once reached. The same applies to holes, and these intervals are also dropped from further iteration. However, the constraints are not applied until a 'burn-in' period corresponding to 50 global iterations has taken place. Convergence is declared when a difference in global composite likelihood between two consecutive iterations is less than 0.01.

Towards a genome-wide LD map.

In a map of n loci there are n-1 estimates of ε, achieved through maximizing the composite likelihood, for which the computation time may be substantial. The computation time depends on a number of factors, but particularly the number of pairs used in map construction. Exclusion of pairs which contain no significant information about a given interval is one approach to reducing computation time. However in maps with many 10s of thousands of loci the exclusion of these pairs is inadequate for constructing maps within an acceptable time frame. We have examined a number of alternatives to reduce computation time, including the construction of maps at adaptively increasing densities. However, this was found to offer only modest speed enhancements. The assembly of maps in overlapping sections, with distances averaged in the overlap region, is much more promising and we here examine the impact of this approach on the quality of the map. For this evaluation the September 2004 release of the HapMap data for chromosome 22 was used. The CEPH sample comprises 9,658 loci from 60 unrelated individuals of Western-European ancestry. We constructed 9 LD maps of chromosome 22 for a range of numbers of segments between 1 (the complete map constructed in one piece) and 200 pieces (Table 2). We computed the error variance for each map by testing the fit of a 'standard' set of pairwise data for each chromosome (with default settings of 500 Kb as maximum window size and 100 as number of intervals). This enabled direct comparison of the relative efficiency of alternative numbers of map segments and the relationship to computing time. For each map the error variance,V, was computed and the efficiency of each map was computed relative to the map constructed in one piece. We also looked at the relative computation time and relative resulting map lengths in the same way.

Results

The LDU map length (Table 2) is rather stable showing a maximum increase in length of ~7% over the range of tests (maps constructed in 1 to 200 segments). The error variances are similarly stable varying over the range 0.841-0.886. The relative efficiency (Figure 1) is therefore high across the range with a maximum loss of information of only ~5% when the map is constructed in 200 segments and much reduced losses for maps built in fewer segments (for example the loss of information

is less than 1% for maps built from segments of 698 loci). Figure 2 plots the contour of the LDU maps constructed as one piece in contrast to the LDU map constructed from 200 segments. Both contours show details of the LD structure of chromosome 22 including two large regions (plateaus) of at least 5 Mb at around 30 and 40 Mb along the chromosome where there is extensive LD. There are also 3-4 regions of rather intense recombination, the most striking of which is around 26 Mb along the chromosome. The contour for the two maps is strikingly similar and the small difference in map length appears to be spread over the whole length of the map. This suggests that the segment approach slightly exaggerates map length because of the loss of information at the ends of segments where there are no flanking SNPs. It is important to balance computational feasibility and number of segments for map construction. Although the LDU maps are stable and the loss of information is minimal the computation times of a SUN V440 server vary enormously over the range. The computing time for the map when constructed in one piece is 14 fold greater than for the map constructed from 200 segments. It is evident that a good compromise between optimal computational times and minimising information loss in the map is achieved in the 100-1000 loci per segment range. The use of ~500 loci per segment seems justified for map construction generally and the construction of maps of the largest chromosomes becomes feasible, even at the higher SNP densities of the later HapMap releases.

Discussion

The results of the analysis show that LD maps are robust to segmental assembly of LD maps with very little loss of information even for the smallest segment sizes. This justifies the construction of a genome-wide LD map using selected pairs and through segmental map assembly. The results demonstrate that genome-wide LD maps are achievable even at the highest marker densities which, on completion of the HapMap project, will include >3 million SNP genotypes, for a range of populations. The mean extent of LD in the CEPH sample is approximately 50 Kb, implying a mean spacing of ~3Kb in a one million SNP map. A pair of SNPs separated by 100 intervals will span ~300 kilobases and therefore imposing this limit will result in little or no loss of information at HapMap densities. However, when applied to a map of 50,000 SNPs, which will be exceeded for the larger chromosomes, there will still be more than 2.5 million pairs for analysis. Because of the computational load, and the modest loss of information we have demonstrated, map assembly using overlapping segments is the most practical approach. As an illustration we constructed LDU maps of chromosome 1 using the September 2004 release of HapMap (28,685 SNPs) and the February 2005 release (53,401 SNPs), Table 4. Reducing Num_intervals to 50 for the higher density data generates somewhat fewer total pairs and, even with map assembly in larger segments of 1000 loci, the computation time is much reduced. Interestingly the use of 86% more SNPs in the February 2005 sample only reduces the number of holes by 28%. Further SNP typing focussing on regions with holes would be much more efficient than the general addition of SNPs randomly over the whole map. The higher density LDU map is 4.8% longer, a proportion of the increase in length presumably reflecting the resolution of holes which are concentrated in recombination high areas (Tapper et al, 2003). Overall the limit of 3 LDUs imposed in these intervals may be conservative. The effective bottleneck time (the effective number of generations over which

recombination has taken place, Zhang et al, 2004) for chromosome 1, which spans 2.865 Morgans (Kong et al, 2004), is 4212.2 / 2.865 = 1470 (~36,750 years at 25 years per generation). The large number of accumulated meioses provides the dramatically higher resolution of the LDU map relative to the linkage map, which is critical for disease gene mapping.

Graphically (Figure 3) there is little differentiation between the LDU map and the much lower resolution linkage map (Kong et al, 2004). When map lengths are compared in 2 Mb sliding windows (Figure 4), LDU and linkage maps show similar broad recombination intense regions in which there is a high density of much narrower recombination hot-spots.
Map assembly in overlapping segments is ideally suited to GRID computing (for example using the Condor program, http://www.cs.wisc.edu/condor/) and might be profitably achieved through a WWW based tool. We are currently considering this possibility.

References.

Collins, A. & Morton, N. E. (1998). Mapping a disease locus by allelic association. *Proc Natl Acad Sci U S A* 95, 1741-1745.

Collins, A., Lonjou, C. & Morton, N. E. (1999). Genetic epidemiology of single-nucleotide polymorphisms. *Proc Natl Acad Sci U S A* 96, 15173-15177.

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

Dawson, E., Abecasis, G. R., Bumpstead, S., Chen, Y., Hunt, S. Beare, D. M., Pabial., J.,
Dibling, T., Tinsley. E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lohmussaar, E., Zernant, J. , Tonisson, N., Remm, M., Magi, R., Puurand, T., Vilo, J., Kurg, A., Zernant, JRice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D. R., Cardon, L. R., & Dunham, I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* 418, 544-548. Epub 2002 Jul 2010.

Gibson J, Tapper W, Zhang W, Morton N and Collins A (2005). Cosmopolitan linkage disequilibrium maps. Human Genomics, in press.

Gomes, I., Collins, A., Lonjou, C., Thomas, N. S., Wilkinson, J., Watson, M. & Morton, N. (1999). Hardy-Weinberg quality control. *Ann Hum Genet* 63, 535-538.

Hill, W. G. (1974). Estimation of linkage disequilibrium in randomly mating populations. *Heredity* 33, 229-239.

International HapMap Consortium, The International HapMap Project. *Nature* 426, 789-796 (2003).

Ke, X., Hunt, S., Tapper, W., Lawrence, R., Stavrides, G., Ghori, J ., Whittaker, P ., Collins, A ., Morris, A. P ., Bentley, D ., Cardon, L. R .& Deloukas, P, (2004). The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum Mol Genet* 13, 577-588. Epub 2004 Jan 2020.

Kong, X., Murphy, K., Raj, T., He, C., White, P. S. & Matise, T. C. (2004). A combined linkage-physical map of the human genome. *Am J Hum Genet* 75, 1143-1148. Epub 2004 Oct 1114.

Lonjou, C., Zhang, W., Collins, A., Tapper, W. J., Elahi, E., Maniatis, N. & Morton, N. E. (2003). Linkage disequilibrium in human populations. *Proc Natl Acad Sci U S A* 100, 6069-6074. Epub 2003 Apr 6029.

Malecot, G. Les Mathematiques de l'Heredite (Maison et Cie, Paris, 1948).

Maniatis, N., Collins, A., Xu, C. F., McCarthy, L. C., Hewett, D. R., Tapper, W., Ennis, S., Ke, X. & Morton, N. E. (2002). The first linkage disequilibrium (LD) maps:

delineation of hot and cold blocks by diplotype analysis. *Proc Natl Acad Sci U S A* 99, 2228-2233. Epub 2002 Feb 2212.

Maniatis, N., Collins, A., Gibson, J., Zhang, W., Tapper, W. & Morton, N. E. (2004). Positional cloning by linkage disequilibrium. *Am J Hum Genet* 74, 846-855. Epub 2004 Mar 2026.

Maniatis, N., Morton, N. E., Gibson, J., Xu, C. F., Hosking, L. K. & Collins, A. (2005). The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. *Hum Mol Genet* 14, 145-153. Epub 2004 Nov 2017.

Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P. Y. & Collins, A. (2001). The optimal measure of allelic association. *Proc Natl Acad Sci U S A* 98, 5217-5221. Epub 2001 Apr 5217.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1994). Numerical recipes in C. Cambridge University Press, Cambridge.

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, Ackerman HC, Campbell SJ, Altshuler D, Cooper R, Kwiatkowski D, Ward R, Lander ES. (2002). Detecting recent positive selection in the human genome from haplotype structure.Nature. 2002 Oct 24;419(6909):832-7. Epub 2002 Oct 09.

Tapper, W. J., Maniatis, N., Morton, N. E. & Collins, A. (2003). A metric linkage disequilibrium map of a human chromosome. *Ann Hum Genet* 67, 487-494.

Zhang, W., Collins, A., Maniatis, N., Tapper, W. & Morton, N. E. (2002). Properties of linkage disequilibrium (LD) maps. *Proc Natl Acad Sci U S A* 99, 17004-17007. Epub 12002 Dec 17016.

Zhang, W., Collins, A., Gibson, J., Tapper, W. J., Hunt, S., Deloukas, P., Bentley, D. R. & Morton, N. E. (2004). Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc Natl Acad Sci U S A* 101, 18075-18080. Epub 12004 Dec 18016.

Table 1 Haplotype frequencies (a, b, c and d) for a pair of SNPs

		SNP_2 alleles		
		1	2	
SNP_1 alleles	1	a	b	Q
	2	c	d	1-Q
		R	1-R	

The table is ordered such that:
Q, 1-Q are allele frequencies at SNP_1, where $Q < (1-Q, R, 1-R)$ and R, 1-R are allele frequencies at SNP_2 and $ad > bc$.
$D = ad-bc$
$\rho = D/Q(1-R)$
$K_\rho = mQ\ (1-R)/R(1-Q)$, where m is the sample size for the pair of SNPs.
$\chi^2 = \rho^2 K_\rho$

Table 2. Maps of chromosome 22 constructed using different numbers of segments.

Number of segments	Loci per segment	Map length LDU	Error variance	Computation time, minutes
1	13959	1017	0.841	2471
2	6980	1017	0.842	2348
6	2327	1022	0.842	859
14	997	1024	0.843	760
20	698	1037	0.847	733
40	349	1040	0.853	509
60	233	1055	0.851	409
100	140	1054	0.870	408
200	70	1089	0.886	177

Figure 1.

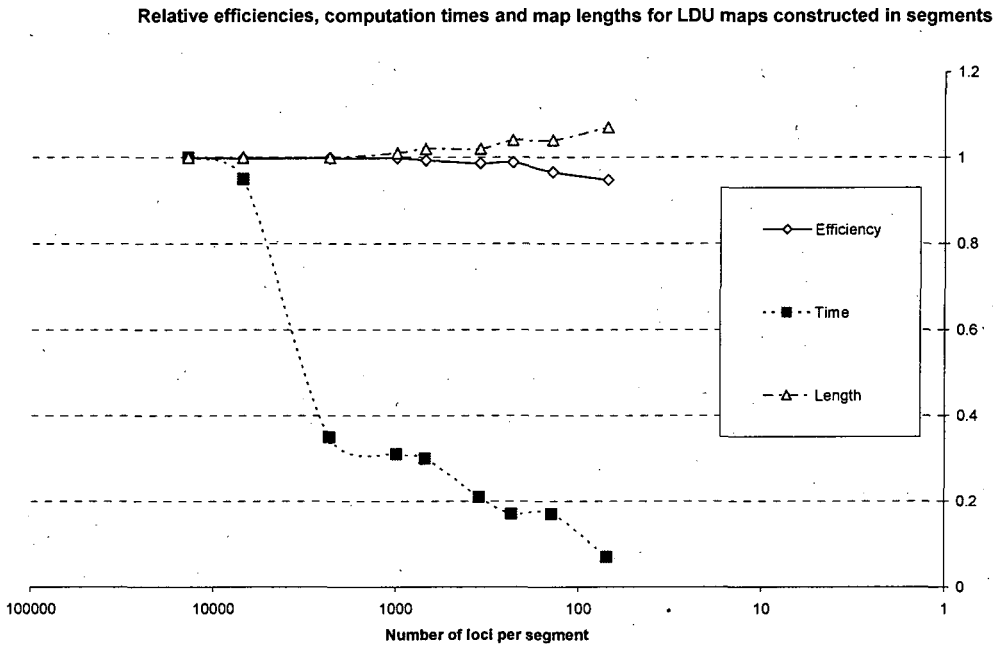


Figure2