UNIVERSITY OF
Southampton

FACULTY OF MEDICINE, HEALTH AND LIFE SCIENCES
School of Medicine

Defining linkage disequilibrium patterns and tracts of extended homozygosity to compare populations and search for disease genes.

by
Jane Gibson

Thesis for the degree of Doctor of Philosophy
May 2008

Abstract

This project aimed to define linkage disequilibrium (LD) patterns and tracts of extended homozygosity in order to compare populations and search for disease genes. SNP genotype data were analysed on a Unix platform, using the programs LDMAP+, for linkage disequilibrium unit (LDU) map creation, and CHROMSCAN-cluster, for association mapping, as well as software written as part of this project, in the C programming language, for determining tracts of homozygosity and for autozygosity mapping. LDU maps were compared over populations showing similarity in LD structure. A cosmopolitan LDU map which represents the LD patterns of different population samples was produced and able to recover 91-95% of the information in the original population specific data. Genome-wide LDU maps were created, compared across populations, and compared with the linkage map to estimate effective bottleneck time (t), the time since the last major bottleneck for each population.

This project also discovered an unanticipated amount of homozygosity in the outbred individuals genotyped in the HapMap project. Large homozygous tracts are expected in inbred individuals and this analysis was able to determine 3 individuals with high levels of homozygosity consistent with recent inbreeding. The relationship between tracts of homozygosity and LD was investigated, using the LDU maps, showing that long tracts of homozygosity are more likely to occur in regions of high LD where the underlying haplotypes are of limited diversity. The relationship shown between LD and homozygosity enabled a more powerful approach to autozygosity mapping of a recessive locus in a consanguineous pedigree affected by Congenital Nephrotic Syndrome. High density SNP genotyping of affected individuals pinpointed regions of homozygosity which segregate with the disease, with the advantage of using few individuals and without the need for statistical inference from linkage. The regions determined were then prioritised on the basis of LDU length, therefore adding weight to regions of true autozygosity over regions of homozygosity associated with high LD. This analysis successfully determined a region containing a strong candidate gene (PLCE1) which has subsequently been shown to be mutated in the affected individuals.

Extending the search for disease genes to complex disease studies, a genome-wide association scan was carried out, using real case-control data with an undisclosed disease and utilising the LDU maps. A combination of the results from the multi-SNP approach of CHROMSCAN-cluster and single SNP results allowed selection of regions for follow up in a multi-stage analysis.

Table of contents

CHAPTER 4 – EXTENDED TRACTS OF HOMOZYGOSITY IN OUTBRED POPULATIONS 62

CHAPTER 5 - AUTOZYGOSITY MAPPING TO SEARCH FOR A CANDIDATE REGION OR GENE FOR CONGENITAL

List of tables, figures and appendices

Declaration of authorship

I, **Jane Gibson**, declare that the thesis entitled,

Defining linkage disequilibrium patterns and tracts of extended homozygosity to compare populations and search for disease genes.

and the work presented in it are my own. 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 as:

1. Tapper W, **Gibson J**, Morton NE, Collins A.
A comparison of methods to detect recombination hotspots.
Human Heredity. 2008 In press.

2. **Gibson J**, Tapper W, Cox D, Zhang W, Pfeufer A, Gieger C, Wichmann HE, Kääb S, Collins AR, Meitinger T, Morton N.
A multimetric approach to analysis of genome-wide association by single markers and composite likelihood.
Proc Natl Acad Sci U S A. 2008 Feb 19;105(7):2592-7. Epub 2008 Feb 11.

3. **Gibson J**, Morton NE, Collins A.
Extended tracts of homozygosity in outbred human populations.
Hum Mol Genet. 2006 Mar 1;15(5):789-95. Epub 2006 Jan 25.

4. Tapper W, Collins A, **Gibson J**, Maniatis N, Ennis S, Morton NE.
A map of the human genome in linkage disequilibrium units.
Proc Natl Acad Sci U S A. 2005 Aug 16;102(33):11835-9. Epub 2005 Aug 9.

5. **Gibson J**, Tapper W, Zhang W, Morton N, Collins A.
Cosmopolitan linkage disequilibrium maps.
Hum Genomics. 2005 Mar;2(1):20-7.

6. Zhang W, Collins A, **Gibson J**, Tapper WJ, Hunt S, Deloukas P, Bentley DR, Morton NE.
Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps.
Proc Natl Acad Sci U S A. 2004 Dec 28;101(52):18075-80. Epub 2004 Dec 16.

Signed: ....
Date: 13/05/08

Acknowledgments

I would like to thank past and present members of the Genetic Epidemiology and Bioinformatics Group, particularly Dr William Tapper, Dr Nikolas Maniatis and Dr Sarah Ennis and my supervisors Prof. A. Collins and Prof. N.E. Morton.

I would also like to thank the collaborators and data providers; Wessex Clinical Genetics Service and Wessex Regional Genetics Laboratory for the Congenital Nephrotic Syndrome collaboration, Perlegen Sciences for providing the genome-wide anonymous case-control dataset, and publicly available data from the Wellcome Trust Centre for Human Genetics used to create a Cosmopolitan LDU map of a region on chromosome 20, and the International HapMap Project SNP data used for analysing homozygosity and creating the genome-wide LDU maps used throughout this project.

Definitions and abbreviations

MAF	Minor allele frequency
HWE	Hardy-Weinberg equilibrium
LDMAP (+)	program for creating LDU maps
CHROMSCAN (-cluster)	program for association mapping
LDU	Linkage Disequilibrium units
cM	centiMorgan
bp	basepairs
Kb	Kilo-basepairs
Mb	Mega-basepairs
SNP	Single nucleotide polymorphism
RFLP	restriction fragment length polymorphism
STRP	short tandem repeat polymorphism
CNV	Copy number variation
IBS	Identity by state
IBD	Identity by decent
GWA(s)	Genome-wide association
CEPH	Centre d'Etude du Polymorphisme Humain
CEU	CEPH Utah residents with ancestry from northern and western Europe
CHB	Han Chinese in Beijing
JPT	Japanese in Tokyo
YRI	Yoruba in Ibadan

Chapter 1 – Introduction and aims

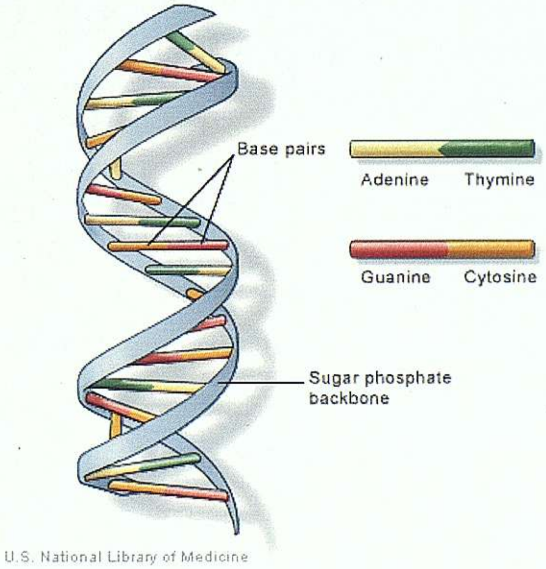
1.1 Introduction

There are three maps currently used to describe the human genome. The physical map, measured as sequenced base-pairs; the linkage map, genetic distance measured in families across a generation; and the linkage disequilibrium map, genetic distance measured in unrelated individuals across the many generations since the founding of the population.

1.1.1 The sequence map

The physical map reflects the structure of Deoxyribonucleic acid (DNA), the molecule responsible for the inheritance of genetic traits. It consists of three parts; a sugar, a phosphate, and a base, and forms a double helix structure (figure 1.1).

Figure 1.1 The double helix structure of DNA showing the base-pairs.



(U.S.National Library of Medicine 2006)

There are 4 types of base Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). Long strands of DNA hold information in the sequence of these base-pairs. There are 46 distinct pieces of DNA in each of our cells. These are the chromosome pairs 1 to 22 and the sex chromosomes X and Y, with one set inherited from each parent. The structure of DNA was first described by Watson and Crick (Watson and Crick 1953). The first base-pair sequencing method was developed by F. Sanger in the 1970's (Sanger and Coulson 1975; Sanger, Nicklen, and Coulson 1977). Several methods were proposed and there has been development of improved high throughput methods making sequencing of the entire human genome possible.

The Human Genome Project (HGP) was formally launched in October 1990 by the U.S. Department of Energy and the National Institutes of Health's National Human Genome Research Institute (NHGRI). The publicly funded Human Genome Project announced the release of a draft sequence of the Human Genome in 2001 (Lander et al. 2001), at the same time as the private venture by Celera (Venter et al. 2001); the 'finished' sequence was announced in April 2003 (Collins, Morgan, and Patrinos 2003; Collins et al. 2003). Since that time, regular updates to the map have reduced gaps in areas which are hard to sequence due to heterochromatin or repeat sequences. The availability of the base pair (bp) sequence of the human genome has provided a wealth of information, which is publicly accessible via several databases such as, the National Centre for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov/) and UCSC genome browser (http://genome.ucsc.edu/). As well as the sequence information the various databases provide further annotation of the sequence such as gene locations and genetic marker positions. The Human Genome Project provides a standard reference sequence (a composite of several individuals), it is estimated that any 2 genomes are 99.9% identical, however every individual genome is unique and 0.1% difference amounts to millions of variations in the 3.2 billion base-pairs of sequence. These differences account for the heritable variation among individuals including susceptibility to disease (Kruglyak and Nickerson 2001).

1.1.2 Genetic variation

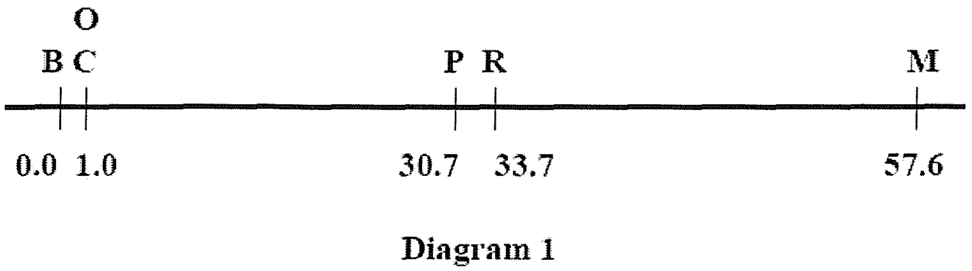
Human variation can be observed in many forms, from the ABO blood groups and serum protein variations to DNA sequence polymorphisms. Examples of DNA polymorphisms are restriction fragment length polymorphism (RFLPs), minisatellites or variable number tandem repeats (VNTRs), microsatellites or short tandem repeat polymorphisms (STRPs) which are the basis of the most linkage maps, small insertion/deletions (indels) (Kidd et al. 2004) and Single Nucleotide Polymorphisms (SNPs), a change in a single base pair at a given location. SNPs are the most frequent type of polymorphism in the human genome; they are generally bi-allelic and have low mutation rates. Large collections of SNPs have now been established and described in public databases such as dbSNP (NCBI 2005) and The SNP Consortium (The SNP Consortium 2005). Approximately 10 million common SNPs are estimated to exist (Botstein and Risch 2003). Non-synonymous SNPs cause a change in the amino acid coded by a gene. They are known to cause disease in many monogenic disorders and are a priority when looking for disease causing mutations. However synonymous SNPs, which do not change the amino acid coded, and SNPs in non-coding regions, perhaps promoter or regulatory regions, may affect the regulation and splicing of genes and also lead to disease. Most SNPs are located in non-coding regions of the genome and although many may not cause disease themselves they are still very useful as genetic markers since they may be associated and inherited with causal polymorphisms, and also may be used as markers for population genetics and evolutionary studies (Celedon 2005).

High throughput genotyping is now standard technology and many investigators use the commercially available genotyping platforms. Affymetrix gene chips use randomly chosen SNPs depending on the restriction enzyme used to cleave the DNA. Illumina have used HapMap data to be more selective in choosing SNPs that offer the best coverage based on LD patterns. There are ongoing modifications to increase quality, call rate and reduce bias. For example, the Dynamic Modelling (DM) algorithm was used by Affymetrix to automatically call genotypes from experimental results. However, it was shown that this particular algorithm has a bias in that missed calls were more often heterozygotes than homozygotes (Rabbee and Speed 2006; Kuruvilla et al. 2006). A new algorithm called BRLMM was introduced to overcome this bias and is the standard algorithm currently used by Affymetrix. There are several quality control procedures carried out on genotypic data such as removing SNPs or individuals with a certain percentage of missing calls, duplicate typing of a percentage of SNPs or individuals to ensure concordance and Hardy-Weinberg equilibrium tests. Overall for large samples obvious genotyping errors can be avoided. Some errors may remain but with error rates much less than 1% they are likely to have negligible effects on most analyses. However it is wise to individually check the cluster plots used for genotype calling and the quality scores associated with genotypes of interest to check the reliability of the information.

1.1.3 The Linkage map

Long before determining a physical map was considered possible or DNA was known to be the inherited coding structure, there was a genetic map. The very first genetic map was constructed in 1913 by Alfred H. Sturtevant an undergraduate student of Thomas H. Morgan at Columbia University. They had been working on 6 sex linked 'factors' in *Drosophila.* The factors were given a linear order based on the length and strength of their association determined by the number of meiotic crossovers between factors, the phenomena of interference was also noted, where one crossover inhibits another close by.

Figure 1.2 The first diagram of a genetic map.

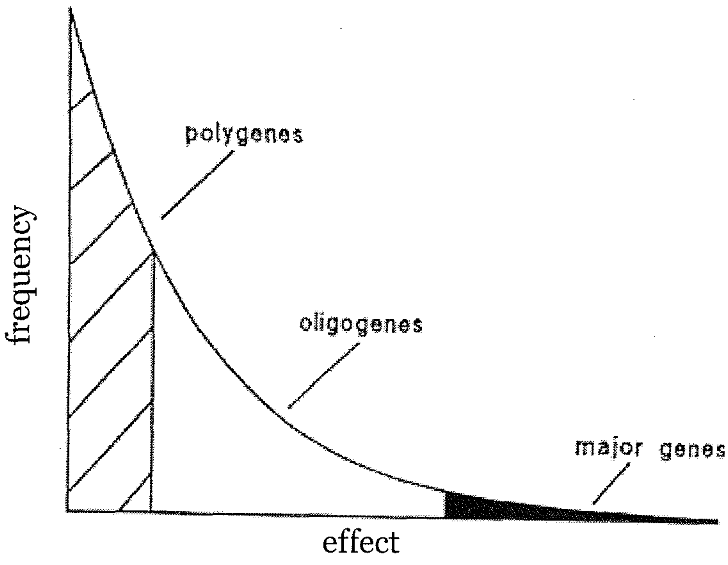


(Sturtevant 1913)

This diagram of factors was the forerunner to the linkage map which measures the recombination rate in centiMorgan (cM) units. The first comprehensive linkage map was produced by (Dib et al. 1996) and the most comprehensive linkage map to date is based on the deCODE pedigrees from Iceland (Kong et al. 2004).

Linkage maps have been valuable in identifying disease-causing ‘major’ genes. In linkage studies chromosomal segments which co-segregate with the disease in families are identified, and predicted to contain the causal variation. It is a powerful method but generally only narrows a region to a few Megabases (Mb), which may include many genes or variants. The availability of the gene annotated base-pair sequence has allowed candidate gene analysis as a method of narrowing down the region of interest (Carlson et al. 2004), where genes with a known function that may biologically affect the phenotype are studied further. Linkage studies have been less successful for complex diseases which are caused by polygenes, variations of small effect in multiple genes.

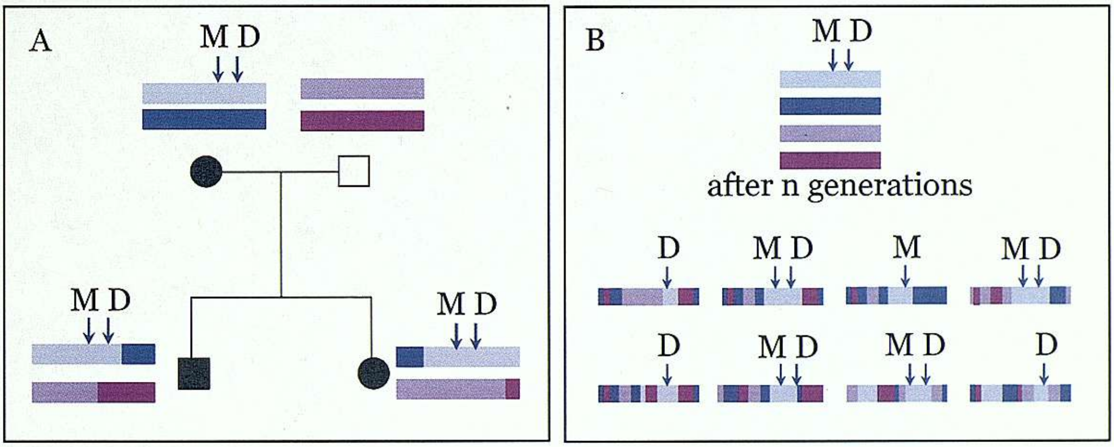
Figure 1.3 Genes and their effects.



1.1.4 Linkage Disequilibrium

Attention has therefore shifted towards linkage disequilibrium (LD) which describes the tendency of linked alleles to be inherited together more often than would be expected under random segregation. The potentially higher resolution of disease mapping using linkage disequilibrium makes it an attractive option. LD is created when a small number of founding individuals and therefore small numbers of haplotypes form a new population corresponding to a bottleneck. The major influence on LD is recombination and the amount of time recombination has had to break up LD since the last major bottleneck, to a lesser extent, mutation, genetic drift and selection also have an effect (Tapper et al. 2003). Linkage mapping tracks a disease (D) and genetic markers (M) through 1-2 generations in a family, limiting the linked region of interest by meiotic recombination events. Association mapping using LD, utilises a similar idea in a population sample, to determine a region of interest by association of the disease with genetic marker alleles, narrowed by historical recombination events determined by LD patterns (figure 1.4).

Figure 1.4 Linkage mapping (A) versus association mapping by LD (B).



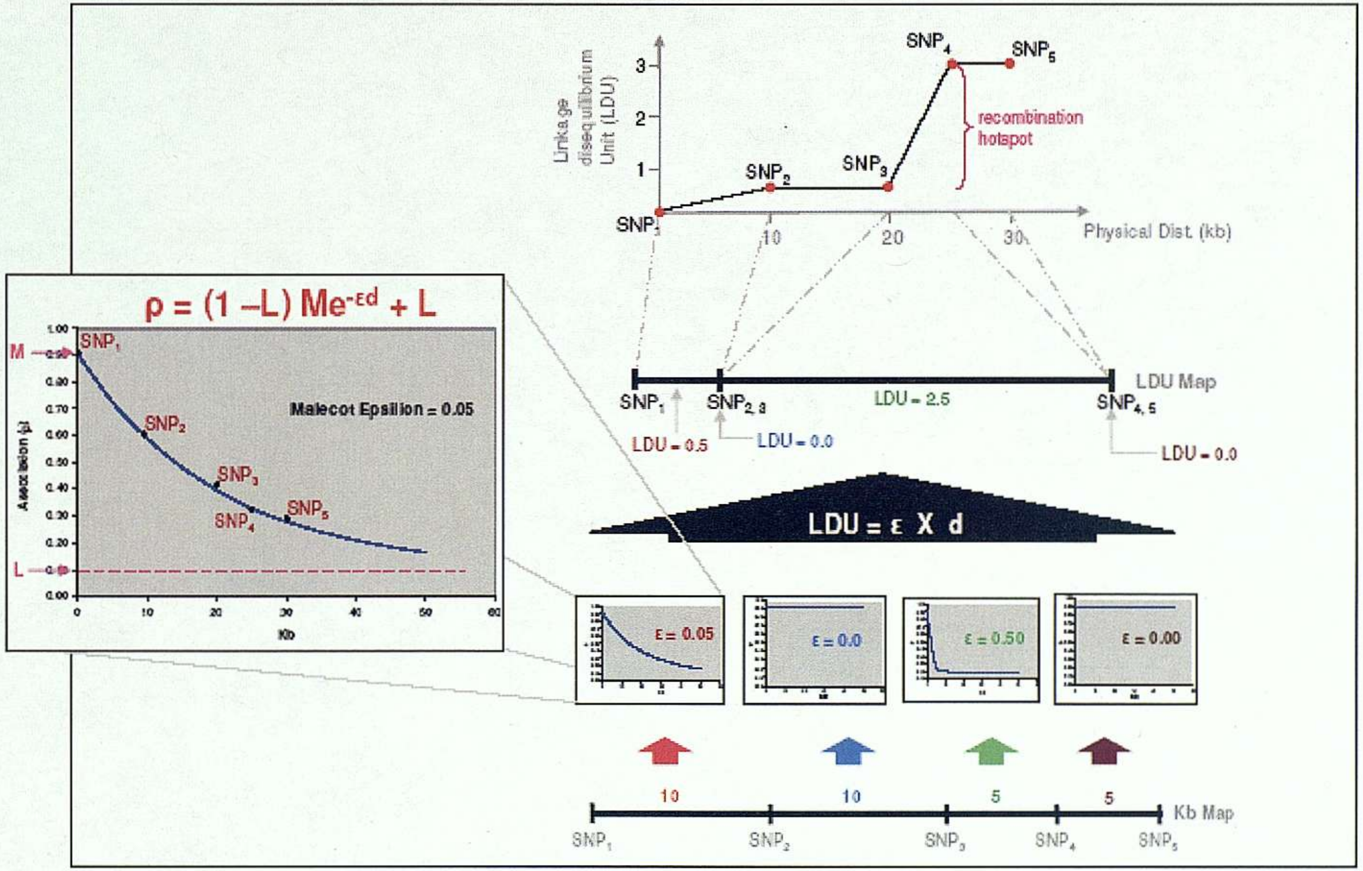
1.1.5 Measures of LD

There are several commonly used measures of LD. D is a simple measure of disequilibrium and is calculated as D= fAB- (fA x fB), with fAB as the observed frequency of the AB haplotype and fA x fB being the expected frequency based on the individual frequencies of the two alleles A and B. D is not of great use for comparing the strength of LD since it has a maximum value (Dmax) which is highly dependent on allele frequency. To account for this Lewontin provides an extension, D' calculated as D'=D/Dmax (Lewontin 1964). With an arbitrary assignment of alleles this value can be positive or negative and is therefore presented as the absolute value |D'|. |D'| =1 shows complete LD but lower values have a less clear interpretation, since D' is dependent on sample size and is inflated in small samples. Another commonly used measure is r^2, which is equal to D^2 divided by the product of the allele frequencies at the two loci. This is more stable to sample size but is again less reliable for low allele frequencies. r^2 is used to determine power in association mapping and predict the sample sizes required, whereas as D' is a measure of LD itself. These pairwise measures, can be plotted as 'heatmaps' using software such as GOLD (Graphical Overview of Linkage Disequilibrium) (Abecasis and Cookson 2000) or Haploview (Barrett et al. 2005). These programs are useful for providing a graphical visualization of pairwise measures between many SNPs but do not allow the creation of a whole linear additive map.

1.1.6 LDU maps

The LDU map uses the association metric rho (ρ) which is a probability and therefore ranges 0-1. Rho is equivalent to |D'| for pairs of SNPs, but not for marker/disease association and is the most robust metric to allele frequency but is still sensitive to sample size (Collins and Morton 1998). This metric is calculated using pairwise data and also modelled by the Malecot equation in the LDMAP program. The theory for the first map of LD patterns was developed by Maniatis *et al.*, it is a map with additive distances in LD units (LDU) analogous to the linkage map in cM (Maniatis et al. 2002). The LDU map is based on the Malecot model which was originally designed for isolation by distance but has been adapted to model the decline of LD over distance (Collins and Morton 1998). The model is, $\hat{\rho} = (1-L)Me^{-\varepsilon d} + L$, and the 3 main parameters are M, L, and ε. M is the association at 0 distance and has an evolutionary interpretation as it reflects the association at the last major bottleneck. L is the association at large distance and reflects background LD levels and the effect of sample size, which is known to affect rho. Epsilon (ε) measures the decline of LD over distance; a large ε reflects a rapid decline of LD whereas a small ε reflects a more gradual decline. LDU is calculated as the product of epsilon and distance in Kb.

Figure 1.5 The method to create an LDU map.



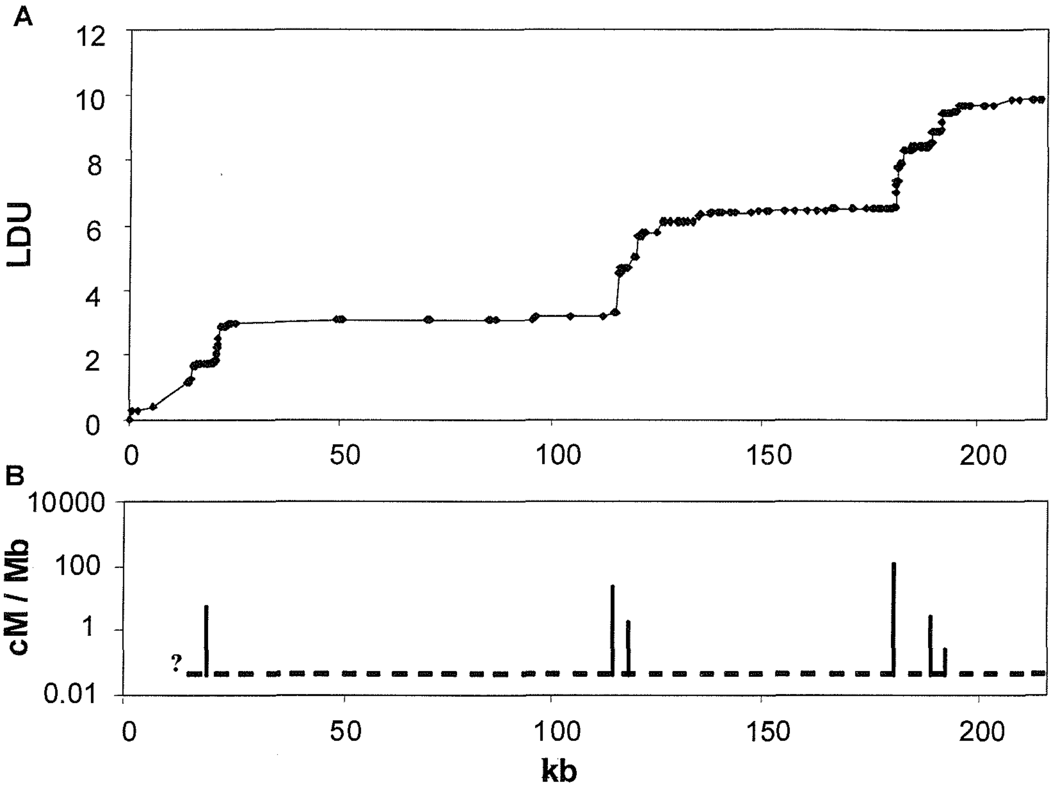
This diagram shows fitting of the Malecot model and calculating LDU from the estimated ε for each pair of SNPs to create an additive LDU map showing plateaus of high LD/low recombination and steps of low LD/high recombination (Collins, Lau, and De La Vega 2004).

For all pairs of SNPs, ρ is calculated using the observed pairwise data and estimated using the model. The estimated and observed values are used to calculate the composite -2 log likelihood = $\sum K(\rho - \hat{\rho})^2$, this is minimised so that the difference between the two rho values is closest to zero. Composite likelihood is a combination of likelihoods, usually of small subsets of data, this reduces computational complexity and allows large datasets and complex models to be handled when a standard likelihood is not feasible. A drawback to composite likelihood is that the summation is over non-independent elements (Zhang et al. 2002). Of the three parameters in the Malecot model, L is not estimated but a 'predicted L' is calculated from the data, as 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 (the weighted mean deviation for a normal distribution). Since L is the asymptote, it is not observed in a small region, and the block structure revealed by a high density of SNPs distorts a direct estimate of L thus predicted L has been shown to give more reliable results than estimating the L parameter (Zhang et al. 2002). Epsilon is iterated for each SNP interval, it is incrementally changed and the magnitude and direction of the change is determined using the Newton-Raphson algorithm for finding the roots of non-linear equations. The -2 log composite likelihood is minimised for each interval to provide the best model fit to the observed data. The parameter M is assumed constant across the whole LDU map but is iterated periodically to minimise the -2 log likelihood. For the creation of an LDU map the epsilon value for each interval is multiplied by the Kb distance to give a value in LDU, beginning with 0 LDU at the p-ter of the map the values are cumulative to give an additive map. A further measure called the swept radius, calculated as $1/\varepsilon$, shows the average extent of 'useful' LD on the kilobase scale. The LDU map can be plotted on a graph opposite the kb map revealing plateaus and steps. The plateaus show a low LDU/Mb ratio, corresponding to a region of high LD or low recombination. The steps show a high LDU/Mb ratio, corresponding to a region of low LD or high recombination (Maniatis et al. 2002).

1.1.7 Properties of the LDU map

Since recombination is the main force behind LD structure, information about recombination can be reliably obtained from the LDU map. This was shown by the remarkable correspondence between the results from Jeffreys *et al.* (Jeffreys, Kauppi, and Neumann 2001), which was a direct measure of meiotic recombination carried out by sperm typing, and the LDU maps of the corresponding region (Zhang et al. 2002) (figure 1.6).

Figure 1.6 A 216-kb segment of class II region of MHC.



An LDU map (A) and the corresponding region analysed by sperm typing (B), showing the agreement between steps in the LDU map and the localisation of the recombination hotspots shown as vertical bars in B (Zhang et al. 2002).

Since the creation of the first LDU maps of small regions, LDU maps have been created of whole chromosomes. Tapper *et al.* (2003) created LDU maps for chromosome 22 for 2 European samples, allowing the LDU map to be compared with the linkage map over a whole chromosome. There was a good correspondence between the two genetic maps despite the comparatively low resolution of the linkage map. The LDU map also allowed the LD patterns in different chromosomal regions to be compared showing the variation in LDU (Tapper et al. 2003).

An LDU map can be used to facilitate positional cloning or association mapping (Maniatis et al. 2004), enhance the resolution of the linkage map, compare populations (Lonjou et al. 2003), and detect selective sweeps and other evolutionary events. The linkage map appears, on limited evidence, not to vary between populations. The LDU map, however, varies with different population histories, principally the 'age' of the population, the time since the last major bottleneck. The most apparent difference is found between African and non-African populations, presumably reflecting the 'Out of Africa' bottleneck (Lonjou et al. 2003; Reich et al. 2001). A small number of individuals representing a small sample of the haplotypes present in Africa at the time founded a new Eurasian population, resetting LD in the new population. In terms of the Malecot model the parameter M would approach 1 and ε would be small reflecting the high level of LD (Lonjou et al. 2003). Although the overall length of the LDU map is longer in African populations, reflecting more recombination, the broad patterns, in terms of plateaus and steps, are aligned. This similarity can be explained by the co-localisation of recombination hotspots in all populations. The intensity of recombination shown in the map in these areas varies due to the differences in time, with more intense hotspots (longer steps) in the African populations causing the increased map lengths (De La Vega et al. 2005). This high correlation in LD patterns could allow a cosmopolitan LDU map to be made incorporating multiple populations; this standard map could then be scaled to represent the LD structure in any single population (Lonjou et al. 2003).

1.1.8 The HapMap project

The idea that the genome can be divided into regions or blocks that have low haplotype diversity (Daly et al. 2001) led to the suggestion that some markers could be used as surrogates for others with which they are in high or complete LD; fewer SNPs would reduce the cost and workload of association studies. A demand for a better understanding of the LD structure of the human genome in order to choose which SNPs to type to get the maximum benefit, prompted the International HapMap Project. There have been many methods and software programs developed to choose these 'haplotype tagging' SNPs, such as Tagger (de Bakker et al. 2005) and HapBlock (Zhang et al. 2005). The International HapMap project aimed to catalogue human variation with the objective of helping investigators choose tagging SNPs.

The International HapMap project began in 2002 as a collaboration between scientists and funding agencies from Japan, the United Kingdom, Canada, China, Nigeria, and the United States. It aimed to genotype over a million SNPs in 4 populations, CEPH Utah residents with Northern and Western European ancestors, Japanese from Tokyo, Han Chinese from Beijing, and Yoruba from Ibadan, Nigeria. In February 2005 Phase I of the project was completed with 1.2 million SNPs genotyped in the 4 populations. After some quality control measures and error fixing the analysis group carried out their initial analysis of this Phase I data (Altshuler et al. 2005). The HapMap Consortium continued genotyping, extending the project to Phase II, with the new goal of approximately 4 million SNPs resulting in, on average, 1 SNP per 600bp throughout the genome. The data are periodically released into the public domain via the website http://www.hapmap.org (International Hapmap Group 2005; National Institutes of Health and National Human Genome Research institute. 2002; National Institutes of Health and National Human Genome Research institute. 2005). This resource provides large amounts of publicly available genotype data on 269 (270 in Phase II) individuals.

The initial analysis of the HapMap phase I data by the HapMap analysis group gave a description of the data collection, genotyping methods and quality control procedures. They also carried out some analysis of the data and discussed the properties of LD in the genome. Many phenomena previously described in smaller samples and smaller regions were confirmed over the whole genome, such as the major determinant of variation in LD being recombination, the block-like structure of LD, and the presence of recombination hotspots. The extent of LD was shown to vary across populations and be less extensive in the YRI sample, more extensive LD was shown on the X chromosome with more long-range haplotypes. Previously reported observations of increased LD towards the telomeres, and reduced LD towards the centromeres, and also a correlation between LD level and chromosome length were also confirmed. The authors describe the projected use of the HapMap resource for directing association studies and selecting tagSNPs, though acknowledge that more data on more populations and data including rare SNPs (Phase I selectively genotyped common SNPs) is needed to fully test the portability of tagSNPs across populations. The paper also highlights the use of the HapMap data for studying natural selection, empirical analyses were carried out determining the most extreme candidate regions for selection using a long-range haplotype method and differences in allele frequencies between populations. However, it is accepted that different types of selection leave different signatures in current genetic variation data and many methods use different approaches, also the SNP ascertainment bias because of a focus on common variation complicates any analysis, and careful interpretation of results are required (Altshuler et al. 2005).

1.1.9. Coalescent Theory

Coalescent Theory models the underlying genealogy which led to current human variation. It is modelled backwards in time as a tree, where 2 different lineages (haplotypes) coalesce to a single ancestor at each level of the tree, going back to a single common ancestor, the most recent common ancestor (MRCA) of the current population with the mutation events which lead to the changes being superimposed on the branches of the tree. To add recombination to the coalescent model, lineages which join (coalesce) when they have the same ancestor can also split (bifurcate) when the same segment has 2 ancestors due to a recombination event. As well as mutation and recombination, other genetic processes such as population size fluctuations, genetic drift and selection can also be incorporated into this kind of model. The model then contains several parameters of interest as well as a genealogy. Simulations are carried out to define the parameters for the model that produces simulated data matching the observed genetic variation data. However, the more complex models have drawbacks due to the computational complexities involved and models may fail to produce a result in reasonable time. Two methods for attempting to resolve this problem are the 'first approximation' which removes the need for an exact match between the simulated and observed data, and the 'second approximation' in which the model itself is simplified (Marjoram and Tavare 2006). Coalescent-based methods are complex and there are many variations, using these models for determining recombination rates and for association mapping problems is a current focus. Any statistical analysis of genome-wide data will face problems due the amount of data, the complexity of the data, the amount of autocorrelation and the rate of false positives unique to such data. 'Historical recombination maps' produced by the LDHAT program are based on a coalescence method (McVean et al. 2004) and have been shown to correlate well with linkage maps over the same regions and recombination rates directly measured from sperm-typing data (Jeffreys, Kauppi, and Neumann 2001). Instead of modelling recombination over the whole genome, the method simulates genealogies while moving along a sequence, dividing the data into subsets and combining likelihood calculations, therefore this is a composite likelihood method. A similar coalescent-based method has been developed to identify recombination hotspots using a program called LDHOT based on the recombination maps produced by LDHAT (Myers et al. 2005).

1.1.10 Extended homozygosity and autozygosity mapping

The data provided by the HapMap project can be used to create LDU maps of the whole genome and the LDU patterns can be compared in different genomic regions and across populations to give a better understanding of LD. LDU maps of these data may also be used to study the demography, and evolutionary events that have shaped current human populations. The HapMap data allows an insight into normal variation and normal levels of homozygosity. When a biallelic marker has identical alleles it is homozygous; this can indicate identity by state (IBS) or identity by decent (IBD). In IBD homozygosity is likely to extend to neighbouring SNPs which are inherited together on the same chromosomal segment, creating a region of homozygosity. Broman and Weber (1999) found long tracts of homozygosity were more common than expected in a CEPH sample using short tandem repeat polymorphisms (STRPs). They found that it was not unusual to find individuals from outbred populations to have long homozygous tracts of >10cM. They examined the role of possible typing error, back mutation of STRPs, gene conversion events and the limitations imposed by locally low marker density in determining the limits of homozygous segments. Although relationships between 'unrelated' individuals in some pedigrees were determined, there remained a degree of autozygosity approaching or exceeding that expected in the progeny of a first cousin mating where relationships were not detected (Broman and Weber 1999). The HapMap data provide an ideal opportunity to look at this phenomenon in high quality and dense SNP data in 4 outbred populations.

Consanguinity is known to be associated with an increased risk of rare recessive disease. The low haplotype diversity means that a rare mutation is more likely to be seen in its homozygous form in a family with some consanguinity. Knowledge of the levels of extended homozygosity in healthy outbred individuals would provide useful information for the mapping of autosomal recessive genes with homozygosity mapping (Lander and Botstein 1987). This method is potentially faster and easier than conventional linkage studies with fewer individuals required. Some groups are beginning to use high throughput genotyping technology such as the Affymetrix chips as a relatively cheap method of genotyping a small number of individuals within an inbred family to detect regions of homozygosity associated with a recessive phenotype on a genome-wide scale (Chiang et al. 2006; Weber et al. 2005). Homozygosity and LD are both determined by the underlying haplotype structure, this may allow homozygosity associated with LD to be distinguished from autozygosity. Prioritising regions of homozygosity with respect to LD structure, using LDU maps, has the potential to increase power in this type of study.

1.1.11 Genome-wide association analyses

Many diseases have an underlying genetic basis, some are caused by 'major' genes which are rare but have a large effect, such as the ΔF508 variant of the CFTR gene in Cystic Fibrosis. To determine a gene with large effect, genetic linkage analysis is very effective. The alleles of polymorphic genetic markers are determined in several generations within affected families and the 'linked' region shared by affected individuals determines the location of the disease causing variant. The regions detectable by this method are large, generally on a scale of several centiMorgans (cM) corresponding to several megabases (Mb) of physical sequence; and linkage analysis has low power to detect common variants with modest levels of disease risk, such as those predicted to give the genetic contribution to many complex diseases. Association analyses are expected to be more powerful in these cases because for modest risk alleles the pattern of allele sharing among individuals within a family is less striking than the pattern of allele sharing between unrelated individuals (Carlson et al. 2004).

The first genome-wide linkage study of a complex human disease was carried out in 1994 for Type 1 Diabetes (Davies et al. 1994). It showed the importance of the HLA region on chromosome 6. Although many genome-wide linkage studies have since been undertaken for common diseases, the disappointing reproducibility of the results and low power mean this has not been as successful an approach as hoped (McKinney and Merriman 2007). More powerful association analyses can be carried out on large cohorts of unrelated cases and controls, to look for differences in the frequencies of alleles between these groups. Genome-wide association studies (GWAs) can scan the whole genome for variants affecting a certain disease without a prior hypothesis of likely candidates or necessarily any knowledge of the disease pathogenesis. This type of analysis will allow the detection of novel pathways and genes that would not be candidates based on current knowledge providing vital new biological insights which may hold the key to novel therapies (Farrall and Morris 2005; McKinney and Merriman 2007).

Association analyses, however, also have problems. Unlike linkage studies, which are carried out in families, association studies can produce spurious results due to underlying population structure. Population stratification in a sample, a mix of 2 isolated groups, one with high disease frequency and one low, may show false positive association between the disease and any marker that shows an allele frequency difference between the two groups (Clark 2003; Helgason et al. 2005). To avoid this problem studies are carried out on a single population sample in which there is no evidence of a recent influx of genes with differing ancestries. However, this is usually based on self identified ethnicity and limited knowledge of ancestry. There are methods available to control for unknown population stratification such as, Genomic Control (GC) methods which apply a correction to the statistical distribution of the association metric used, based on a measure of the variability of genotypes (Devlin and Roeder 1999). Another approach is to identify outlying individuals or assign individuals to various population clusters and carry out separate association tests on the population stratified groups. An example of this type of method is, Multi Dimensional Scaling (MDS) within the PLINK analysis toolset, which is carried out on the basis of the genome-wide average proportion of alleles shared identical by state (IBS) between any two individuals (Purcell et al. 2007). Population admixture can be useful for a method known as admixture mapping, which has been successfully used for mapping hypertension loci (Zhu et al. 2005). However, the usefulness of this method remains to be determined by further examples and there are several issues to overcome. For example, the alleles in the parental populations are required to be relatively homogenous and the allele frequencies must differ substantially. Furthermore admixture in human populations seldom happens at a specific point in time but over a period and the parental populations may not be available for study or known precisely (Jorde 2000; McKeigue 2005).

Association mapping also relies on careful ascertainment of samples and accurate phenotype measures. It is important to ensure that case and control samples have been processed in the same way and there is no systematic bias, which would produce misleading results. With common diseases it is possible for a proportion of the controls to become cases in the future, which would reduce power. However, control samples can be enriched using 'hypercontrols' ie. individuals much older than the normal age of onset or at the extreme lower end of the disease spectrum; cases can also be enriched by sampling individuals with strong family history or particularly extreme phenotypes. This should result in an increase in power to detect real genetic effects. It is also very important in retaining power that the cases are stringently phenotyped in a uniform way. Genetic heterogeneity, different genotypes causing the same disease, may be partly addressed by stratifying cases based on previously determined susceptibility alleles.

There are different approaches to choosing SNPs to genotype for an association study. A direct approach uses candidate genes for a particular disease based on functional evidence or suggestive linkage results and coding SNPs within those genes with the hope of genotyping the causal variant. It is possible to prioritise non-synonymous SNPs which alter an amino acid these are implicated as high risk alleles in many mendelian disorders. Although many identified variants for complex diseases are in non-coding regions and are thought to have regulatory interactions with other genes. Therefore this may not be useful for common diseases with moderate risk alleles. An indirect approach is to genotype a high number of SNPs genome-wide with the hope of genotyping a variant that is in LD with the causal locus. Most investigators use the commercially available genotyping platforms, which vary in the coverage attained. Affymetrix gene chips use randomly chosen SNPs determined by the restriction enzyme used to cleave the DNA. Illumina have used HapMap data to be more selective in choosing SNPs that offer the best coverage based on LD patterns. There are therefore some regions of the genome not well covered by Affymetrix. However should genotyping fail on a certain SNP tagging a large region the Illumina platform would not necessarily have a nearby SNP able to cover the region. Since these platforms offer by far the cheapest strategy for genotyping large numbers of SNPs in large samples they are of great value particularly for a 2 stage analysis where more targeted genotyping can be carried out on a smaller scale in the second stage.

1.2 Aims

The main aim of this project is to define linkage disequilibrium patterns and tracts of extended homozygosity in order to compare populations and search for disease genes.

LDU maps will be created using in-house software (LDMAP+) and used to investigate the similarities and differences in patterns of LD across populations, and determine the properties and utility of a cosmopolitan LDU map. The recent release of whole genome genotype data provided by the International HapMap project will allow creation of genome-wide LDU maps in different populations and an investigation of evolutionary history of these populations by estimating the Effective Bottleneck Time (t).

The HapMap data will also be used to investigate homozygosity in the human genome determining the amount and location of tracts of homozygosity and their relationship with LD patterns as described by LDU maps. Knowledge of levels of homozygosity in healthy individuals and the relationship with LD will add power to homozygosity mapping methods. This will be exploited using data on individuals from a consanguineous family, affected by Congenital Nephrotic Syndrome, to localise a candidate gene or region responsible for the disease.

The use of LDU maps for association mapping of genes and variants involved in complex diseases will then be investigated, with a genome-wide association scan of anonymous data. In-house software (CHROMSCAN-cluster) will be tested with genome-wide data and simple single SNP chi square tests will also be considered. The aim of this initial scan will be to determine regions for follow up in more detail in a second stage. Whole genome analyses are the basis of new and innovative approaches to discovering disease genes and an accurate and informative description of levels of homozygosity and patterns of LD for this type of data will be invaluable.

Chapter 2 - Cosmopolitan LDU maps

2.1 Introduction

Linkage maps have been an invaluable tool for mapping major genes through linkage analysis. Linkage analysis tracks the segregation of a disease and a marker through a single generation within families. Candidate regions can be narrowed to a few cM in genetic distance which corresponds to a few Megabases (Mb) in physical distance. These regions are narrowed by the recombination events that take place during meiosis. Linkage disequilibrium (LD) differs since it is influenced by recombination events that take place over many generations since the founding of the population. The higher number of recombination events allows the candidate region to be narrowed much further allowing higher resolution fine scale mapping of disease genes and causative variants. Many methods of performing disease mapping using LD are currently being developed, investigated and validated (Maniatis et al. 2004; Zaykin, Meng, and Ehm 2006; Morris et al. 2003). LD is a major focus for investigators in their mission to locate moderate risk genes involved in complex human disease. Knowledge of the background variation and structure of LD across the whole human genome would be an invaluable tool to this end, in the same way that the linkage map has been useful for the mapping of high risk 'major' genes. With the advent of high density Single Nucleotide Polymorphism (SNP) panels for whole chromosomes, LD structure over larger areas can be determined. Different human populations have different population histories, such as differences in time since the last major bottleneck, which affect LD. Recombination patterns are thought to be the similar in all populations, although critical evidence is lacking (Jorgenson et al. 2005). Over 1-2 generations there is no detectable effect of drift, selection, mutation and therefore linkage maps are assumed to be similar irrespective of the population studied. The linkage map effectively represents current patterns of recombination whereas the LDU map is mostly determined by historical patterns. However, the LDU map is also affected, to a lesser degree, by selection, mutation and drift.

Patterns of LD across large regions can be described by blocks of high LD separated by small regions of high recombination. Sperm typing data produced by Jeffreys *et al.* (Jeffreys, Kauppi, and Neumann 2001) supports this finding with directly observed meioses in sperm. Long range patterns of LD tend to be conserved across populations and differences due to duration can be modelled. Lonjou *et al.* analysed small data samples ranging 120Kb-1.3Mb with low SNP density (18 SNPs across 1.3Mb) and several samples of higher density (1 SNP per 2Kb) in small regions of average size 250Kb. A cosmopolitan map created from combining samples was able to recover 95% of the information in different population maps by appropriate scaling (Lonjou et al. 2003). The similarity of LD structure across populations has been shown by several studies. Shifman et al. compared LD (D' and r^2) in 3 types of population, admixed (African American), outbred (Caucasian) and isolated (Ashkenazi Jews). They found very similar allele frequencies between the Caucasian and the Ashkenazi Jew populations which both differed from the African Americans, and an average decline of LD of a similar rate in the Caucasians and Ashkenazi Jews but a more rapid decline in African Americans. They also found that LD was highly correlated across populations (Shifman et al. 2003).

The major difference between populations has been found between African and non-African populations reflecting the presumed 'out of Africa' bottleneck. This would have restricted the diversity of haplotypes founding the non-African populations effectively resetting LD at this point. Differences in the time recombination has had to break up founding haplotypes and therefore the amount of LD between populations can be modelled linearly by scaling, while the underlying structure of LD remains intact.

2.2 Aims

The aim of this chapter is to investigate the possibility and feasibility of developing a standard LDU map that is useful and informative for multiple populations. The similarity of patterns of LD in different populations over a large region of chromosome 20 will be determined. Previous work has shown that LD patterns across populations are very similar even though a difference in the time that recombination has had to accumulate since an 'effective bottleneck' creates different scales to the LDU maps (Lonjou et al. 2003; Zhang et al. 2004). The aim is to create a cosmopolitan LDU map by combining genotype data for 4 populations on a 10Mb region of chromosome 20 and then determine how well this map represents the information of each of the 4 populations separately.

2.3 Methods

2.3.1 Data

The data analysed consist of 5,954 Single Nucleotide Polymorphisms (SNPs) genotyped over a 10,098Kb region of chromosome 20q12-13.2. The data were previously published and made available by Ke *et al.* (2004). The genotype data are for 282 individuals across 4 populations; 97 African Americans, 96 UK Caucasians, 47 Utah individuals from the Centre d'Etude du Polymorphisme Humain (CEPH) panel and 42 East Asians (32 Japanese and 10 Chinese). The data were screened for quality and no significant deviations from Hardy-Weinberg equilibrium were detected. Five SNPs were removed because they were rare with a minor allele frequency <0.05. This left a total of 5949 SNPs, not all of which were genotyped in all populations (table 2.1). The alleles for each SNP were coded as 11, 22, 12 with 00 denoting missing data.

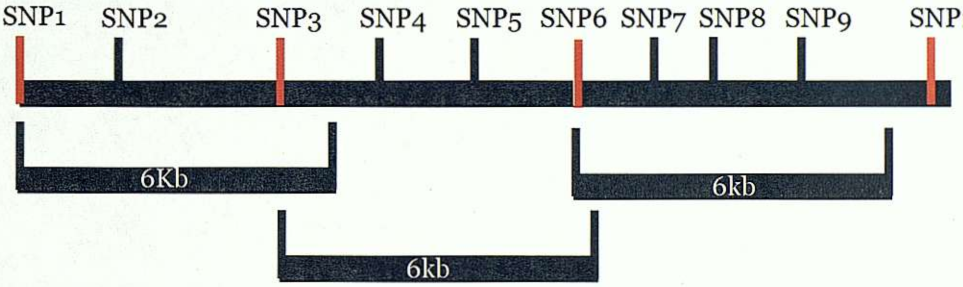
Table 2.1 Chromosome 20 data sample.

Population sample	No. individuals	No. SNPs
AF (African Americans)	97	4938
CA (UK Caucasians)	96	4427
CE (Utah CEPH)	47	5309
AS (East Asians; Japanese and Chinese)	42	4160

2.3.2 Selecting SNP densities

Constructing LDU maps by analysing pairwise data for thousands of SNPs is computationally intensive. To reduce the computational burden the SNP density was reduced, providing an ideal opportunity to investigate the effects of different SNP densities on the quality of LDU maps. The average SNP density of the whole sample was 1 SNP every 2Kb. This was reduced to a density of 1 SNP every 6Kb, similar to that of the initial target of the International HapMap project, and then further to 1 SNP every 8, 10, 12 and 15Kb and constructed corresponding LDU maps. This reduction in density was performed using an algorithm, designed to achieve a uniform spacing of SNPs on the physical map, avoiding large gaps (figure 2.1). Starting from the end of the map closest to the p telomere the first typed SNP was designated the 'starting SNP' and two other SNPs were identified that were either side of a position a selected number of Kb away. The SNP closest to that position was chosen. The chosen SNP then became the new 'starting SNP' and the process was continued along the length of the map. In the case that the 2 selected SNPs were of equal distance from the chosen position the SNP closest to the 'starting SNP' was chosen. The length of the region (10,098 kb) was then divided by the number of SNPs selected to calculate the average density over the region. The process was repeated using a range of Kb distances until the desired mean density was achieved.

Figure 2.1 Diagram showing the algorithm to select SNPs at reduced densities. Here selecting the red SNPs (1,3,6 and 10) at a 6Kb average density.



A total 1694 SNPs from the whole sample i.e. 1 SNP every 5.96Kb, were chosen for the 6Kb map. Due to differences in the SNPs genotyped in the different population samples, the actual densities vary ranging from 1 SNP every 6.6-8.8 Kb (table 2.2).

Table 2.2 Number of SNPs at a 6Kb density.

Population sample	No. SNPs	Average Density
AF (African Americans)	1338	7.5
CA (UK Caucasians)	1211	8.3
CE (Utah CEPH)	1518	6.6
AS (East Asians; Japanese and Chinese)	1153	8.8

2.3.3 Creating LDU maps

LDU maps (Maniatis et al. 2002) are based on the Malecot model,

$$\rho = (1\text{-}L)Me^{-\varepsilon d} + L$$

which describes the decline in association ρ as a function of physical distance d (in Kb). The parameters of the model are M, the maximum association at zero distance, reflecting association at the last major bottleneck. L, the residual association at large distance and ε, the exponential decline of ρ with distance. The Malecot parameters ε and M are estimated by fitting multiple pairwise association probabilities, ρ, and corresponding information, Kρ, using composite likelihood. We used the predicted L (Lp) (Morton et al. 2001), rather than the estimate of the L parameter since Lonjou et al (2003) found that estimating L can leading to distortions in the LD map through the creation of 'holes' between adjacent SNPs. The LDMAP program,(http://cedar.genetics.soton.ac.uk/pub/PROGRAMS/LDMAP/) computes ε for each interval between pairs of SNPs. The parameters are estimated to maximise the composite likelihood, then the length of the ith interval, in LDU, is given by $\varepsilon_i d_i$. The LDU values are summed to give an additive map. An upper limit of 3 LDUs is imposed to maintain the integrity of the map, and intervals of 3 LDUs are termed 'holes'. These areas of the genome have been shown to have high recombination and require more densely genotyped SNPs to resolve the holes (Tapper et al. 2001), though it is possible that holes in intense recombination hotspots are impossible to resolve. To construct a cosmopolitan (COS) LDU map, pairwise SNP haplotype frequencies were converted to counts using the algorithm described by Hill (Hill 1974). The haplotype counts were then combined across populations by summing among matching pairs of loci. Markers were coded consistently so that an allele coded 1 in one population was also coded 1 in all other populations where the SNP appeared. Pairwise association probabilities (ρ) and the corresponding information (Kρ) were computed from the haplotypes counts to create a cosmopolitan map (Collins, Lonjou, and Morton 1999). LDU maps were created for each of the 4 population samples and 'cosmopolitan' maps were created at various SNP densities.

2.3.4 Evaluating the maps

The different density cosmopolitan LDU maps were evaluated by comparing Malecot parameters, map lengths, swept radii and the number of holes. The population specific maps were compared in the same way. To investigate the SNP density required to resolve a hole in the map, SNPs were added to intervals of 3 LDU (holes) where they existed in the full dataset, and the effect on map length was determined.

The cosmopolitan map was compared to the Kb map and population-specific LDU maps by fitting the multiple pairwise data to the cosmopolitan map, using the Malecot model with kb or LDU as the distance and maximising the composite likelihood. The error variances when fitting the pairwise data for a given population to the kb map (V_{Kb}), to the population specific LDU map (V_{POP}) and to the cosmopolitan map (V_{COS}) and the degrees of freedom were calculated in the following way. The degrees of freedom were computed as N – (m-1) – r, where N is the number of pairs, m is the number of loci (therefore m-1 intervals in which ε may be estimated) and r is the number of additional parameters estimated. N_i and N_c are defined as the number of pairs of SNPs (pairwise association probabilities) in the ith population data sample and cosmopolitan data sample respectively. The number of SNP markers in the ith population sample and cosmopolitan sample respectively are m_i and m_c.

$V_{Kb} = -2\ln L / (N_i - 2)$, where ε and M are estimated.

V_{LDU} = -2lnL / (N_i – (m_i - 1) -1), where M is estimated and ε is estimated in each map interval.

V_{COS} = -2lnL / (N_i – (N_i / N_c)(m_c – 1) – 2), where m_c-1 intervals in the cosmopolitan map have been previously computed using the proportion of data represented by the ith population sample as N_i/N_c, and ε and M are estimated.

The relative efficiency (RE) of the cosmopolitan maps was calculated, to determine how much of the information was recovered, as RE = V_{POP} / V_{COS}. The ratio of the ε value estimated when the population specific data is fitted to the cosmopolitan map, and the ε value for the cosmopolitan map itself, provides the scaling factor.

2.4 Results

2.4.1 Evaluation of the cosmopolitan LDU maps at different SNP densities

The map lengths range from 187-204 LDU and the number of holes ranges from 2-7. The number of holes is generally larger in the longer maps, although this is not always the case as the ‘15Kb’ map has one more hole than the ‘12Kb’ map but is slightly shorter. The small number of holes relative to the number of intervals (m+1) shows that the LD patterns are well characterised and the SNP density and coverage is adequate.

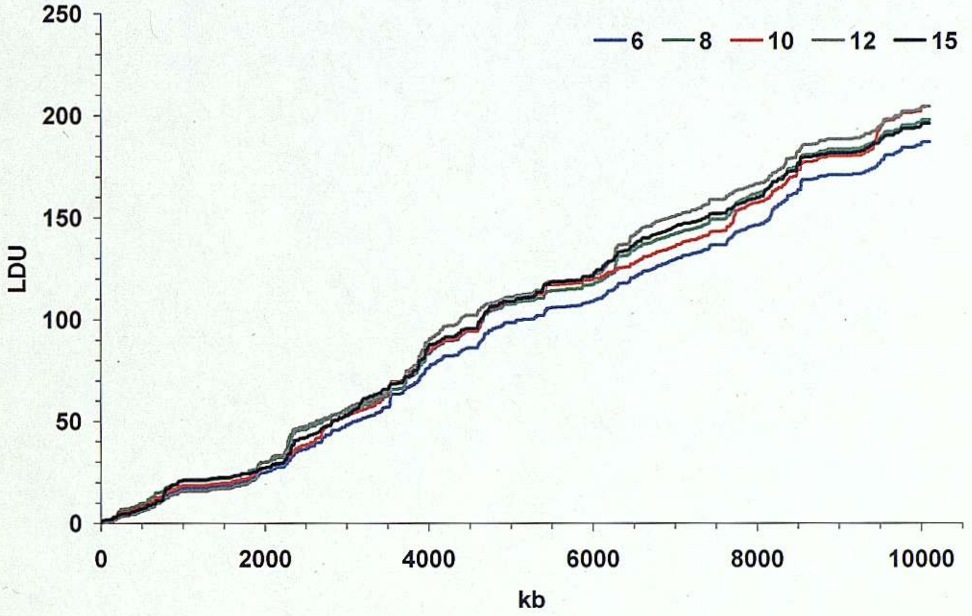
Table 2.3 Cosmopolitan LDU maps at different Kb marker densities.

Density	N	m	ε	M	Lp	-2lnL	df	V_{LDU}	No. LDUs	No. holes
6	132171	1691	1.1521	0.894	0.091	179822	130480	1.378	187.15	2
8	76236	1289	1.1609	0.877	0.092	105937	74947	1.413	198.02	5
10	45221	992	1.1399	0.895	0.090	61091	44229	1.381	204.41	5
12	31497	833	1.1331	0.897	0.091	40581	30664	1.323	204.56	6
15	20483	670	1.1381	0.870	0.090	28439	19813	1.435	196.28	7

N - number of pairs, m - number of loci, ε/M/Lp – Malecot parameters, -2lnL – composite -2 log likelihood, df – degrees of freedom, V_{LDU} – residual error variance for the LDU map.

Over the different densities the maps remain relatively consistent in overall length, and the graph (figure 2.2) of the maps shows that the contours are also well conserved. This shows that the broad patterns of LD are retained even at low densities and the LDU map is robust to such changes in SNP density.

Figure 2.2 Graph of cosmopolitan maps at different densities.



It was decided that a density of 1 SNP every 6Kb representing approximately 500,000 SNPs genome-wide, was suitable for evaluating the cosmopolitan map.

2.4.2 Evaluating the fit of the population-specific and cosmopolitan pairwise data to the Kb and LDU maps

The pairwise data were fitted to the Kb map for each population and the cosmopolitan map. The swept radii which show the average extent of LD, range from 80-105 Kb with the AF population having the least extensive LD. AF also has the lowest M value, 0.66, reflecting a larger effective population size. The COS map has values intermediate between the AF and other populations.

Table 2.4 Fitting the pairwise data to the physical (Kb) map.

Population	N	m	ε	M	Lp	-2lnL	df	V_{kb}	Swept radius(kb)
COS	132171	1691	0.01024	0.738	0.091	339109	132169	2.566	97.6
AF	87135	1338	0.01243	0.661	0.136	114123	87133	1.310	80.4
CA	71097	1211	0.01043	0.877	0.135	109046	71095	1.534	95.9
CE	111067	1518	0.00953	0.805	0.197	102478	111065	0.923	104.9
AS	64586	1153	0.01117	0.861	0.204	52781	64584	0.817	89.6

N - number of pairs, m - number of loci, ε/M/Lp – Malecot parameter estimates, -2lnL – composite log likelihood, df – degrees of freedom, Swept radius – 1/ε, V_{kb} – residual error variance on fitting pairwise data to the kb map.

The data were then fitted to the LDU maps. LDU map lengths for the 4 population samples range from 204-272. The AF population has the longest map showing less LD overall. The COS map is 187 LDU which is shorter than the intermediate value that might have been expected. However, there are only 2 holes in the COS map and 9-17 holes in the population specific maps. The number of holes in the population specific maps was reduced by adding SNPs which were genotyped in the original high density data where available. The figures in brackets show the map lengths and number of holes when extra SNPs were added. The number of holes reduces but not substantially, and the maps in general become marginally shorter, except the CE map.

Table 2.5 Fitting the pairwise data to the LDU maps.

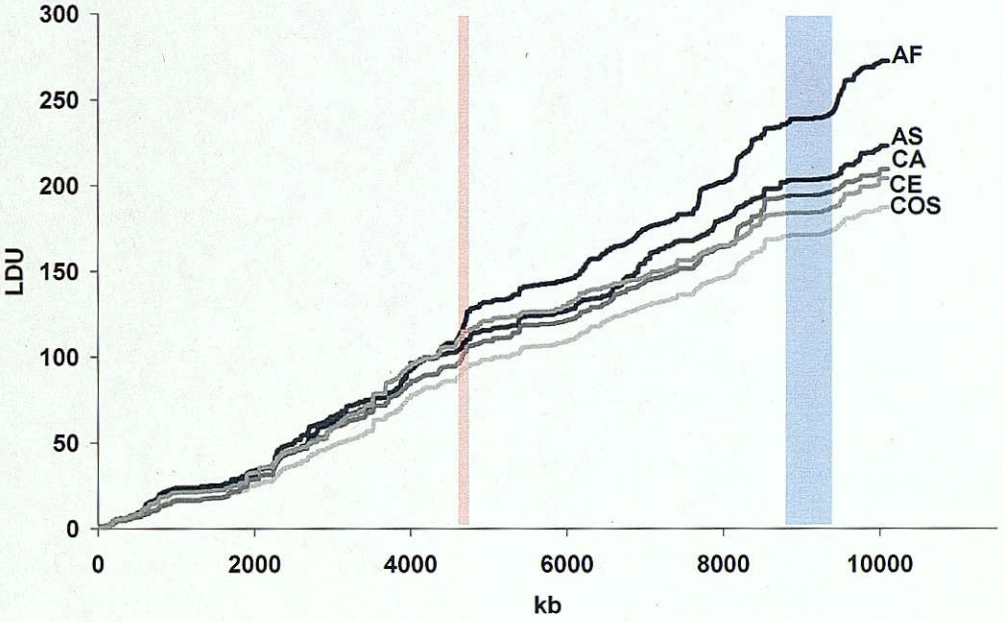
Population	ε	M	-2lnL	df	V_{LDU}	No. LDUs*	No holes*
COS	1.1521	0.894	179822	130480	1.378	187.15	2
AF	1.1661	0.842	77916	85797	0.908	272.49 (268.22)	13 (10)
CA	1.0754	0.957	54057	69880	0.774	209.62 (208.06)	9 (8)
CE	1.1290	0.924	67225	109549	0.614	204.19 (204.66)	12 (9)
AS	1.0811	0.923	33777	63434	0.532	223.20 (222.29)	17 (13)

ε/M – Malecot parameter estimates, -2lnL – composite log likelihood, df – degrees of freedom, V_{LDU} – residual error variance for the LDU map.

* values in brackets have extra SNPs added to the maps where available in the original data.

The LDU maps for each population sample and the cosmopolitan map were plotted against the Kb map (figure 2.3). The maps vary in length with the AF map being the longest, the COS map being the shortest and the CE, CA and AS maps of similar lengths in between. There is general agreement in the contours of the maps, which seem to have blocks of high LD in the same locations (Blue area, figure 2.3) and steps of high recombination in the same locations (Red area, figure 2.3). The size of the steps seems to be the factor that varies most, altering the overall length of the maps.

Figure 2.3 A graph of all populations.



2.4.3 Fitting population-specific pairwise data to the cosmopolitan LDU map

The pairwise data for each population was fitted to the COS map in turn. Again the AF population has the highest epsilon and lowest M values.

Table 2.6 Fitting the data for each population to the COS map.

Population	ε	M	-2lnL	df	V_{COS}
AF	1.5323	0.811	82334	86019	0.957
CA	1.0659	0.968	59172	70186	0.843
CE	1.0231	0.927	70952	109645	0.647
AS	1.1859	0.931	37198	63758	0.583

ε/M – Malecot parameter estimates, -2lnL – composite log likelihood, df – degrees of freedom, V_{COS} – residual error variance for the individual population data fitted to the COS map.

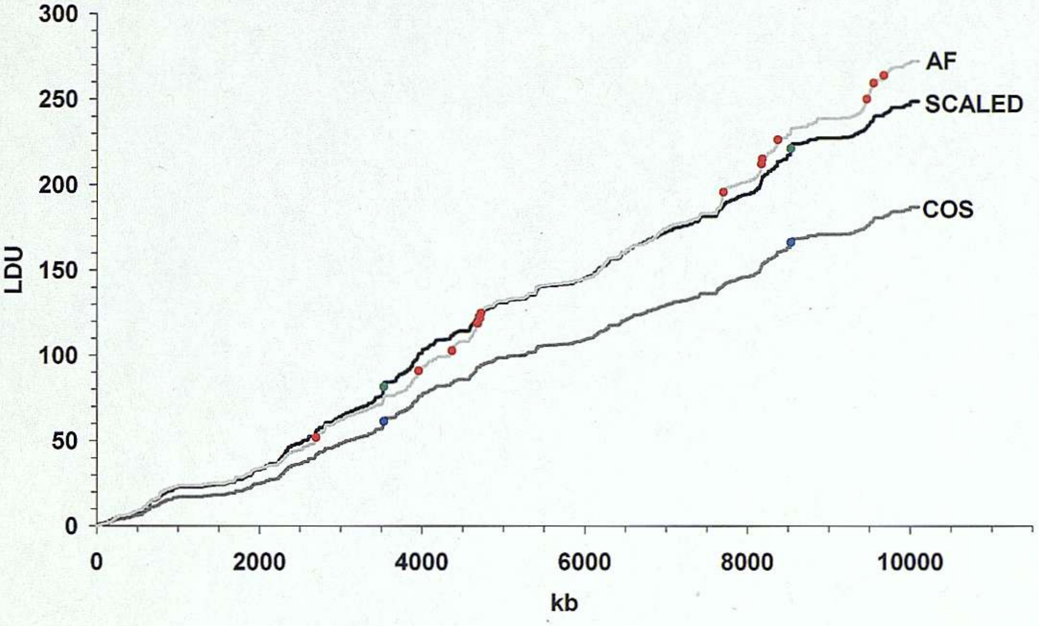
The scaling factors are calculated as the ε value estimated when the population specific data is fitted to the cosmopolitan map divided by the ε value for the cosmopolitan map. The relative efficiency for each map is calculated as a ratio of V_{LDU}/V_{COS}. The values range from 91-95% showing the proportion of the information which is recovered by scaling using the appropriate scaling factor.

Table 2.7 Relative efficiency of different maps and scaling factors for each population.

Population	V_{kb}	V_{LDU}	Vcos	Relative efficiency of kb map (V_{LDU}/V_{kb})	Relative efficiency of COS map (V_{LDU}/V_{COS})	Scaling factor relative to COS map
AF	1.310	0.908	0.957	0.693	0.949	1.330
CA	1.534	0.774	0.843	0.504	0.918	0.925
CE	0.923	0.614	0.647	0.665	0.949	0.888
AS	0.817	0.532	0.583	0.651	0.913	1.029

Figure 2.4 shows scaling of the COS map with the AF scaling factor. The scaled map is shorter in length but has only 2 holes whereas the AF map has 13 holes, which are known to inflate LDU map lengths. The 2 holes remaining in the scaled map are coloured in green (blue in the COS map) and the 13 holes in the AF map are coloured red.

Figure 2.4 A graph of the AF and COS LDU maps with the COS map scaled by the AF scaling factor.



2.5 Discussion

A suitable SNP density to create and analyse the cosmopolitan LDU map was determined, taking into account the trade-off between the quality of the map and the computational time taken to produce the map. With the version of LDMAP used a 6Kb density was viable. To determine the effect of reducing SNP density, cosmopolitan maps were also made with 1 SNP every 8, 10, 12 and 15 Kb. Reducing the SNPs density dramatically, and therefore, reducing the information available to make an LDU map, would result in a lower quality map. However, it seems that over the range of 1 SNP every 6 to 15 Kb the maps are relatively robust to density changes. The LDU maps have similar lengths (187-204LDUs) and have few holes relative to the large number of intervals considered. Holes generally appear to make maps longer in length, however sometimes adding SNPs does not resolve a hole as with COS maps of 12Kb and 15Kb density. The 15Kb map is actually slightly shorter even though it has 1 more hole. This is a case where 1 hole which is not resolved by the addition of a SNP becomes 2 holes. It has been shown that holes represent areas of the genome which have high levels of recombination and therefore tend to require a high number of SNPs to resolve them (Tapper et al. 2003). A density of 6Kb was the best map for the purposes of this study, it also reflects the approximate initial target of the HapMap project of 500,000 SNPs genome-wide (The International HapMap Consortium 2003). Although adding SNPs adds information and therefore would result in a map of higher quality and accuracy, low density maps are still useful. They provide a basic and robust description of the broad LD patterns and can be added to with information provided by more SNPs or more individuals at a later stage.

LDU maps were made for each population at this density, and the maps were compared. For each sample the pairwise data were fitted to the Kb and the LDU maps (tables 2.4 & 2.5). The fit of the data was better for the LDU map than the Kb map as shown by the lower residual error variance. This is expected since the Kb map does not reflect patterns of LD. The M parameter in the Malecot Model has an evolutionary interpretation, and reflects the haplotype diversity at an 'effective' bottleneck. The lower value of M in the AF population can therefore be explained by the longer time since the last major bottleneck in that population compared to the more recent 'Out of Africa' bottleneck in the other populations. The 'Out of Africa' hypothesis suggests that a small number of individuals left Africa to populate the other continents. This bottleneck caused the effective resetting of LD at this point due to low haplotype diversity in the relatively small number of founders. The swept radii show the extent of LD and the lower value in the AF population shows reduced LD, as does the longer length of the AF LDU map relative to the maps of the other samples. More recombination results in more or longer steps in the LDU map and therefore a longer map overall. Although, the overall lengths of the LDU maps vary, with the major difference being between the AF and the other populations, the pattern of LD shown by the contours in figure 2.3 remain very similar across populations. This shows that the structure of LD is common across populations even though the intensity changes. Recombination is not uniformly distributed across the human genome but is concentrated at various locations called recombination hotspots, which tend to be small regions of 1-2Kb (Jeffreys, Kauppi, and Neumann 2001). The co-localisation of these hotspots creates the similar structure of LD across populations. The intensity of recombination at these hotspots generates the differences between populations of different ages (time since last major bottleneck) and is shown in the LDU maps as differing lengths.

Holes in the maps tend to exaggerate the length of LDU maps. SNPs were added to holes in the population specific LDU maps in areas where SNPs were available in the original high density data. The addition of SNPs resolved some of the holes and had the general effect of reducing map length. However, the CE map became marginally longer (204.19 to 204.66LDUs) even though 3 holes were resolved. Holes which are the result of intense recombination hotspots rather than low local SNP density require particularly high SNP density to resolve them, suggesting more SNPs are needed in these holes to have a more dramatic and predictable effect.

The relative efficiency is calculated as a ratio of the residual error variances when fitting the pairwise population specific data to the COS map and the residual error variance of the COS map itself (table 2.7). The relative efficiencies for the 4 populations ranged from 0.91 -0.95, showing that 91-95% of the information represented in the population specific data can be recovered from the COS map with appropriate linear scaling. A loss of between 5 and 9% is tolerable and shows that the COS map could indeed be very valuable for a wide range of populations. The epsilon value for each population relative to that for the COS map was calculated to provide an appropriate scaling factor for each population. There is very good correspondence between the scaled and original AF LDU maps, shown by plotting against the Kb map (figure 2.4). The scaled map had a reduced number of holes (2) due to the increased information provided by the extra SNPs and individuals present in the COS map. The original AF map has 13 holes, the positions of these holes along the map were also plotted and it is evident that the holes towards the end of the map are likely to explain the longer length of the original AF map relative to the scaled map.

There are several benefits to the idea of a cosmopolitan map. One standard map can be used for many populations and using data from multiple populations to create the map results in a map of higher quality and accuracy since more individuals and more SNPs are used, providing a higher resolution map with fewer holes. A standard map which is made available to other researchers would reduce the costs involved in association mapping of complex disease, by reducing the need to genotype so many individuals for the purposes of defining LD in the region under study. LDU maps are useful directly for association mapping or for increasing the resolution of linkage maps, as well as investigations into population history, such as inbreeding, and evolutionary forces, such as selection. Genome-wide LDU maps are integrated into Linkage Disequilibrium DataBase (LDDB) and publicly available. A database of scaling factors may also be required or investigators could use a population specific LDU map in a particular region to calculate a scaling factor which could then be applied across the whole genome. There should be no difference in the scaling factors required for different chromosomes within a population, since they have been under the same evolutionary and recombination conditions. Similar ratios of LDU map length across a population over chromosomes would show this.

Genotyping technology has moved forward with such speed over the last few years that modifications will be required to cope with the magnitude of data proposed to be made available by the HapMap project. Updates to the LDMAP program in terms of the way the algorithm is run and how the data files are handled, including the possibility of parallelising the program, will mean that genome-wide LDU maps and corresponding cosmopolitan maps are a feasible proposition. HapMap data will allow the extension of comparisons of LD patterns across populations to a genome-wide scale, allowing validation of the use of a single scaling factor across chromosomes.

2.6 Conclusion

This work supports and extends the findings of Lonjou *et al.* (Lonjou et al. 2003), showing that the careful modelling of LD patterns in humans can show the similarity in LD patterns across populations. The tendency for recombination hotspots to be restricted to particular locations which are co-localised across populations explains the remarkable correspondence of the broad LD structure as shown by the contour of the LDU maps. Cosmopolitan or composite LDU maps are therefore a feasible alternative to the costly genotyping of many SNPs and individuals in every population. Thus a standard map and a set of scaling factors will be a valuable tool for association mapping in many populations.

Chapter 3 - Creating and analysing genome-wide LDU maps of multiple populations

3.1 Introduction

An LDU map is a powerful tool for describing the structure and intensity of Linkage Disequilibrium (LD) in the genome (Maniatis et al. 2002). LDU maps have been made for various small regions of the genome (Lonjou et al. 2003; Zhang et al. 2002), a 10MB region of chromosome 20 (Gibson et al. 2005) and also a whole chromosome, chromosome 22 (Tapper et al. 2003). Maps of this kind allow us to gain a picture of the structure of LD, a description of the human genome in Linkage Disequilibrium Units is useful in various ways. LD patterns can be used to determine the most informative SNPs for association mapping, and LDU maps can be used for association mapping of complex traits in the same way that a linkage map has been used for linkage mapping of major genes with great success. Genome-wide LDU maps have higher resolution which should allow a disease gene or variant to be located to a much smaller region than by linkage (Maniatis et al. 2004). Many aspects of population history and demography can be investigated using LDU maps including a measure of the age of a population in terms of the effective bottleneck time. This is a measure of time since the last major bottleneck taking into account the cumulative effects of successive bottlenecks. Processes that have occurred independently in different population groups such as the response of different populations to different environmental factors, known as selection, can also be studied using LDU maps.

Tagging haplotypes can reduce the number of SNPs needed to describe a region by choosing the most informative SNPs for association mapping based on a genome-wide description of LD. This idea and recent technological advances in SNP genotyping have allowed many SNPs to be genotyped at costs which are much less prohibitive than in recent years prompting the International HapMap Project (Daly et al. 2001). The International HapMap project set out to catalogue human variation and began releasing genotype data into the public domain in December 2003 (The International HapMap Consortium 2003). This data is high quality, increasingly high resolution and provides the ideal opportunity to create LDU maps of the whole human genome.

The first release of HapMap data to contain sufficiently high density data for creating LDU maps, on all chromosomes, was release 11 (Sept 2004) although this only contained data for the CEPH population. A total of 665,335 genotypes were downloaded from the release 11, September 2004 public release of the data. These data were filtered and 25.8%, 171,927 SNPs were removed for Hardy-Weinberg deviations and very rare alleles (MAF <5%). 493,408 SNPs remained for creating LDU maps of all 23 chromosomes, giving an average SNP density of 1 SNP per 5.6Kb. The LDMAP program creates LDU maps from pairwise genotype data, this is a computationally intensive process which becomes a problem for creating LDU maps of large regions, or whole chromosomes, at high SNP density. Modifications to the LDMAP program by more efficient file handling have allowed LDU maps to be created from high density data. Further modifications to the LDMAP program allow LDU maps to be created in segments which are then rejoined to create a complete map. The default settings of 500 loci per segment and a 25 SNP overlap were used for the HapMap release 11 data. The Malecot model is fitted for each SNP interval, taking into account surrounding SNP pairs containing that interval. However, in such high density data not all pairs of SNPs are used, with increasing distance LD declines and SNP pairs are less informative and at extremes can introduce background noise to the model. For the HapMap data a maximum of 100 surrounding pairs within a 500Kb distance was determined to be appropriate, giving little loss of information whilst producing a high quality map more rapidly. These changes to LDMAP and the data produced by the HapMap project have made it feasible to create the first genome-wide LDU map of the human genome (Tapper et al. 2005).

A comparison of the linkage map (Kong et al. 2004) and the release 11 LDU map, of the CEU sample, showed a remarkable correspondence with 96.8% of the variance in the linkage map explained by LDU, calculated by regression across chromosome arms for the whole genome. Since the linkage map shows recombination over a single generation and LDU maps show recombination over many generations (ignoring the small contribution of stochastic variation) the ratio of the two values gives the number of generations since the last major bottleneck, taking into account the effects of multiple subsequent bottlenecks. This is termed the "effective bottleneck time" (t). In this data t was calculated as 1,435 generations, multiplied by 20 or 25 years per generation, this gives 28,700 or 35,875 years. Since human chromosomes have undergone the same evolutionary history in terms of opportunity for recombination (except the special case of the X chromosome) the values of t should be constant across the chromosomes. However, there is a small amount of variation in t showing a trend for smaller values of t in the smaller chromosome arms. A process known as chiasma interference, means that a cross-over at one location prevents further crossovers in close proximity. The linkage map has a function to account for this in the final map, however it is a genome-wide measure applied to all chromosomes and there is evidence that interference is more intense in the small chromosomes (Broman et al. 2002). This may lead to an inflation of the linkage map length in the smaller chromosomes and explain the trend for lower values of t (Tapper et al. 2005).

An early release of the HapMap data with genotypes on all 4 samples (November 2004, release 13), allowed me to carry out a preliminary comparison across populations. Four chromosomes, 20, 19, 13 and 10 had sufficient SNP density over all 4 populations to create LDU maps. A filtering process removed duplicate SNPs, very rare markers with an MAF of <5% and those that dramatically deviated from Hardy-Weinberg Equilibrium (chi squared >10). The total numbers of SNPs across the 4 chromosomes used for LDU map construction were, 55,774 for the YRI sample, 69,956 for CEU, 49,068 for CHB, and 48,578 for the JPT sample. The high correspondence between the linkage and LDU maps showed that 96% of the variance in the LDU map is explained by recombination as shown by the linkage map. The effective bottleneck times were estimated using chromosome arm data (table 3.1).

Table 3.1 Effective Bottleneck Times (t) for 4 populations based on 4 chromosomes

Population	Effective bottleneck time t=LDU/Morgans	Time in years (assuming a generation =25 years)
YRI (Yoruba in Ibadan)	1977	49,425
CEU (CEPH Utah residents with ancestry from northern and western Europe)	1559	38,975
CHB (Han Chinese in Beijing)	1772	44,300
JPT (Japanese in Tokyo)	1506	37,650

The African YRI sample was the oldest, consistent with an out of Africa model and population comparisons on the chromosome 20 region (chapter 2) and other studies of multiple populations (De La Vega et al. 2005; Gibson et al. 2005). The HapMap project has continued to release data at higher SNP densities across 269 individuals over the 4 population samples. This data allows genome-wide LDU maps to be created and compared across populations.

3.2 Aims

The aim of this chapter is to use the new segmented version of LDMAP (LDMAP+) to create LDU maps across the whole genome extending previous work to the more complete Phase I release of the HAPMAP data, approximately 1 million SNPs for all 4 populations. The properties of the maps over the different populations will be investigated, and comparison with the linkage map will allow estimation of the Effective Bottleneck Time (t) for each sample.

3.3 Methods

3.3.1 Data

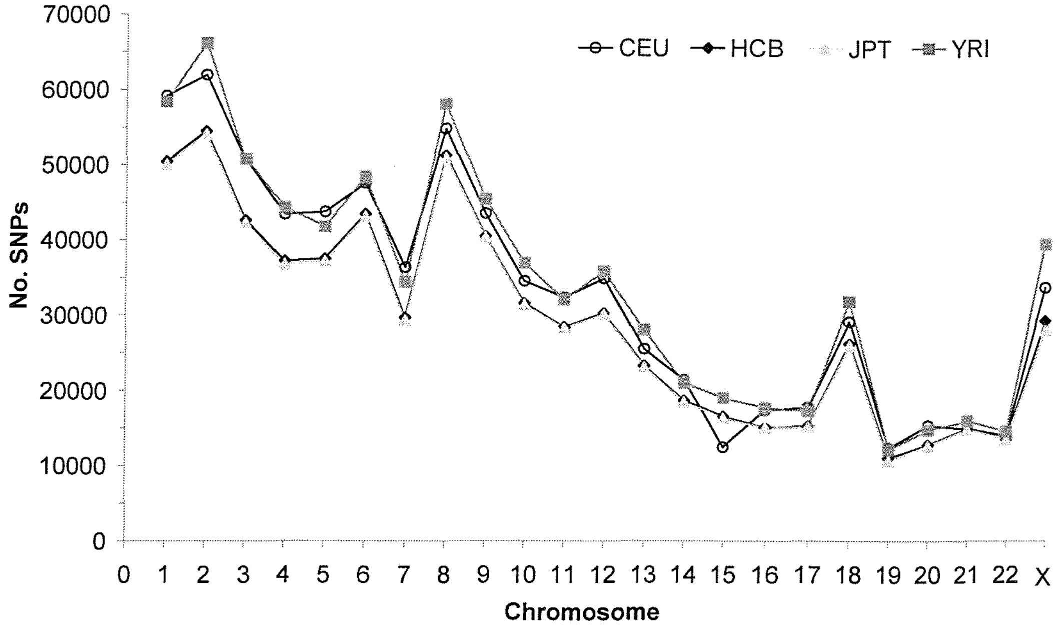
Phase I (release 16, March 2005) of the HapMap data was the first to contain high density genotyping on all 4 populations across the 22 autosomes and the X chromosome. These data were downloaded from the bulk download page of the HapMap website (www.hapmap.org). A "filtered non-redundant" file was downloaded for each chromosome. The files were converted to the .dat file format required by LDMAP and were further filtered to remove Hardy-Weinberg deviations of >10 chi squared and rare SNPs of <5% minor allele frequency.

Table 3.2 The number of SNPs used for map construction for each population.

Population	No. Individuals	No. SNPs over the genome
YRI (Yoruba in Ibadan)	60 founders	783,366
CEU (CEPH Utah residents with ancestry from northern and western Europe)	60 founders	756,065
CHB (Han Chinese in Beijing)	45 unrelated	673,232
JPT (Japanese in Tokyo)	44 unrelated	667,370

The number of SNPs decreases as the chromosome size reduces, with an average of approximately 200-250 SNPs per Mb (1 SNP per 4-5 Kb).

Figure 3.1 Number of SNPs per chromosome.



3.3.2 LDU map creation

A new version of the LDMAP program, LDMAP+, was used to create LDU maps of the whole genome for all 4 populations. The LDMAP+ program creates LDU maps from pairwise data using the Malecot model and composite likelihood as previously described. The maps are created in segments from intermediate files that are also created from segments of the genotype data file. Making maps in segments and reassembling the segments to form a whole map dramatically reduces the computational load caused by handling large files. LDMAP+ requires the user to input the segment size and the overlap between segments. The default values of 1000 loci per segment and an overlap of 25 loci were used for the map creation in this case. A small overlapping section at the boundaries of each segment ensures that there is no "end effect" at each rejoin point. The LDU distance for each SNP interval is calculated from information from surrounding SNP intervals. An "end effect" occurs when there is little or no information about a SNP interval on one side because the end of the region is reached. Therefore there is less information at either end of the region compared to the middle. Further options, which determine how many surrounding SNP intervals are taken into consideration when fitting the Malecot model to each SNP pair, are also required. The defaults of 500Kb maximum distance and 50 intervals were used in this case. Pairwise data included from pairs separated by large distances are essentially uninformative because LD has declined to background levels. Including such pairs adds no information but the computational load increases dramatically.

3.3.3 Fitting the data to the finished map

LDMAP+ creates LDU maps in segments and therefore fits the Malecot model to each segment at a time to maximise the likelihood. The data can then be fitted to the completed LDU map of the whole chromosome. Malecot parameters and genome LDU map lengths can be compared across populations for the whole genome and also investigated chromosome by chromosome.

3.3.4 Comparing LDU maps across populations

Using the completed LDU maps the structure of LD across the 4 populations was assessed using a linear regression of LDU values for SNPs common to all four populations. Regression of each population against each of the other populations gives 6 comparisons. Regression analysis requires a value for each unit of analysis to carry out the highest resolution analysis, SNPs common to all 4 populations were used. A high number of SNPs are common to all 4 populations ~70%, across all chromosomes.

3.3.5 Comparing the LDU and linkage map and calculating Effective Bottleneck Time (t)

The linkage map is not known to vary across populations and it is not influenced by stochastic historical effects since it reflects only 1-2 generations. The linkage map is of lower resolution and generally markers do not extend as far towards the telomeres and the centromere as in the LDU maps. However, comparisons were made over the shared regions. The LDU maps and linkage maps were split into units of analysis of chromosome arms (n=41) and at a higher resolution deciles (chromosome arm/10 on the Kb scale) (n=410). A regression of LDU against Morgans (w), weighted for size by Morgans, allows the linkage and LDU maps to be compared and gives the effective bottleneck time in generations (t).

3.3.6 Comparing values of t across chromosomes

The release 11 HapMap data on the CEPH sample showed a trend for a smaller t value on the smaller autosomes. Since the chromosomes have the same history since the last major bottleneck the value of t is expected to be constant across autosomes within a population. This is investigated using the chromosome arm values of t (LDU/Morgans) and the chromosome arm length described as Megabases per Morgan. The X chromosome is a special case because, apart from the pseudo-autosomal regions, it does not recombine when it is in the XY state in males. For this study the pseudo-autosomal regions were removed prior to creating the LDU maps. Since the LDU map represents recombination over many generations, a third of the time the X chromosome has been in a non-recombining state in males. The X chromosome LDU map was multiplied by 3/2 extending the map by a half to account for this lack of recombination and make the X chromosome comparable to the autosomes, when calculating t.

3.3.7 Fine-scale differences between populations

Genome-wide LDU maps could be used for detecting signatures of selection. A selective sweep is a reduction in variation over a region due to recent positive selection. Neutral variation surrounding the selected gene is lost due to the over representation of the haplotype carrying the positively selected variant. Positive selection in one population but not another (i.e. recent positive selection) would be evident as a difference in LD structure in the selected region. This difference would show as a lack of variation represented by a block of high LD in one population and an LDU "step" in the other. To test this theory a large inversion discovered on chromosome 17 and published in 2005 by DeCODE Genetics (Stefansson et al. 2005) was investigated. The inversion created two distinct lineages called H1 & H2. H2 is common in Europe but found in only ~10% of

Africans. The predicted old age of the H2 lineage and the fact that it is very homogeneous, suggests that it is under positive selection in Europeans. The LDU maps of the YRI and CEU populations were compared in windows of size 500Kb with a 250Kb slide by a ratio of the LDU/Mb in each window.

3.4 Results

3.4.1 Properties of the finished maps

The properties of the LDU maps were compared across population samples for the genome-wide maps and also by chromosome.

Table 3.3 Properties of the genome-wide LDU maps.

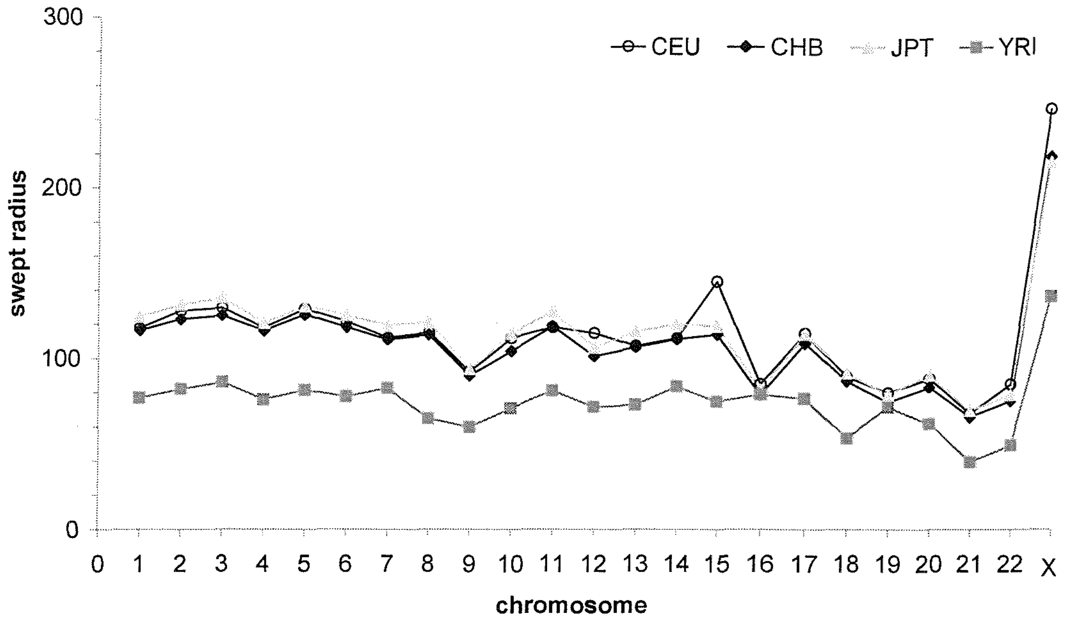
	Av. Swept radius (Kb)	Av. epsilon	Av. L	Av. M	Total LDU length	Total No. holes
CEU	114.133	1.040	0.155	0.974	56250	2911
CHB	107.988	1.030	0.176	0.976	62687	4879
JPT	114.024	1.036	0.177	0.975	56656	3731
YRI	73.996	1.111	0.165	0.908	79499	2958

*Values for individual chromosomes in appendix 1.

Over the whole genome, the YRI sample has the shortest swept radius, longest map, largest epsilon and the smallest M value. The CHB sample has the most holes overall. Values for each chromosome are given in appendix 1.

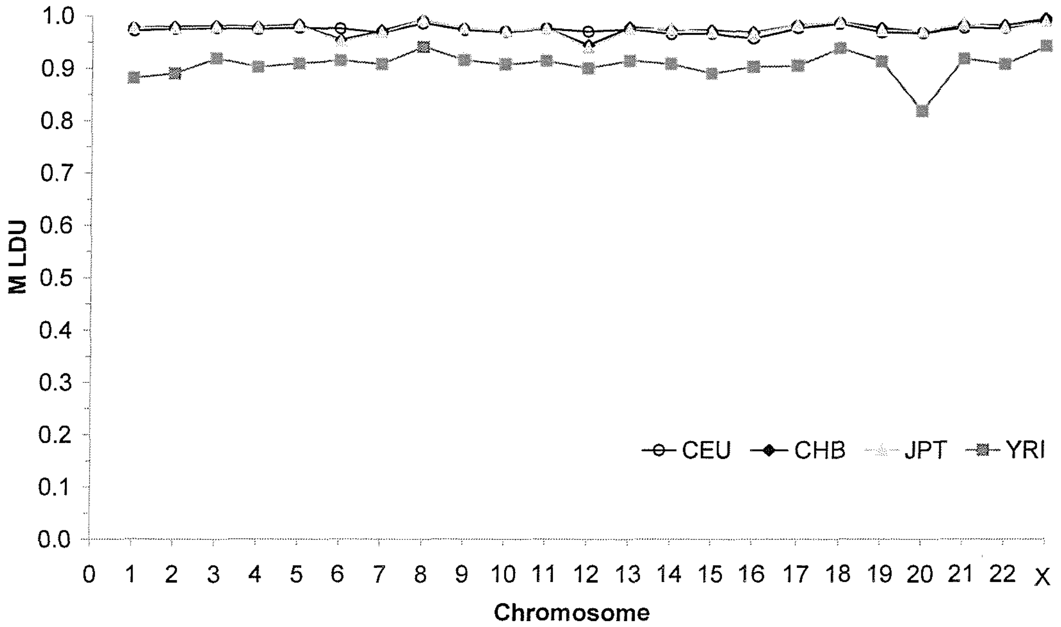
The swept radius reduces slightly towards the smaller chromosomes. The X chromosome has an increased swept radius (figure 3.2).

Figure 3.2 The swept radii by chromosome.



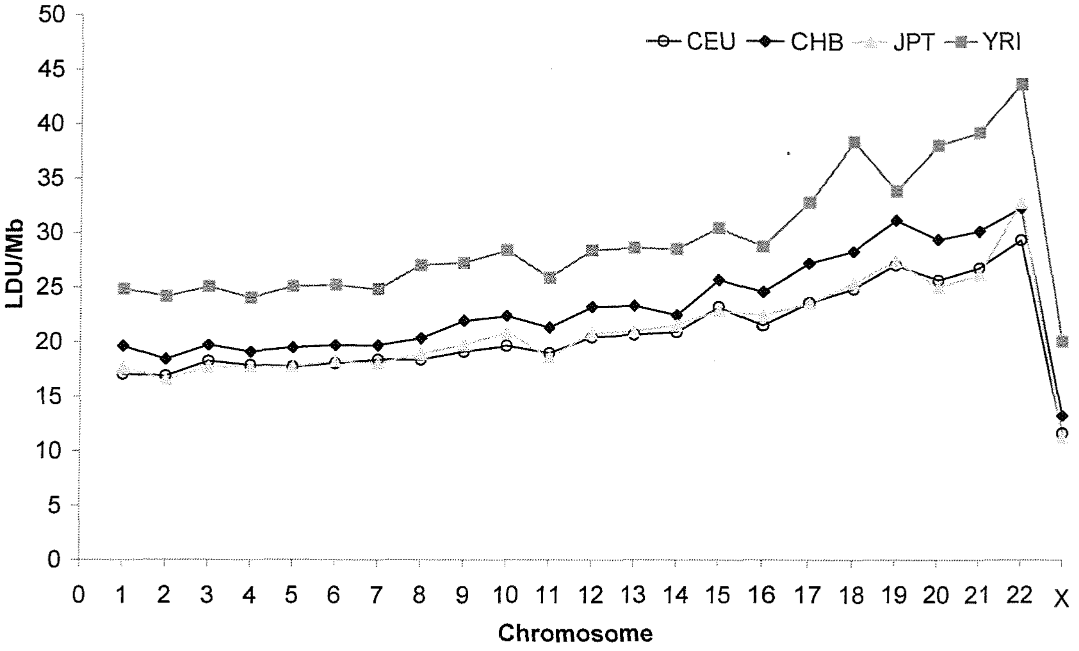
M approaches 1 in all cases but is consistently smaller in the YRI population which is consistent with a larger number of founding haplotypes (polyphyletic origin). There is no trend across chromosomes; however, chromosome 20 in the YRI population is smaller than might be expected, but the difference is small (figure 3.3).

Figure 3.3 The M parameter in the Malecot model across chromosomes.



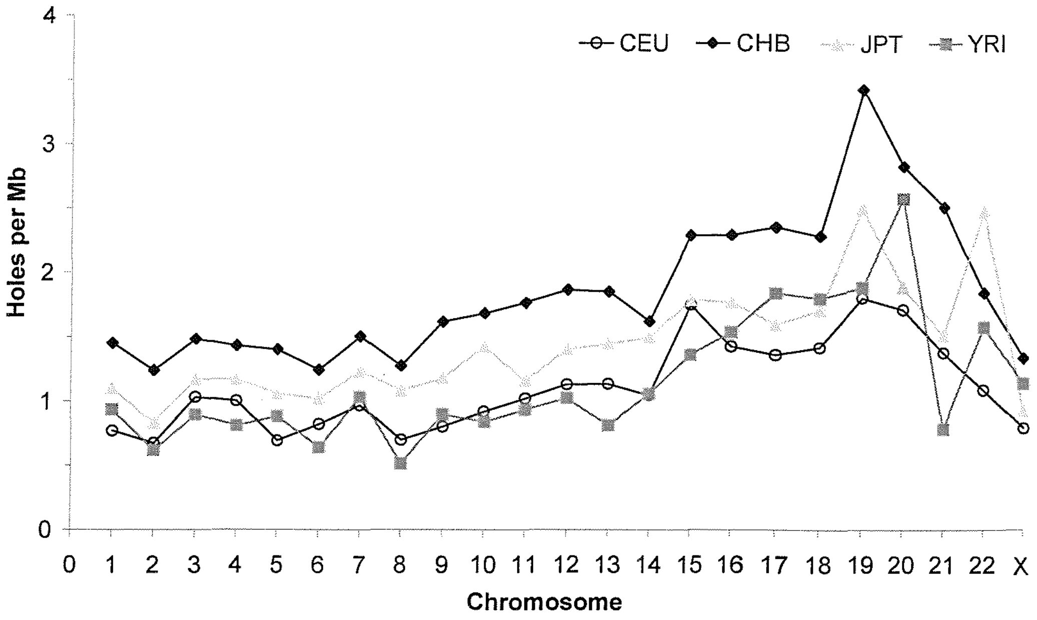
LDU/Mb is higher in the YRI population, there is also a trend for higher LDU/Mb values in the smaller chromosomes. This is consistent with the reported higher recombination intensity due to size dependent control of meiotic recombination (Kaback 1996). The X (23) chromosome has a reduced LDU/Mb measure (figure 3.4).

Figure 3.4 LDU/Mb by chromosome.



There are more holes in the larger chromosomes presumably due to increased length. However, when the length is taken into account as here, holes per Mb show a trend for greater hole density in the shorter chromosomes, perhaps due to the increased recombination in these chromosomes which require more SNPs to resolve the holes. The CHB population generally has more holes than the other populations, this may reflect differences in marker spacing in critical regions.

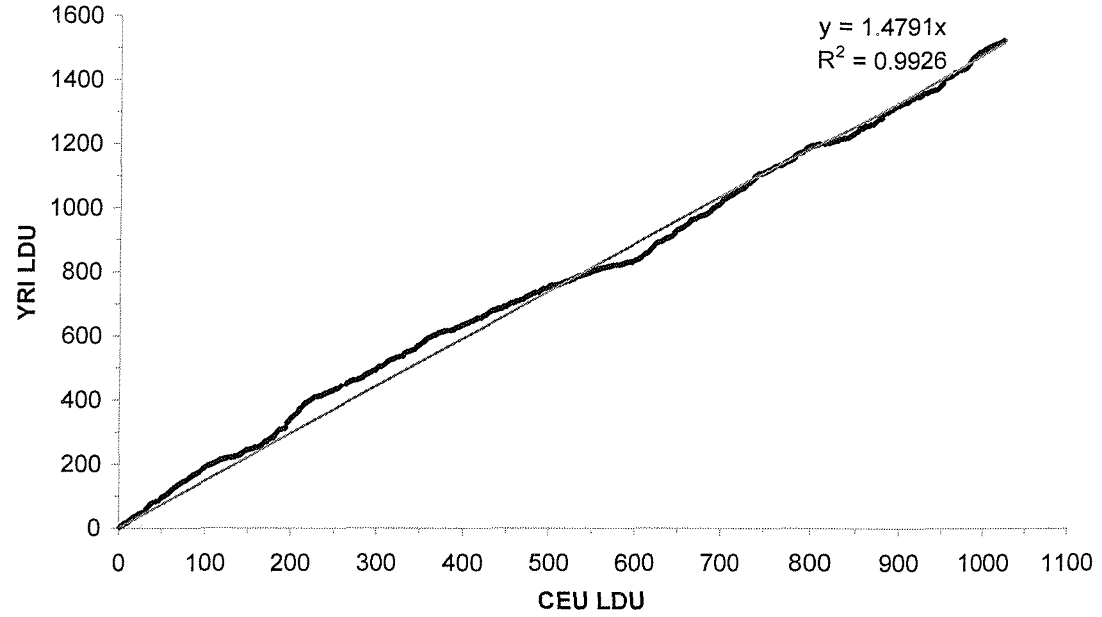
Figure 3.5 Holes per Mb by chromosome.



3.4.2 Comparing LDU maps across populations

LDU maps were made using all available SNPs however not all SNPs are present in all populations. To enable comparison only the SNPs that were common to all four populations were selected. A linear regression between the YRI and CEU LDU locations of SNPs common to both maps, for chromosome 22, gives an R^2 value of 0.9926 and a regression co-efficient of 1.4791 showing the YRI map to be 1.4791 times longer than the CEU map (figure 3.6). R^2 values for all other population comparisons and other chromosomes ranged 0.9926-1 showing an extremely high correspondence.

Figure 3.6 Chromosome 22 CEU LDU versus YRI LDU.



3.4.3 Comparing the LDU map with the linkage map and calculating Effective Bottleneck Time (t)

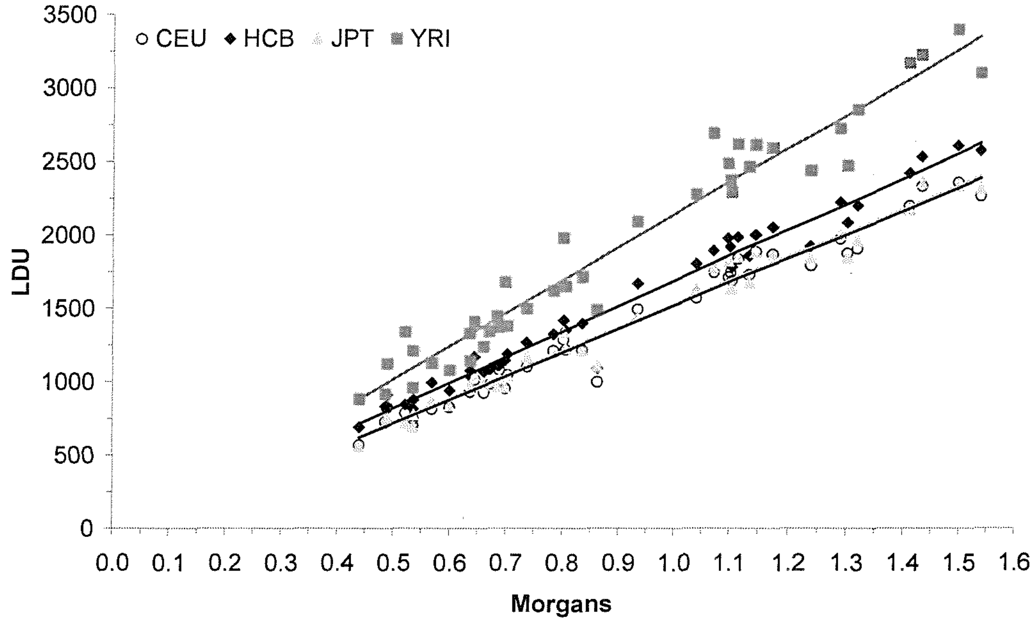
There is a high correspondence between the linkage map in Morgans and the LDU maps of chromosome arms. A regression analysis shows that 99% of variation in the LDU map is explained by the linkage maps for chromosome arms and 97% in deciles (figure 3.7). These values are calculated as the average of the results for each population sample, the values are consistent across populations. A ratio of LDU and Morgans (w), weighted by Morgans, gives the effective bottleneck time in generations (t) (table 3.4).

Table 3.4 Effective Bottleneck time (t) for each population based on the whole genome.

Population	Effective bottleneck time (t=LDU/Morgans)	Time in years with 1 generation =25 years
YRI	2073	51825 (41460)
CEU	1472	36800 (29440)
CHB	1648	41200 (32960)
JPT	1483	37075 (29660)

*values in brackets represent a generation time of 20 years.

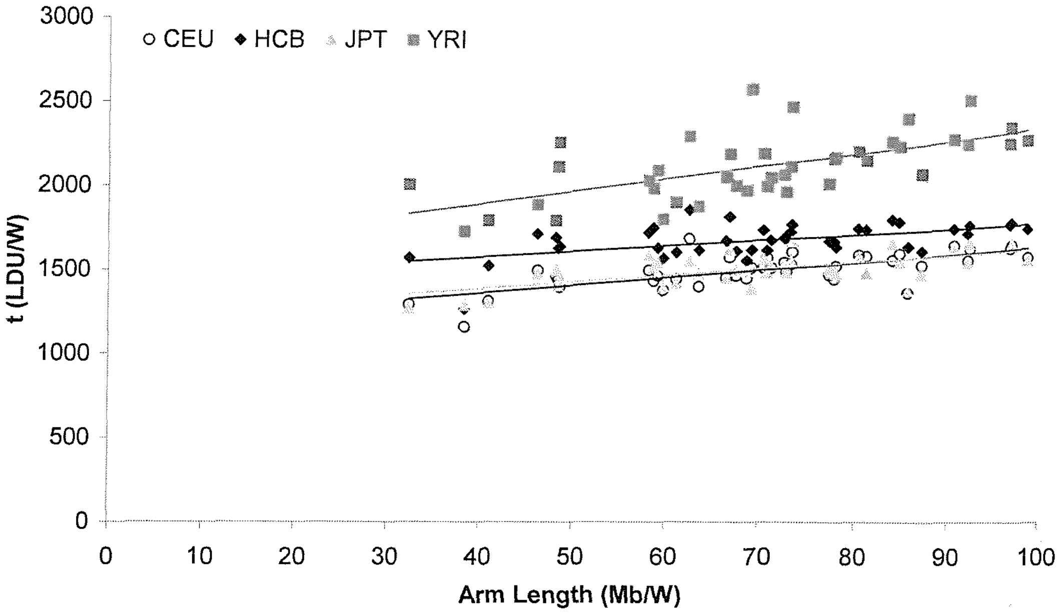
Figure 3.7 Graph of Morgans versus LDU over all 4 populations (each data-point representing a chromosome arm, n=41)



3.4.4 Comparing values of t across chromosomes.

The autosomes have undergone the same history since the last major bottleneck in terms of recombination; therefore the value of t should be constant across chromosomes within a population. However, there is slight variation in t across chromosomes, with a trend for t to be larger in the larger autosomes, this is consistent across population samples (figure 3.8).

Figure 3.8 t values across chromosome arms (W=Morgans).



3.4.5 Fine-scale differences between populations.

To look at the possibility of using genome-wide LDU maps for detecting signatures of selection, a previously published example of a large inversion thought to be under selection on chromosome 17 was investigated (Stefansson et al. 2005). The LDU maps of the YRI and CEU populations were compared in windows of size 500Kb with a 250Kb slide using a ratio of the LDU/Mb in each window between the two populations. When these ratios are plotted against the Kb scale the 900kb region of the inversion, assumed to be under selection, is clearly identified as a peak (figure 3.9). A more detailed look at the region identified by the peak shows that the YRI LDU map has a step where the CEU map has a block (figure 3.10).

Figure 3.9 Ratios of CEU and YRI LDU/Mb across chromosome 17, in 500 Kb windows. The dashed vertical lines indicate the location of the 900Kb inversion.

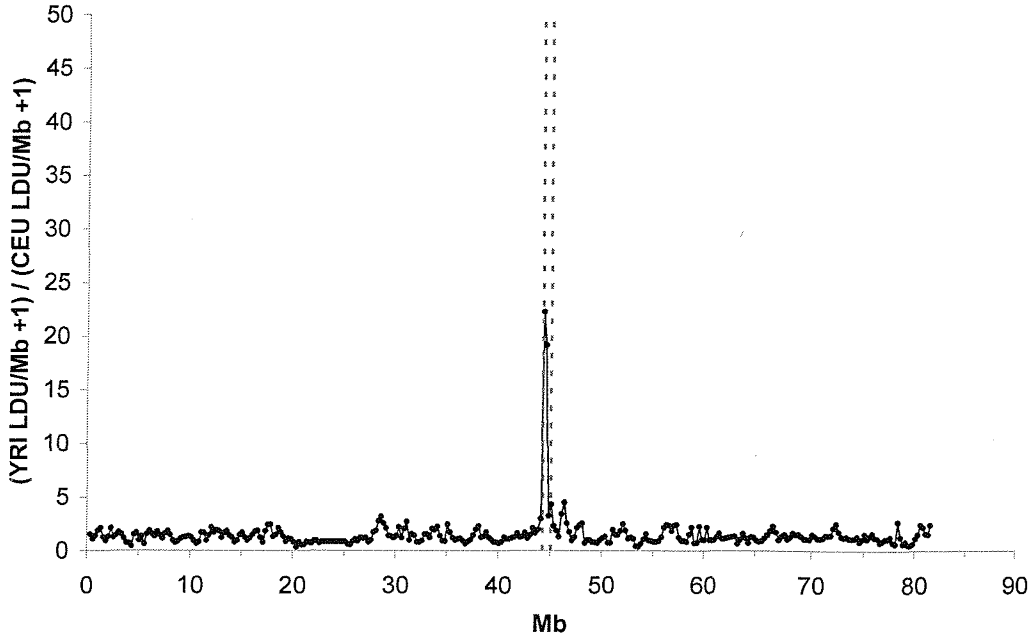
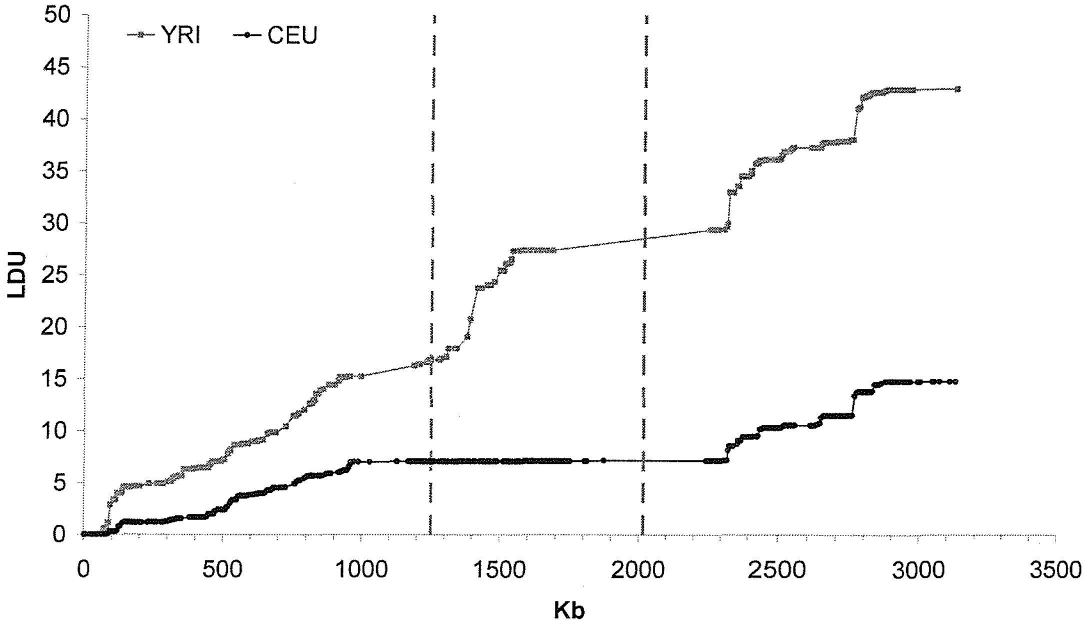


Figure 3.10 LDU maps of the 3.2 Mb region around the peak for the YRI and CEU samples, the dashed lines show the location of the inversion.



3.5 Discussion

LDMAP+ and the genotype data produced by the International HapMap project has allowed genome-wide LDU maps in 4 populations to be created and analysed. The properties of the LDU maps were considered, across the whole genome and by chromosome. The maps vary in LDU length between populations, with the YRI sample having the longest map overall and for each chromosome. This is the same trend as published on chromosome 20 data and other samples (Gibson et al. 2005; De La Vega et al. 2005) with the sample with African ancestry having the longest LDU map. This is consistent with the YRI population being older in terms of time since the last major bottleneck, allowing more recombination to accumulate extending the length of steps in the map creating a longer map overall. The overall length of the CHB genome LDU map is intermediate between the YRI population and the JPT and CEU populations (which are very similar in length), though it may be expected to be more similar to the CEU and JPT maps. The total number of holes in the LDU maps is small relative to the number of SNP pairs considered. Holes are partly due to a lack of information at particular SNP intervals, i.e low local SNP density. If there is insufficient information to keep the map intact, the upper limit of 3 LDUs is given. With high density data such as used here, a value of 2.5 LDUs or greater is considered a hole. The number of holes is also affected by recombination hotspots, there is evidence that holes occur in parts of the genome with particularly high recombination rates (Tapper et al. 2003). The CHB population has the largest number of holes overall, and this trend is also present when the maps are considered chromosome by chromosome. If holes are due to low local SNP density they generally lengthen an LDU map. The excess number of holes in the CHB population may explain why the CHB LDU map lengths are longer than those of the CEU and JPT populations. Accounting for chromosome length, there is a trend for more holes per Mb in the smaller chromosomes. The more intense recombination on these chromosomes as shown by LDU/Mb may account for this. It is also possible that the higher number of holes may be the cause of the higher LDU/Mb on the smaller chromosomes due to artificial lengthening of the maps. However this data is of high and relatively consistent SNP density across chromosomes, therefore it is likely that the holes represent regions of high recombination and not insufficient local SNP density particular to the smaller chromosomes. Although a very high SNP density may resolve a hole in a recombination hotspot, it is also possible that some holes may remain even when all SNPs are typed (Tapper et al. 2005).

The M parameter in the Malecot model has an evolutionary interpretation and shows the amount of LD at the last major bottleneck the value is approaching 1 in the "Out of Africa" populations, and slightly lower in the YRI population consistent across chromosomes. The lower value of M in the YRI population shows its older history and polyphyletic origin, in that recombination had already begun to accumulate in this population at a time when the other populations were going through a bottleneck. The average epsilon values across populations are very similar in the CEU, JPT and CHB samples but larger in the YRI sample, which shows a more rapid decline of LD over distance. The swept radius is calculated as 1/epsilon so it follows that the YRI population has the smallest swept radius showing less extensive LD as compared to the other samples. LDU/Mb is also larger in the YRI population, again showing the lower amount of LD in that population compared to the others. There is a small amount of variation in swept radius across chromosomes with a smaller swept radius in the smaller chromosomes. This indicates that LD extends less in the smaller chromosomes, which is consistent with the known increased recombination in the smaller chromosomes (Kaback 1996). This is also shown by the larger LDU/Mb in the smaller chromosomes. The particularly high swept radius and low LDU/Mb shown on the X chromosome shows that there is more extensive LD on this chromosome. This is unsurprising given that the X chromosome is unable to recombine when it is in males in the XY form. Any SNPs present in the recombining pseudo-autosomal regions were removed from the data prior to creating the LDU map. LD on the X chromosome is therefore broken up by recombination at a much reduced rate (only when present in females) compared to the autosomes.

The whole chromosome LDU maps were investigated to compare the LD structure across populations. The amount of LD varies across populations, with more LD in the younger populations where recombination has yet to break it up. The broad patterns of LD, however, are shared by populations and shaped by the co-localisation of recombination hotspots (Gibson et al. 2005; De La Vega et al. 2005). This is the premise that allows construction of a cosmopolitan or standard LDU map that can be scaled to be applied to various populations (Gibson et al. 2005). Due to computational constraints, it was not feasible to construct and compare cosmopolitan maps by the method of combining haplotype counts across populations and creating LDU maps from the combined data. However, a linear regression of one population against another, using the LDU values for all SNPs common to both populations, was carried out to determine the similarity of LD structure. Figure 3.6 shows the regression line for chromosomes 22 with YRI LDUs against CEU LDUs. The regression coefficient (1.479) shows the relative scale of the YRI map to the CEU map. This is consistent with the value of 1.43 between the African American and the European samples calculated for the chromosome 20 data presented in Chapter 2 (Gibson et al. 2005). The high R^2 value (0.993) shows the remarkable similarity of LD structure across populations, even between African and non-African populations. This similarity was consistently high across chromosomes. Removing SNPs that were not present in all populations for this analysis will have removed some of the variation between populations. However, all SNPs were used to create the maps and thus SNPs that were removed for this purpose would still have had an effect on the LDU values of surrounding SNPs. The total number of SNPs removed for this purpose was relatively small (~30%) and the SNP density remained high. It is possible that the extremely high R^2 values are inflated, but only by a small amount.

The linkage map measures recombination, which is the main force in defining LD patterns. The linkage map and the LDU maps for all samples were compared. The linkage map, made using CEPH family data, does not vary across populations because it does not measure historical recombination. It can therefore be used to compare current and historical recombination in all population samples. Due to the low resolution of the linkage map in comparison to the LDU map, larger units of analysis were required. The genome was split into chromosome arms and, for higher resolution analysis, deciles (chromosome arm/10 on the Kb scale). It has been shown previously that LDU maps correspond well to linkage maps of the same region, since recombination is the major influence on LD (Zhang et al. 2002; Tapper et al. 2003). In this case linear regression showed that 99% of the variance in LDU across the data is explained by recombination as shown by the linkage map when chromosome arms are analysed (figure 3.7). The corresponding value when chromosome deciles are analysed is 97%. These values are the average over the 4 population samples which showed very similar and consistent results.

An additive LDU map has great value in association mapping, and can also give insights into population history. Linkage maps show current recombination over a single generation, whereas LDU maps show recombination over many generations since the last major bottleneck. Using this relationship it is possible to make an estimate of "effective bottleneck time" (t) in generations. This is a measure of time since the last major bottleneck taking into account the cumulative effects of subsequent bottlenecks. Estimates of t were calculated for each of the 4 populations. Using the data for chromosome arms, t = 51,825 years for the YRI population and 36,800 years for the CEU population, 41,200 years for the CHB population and 37,075 years for the JPT population (assuming a generation time of 25 years) (table 3.4). The CEU value compares well with the previous estimate of 35,857 years, from the lower density HapMap data (release 11) (Tapper et al. 2005). The value of t is expected to be consistent across chromosomes, since the chromosomes have undergone the same history in terms of opportunity for recombination. However, there is a slight trend for t to be larger in the larger autosomes consistent across samples and previously noted in the CEU population (Tapper et al. 2005). The phenomenon of interference occurs when one crossover event at meiosis inhibits the presence of another in close proximity. The linkage map is constructed with a single genome-wide measure (Kosambi function) to account for interference on all chromosomes, although the effect is more pronounced on the smaller chromosomes (shown in mice but not conclusively in humans) (Broman et al. 2002). It is possible that this leads to the inflation in size of linkage maps in the smaller chromosomes. This difference may account for the small variation in t seen here. The X chromosome spends a third of the time in males and this recombination difference was accounted for when calculating t. The t value in the X chromosome arms was slightly below the value expected in the analysis of the CEU population using release 11 HapMap data (Tapper et al. 2005), although this deviation is not seen here. The higher density of the Phase I data may explain this discrepancy.

Lastly the fine-scale differences in LDU maps between populations were investigated for a specific case of a large inversion, under selection, discovered on chromosome 17 (Stefansson et al. 2005). The inversion has created two distinct lineages known as H1 and H2. H2 is only found in ~10% of Africans but is much more common in Europeans. The H2 lineage is predicted to be older and has been shown to be very homogeneous, an explanation for this is that H2 is under positive selection in Europeans. A difference in selection, in a specific region, between populations would result in different LDU patterns at that location, due to reduced variation around the selected locus. The LDU maps of the YRI and CEU populations were compared in windows of 500Kb (with a 250Kb slide) by a ratio of the LDU/Mb in each window. When these ratios are plotted against the Kb scale the 900Kb region of the inversion, assumed to be under selection, is clearly identified as a peak on the graph (figure 3.9). The peak is caused by a large block of high LD in the CEU population where there is a step in the YRI population, suggestive of a selective sweep in the CEU population (figure 3.10). The peak could be a signature of selection in one population, but it is also possible that it could be a result of the inversion itself. Structural variants that change the location of SNPs relative to one another, such as inversions, may have unpredictable effects on the LDU map. Genome-wide LDU maps provide an opportunity to scan the whole genome for other similar peaks, but when analysing such large data sets many peaks will appear by chance and there is a multiple testing problem to overcome when applying a significance level to this type of analysis. Also using windowing of data requires arbitrary definitions of window size and of any overlap, both of which will have an effect on the size of the regions and the minimum differences in LDU/Mb detectable. However, this striking result shows that fine-scale comparison of LD patterns by comparing genome-wide LDU maps may reveal evidence of selection, though results must be carefully interpreted.

3.6 Conclusions

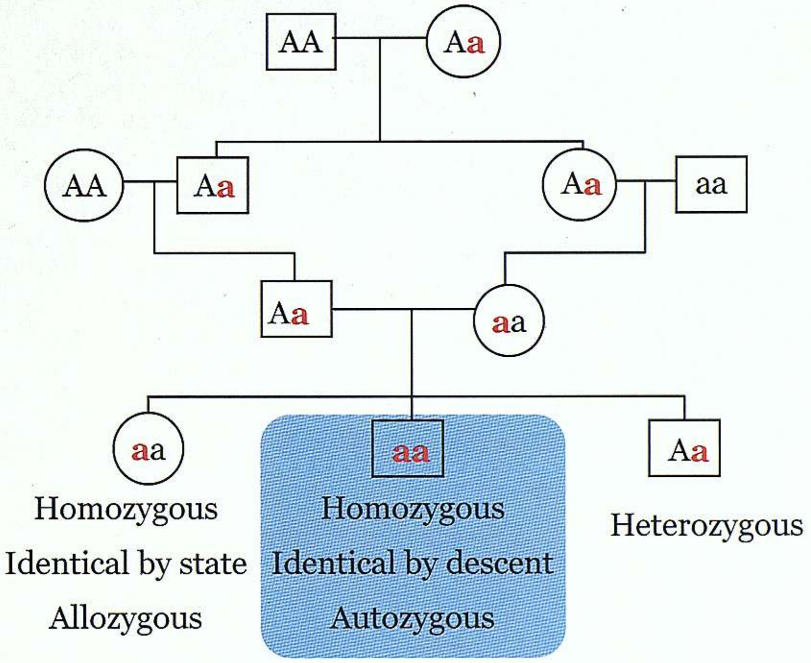
Genome-wide LDU maps were created from the HapMap Phase I high resolution data on 4 population samples. The LDMAP+ program is able to handle such high density data and produce good quality maps in reasonable time. The properties of the maps follow previous analyses on the CEU sample and preliminary data on 4 chromosomes from an early HapMap release. These results show the similar broad structure of LD across the whole genome, backing up the results on chromosome 20 presented in chapter 2. The genome-wide LDU maps allow values of effective bottleneck time (t) to be estimated allowing estimates of population age. The fine-scale difference in LD patterns between populations, detected by comparison of LDU maps, shows a possible novel method for detection of selective sweeps.

Chapter 4 – Extended tracts of homozygosity in outbred populations

4.1 Introduction

An individual has two sets of chromosomes, one from their mother and one from their father, and therefore has two alleles at each locus on the autosomes. The two alleles can be different, called heterozygous, or they can be the same, called homozygous. There are two types of homozygosity; when the two alleles are identical by state but arise from two different sources this is called allozygous, or when the alleles are identical by decent and thus come from the same source this is called autozygous (figure 4.1). Autozygosity will not affect just a single marker but will extend to neighbouring markers on the chromosomal background that is inherited. This results in some individuals having long tracts where homozygous markers occur in an uninterrupted sequence.

Figure 4.1 Pedigree illustrating allozygous, autozygous and heterozygous.



Since recombination interrupts long chromosome segments over time, the length of a homozygous segment depends in part on the time since the last common ancestor of the parents, the source of the chromosomal segment that is identical by decent. It is therefore expected that longer tracts of homozygosity would be found in inbred populations as opposed to outbred populations. However, long tracts of homozygosity have been recorded previously in CEPH individuals (Broman and Weber 1999). 8,000 short tandem-repeat polymorphisms (STRPs) in CEPH families were analysed and several families with long homozygous segments exceeding 10 centiMorgans (cM) in length were identified. The authors examined the roles of possible typing error, back mutation of STRPs, gene conversion events and the limitations imposed by locally low marker density in determining the limits of homozygous segments. In some pedigrees they were able to determine relationships between apparently unrelated individuals, but there remained a degree of autozygosity approaching or exceeding that expected in the progeny of a first cousin mating where relationships were not detected.

The International HapMap project, which provides very densely genotyped Single Nucleotide Polymorphism (SNP) markers across the whole genome in 4 outbred populations, provides an ideal opportunity to investigate tracts of homozygosity. Single nucleotide polymorphisms are thought to be of more ancient origin than STRPs. We might therefore expect to see, in comparison, fewer and shorter homozygous tracts in SNP maps. However, this is partly offset by the relatively reduced mutation rate of these markers which might allow the longer tracts to remain unbroken over more generations. In addition to autozygosity another explanation for long tracts of homozygosity is that they appear in parts of the genome where there is relatively little recombination, due to high linkage disequilibrium (LD). LD is the tendency for alleles to be inherited together more often than would be expected under random segregation. In the human genome there are regions of strong LD broken up by small regions of intense recombination (Jeffreys, Kauppi, and Neumann 2001). Blocks of LD represent regions of the genome where a small number of haplotypes account for most of the variation. An individual inheriting two copies of a common haplotype in a particular location would be homozygous over that region. LD as represented by LDU maps (Maniatis et al. 2002) shows the contour of the LD patterns and identifies steps which correspond to regions of high recombination and plateaus which reflect recombination cold regions (high LD) (Zhang et al. 2002). LDU maps of the 22 autosomes and the X chromosome based on the Phase I HapMap data have been constructed using the LDMAP+ program. These maps are described in chapter 3 and available in the Linkage Disequilibrium DataBase (LDDB) (Genetic Epidemiology and Bioinformatics group 2008). They can be used to investigate the extent to which high LD corresponds to tracts of homozygosity.

Extended tracts of homozygosity in a particular region of the genome, common among individuals within a population, may indicate a selective sweep. An example of a gene suggested to be under positive selection is the lactase gene on chromosome 2. The *LCT* gene encodes the enzyme lactase-phlorizin hydrolase. There is a great deal of epidemiological data in favour of recent positive selection at this locus. The ability to use this enzyme to digest lactose during adulthood varies dramatically across worldwide populations, with particularly high rates among northern Europeans. A high rate of lactase persistence in European populations can be explained by positive selection resulting from increased nutrition from dairy, the only dietary source of lactose; and the geographic distribution of lactase persistence matches the distribution of dairy farming (Bersaglieri et al. 2004).

Various factors may influence the length, abundance and location of homozygous tracts including, mutation rate, population structure, uniparental disomy (UPD), natural selection, recombination, and linkage disequilibrium patterns. The extremely dense SNP genotyping in the HapMap sample allows examination of the distribution, size and location of homozygous tracts and their relationship to recombination and linkage disequilibrium patterns and also consideration of other mechanisms.

4.2 Aims

Relatively short segments of homozygosity in the apparently outbred HapMap populations would be expected, and longer tracts would be expected to be uncommon and restricted in length. The aim of this work is to characterise extended tracts of homozygosity (>1Mb) in the human genome as represented by the HapMap data and examine the relationship between the location and size of long tracts of homozygosity and the role of recombination and linkage disequilibrium patterns; and also examine evidence for recent inbreeding having a role in the formation of long tracts of homozygosity in some HapMap individuals.

4.3 Methods

4.3.1 Data

The data examined were produced by the International HapMap Consortium and released into the public domain via their website (International Hapmap Group 2005). Phase I of the HapMap data provides over a million SNPs genotyped in 209 unrelated individuals; 60 CEPH Utah residents with ancestry from northern and western Europe (CEU), 45 Han Chinese from Beijing (CHB), 44 Japanese from Tokyo (JPT) and 60 Yoruba from Ibadan, Nigeria (YRI). The average SNP density is 1 SNP every 5kb (The International HapMap Consortium 2003).

The HapMap Phase I data (non-redundant files that have passed quality control) were downloaded from the HapMap website and include 3,970,277 genotypes across the 22 autosomes over all four populations. These data were further filtered to remove genotypes with significant deviation from Hardy-Weinberg (Chi square >10) and minor allele frequencies below 0.05. The total number of genotypes from each population were 728,353 genotypes from the CEU sample, 744,006 genotypes from the YRI sample, 644,060 genotypes from the CHB sample and 639,460 genotypes from the JPT sample.

4.3.2 LDU maps.

Genome-wide LDU maps were constructed for the 4 populations of the HapMap data following Maniatis *et al.* (2002) and Tapper *et al.* (2005). LDU maps were constructed, using the LDMAP+ program, from multiple pairwise association data using a model which describes the decline in association, ρ, with distance: $\rho= (1\text{-}L)Me^{-\varepsilon d} + L$. The LDU distance is calculated as $\varepsilon_i d_i$ for each interval i of d kilobases between a pair of SNPs and LDU locations are computed by summation over intervals. The additive maps produced are available from LDDB and described in chapter 3.

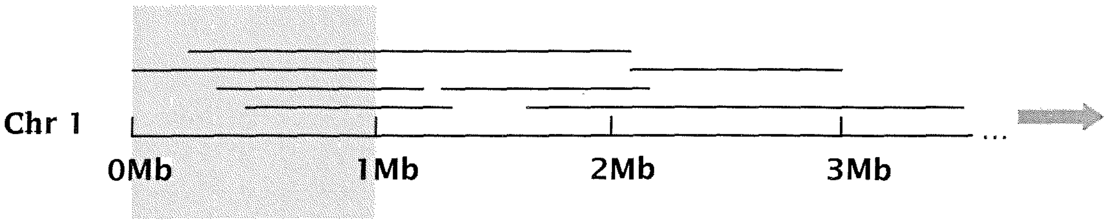
4.3.3 Definition of extended homozygous tracts.

The genotypes were coded 11, 12, or 22 with 11 and 22 being the homozygotes. For each individual starting from the p telomere of chromosome 1 each SNP was identified as either homozygous or heterozygous. An extended homozygous tract was defined as an uninterrupted sequence of homozygous SNPs spanning at least 1Mb in a single individual. For each extended homozygous tract, the starting SNP and kb location, the ending SNP and Kb location, the number of SNPs it contained and the starting and ending LDU locations were recorded. SNPs with missing data ('NoCall's) were ignored. The average SNP density across all populations is approximately 1 SNP every 5 Kb. Since a locally low SNP density may artificially extend a homozygous tract, tracts with an average SNP density of less than 1 SNP per 5Kb (200 SNPs per Mb) were excluded. Also omitted were the centromeric regions and acrocentric p-arms for the same reason.

4.3.4 Examining the extended homozygous tracts

The number of tracts (>1 Mb) and maximum tract length were determined for each population sample. To determine the relationship between the amount of LD and the number of homozygous tracts present, for each sample, the whole genome was analysed in 1 Megabase segments. Each chromosome was split into 1 Mb segments, the remaining shorter segments at the end of each chromosome were also included. The LDU/Mb ratio was calculated for each 1 Mb segment using the LDU map for that population, and the mean tract coverage in kilobases obtained. For each population this was computed by summation, over all individuals, of the length (Kb) of homozygous tracts covering each 1 Mb segment and dividing by the number of individuals (figure 4.2).

Figure 4.2 Calculating tract coverage for each 1Mb segment in turn.



To determine whether the location of homozygous tracts is related to LD structure the correlation between tract coverage and LDU/Mb was obtained and linear regression was performed using LDU/Mb as the dependent variable and tract coverage as the independent variable. The units of analysis were 1 Mb segments of the genome. Analyses were carried out for each population sample separately and for the concatenated sample. To confirm that the location of homozygous tracts is directly related to the recombination pattern the same analysis was carried out but with the linkage map in cM/Mb replacing LDU/Mb. This is important because, although at a lower resolution, the linkage map is based on an entirely independent sample, whereas the structure of the LDU map created from the HapMap data must partly reflect the presence of homozygous tracts in that sample. The linkage map used (Kong et al. 2004) comprises 14,759 polymorphic markers.

A correlation analysis between tract coverage values, in each megabase, for all population pairwise combinations allowed assessment of whether the distribution of tracts across the genome was similar in all populations. i.e. If a region of the genome with a large number of tracts in one population also had a large number of tracts in the other populations.

Next the frequency of tracts for each individual was observed, and the average tract counts, per individual, for each population were calculated. The amount of LD in regions where homozygous tracts occur was investigated for each individual and compared to the genome-wide average LDU/Mb. Three individuals were highlighted as having more numerous and longer tracts than others in their respective population samples. To examine if the LDU/Mb in the homozygous regions of these three individuals is significantly different from the levels in other individuals from the same sample, a regression model weighted by physical size in Mb was used. LDU/Mb was the dependent variable and x was the independent variable with x=1 for individual NA12874 and x=0 for other individuals in the CEU sample. The same model was used with 2 variables (x, x1) for the 2 outliers in the JPT sample.

To gain a preliminary look at the relationship between homozygosity and selection the locations of tracts were investigated. The numbers of individuals (CEU sample) with a tract in each 1Mb segment across chromosome 2 were plotted against the physical map. Chromosome 2 was chosen as it contains the LCT gene which has been shown to be under selection, and positively selected in Caucasian populations (Bersaglieri et al. 2004).

4.4 Results

Across the four populations a total of 1393 homozygous tracts met the criteria defined in the methods section. The longest tract over all populations was 17.9 Mb in an individual from the JPT sample (table 4.1). This tract comprises 3922 consecutive homozygous SNPs.

Table 4.1 Number and maximum length of homozygous tracts identified.

HapMap Population Sample	No. Unrelated Individuals	No. Tracts	Max. Length Mb
CEU (CEPH Utah residents with ancestry from northern and western Europe)	60	498	6.48
CHB (Han Chinese Beijing)	45	263	2.63
JPT (Japanese Tokyo)	44	370	17.91
YRI (Yoruba Ibadan Nigeria)	60	262	11.14
ALL	209	1393	17.91

Correlation and linear regressions between LDU/Mb and tract coverage allowed the relationship between homozygosity and LD structure to be determined. Analyses were carried out for each population sample separately and for the concatenated sample. All of the correlations were significant at $p<0.0001$ with correlation coefficients of around -0.3 for all samples. The regression analyses were also all significant ($p<0.0001$) with R^2 values ranging from 8-10% (table 4.2).

Table 4.2 Correlation and regression between LDU/Mb and tract coverage.

Population sample	Correlation Coefficients	Regression R^2
CEU	-0.32	0.10
CHB	-0.29	0.09
JPT	-0.30	0.09
YRI	-0.28	0.08
ALL	-0.29	0.08

The same analysis was carried out using the linkage map (cM/Mb) instead of the LDU map (LDU/Mb). This confirms the relationship with LD, or in this case recombination, since the linkage map is based on an entirely independent sample. Again all the results were highly significant ($p<0.0001$) with correlation coefficients of -0.2 and R^2 values ranging 4-5% (table 4.3).

Table 4.3 Correlation and regression between cM/Mb and tract coverage.

Population sample	Correlation Coefficients	Regression R^2
CEU	-0.23	0.05
CHB	-0.21	0.04
JPT	-0.21	0.04
YRI	-0.21	0.04
ALL	-0.21	0.04

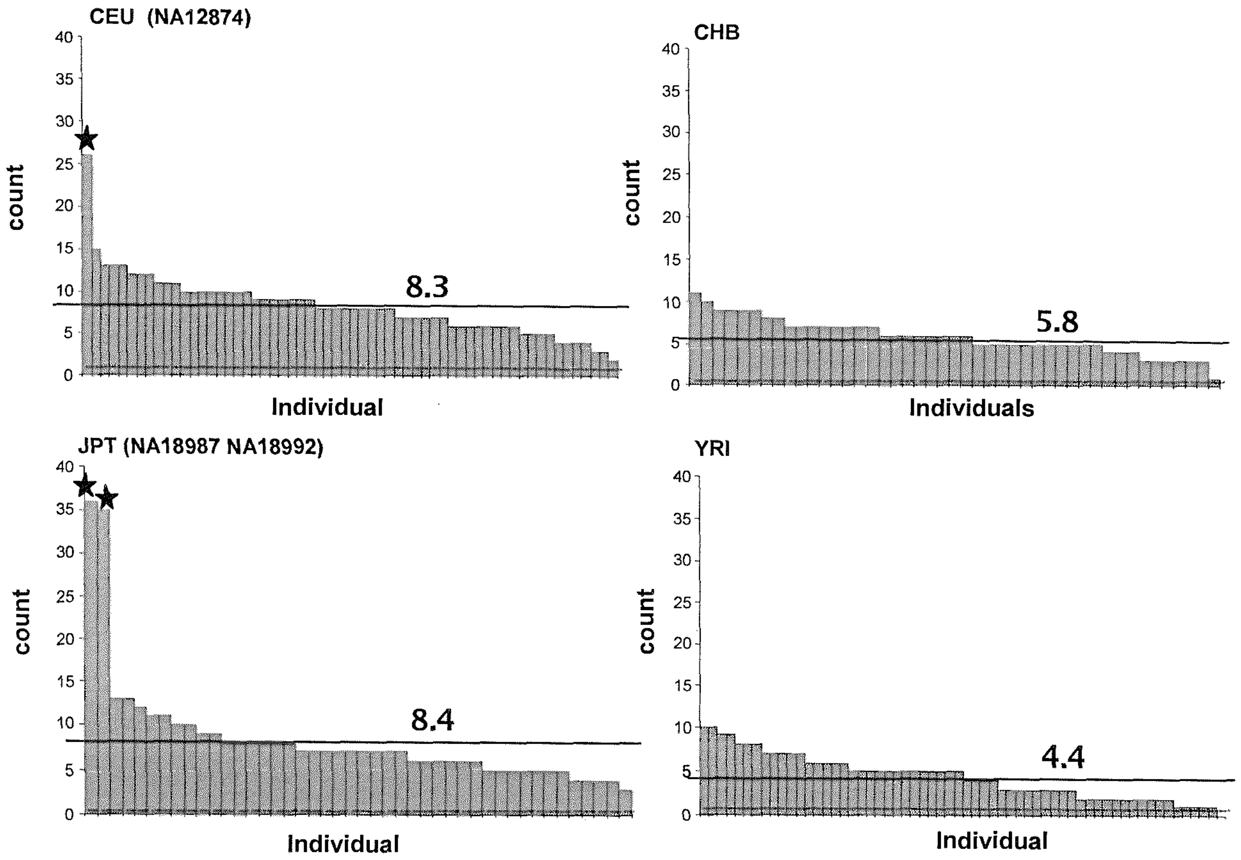
Next the correlation between tract coverage values across populations was examined to determine if the distribution of homozygous tracts in the genome was similar in the different populations. The results showed that all correlations were significant (p<0.0001) with correlation coefficients ranging 0.27-0.68. The YRI and CEU samples are the least similar and the CHB and JPT samples are the most similar (table 4.4).

Table 4.4 Correlation of 'tract coverage' values across all populations

	CHB	JPT	YRI
CEU	0.51	0.46	0.27
CHB		0.68	0.30
JPT			0.30

The distribution of tracts per individual across the 4 populations was also examined. The average tract count per individual ranged from 4.4 - 8.4 for the 4 populations. The YRI sample had the fewest homozygous tracts per individual and the JPT sample had the most. Three individuals were found with particularly high tract counts, one in the CEU sample (NA12874) and two in the JPT sample (NA18992, NA18987) (figure 4.3).

Figure 4.3 A bar graph for each population.



Each bar on the graph represents an individual and the Y-axis (0-40) shows a count of the number of tracts in that individual (individuals ordered by magnitude). The horizontal line and the figure on the graph show the mean average tract count for each population. The three stars show the three individuals with particularly high tract counts.

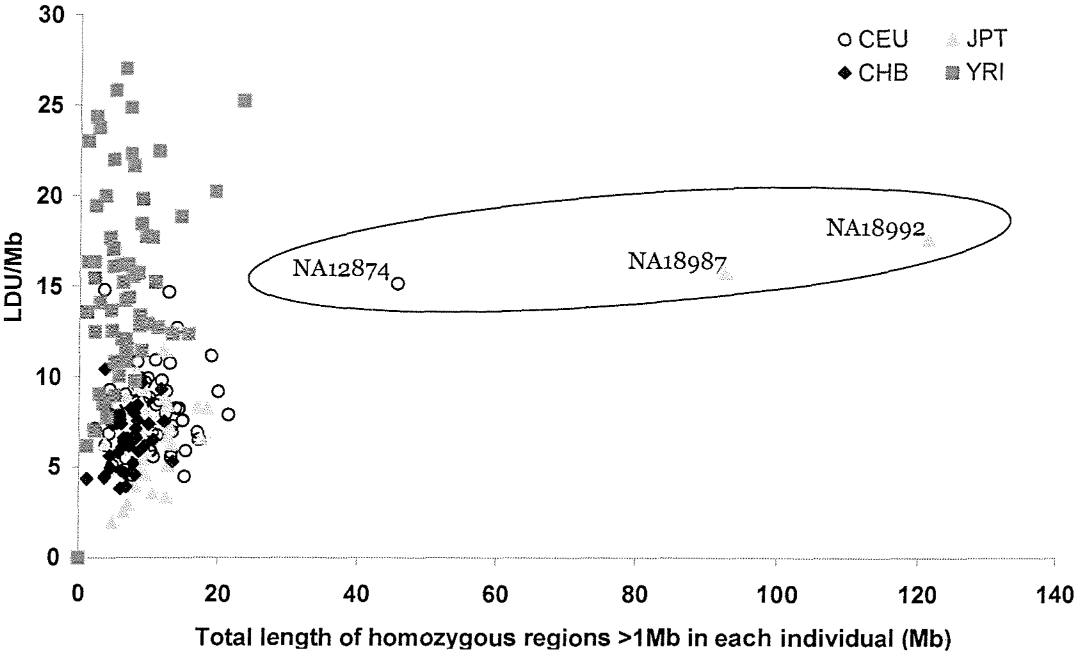
The amount of LD in regions where homozygous tracts occur was investigated for each individual, averaged for each population, and compared to the genome-wide average LDU/Mb for each population. This confirmed the correspondence between long homozygous tracts and regions of strong LD shown previously, since the mean LDU/Mb in regions containing homozygous tracts is much lower than the genome average (table 4.5).

Table 4.5 LDU/Mb in tract regions and genome-wide for each population.

Population	Genome-wide LDU/Mb	Tract regions LDU/Mb
CEU	20.2	8.3
CHB	22.6	6.7
JPT	20.4	7.3
YRI	28.4	15.4

However, the three individuals highlighted in the previous analysis stand out as outliers as they have tracts in areas that do not have particularly high LD, in fact, have levels of LD approaching the genome average for their populations, ranging from 15.1-17.6 LDU/Mb for the 1 CEU and 2 JPT individuals (figure 4.4).

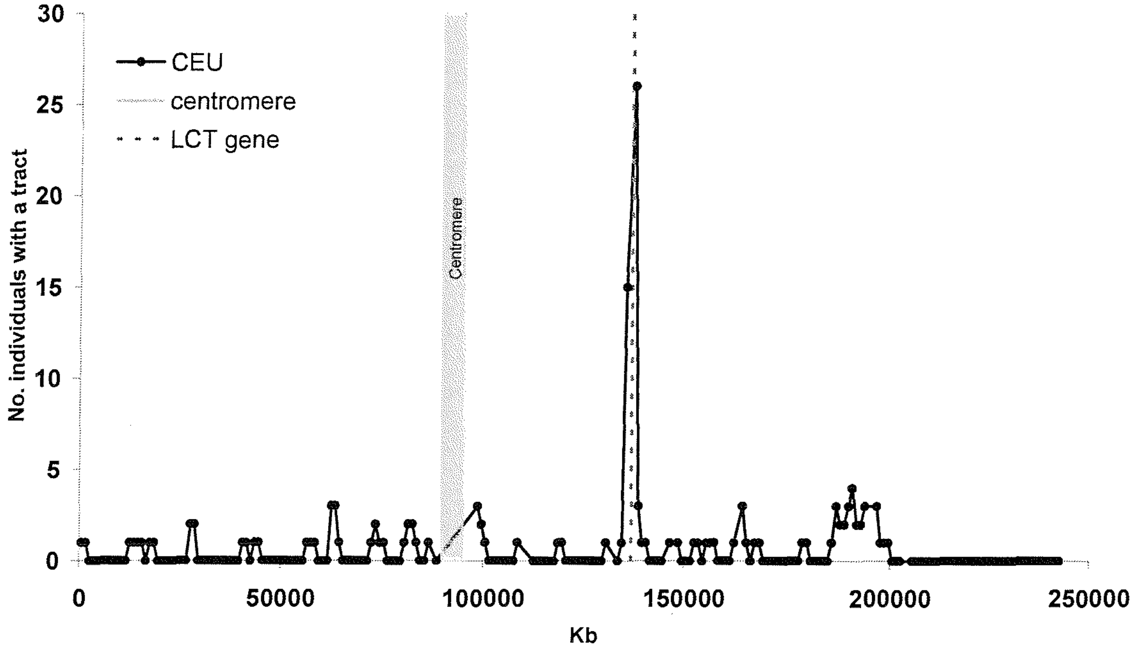
Figure 4.4 Graph showing LDU/Mb for each individual, the 4 populations shown. Circled are the 3 individuals with particularly high tract counts and higher LDU/Mb, in tract regions, than the rest of the individuals in their populations.



The three individuals, highlighted above, NA12874, NA18992 and NA18987 have tracts in regions of significantly higher LDU/Mb (less LD) than is typical for homozygous tract regions. Regression analysis shows that NA12847 explains 42% of the variance in LDU/Mb in the tract regions of the CEU sample and NA18992 and NA18987 together explain 89% of the variance in the JPT sample.

Chromosome 2 and the CEU sample, were chosen to highlight the relationship between selection and homozygous tracts, since chromosome 2 contains the LCT gene known to have been under positive selection in Caucasians (Bersaglieri et al. 2004). The number of individuals with a tract in each 1Mb segment across chromosome 2 was plotted. The graph clearly shows a peak where there is a particularly large number of people within the population with a tract, 26 out of 60 (43%). This peak aligns with the location of the lactase gene at ~136Mb (figure 4.5).

Figure 4.5 Graph of chromosome 2 with the number of individuals with a tract for each 1Mb segment plotted and the location of the lactase gene, which aligns with an obvious peak.



4.5 Discussion

This work has shown that homozygous tracts are remarkably common and long even in unrelated individuals from the apparently outbred populations represented in the HapMap data. The evidence indicates that homozygous tracts are generally found in regions of relatively extensive LD and locally low rates of recombination. The presence of relatively short haplotype 'blocks', regions of low haplotype diversity, has been well known for several years (Daly et al. 2001). However, the presence of much longer homozygous tracts (lengths sometimes greatly in excess of one megabase) was not widely anticipated.

Homozygous tracts can occur when a child inherits the same chromosomal segment from both parents, who themselves inherited it from a common ancestor. There are two broad mechanisms by which this could happen. One explanation is that the parents have a relatively recent common ancestor so there has been little opportunity for recombination to break up the segment. A second possibility is that any relationship between the parents is distant but a lack of recombination in the region (i.e. a region of high LD) has enabled the ancestral segment to persist intact.

This study considers only homozygous tracts that exceed 1 Mb in length, but there are numerous smaller segments many of which must contribute to the low haplotype diversity characterised in blocks. Given this, the true level of homozygosity in the genome is likely to be much greater than indicated by this analysis of the more extreme examples. These results show that extensive LD in a region correlates with a higher proportion of homozygous tracts in that region. This is because genomic regions with low recombination (high LD) allow particularly long chromosome segments to remain intact over time, increasing the chance that they come together in an individual as a homozygous tract.

Analysis of tract coverage between populations shows that tracts tend to be co-localised in all populations. Patterns of LD are also highly similar across human populations (De La Vega et al. 2005; Gibson et al. 2005) and follow the same trends as our analysis here, with JPT and CHB having the most similar and YRI and CEU having the least similar LDU maps. The co-localisation of recombination hot-spots in all human populations allows long homozygous tracts to persist in the shared intervening regions which have low meiotic activity. As might be expected, the YRI population has the fewest long tracts per individual (4.4), reflecting the longer time over which recombination has been breaking haplotypes in this sub-Saharan African population.

It is assumed that the four HapMap samples are representative of relatively outbred human populations and that homozygosity is not particularly exaggerated due to a limited number of haplotypes or 'atypical' individuals represented in the sample. However, there is little information about the individuals that contributed to the samples; sample sizes of 44 to 60 unrelated individuals are fairly small and it is conceivable that the samples are not truly representative of the whole of the population in each case. Three individuals stood out in this analysis as having particularly long tracts and high tract counts, one in the CEU sample and two in the JPT sample. This study has shown that the tracts in these individuals are not associated with regions of elevated LD, in contrast to tracts in other individuals. Therefore it is reasonable to suggest a different mechanism accounts for the tracts in these cases, such as recent inbreeding, suggesting that the parents of these individuals have an unknown relationship which goes back a comparatively small number of generations. In the most extreme case (NA18992) long homozygous tracts cover ~ 4% of the genome of that individual. Since only contiguous tracts longer than 1Mb are included here, the total amount of homozygosity is likely to be much higher. Autozygosity of 6% is to be expected in the offspring of a first cousin mating (Broman and Weber 1999). The same two Japanese individuals have been identified independently by the HapMap Consortium as showing 'an above average degree of cryptic relatedness' (Altshuler et al. 2005). The same analysis did not identify the CEU individual (NA12874), however. It seems reasonable to assume that the impact on the LDU map by including these three individuals is modest although this was not tested directly.

Aside from inbreeding and LD, there are other mechanisms which might contribute to the observed extent of homozygosity. There are different types of UniParental Disomy (UPD); isodisomy is the form where a child inherits two copies of the same chromosome from the same parent. This results in the child being homozygous at all loci. Segmental isodisomy can occur when a part, but not the whole, chromosome is affected. UPD can cause various diseases when it occurs in a region with imprinted genes, and can also cause rare recessive disorders. A case of maternal UPD of chromosome 1 was found by chance (Field et al. 1998) as there were no apparent phenotypic effects. This suggests that as well as isodisomy for rare recessive genes or UPD in imprinted regions, some cases of UPD may be asymptomatic and perhaps quite common. The phenomenon has been little studied where not associated with a disease, however, scanning methodologies to detect UPD are being developed, initially as a diagnostic tools, but could be used to answer this question in the future (Bruce et al. 2005; Altug-Teber et al. 2005).

Heterozygous deletions can sometimes be detected by apparent homozygosity over an extended region. For example, Huie *et al.* found a novel 8Kb deletion in a patient with glycogen storage disease type II. Apparent homozygosity may serve as an indicator of the presence of a heterozygous deletion but other molecular techniques are required to be definitive (Huie et al. 2002). SNPs with missing data were not included in this analysis, however, when a run of markers with missing alleles is detected this may indicate a homozygous deletion. Deletions might account for a few homozygous tracts but we presume that none of the long tracts we have examined here reflect cryptic deletion.

Particularly long haplotypes that are common in a population may be evidence of a region that has undergone selection, or may indicate a region that is a cold spot for mutation and recombination. The lactase gene has been shown to be under positive selection and aligns with a region of the genome that has a high number of individuals with a homozygous tract within the CEU population sample. It seems that selection may contribute to the homozygous tracts in some regions of the genome particularly when tracts are common in individuals within a population. Recombination as shown in the LDU map is likely to be having a larger effect in general. The amount of homozygous tracts associated with selection and the extent to which selection is detectable by homozygous tracts requires further study.

The HapMap data undergoes extensive QC procedures but genotyping or reporting errors are still possible. The recent release of Phase II data, allowed confirmation that the longest tract detected (17.9Mb) was still present and was not the result of a problem with the original data. The Phase II data comprises 3,902,623 genotypes that passed QC for the JPT sample. The 17.9 Mb homozygous tract in individual NA18992 was identified, the same region, which had 3,922 SNPs in Phase I, had 12,778 SNPs in Phase II. 11 heterozygotes break the tract into 12 pieces, the largest of which was 5.619 Mb. No two heterozygotes were adjacent. The presence of only 11 heterozygotes in a contiguous tract of 12,778 otherwise homozygous SNPs spanning 17.9 Mb suggests that these 11 comprise typing errors and/or relatively recent mutations. It seems therefore that the much higher density genotyping in Phase II will break some of the very long tracts but the strong relationships to the LD structure and evidence for inbreeding will be preserved.

4.6 Conclusion

This work has shown that homozygous tracts are common, in some cases very long, and, in a few cases reflect recent inbreeding within the pedigree. In general, long homozygous tracts reflect the presence of long ancestral haplotypes that remain intact because of locally low rates of recombination or, more rarely, other mechanisms such as UPD, deletions and in particular locations, selection. Since only homozygous tracts >1Mb were considered, the degree of homozygosity characterized is likely to be conservative. It is conceivable, that the abundance of homozygous regions and their contribution to long regions of high LD will significantly reduce our ability to fine map disease genes using association, and affect the interpretation of autozygosity mapping studies.

Chapter 5 - Autozygosity mapping to search for a candidate region or gene for Congenital Nephrotic Syndrome with Diffuse Mesangial Sclerosis (DMS).

5.1 Introduction

DMS is a rare form of Congenital Nephrotic Syndrome, a rare inherited disorder characterised by protein in the urine (proteinuria) and swelling of the body which leads to kidney failure. The age of onset and symptoms vary and overlap between forms. Some can be managed with medication, but different forms and individuals respond differently to treatment; CNS can be severe and lead to death in early childhood (MedlinePlus 2007). In this study the prognosis is poor for the affected individuals, there is no specific therapy and end stage renal failure can be expected by the age of 5. The index case Pedigree ID3 died before the age of 2. There are several forms of congenital nephrotic syndrome in which causal genes have been identified. Based on phenotype these genes are ruled out as a cause, but different mutations in these genes need to be considered.

Table 5.1 A summary of genes involved in congenital nephrotic syndromes.

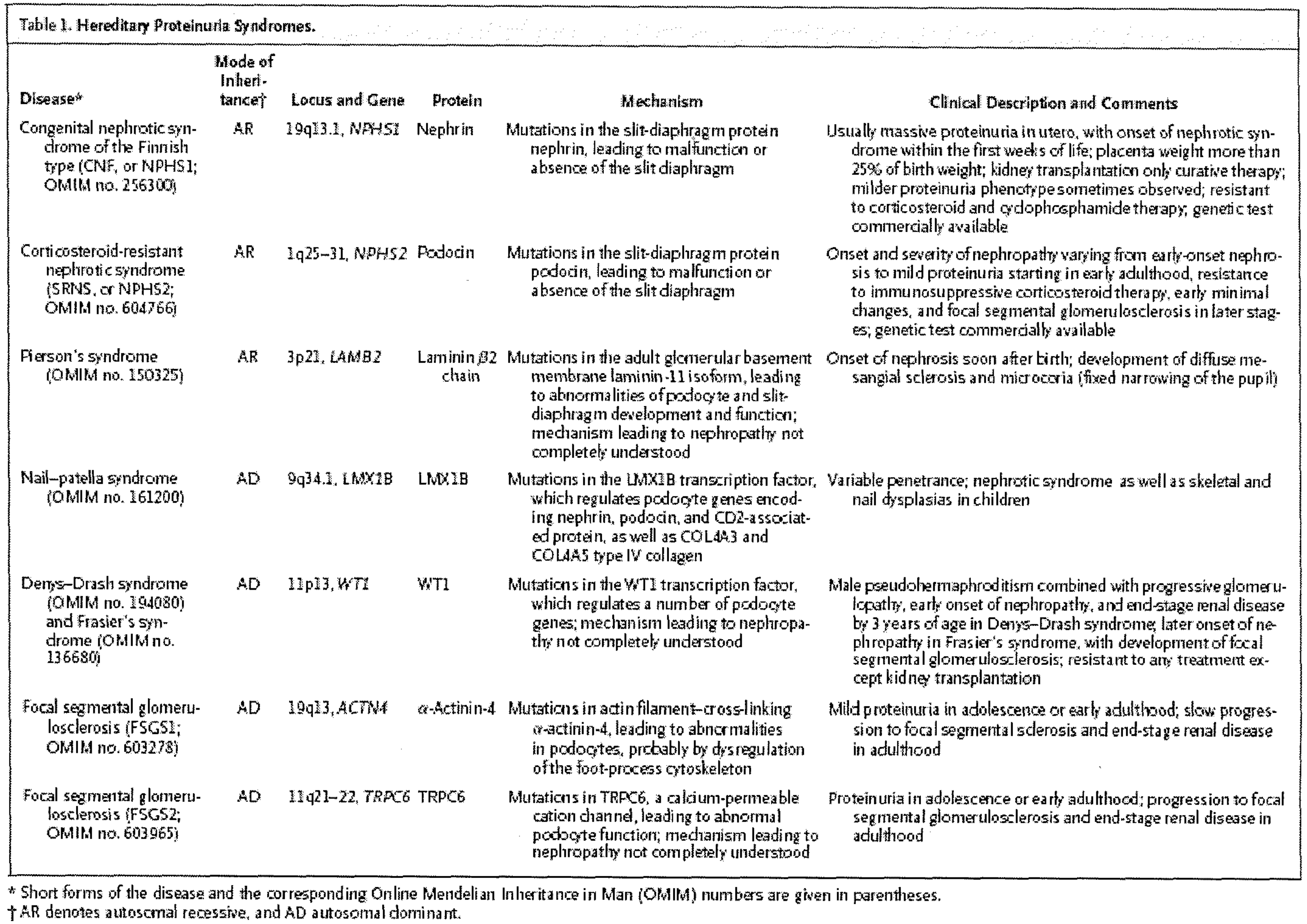


Table 1. Hereditary Proteinuria Syndromes.

Disease*	Mode of Inheri-tance†	Locus and Gene	Protein	Mechanism	Clinical Description and Comments
Congenital nephrotic syndrome of the Finnish type (CNF, or NPHS1; OMIM no. 256300)	AR	19q13.1, *NPHS1*	Nephrin	Mutations in the slit-diaphragm protein nephrin, leading to malfunction or absence of the slit diaphragm	Usually massive proteinuria in utero, with onset of nephrotic syndrome within the first weeks of life; placenta weight more than 25% of birth weight; kidney transplantation only curative therapy; milder proteinuria phenotype sometimes observed; resistant to corticosteroid and cyclophosphamide therapy; genetic test commercially available
Corticosteroid-resistant nephrotic syndrome (SRNS, or NPHS2; OMIM no. 604766)	AR	1q25–31, *NPHS2*	Podocin	Mutations in the slit-diaphragm protein podocin, leading to malfunction or absence of the slit diaphragm	Onset and severity of nephropathy varying from early-onset nephrosis to mild proteinuria starting in early adulthood, resistance to immunosuppressive corticosteroid therapy, early minimal changes, and focal segmental glomerulosclerosis in later stages; genetic test commercially available
Pierson's syndrome (OMIM no. 150325)	AR	3p21, *LAMB2*	Laminin β2 chain	Mutations in the adult glomerular basement membrane laminin-11 isoform, leading to abnormalities of podocyte and slit-diaphragm development and function; mechanism leading to nephropathy not completely understood	Onset of nephrosis soon after birth; development of diffuse mesangial sclerosis and microcoria (fixed narrowing of the pupil)
Nail–patella syndrome (OMIM no. 161200)	AD	9q34.1, *LMX1B*	LMX1B	Mutations in the LMX1B transcription factor, which regulates podocyte genes encoding nephrin, podocin, and CD2-associated protein, as well as COL4A3 and COL4A5 type IV collagen	Variable penetrance; nephrotic syndrome as well as skeletal and nail dysplasias in children
Denys–Drash syndrome (OMIM no. 194080) and Frasier's syndrome (OMIM no. 136680)	AD	11p13, *WT1*	WT1	Mutations in the WT1 transcription factor, which regulates a number of podocyte genes; mechanism leading to nephropathy not completely understood	Male pseudohermaphroditism combined with progressive glomerulopathy, early onset of nephropathy, and end-stage renal disease by 3 years of age in Denys–Drash syndrome; later onset of nephropathy in Frasier's syndrome, with development of focal segmental glomerulosclerosis; resistant to any treatment except kidney transplantation
Focal segmental glomerulosclerosis (FSGS1; OMIM no. 603278)	AD	19q13, *ACTN4*	α-Actinin-4	Mutations in actin filament–cross-linking α-actinin-4, leading to abnormalities in podocytes, probably by dysregulation of the foot-process cytoskeleton	Mild proteinuria in adolescence or early adulthood; slow progression to focal segmental sclerosis and end-stage renal disease in adulthood
Focal segmental glomerulosclerosis (FSGS2; OMIM no. 603965)	AD	11q21–22, *TRPC6*	TRPC6	Mutations in TRPC6, a calcium-permeable cation channel, leading to abnormal podocyte function; mechanism leading to nephropathy not completely understood	Proteinuria in adolescence or early adulthood; progression to focal segmental glomerulosclerosis and end-stage renal disease in adulthood

* Short forms of the disease and the corresponding Online Mendelian Inheritance in Man (OMIM) numbers are given in parentheses.

† AR denotes autosomal recessive, and AD autosomal dominant.

(Tryggvason, Patrakka, and Wartiovaara 2006)

In this study there are 5 affected individuals who have non-syndromic DMS which appears most similar to Pierson's syndrome but without eye symptoms (microcoria), DMS is also a feature of Denys-Drash Syndrome. Therefore, the regions around LAMB2 and WT1 are of particular interest as are genes known to interact with these loci (table 5.1). The PLCE1 gene has also recently been found to be associated with nephrotic syndrome in individuals with non-syndromic DMS histology offering a further gene of interest (Hinkes et al. 2006). The phenotype described by Hinkes *et al.* best matches the phenotype in this study and is therefore the strongest candidate.

The family involved originates from Pakistan and has a complex and incomplete pedigree with a large degree of consanguinity (appendix 2). The affected individuals are all from the same, most recent, generation and originally consisted of 3 males and 1 female. The female was less severely affected, which, it was speculated, may be due to early treatment and a better response. However, it was later determined, after withdrawal of treatment, that the female was no longer suffering from the disease. The disease is assumed to be autosomal recessive due to the patterns of inheritance in the pedigree and X-inactivation work carried out to rule out an X-linked pattern (personal communication Prof. D. Robinson, Wessex Regional Genetics Laboratory). Urine tests were also carried out on all available family members to check for proteinuria and detect any other mildly affected individuals but no additional cases were found. The consanguinity in this pedigree and the presence of just a few affected individuals within one family means that autozygosity mapping is likely to be the most effective and efficient approach to finding the gene involved.

The concept behind autozygosity mapping assumes that a large region of homozygosity shared among affected individuals is likely to contain the disease variant (Lander and Botstein 1987). Consanguinity in the pedigree means that the affected child is likely to have inherited the same mutation on the same haplotype from both parents, who in turn, inherited it from a common ancestor. If the mutation occurred recently within the family the region of homozygosity is expected to be large since there would not have been time for recombination substantially break up the ancestral causal haplotype. There are expected to be many large regions of homozygosity within individuals in an inbred pedigree since shared haplotypes will have been inherited through several lineages. Homozygous regions shared by all affected individuals should narrow the search substantially, and comparing these regions with unaffected individuals from the same family should help to narrow the region of interest further.

The strategy of autozygosity mapping, followed by identification of conserved haplotypes and mutation analysis of candidate genes, has proved to be successful in the characterisation of recessive disease genes and in understanding the biology of disease as well as normal processes. There are several examples of successfully mapped novel loci.

An example of a causal gene identified by autozygosity mapping is MKS3 in Meckel-Gruber Syndrome, a rare autosomal recessive condition. Eight consanguineous families, with 9 affected individuals originating from the Indian sub-continent were studied by Morgan *et al.* (Morgan et al. 2002). A genome wide analysis using 200 microsatellite markers in the affected individuals revealed a region of homozygosity ~25cM in length. Two candidate genes in the region were sequenced for mutation detection without success, however a heterozygous SNP in one of the genes narrowed the region further to 15cM. This region contained >50 genes but with no strong candidates.

Further work by Smith *et al.* (Smith et al. 2006) identified the gene. A 10K Affymetrix Chip was typed in 5 affected individuals reducing the region again. 22 of 66 genes in this region were sequenced but no mutations were identified. A rat model with a similar phenotype was investigated and the human ortholog of the rat causal gene, which was present in the identified region, was sequenced. Different mutations were identified in the 5 families. The mutations were all homozygous consistent with consanguinity and segregated with the disease from both parents. The mutations were not found in >120 controls showing that they were not common polymorphisms. Searching for regions of homozygosity was integral to the search for this gene, and a 10K SNP array was required to narrow the region detected by microsatellites. However, other approaches were required for final identification of the gene.

Other examples include, the identification of a mutation in WNT10A in Ectodermal Dysplasia cases (Adaimy et al. 2007); BLOC1S3 mutations in Hermansky-Pudlak Syndrome (Morgan et al. 2006); mutations in RAB3GAP in Warburg Micro Syndrome (Aligianis et al. 2005). These studies all used an approach which began with genome wide typing of microsatellites in the affected individuals to determine a common region of homozygosity. This was followed up by typing more microsatellite markers in the region and in more individuals. Linkage programs were then used to provide a LOD score for the significance of the region. Examples of programs available are MAPMAKER/HOMOZ (Kruglyak, Daly, and Lander 1995), LINKAGE/FASTLINK (Cottingham, Jr., Idury, and Schaffer 1993; Schaffer et al. 1994) and Easylinkage (Lindner and Hoffmann 2005). These programs require assumptions about penetrance, disease gene frequency, pedigree completeness (inbreeding coefficients), marker allele frequency in particular populations, and also require data from several affected and unaffected family members. The regions defined in this way can be very large, 10-20cM depending on the density of microsatellites used, and may contain many genes. Candidate genes can be sequenced for mutations but if none are found or there are no strong candidates further narrowing of the region is required. Fine mapping of regions has been carried out using a 10K Affymetrix SNP chip on affected individuals, for example in the MKS3 and BLOC1S3 studies (Smith et al. 2006; Morgan et al. 2006).

The development of high density SNP genotyping technologies and the relatively low costs involved when only a few individuals need be typed, mean that many studies can now use high density SNP arrays to carry out autozygosity mapping. High density typing means that regions of homozygosity can be identified and visualized without the need for statistical inference or LOD scores. This technique has the advantage of speed and resolution. For autozygosity mapping, the assumption that the disease is caused by a homozygous mutation inherited from a relatively recent ancestor must be correct and the gain in resolution is in part dependent on the pedigree. If the affected individuals are closely related and the mutation occurred very recently, the region of homozygosity harbouring the mutation is likely to be large, whereas if the affected individuals have a higher degree of separation and the mutation occurred less recently, the region will be smaller and the advantage of high resolution SNP genotyping will be greater (Gibbs and Singleton 2006).

There are several examples of disease loci identified using high throughput genotyping technology such as the Affymetrix chips and Illumina bead arrays. Genome wide screening in 2 affected Kartagener Syndrome patients (Gutierrez-Roelens et al. 2006) used microsatellites to identify 10 regions of autozygosity and 26 uninformative regions. Additional microsatellite markers in this small family were not informative, so higher density screening was carried out using the 10K SNP array. This refined the candidate regions to a 44.6Mb region on chromosome 1 and a 13.7Mb region on chromosome 7. The higher density of the SNP array over the microsatellite panel allowed these regions to be discovered, however, the regions are still large and although higher density SNP array data may refine the locations more individuals from more families would be required to substantially narrow the regions of interest.

The search for genetic variants causing the autosomal recessive form of Severe Congenital Neutopenia (Kostmann syndrome) was attempted by two different methods (Melin et al. 2007; Klein et al. 2007). Melin *et al.* used a 10K SNP array in 4 affected individuals from one family. Software was written to define regions of homozygosity in 3 ways. Firstly, regions greater than 1Mb in size. Secondly, regions containing >= 20 consecutive homozygous SNPs in all 4 affected individuals, and lastly, regions containing >=30 consecutive homozygous SNPs in 3 of the 4 affected individuals systematically removing one individual. The regions defined in this way were further analysed by microsatellites in all members of 2 families. The 10K analysis found no regions that could be confirmed by microsatellites. Higher density analysis using the 100K array identified 30 regions, one of which was confirmed in 3 of the 4 affected individuals. The region was 1.8Mb and was further confirmed by 2 affected individuals from a second family. The presence of the same haplotype narrowed the candidate region to 1.2 Mb containing 37 known genes. The 10K data confirms this result, although on the original screen the region was not detected because it contains only 4 homozygous SNPs and several hundred such regions were found in the 10K array analysis. This shows that the 10K SNP array may not be of sufficient density for this type of study.

At the same time as this study, Klein *et al.* approached the identification of the genes responsible for Kostmann syndrome with genome wide genotyping of 217 microsatellites carried out on 4 affected and 4 unaffected individuals from 3 families. Only 1 of these markers was homozygous in all 4 cases, all available family members were genotyped for microsatellites in the region and a peak LOD of 4.15 was obtained. This approach gave an interval of 34.4Mb containing 275 genes, a much larger region than that defined by Melin *et al.* Prioritizing potential candidates led to mutation screening of HAX1 and identification of a causal mutation. 15 of 63 further patient samples had the mutation and 200 healthy controls did not. The 2 groups collaborated and both authors are listed on both papers. They used different methods and different patient samples for the initial screen, however, the region defined by Melin *et al.* using the 100K SNP array was much smaller and gave fewer candidate genes to investigate in the next stage of analysis.

A study by Chiang *et al.* identified TRIM32 as the 11th locus for Bardet-Biedl syndrome (BBS11) (Chiang et al. 2006). An initial genome wide microsatellite screen using 400 markers was uninformative and failed to identify any regions homozygous in the 4 affected individuals studied. The use of a 50K SNP array identified 14 regions with >= 25 SNPs in the 4 cases. Typing microsatellites in all available family members in these regions excluded all but one, the largest region detected, which was 2.4Mb. This region had no microsatellites in the original screen so would have been impossible to identify. Mutation screening in the candidate region revealed a mutation in TRIM31 which was confirmed by its absence in 184 controls.

Puffenberger *et al.* used the 10K SNP array to investigate the cause of symptomatic epilepsy syndrome in a group of 7 distantly related Mennonite children. Surprisingly analysis did not show any large blocks of homozygosity common to all 7 patients. To explain this lack of autozygosity, both locus and mutation heterogeneity were considered but with no success. Regions of the 10K array which have low SNP coverage were then investigated. A chance observation of a single SNP which produced 'NoCalls' for all 7 patients despite an average call rate of 98.5% led to further investigation which showed that all 14 parents of the cases were called as homozygous for this particular SNP. It was determined that the parents were actually hemizygous for the SNP and all 7 cases were homozygous for a deletion in a 7Kb region around the SNP in the LYK5 gene (Puffenberger et al. 2007). The lack of local SNP coverage on the lower density SNP arrays in particular regions is another reason high density arrays are required for autozygosity mapping. This study also shows that identifying regions where 'NoCalls' are present, consistent with a small causal deletion, must be considered in the analysis of these data.

There are some analysis and visualisation programs designed for this type of data. Examples include, Scamp (Forshew and Johnson 2004), ExcludeAR (Woods et al. 2004), AutoSNPa (Carr et al. 2006), IBDfinder (Carr, Sheridan, and Bonthron 2007), and PLINK (Purcell et al. 2007). However, Scamp and ExcludeAR are based on a Microsoft excel spread sheet format, while Scamp only analyses microsatellites, ExcludeAR analyses SNP data but both are limited to the amount of data they can handle and neither can cope with more than the 10K Affymetrix array (Forshew and Johnson 2004; Woods et al. 2004). AutoSNPa analyses SNP array data and can load data from the 250K array but this will increase computational time. It also requires pedigree data, and is primarily designed to visually analyse results, from which regions chosen by eye can be exported to text or excel files for further scrutiny. IBDfinder provides a less restricted qualitative approach to the identification of identity by decent (IBD) regions. It ignores pedigree structure, thereby allowing the analysis of singletons and groups of unrelated individuals. It is designed to handle Affymetrix format data and includes an error rate allowance and a SNP density adjustment. It effectively scores each marker in each individual based on the number of adjacent homozygous SNPs, then combines this information across individuals in windows of 0.125Mb (or cM), as the number of individuals with or without IBD in that window. It takes no account of linkage disequilibrium and, like AutoSNPa, the results are visualised and would require examination of interesting regions to define them more precisely (Carr et al. 2006; Carr, Sheridan, and Bonthron 2007). PLINK offers detection of runs of homozygosity in windows of user defined SNP number or Kb size. It also allows a user defined number of heterozygotes or NoCalls in each window. The program allows pooling of regions of homozygosity across individuals allowing a threshold to be set for the amount of allelic identity, but the genome-wide output is extensive and thus difficult to interpret. Measuring homozygosity in a sliding window approach does not provide accurate definitions of regions and this method does not take into account LD. A method for detection of shared extended haplotypes of IBD is in progress but documentation has yet to be produced (Purcell et al. 2007). Visualization methods are intuitive but become more difficult to interpret the more data and more individuals involved.

5.2 Aims

The aim of this work is to use densely genotyped SNP data to determine regions of homozygosity in inbred individuals affected with Congenital Nephrotic Syndrome and determine a homozygous region associated with the rare autosomal recessive disease. The program written for searching for homozygous regions in HapMap individuals is to be extended to incorporate data from several individuals at once. Regions will be determined on both the LDU and Kb scale. Knowledge of homozygosity in outbred populations from previous work (chapter 4) shows that regions which are in LD blocks are more likely to be homozygous by chance (because of low haplotype diversity), this may be important in determining the most likely candidate regions. Genes in the selected candidate regions will be compared with genes involved in known kidney diseases, with kidney related functions or known interaction with other candidate genes. Further to this project, mutation screening in candidate genes and the identification of a causal locus would give the possibility of anti-natal and carrier testing assisting in genetic counselling in this family.

5.3 Methods

5.3.1 The cases

The pedigree shows complex consanguinity and an autosomal recessive inheritance pattern with all the affected individuals in one generation (appendix 2). An ancestor common to the affected individuals (ID58) is a possible the route of inheritance. Originally the presence of an affected female gave weight to the rejection of an X-linked pattern, however the relatively small number of affected males in the pedigree and X-inactivation work have also helped to rule this out. Two of the cases (ID25 and ID114) were diagnosed with a mild phenotype, with levels of protein in the urine higher than the normal range, but not extreme. In ID25 this proteinuria resolved after treatment with ACE inhibitors. ID114 was also diagnosed with a mild phenotype on the basis of higher than the normal, but not extreme, levels of proteniuria. ID114 seems to be following the same disease pathway as ID25 but is currently only a few months old and has just started treatment. Three of the cases (ID3, ID4 and ID19) were diagnosed with a severe phenotype with extreme levels of proteinuria and were not responding well to treatment. All 3 have had renal biopsies and DMS histology confirmed (Personal communication via project meeting, Dr R. Gilbert, Consultant Paediatric Nephrologist, Southampton General Hospital).

5.3.2 Data

This analysis was based on 3 datasets. The initial dataset was genotype data from the 50K Affymetrix SNP microarray, the second was the higher density 500K Affymetrix SNP microarray, and the final dataset was the Illumina humanhap550 bead array. A total of 7 individuals were genotyped on one or more of the platforms. A range of genotyping efforts were undertaken due to an unanticipated lack of homozygosity common to the affected individuals in the initial analysis and concerns over genotyping accuracy. However, the phenotype status of some individuals changed during this study, leading to a different interpretation of the results.

Table 5.2 The 7 genotyped individuals

Pedigree ID			sex	Phenotype; unaffected (0) mild (1) affected (2)	Genotyped?		
individual	father	mother			Affymetrix 50K	Affymetrix 500K	Illumina 550K
3	2	1	male	2	yes	yes	yes
4	2	1	male	2	yes	yes	yes
19	98	65	male	2	yes	yes	yes
114	64	13	male	1	no	no	yes
25	78	15	female	1	yes	yes	yes
17	98	65	male	0	no	yes	no
18	98	65	male	0	no	yes	no

5.3.3 Investigating data quality

The call rates for all genotypes over all individuals called was high. However, ID3 in the Affymetrix 50K dataset, ID19 in the Illumina dataset had lower genotype call rates than the rest.

Table 5.3 Genotype call rates (%) over all data for all individuals genotyped on the 3 platforms.

	Affymetrix 50K	Affymetrix 500K	Illumina550K
ID3	94.92	96.83	99.47
ID4	99.49	98.85	99.49
ID19	98.95	98.68	90.07
ID114	-	-	97.93
ID25	98.53	98.13	99.51
ID17	-	97.82	-
ID18	-	98.43	-

The data were organised by chromosome and location on the physical map in Kb (NCBI build 36.1, UCSC build 18, Mar06). SNPs on the X, Y and 'unknown' chromosomes or those which could not be located on the current sequence were removed. The numbers of SNPs available for analysis were 57,179 in the Affymetrix 50K data, 440,734 in the Affymetrix 500K data and 547,475 in the Illumina 550K data.

The Affymetrix 500K chip and the Illumina 550K bead array data have 76,116 SNPs in common. In the 4 individuals genotyped on both platforms there were 11,981 'NoCalls', leaving 292,483 successfully typed genotypes. All possible combinations of genotype calls on the 2 platforms in each individual were recorded and different classes of discrepancies were detected. The 2 platforms use different methods to code SNP genotypes so in a proportion of cases an Affymetrix AA call is the same as a BB Illumina call. However, discrepancies where one platform called a heterozygote and the other a homozygote, indicates an error in one of the genotype calls. This would have a significant impact on this analysis since a questionably typed heterozygote could break up an otherwise long homozygous region. The two platforms provide confidence scores for each genotype call but the confidence scores have different interpretations. Affymetrix scores range 0-1 with scores closer to 0 having a higher confidence. A default threshold of 0.5 is applied and genotypes with a score above this are not called. Illumina scores range from 0-1 with scores closer to 1 having a higher confidence. Genotypes with a score below the default threshold of 0.25 are not called. The average scores for heterozygous calls and the 2 homozygous calls were calculated and average confidence scores for each class of discrepancy were also calculated.

To investigate confidence score thresholds to optimise data quantity and accuracy, genotypes with the lowest scores were removed using several percentile cut offs, 5, 10, 50, 80, 85 and 90. Each threshold was applied and the number of discrepancies for each class was recounted. An optimal reduction in data, based on scores, was defined and all known discrepancies were also removed before analysis.

5.3.4 Checking for small deletions

To check if any small homozygous deletions were picked up in the data, software was written in C to count the number of genotypes given a 'NoCall' in all affected individuals and detect runs of consecutive 'NoCalls'.

5.3.5 Defining regions of homozygosity

Software was written in C to search through the data, SNP by SNP, and detect regions of consecutive homozygous SNPs flanked by heterozygotes. Firstly regions of homozygosity were detected for each individual and then regions where all the affected are homozygous for the same alleles at consecutive markers were recorded. These regions are flanked by SNPs where at least one of the 4 individuals is heterozygous or homozygous for the opposite allele. Therefore all affected will be homozygous for the same haplotype over each region detected. Centromeric and heterochromatic regions were excluded from analysis.

Long regions of homozygosity common to all 4 affected may be adjacent but broken by a single heterozygote call. To identify this, a count of homozygous SNPs following the SNP that ends a region was also recorded. This is the number of SNPs in the next homozygous region if the two regions are separated by a single marker.

5.3.6 Prioritising and selecting regions for follow up

Regions of interest were prioritised by the genetic length of the region in LDUs. The genetic length is the most informative since it takes into account the linkage disequilibrium across the region. The physical length in Kb and the number of homozygous SNPs in the region were also recorded. The number of SNPs in a region is useful as a measure of the amount of information in the region, a lower limit of 5 consecutive homozygous SNPs was applied.

A database of functionally relevant candidates, known to be involved in other forms of CNS, involved in other kidney disease, or known to interact with candidates, was created using data from the literature and the Human Kidney Gene DataBase (Human kidney Gene DataBase 2004; Renshaw et al. 2004; Tryggvason, Patrakka, and Wartiovaara 2006; Hinkes et al. 2006).

A list of functionally relevant candidate genes are given in appendix 3.

5.4 Results

5.4.1 Investigating data quality

The average confidence scores are shown (table 5.4) for heterozygotes and homozygotes in the Affymetrix 500K and Illumina 550K datasets (information is not available for the Affymetrix 50K dataset). The confidence is slightly higher (closer to 0) for heterozygous calls than for homozygotes calls for the Affymetrix data. The opposite is true for the Illumina data with homozygote calls given a slightly higher score (closer to 1), although all scores are high.

Table 5.4 Mean confidence scores for homozygous and heterozygous genotype calls.

	Affymetrix 500K		Illumina 550K	
	heterozygous	homozygous	heterozygous	homozygous
ID3	0.053	0.085	0.860	0.863
ID4	0.032	0.059	0.862	0.863
ID19	0.031	0.054	0.829	0.837
ID114	-	-	0.845	0.863
ID25	0.036	0.058	0.862	0.863
ID17	0.039	0.070	-	-
ID18	0.035	0.065	-	-

Comparing the number of discrepant genotype calls between the Affymetrix and Illumina data indicates an error rate of 0.63%. Analysis of discrepant genotype calls between Affymetrix 500K and Illumina 550K datasets showed that in cases where a discrepancy between a heterozygote and a homozygote call was detected, the average confidence score was lower for the platform calling a heterozygote than for other classes of discrepancy.

Table 5.5 Comparison of genotype calls for the 4 individuals genotyped on both the Affymetrix 500K array and Illumina 550K array.

Genotype call									Mean confidence scores	
Affymetrix	Illumina	ID3	ID4	ID19	ID25	All	%	Indicates an error?	Affymetrix	Illumina
AA	AA	12072	12300	11207	12477	48056	16.43	No	0.0523	0.8618
AA	BB	13431	13971	12561	14018	53981	18.46	No	0.0693	0.8594
BB	AA	10733	10738	10016	10949	42436	14.51	No	0.0495	0.8666
BB	BB	14063	14644	13098	14578	56383	19.28	No	0.0730	0.8554
AB	AA	166	69	87	94	416	0.14	Yes	0.2344	0.8717
AB	BB	373	63	81	196	713	0.24	Yes	0.2452	0.8656
AA	AB	32	24	237	24	317	0.11	Yes	0.1109	0.6971
BB	AB	40	27	302	27	396	0.14	Yes	0.1102	0.7073
AB	AB	22870	23322	21171	22422	89785	30.70	No	0.0345	0.8586
Total		73780	75158	68760	74785	292483	100.0			

* Lowest-confidence scores for each platform in bold. Affymetrix scores 0-1, with 0 indicating an accurate call. Illumina scores 0-1, with 1 indicating an accurate call.

Removing a percentage of the data based on confidence scores showed that to remove all known discrepancies from the data required a cut off of 90% in both the Affymetrix and Illumina data, leaving only 1% of the comparable genotypes for analysis (table 5.6). Removal of the lower 10% reduced the percentage of known discrepancies to 0.23% (from 1842 to 566) in the remaining 83.05% of the comparable data.

Table 5.6 The number of discrepancies and percentage of data remaining after the quality thresholds are altered to remove a percentile of the data on both platforms.

Percentile	Illumina threshold	Affymetrix threshold	% of genotype comparisons remaining	No. discrepancies
0	0.50	0.250	100	1842
5	0.69	0.209	91.33	819
10	0.77	0.142	83.05	566
50	0.87	0.031	25.90	73
75	0.91	0.012	6.59	15
80	0.92	0.009	4.25	10
85	0.93	0.006	2.41	6
90	0.95	0.004	1.05	0

Analysing the discrepancies between SNPs typed on both the Affymetrix 500K chip and the Illumina 550K array allowed a reasonable 10% cut off based on confidence scores to be applied. Known discrepancies were also removed in both datasets.

5.4.2 Detecting deletions

Table 5.7 The number of 'NoCalls' for all confirmed affected individuals.

	No. NoCalls for all affected	% of total SNPs genotyped	No. of these successfully typed on another platform
Affy50K	29	0.05	14
Affy500K	107	0.03	19
Illum550K	316	0.06	38

None of these SNPs are adjacent therefore there are no runs of the sort that might indicate a small deletion.

5.4.2 Homozygosity

The level of homozygosity in each dataset was detected at the SNP level. The number of SNPs homozygous for the same allele in all affected was recorded and the percentage of the total calculated.

Table 5.8 The number of SNPs which are homozygous in all confirmed affected individuals.

	No. SNPs homozygous in all affected	% of total
Affy50K	26646	46.60
Affy500K	200579	47.73
Illum550K	214090	39.11

Regions of homozygosity in each individual were then identified. The maximum size in Mb and the number of regions greater the 1Mb were recorded.

Table 5.9 Homozygous regions in each individual for the different datasets, the number of regions >1 Mb and the maximum region length in Mb.

	Affy50K		Affy500K		Illum550K	
	>1Mb	Max length	>1Mb	Max length	>1Mb	Max length
ID3	198	12.48	73	7.41	46	29.97
ID4	176	20.80	56	12.37	37	21.28
ID19	172	28.55	92	11.19	97	6.59
ID114	-	-	-	-	46	19.89
ID25	165	31.39	97	8.45	42	31.27
ID17	-	-	79	10.37	-	-
ID18	-	-	79	16.50	-	-

For comparison the maximum number and size of regions defined in the HapMap samples are presented.

Table 5.10 Homozygous regions defined in the HapMap samples.

Sample	No. Individuals	Mean No. tracts per individual	Max No. tracts per individual	Max Length (Mb)	No. SNPs in sample
CEU	60	8.30	26	6.48	728353
CHB	45	5.84	11	2.63	644060
JPT	44	8.41	36	17.91	639460
YRI	60	4.37	10	11.14	744006

5.4.3 Initial analyses of regions of homozygosity common to the affected individuals.

Initial analysis included ID25 as an affected individual, and defined regions common to all 4 affected individuals. The maximum genetic length was 70.5 LDU in a region containing only 5 SNPs in the Affymetrix 50K dataset.
An increase in SNP density using the 500K Affymetrix chip gave maximum genetic length of 17.29 LDU in a region containing 9 SNPs. The results of these

analyses did not provide an expected single particularly long region in all affected individuals and no strong candidate genes were present in the longest regions detected. It later became known that ID25 had a milder and possibly different phenotype which resolved after treatment, it was therefore unsurprising that the analysis did not give definitive results.

5.4.4 Analyses of regions of homozygosity common to the confirmed affected individuals

New data on the Illumina 550K bead array, were provided for the affected individuals including one new affected born into the family, ID114. ID25 has a different (milder) phenotype to ID3, ID4 and ID19, thus not considering her as one of the affected individuals is likely to have the biggest impact on the results. Again regions were defined using only affected individuals, leaving 3 in the Affymetrix 50K and 500K data and 4 in the Illumina data. The newest individual (ID114), typed only in the Illumina sample has a mild phenotype similar to the phenotype observed in ID25, and is too young to have had the diagnosis confirmed by renal biopsy. For this reason analysis was carried out with and without ID114 in the Illumina sample. The Illumina dataset has higher density typing, generally higher call rates (except ID19), and the analysis of homozygous regions by individual, shows longer regions implying less erroneously typed heterozygotes than the Affymetrix 500K dataset. It also has data on all confirmed and unconfirmed affected individuals. Therefore the main analysis was carried out using the Illumina dataset, and results were then confirmed using the Affymetrix datasets.

Analysing the Illumina 550K array dataset with ID114 gave a maximum region length on the genetic scale of 12.52 LDU containing 14 SNPs. No single region stood out as particularly long or was backed up by good evidence in the form of a high SNP number. Reanalysing the Illumina 550K array dataset without ID114 gave a maximum region length on the genetic scale of 54.81 LDU for a region on chromosome 13 containing 787 SNPs. Several adjacent regions on chromosome 13 are also detected, suggesting the possibility of a single long region broken by erroneous genotype calls. A region on chromosome 10 was the second longest on the LDU scale but longest on the physical scale and contains the largest number of SNPs. This region is of particular interest since it contains the PLCE1 candidate gene.

Table 5.11 Illumina 550K array, top 10 regions ordered by genetic length on the LDU scale. A) Regions common to ID3, ID4, ID19 and ID114. B) Regions common to ID3, ID4 and ID19 (see appendix 4 & 5 for Affymetrix 50K and 500K results)

		Location (Kb)					No. SNPs in
				Kb	LDU	No.	following
	Chr	Start	End	length	length	SNPs	region
A)	10	45468.76	47063.96	1595.2	12.52	14	0
	3	114797.7	114806.5	8.77	12.09	7	3
	2	86902.21	88092.11	1189.89	11.62	21	9
	10	34234.54	34288.87	54.33	9.97	27	2
	1	230117.3	230223.4	106.11	9.92	56	2
	11	126346.8	126415.8	69.01	9.91	48	5
	16	86431.64	86521.22	89.58	9.53	13	2
	1	98726.83	98752.58	25.75	9.49	11	3
	2	199350	199374.4	24.38	9.29	7	0
	19	36071.87	36084.21	12.34	9.08	6	4
B)	13	25046.97	26660.55	1613.59	54.81	784	70
	10*	95364.46	98420.58	3056.12	42.55	1034	35
	13	23512.1	24130.55	618.45	34.71	339	343
	13	24132.95	25045.65	912.7	22.23	343	784
	17	74075.91	74116.72	40.81	18.45	19	0
	5	150883.7	150964.4	80.73	16.13	43	2
	15	20606.43	21218.85	612.43	13.17	17	0
	10	45468.76	47063.96	1595.2	12.52	14	0
	3	114797.7	114806.5	8.77	12.09	7	3
	2	86902.21	88092.11	1189.89	11.62	21	9

* Region containing PLCE1 gene.

Reanalysing the Affymetrix 50K and 500K array datasets, both showed long regions of homozygosity common to the confirmed affected. The regions are longer on both the genetic and physical scale than those detected including ID25 and have greater evidence in the form of high SNP number. The regions on chromosome 10 and 13 appeared in all 3 datasets and are thus the most convincing candidate regions.

5.4.5 Candidate regions

Investigating these candidate regions further, the homozygous regions falling partially or wholly within these candidate regions were examined in all individuals. The candidate region on Chromosome 10 defined by the Illumina data is located between 95364.46-98420.58 Kb. the region defined by the Affymetrix 500K data is located between 95282.73-96261.83 Kb. The candidate region on chromosome 13 is located between 25046.97-26660.55 Kb in the Illumina data and 23906.63-25639.08 Kb in the Affymetrix 500K data. Table 5.12 shows that the 3 confirmed affected individuals have long homozygous regions across the length of the candidate region on chromosome 10 and the mildly affected (ID25 and ID114) and the unaffected (ID17 and ID18) have many much smaller regions of homozygosity. The same pattern is seen in the chromosome 13 region, with the exception of the unaffected ID17 which is homozygous across the candidate region. The affection status of ID17 is not thought to be ambiguous therefore this result seems to rule out the chromosome 13 region as a direct cause of this condition.

Table 5.12 Regions within or overlapping the chromosome 10 candidate region, in the Illumina 550K and Affymetrix 500K datasets. Showing regions of homozygosity in all 7 individuals genotyped.

		Phenotype; unaffected (0) mild (1) affected (2)	location (Kb)		KB length	LDU	SNPs	SNPs in following region	Other regions within or adjacent to the candidate region
			Start	End					
Illumina550K	ID3	2	72628.49	98526.09	25897.61	448.35	8958	2	
	ID4	2	95364.46	98526.09	3161.64	47.23	1070	2	
	ID19	2	95364.46	98420.58	3056.12	42.55	1034	35	
	ID114	1	97725.08	97988.57	263.49	4.25	101	0	+69 smaller regions
	ID25	1	96284.84	96459.62	174.78	0.47	23	1	+107 smaller regions
Affy500K									+18 large adjacent regions
	ID3	2	95226.92	97016.52	1789.60	24.90	239	310	
	ID4	2	95282.73	96261.83	979.10	19.00	165	377	
	ID19	2	95282.73	97016.52	1733.79	21.97	232	310	
	ID25	1	96127.14	96359.42	232.28	0.11	12	4	+20 smaller regions
	ID17	0	95772.92	95956.76	183.84	3.31	29	2	+9 smaller regions
	ID18	0	95772.92	95956.76	183.84	3.31	29	4	+10 smaller regions

Table 5.13 Regions within or overlapping the chromosome 13 candidate region, in the Illumina 550K and Affymetrix 500K datasets. Showing regions of homozygosity in all 7 individuals genotyped.

		Phenotype; unaffected(0) mild (1) affected (2)	location (Kb)		KB length	LDU	SNPs	SNPs in following region	Other regions within or adjacent to the candidate region
			Start	End					
Illumina550K	ID3	2	23104.63	27358.66	4254.03	145.68	2002	3	
	ID4	2	23104.63	27358.66	4254.03	145.68	2002	3	
	ID19	2	25046.97	26660.55	1613.59	54.81	784	70	+6 other large regions
	ID114	1	24938.85	25207.14	268.29	3.40	115	0	+112 smaller regions
	ID25	1	24229.03	24469.23	240.21	1.94	72	0	+107 smaller regions
Affy500K	ID3	2	23906.63	25639.08	1732.45	42.64	322	177	+3 other large regions
	ID4	2	23906.63	27103.08	3196.45	89.00	635	69	+4 other large regions
	ID19	2	23906.63	26863.54	2956.91	86.16	582	7	+3 other large regions
	ID25	1	24226.68	24467.35	240.67	1.69	29	0	+61 smaller regions
	ID17	0	23906.63	27103.08	3196.45	89.00	635	313	+8 other large regions
	ID18	0	24933.66	25181.31	247.65	3.02	55	3	+39 smaller regions

5.5 Discussion

5.5.1 Phenotype ambiguity

There are several issues to be tackled in these data the most important being ambiguity in phenotype. ID25 was originally assigned as affected but was subsequently assigned as mildly affected. Removing this individual from the analysis made a large impact on the results. ID114 is a newborn with a mild phenotype, carrying out the analysis with and without this individual was the best way to insure against the possibility of a change in phenotype status of this individual. Only ID3, ID4 and ID19 have had their diagnosis of DMS confirmed by renal biopsy. Urine tests of all available adults in the family found no proteinuria, making it unlikely that there are other mildly affected family members. This evidence and the inheritance pattern in the pedigree show the disease is most likely to be an autosomal recessive condition. It is also sensible to consider the possibility of mutation heterogeneity, for example ID25 and ID114 were diagnosed with milder symptoms than the other individuals and it is possible that a different mutation is causing a different phenotype in these cases and further justifies analysis excluding these individuals.

5.5.2 Genotype quality

The form of the data available added complexity to this analysis. There are 3 different data sets, with 7 individuals genotyped on one or more platform. However, using several datasets has the advantage of extra SNP coverage and replication of results on an independent genotyping platform. The Illumina array has a more even distribution of SNPs on the genetic map (using a haplotype tagging approach to choose their SNPs) and the highest SNP density which should be helpful in this analysis. Melin *et al.* (2007) showed that the 10K Affymetrix array failed to detect a homozygous region, later detected by the Affymetrix 100K array, due to insufficient local SNP density. There is also, however, a disadvantage of very high SNP density in this type of analysis because as SNP density increases so does the expected number of wrongly called heterozygous SNPs, even though the percentage remains small. This can have a large effect on results by breaking up long homozygous regions, whereas miscalled homozygotes would have a much smaller effect on results. Multiple datasets allowed the comparison of genotypes and inferences to be made about the accuracy of the genotype calling. There was a lack of long regions and thus promising results in the original analysis of the Affymetrix 50K and 500K data, including ID25. In such an inbred family this led to questions about the frequency of mistyped heterozygous genotypes.

Genotype call rates for all platforms were high but with notable lower rates for ID3 and ID19 in the Affymetrix and Illumina datasets respectively (table 5.3). Lower call rates do not necessarily imply that the genotypes which were called are questionable since a quality score threshold is applied to all data. However, examination of confidence scores showed that ID3 in the Affymetrix dataset and ID19 in the Illumina dataset have slightly lower mean scores for both genotype calls (table 5.4). Overall confidence was slightly higher for heterozygous over homozygous calls in the Affymetrix dataset but the opposite in the Illumina dataset, although the differences were very small. The availability of genotype calls for 76,116 SNPs in 4 individuals on 2 platforms allowed discrepant results to be investigated. Table 5.5 shows the discrepancies between datasets, ID19 has more cases where Illumina calls a heterozygote and Affymetrix calls a homozygote and ID3 has more cases where the opposite calls are made. This suggests that the lower call rate and average confidence scores for these individuals in these datasets may increase the number of erroneously called heterozygotes. The mean confidence scores for heterozygote Illumina calls when Affymetrix calls a homozygote are lower (0.6971 and 0.7073) than other classes of discrepancy. When the opposite calls are made the Affymetrix scores are lower (0.2344 and 0.2452). Suggesting that the heterozygote calls rather than the homozygote calls are the more questionable.

Removing different percentages of the data based on the confidence scores shows that only by removing 98.95% of the data (90% from Affymetrix and 90% from Illumina) were all the discrepancies removed (table 5.7). A 10% cut off leaves a discrepancy rate of 0.23% down from 0.63% and retains 83.05% of the data which seemed a reasonable balance between data quality and quantity. All known discrepancies were removed from both datasets since it cannot be reliably judged which genotype call is correct.

5.5.3 Detecting deletions

A small percentage of SNPs are NoCalls for all affected individuals although some of these are successfully genotyped on another platform (table 5.7). There were no runs of consecutive NoCalls in all affected individuals which might indicate a deletion. However, the LYK5 paper (Puffenberger et al. 2007) found only one NoCall common to all affected individuals, but it happened to be in a region of the genome poorly typed on the Affymetrix 10K array. The high density genotyping in this study on 3 platforms means failure to detect a deletion is unlikely, but small deletions are still possible.

5.5.4 Homozygosity

Table 5.8 shows that 39.11-47.73% of SNPs typed in the different samples are homozygous for all affected individuals, showing the high level of homozygosity in this inbred pedigree. The summary of homozygous regions detected in the HapMap samples (table 5.10) cannot be directly compared with the results in this study (table 5.9) since the size and number of regions depends to a great extent on SNP density and the HapMap sample had approximately 700,000 SNPs. However, it is clear the size and number of homozygous regions in these samples exceed those in the HapMap samples, which is expected given the consanguinity in the pedigree. The regions detected on the Affymetrix 500K array are a little shorter than expected given the Illumina results, but they are more numerous suggesting that the regions are broken up by isolated and perhaps erroneously called heterozygotes. Table 5.9 also shows shorter regions for ID3 in the Affymetrix datasets compared to the other samples and shorter and more numerous regions for ID19 in the Illumina sample. Both of these samples have lower call rates and average confidence scores than other individuals genotyped on the same platforms further suggesting the presence of miscalled heterozygotes breaking up otherwise long regions of homozygosity.

5.5.5 Regions of homozygosity common to the affected individuals

Only the affected individuals were considered to determine common regions of homozygosity. To require the normal individuals to be heterozygous or homozygous for the opposite allele across the entire region is too strict a criteria, especially since there are many SNPs where only one allele is present in this sample. Therefore a SNP could be homozygous in the affected and unaffected family members without indicating that the region does not segregate with the disease. Instead these SNPs may just be uninformative in this population or sample, since accurate information on allele frequencies in this particular population is not available. It has been shown that regions of homozygosity are more often found in regions of high LD where there has been little recombination and haplotypes inherited through both parents from a common ancestor are more likely to remain intact (Gibson, Morton, and Collins 2006). For this reason the homozygous regions were prioritised on the LDU scale which takes into account linkage disequilibrium, giving more weight to regions of lower LD and high LDU than regions of higher LD which have fewer LDU.

Initial results using the Affymetrix data and including ID25 did not give definitive results. Both Affymetrix datasets showed some large regions but each containing few SNPs and therefore no compelling evidence making it difficult to prioritise these regions (appendix 4). Analysing the Illumina data with the new affected individual ID114 also failed to give an expected long region with a large SNP number. However, when ID114 and ID25 were removed from the analysis, limiting to only the confirmed affected, the results changed dramatically. Two candidate regions were identified one on chromosome 10 and one on chromosome 13. The number of SNPs indicated a high level of information in these regions and the number of SNPs in the following region showed that there are 3 regions on chromosome 13 that are separated by single markers, and it is therefore possible that genotyping errors split one long region into 3 smaller ones (table 5.11). Removing ID25 from analysis in the Affymetrix 50K and 500K data confirmed these 2 regions (appendix 5).

5.5.6 Candidate regions on chromosome 10 and 13

These results are particularly interesting because the regions are longer and more convincing (due to higher SNP number) than previous results and the region on chromosome 10 contains the gene PLCE1. This is the strongest candidate gene as it was detected by Hinkes *et al.* (Hinkes et al. 2006) in individuals with the same phenotype (published after the initial analysis). The candidate regions were investigated further to determine the size of homozygous regions in each individual which lie within the common region detected. The evidence in support of the chromosome 10 region relies on the phenotype status of ID114 being 'unaffected'. The mild and possibly treatable condition of ID25 and ID114 may be caused by a different mutation. It does not seem likely that ID25 and ID114 are simply heterozygous for the mutation and therefore less severely affected, since they have a mixture of heterozygote and homozygote genotype calls across the candidate region. The region on chromosome 13 is more complicated, since ID17 an unaffected individual also shares the region of homozygosity, and seems to rule out this region as a direct cause. Other possibilities are a more complex cause of the disease involving 2 interacting genes; the family carries another undetected autosomal recessive condition caused by the chromosome 13 region; or the chromosome 13 region is shared by chance, which is perhaps not unlikely in such a consanguineous family. The presence of the PLCE1 candidate in the chromosome 10 region means a mutation in this gene is a possible cause of the disease and worthy of more investigation.

Future work will include sequencing the 34 exons of PLCE1, and looking for known or new mutations. If mutations are found in the affected individuals this would allow screening of other family members and possibly classification of milder cases that may resolve with treatment, like ID25 and ID114, and those that are more severe. If no mutations are detected, it is possible that expression analysis using biopsy samples may be carried out to see if the gene product is expressed. If the results do not implicate the PLCE1 gene, other genes in the candidate regions will have to be examined by function and then by sequencing.

5.6 Conclusion

Analysis and the presence of the strong candidate gene, PLCE1, indicate that this gene in the chromosome 10 region is most likely to be causal. Further experimental work will be needed to confirm the presence of a mutation in this gene. The chromosome 13 region is more complex since it is shared by an unaffected individual but may have a role in modifying the effect of mutations in the PLCE1 gene (if this is confirmed). There is still no answer to the cause of the milder and possibly treatable condition for ID25 and ID114. However, if this gene allows distinction between the mild and severe cases, mutation screening and carrier testing, it would be invaluable to this family and potentially other cases of congenital nephrotic syndrome with diffuse mesangial sclerosis.

Chapter 6 - A genome-wide association mapping study using an anonymous data sample

6.1 Introduction

Association has overtaken linkage as the most promising method for genome wide studies to determine genes involved in common diseases. Association promises higher resolution and the ability to locate variants to a much smaller interval on the Kb scale, as well as higher power when variants of modest risk are sought. Large unrelated case control cohorts are easier to recruit than multiply affected families, especially for diseases of late onset like many complex diseases. However, association analyses require a higher density of genetic markers than linkage analysis. There are several advances which have made whole genome association analyses feasible. Firstly, the availability of the reference sequence of the human genome with fewer gaps and increasing annotation (UCSC Genome Browser 2007). Secondly, a database of human genetic variation (SNPs) made available by the genotyping efforts of the international HapMap project and their recently released phase II data (International Hapmap Group 2005; Frazer et al. 2007). This provides approximately 4 million SNPs in each of 270 individuals from 4 populations, providing a genome-wide average SNP density of 1 SNP every 600bp. LDU maps have been updated to include the Phase II data, and are publicly available (Lau et al. 2007; Kuo, Lau, and Collins 2007). Also increasing the feasibility of genome wide association studies are the continuing advancements in high throughput genotyping technologies. For example, both Affymetrix and Illumina have large single chip or bead based array systems able to analyse upwards of 500,000 SNPs on a single array. The falling price of this technology has made it a viable option for smaller groups as well as international consortia.

There have been many examples of association projects successfully identifying an association, such as, the lymphotoxin-alpha gene in myocardial infarction (Ozaki et al. 2002), complement factor H in age-related macular degeneration (Klein et al. 2005), PARK10 locus in Parkinson disease (Maraganore et al.

2005), and a variant near the INSIG2 gene associated with obesity (Herbert et al. 2006). However, results in this type of analysis are plagued by lack of replication particularly since a positive association is more likely to be detected if the association is over represented in the sample tested, meaning that a larger sample would be needed to replicate the results (Chanock et al. 2007). This becomes less of a problem when very large sample sizes are analysed. Many international consortia are combining resources to conduct such large scale genome-wide association studies with >1000 cases and controls. Recently the Wellcome Trust Case Control Consortium (WTCCC) published results of the largest study so far, detecting 24 strong association signals and a further 58 moderate association signals over 7 diseases (WTCCC 2007). The high power of these studies promises the most comprehensive results, however independent replication should always be required. The WTCCC results represent replication of several previously reported associations and many of the new associations detected have been replicated in independent studies (Todd et al. 2007; Zeggini et al. 2007; Frayling et al. 2007).

Most association mapping studies rely on a single SNP approach, for example the single SNP chi square test for association at each marker, looking for any SNPs that have a chi square above a certain significance threshold. This has a considerable 'multiple testing' burden especially when using high density genotype array technologies. The p values should be corrected for the number of tests carried out to adjust for the increased probability of obtaining a significant result by chance the more tests carried out. The standard correction is the Bonferroni adjustment which reduces the threshold p value for significance i.e. 0.05 (5%) becomes 0.05/number of tests. Most multiple testing adjustments assume independent SNPs, however, due to LD, this is not the case, making the Bonferroni adjustment ultra conservative. There are other methods such as the False Discovery Rate (FDR) which controls the proportion of type 1 errors in the significant results and is less conservative (Storey and Tibshirani 2003). It is difficult to determine the best strategy for determining a threshold for significance and for this reason permutation tests are often used. This involves shuffling the case control status of the samples and determining an empirical significance threshold from this distribution.

Due to the difficulty of determining a genome wide significance level many researchers ignore this problem since the need for this type of adjustment is debated (Perneger 1998). Performing a second stage analysis or replication is seen as sufficient. The results of the first stage direct the choice of SNPs and regions for a more concentrated second stage where fewer tests would mean less of a multiple testing problem. It may not be necessary to determine formal significance levels in the first stage of a multi stage analysis, instead choose all regions around markers with an uncorrected significant P value for follow-up. It is also possible to rank results according to several criteria such as proximity to candidate genes, or in terms of possible functional significance. Another approach is Bayesian analysis where prior information is used to predict posterior outcomes (Morris, Whittaker, and Balding 2000; Morris 2006). Such prior information can include candidate genes, and previous associations or linkage results.

As well as single SNP tests genotypes can be combined to form haplotypes thereby considering multiple genotypes in association with a disease. Haplotypes have greater power than single SNPs if the causal SNP is not tested or if there are multiple mutations in a gene or region. However, since genotypes in unrelated individuals are unphased, the actual haplotypes inherited from the parents are unknown, statistical algorithms are used to infer the most likely common haplotypes. The EM algorithm and several modifications are implemented in computer programs such as SNPHAP (Clayton 2002). A coalescent based Markov chain model is used in the program PHASE (Stephens and Donnelly 2003). The accuracy of haplotype prediction is good, but limited as longer haplotypes are considered, since there is increasing error and also a massive computational load. Simple 2-5 SNP haplosets may be more powerful for association within a candidate gene or region but the difficulties of determining haplotypes mean interpretation of genome-wide haplotype analysis is complex.

CHROMSCAN-cluster analyses the genome using a region by region scan, multiple SNPs in a region are simultaneously tested by composite likelihood to model association. This reduces the number of tests required and thus the multiple testing burden. The regions analysed are determined by LDU length and SNP density. CHROMSCAN-cluster is a high throughput version of the program CHROMSCAN and analyses multiple regions in parallel subsequently combining the results, in this way CHROMSCAN-cluster can handle large datasets without difficulty. CHROMSCAN-cluster, like LDMAP, is based on the Malecot equation. However, instead of describing the exponential decline of association between markers of known physical location, it describes the decline of association between SNPs and an unknown causal location. CHROMSCAN also accounts for the autocorrelation between SNPs due to LD by using a rank-based permutation test under Ho to determine significance (Morton et al. 2007; Collins and Lau 2007). CHROMSCAN has been used to analyse the CYP2D6 region on chromosome 22. The CYP2D6 gene is responsible for the metabolism of 20% of drugs, mutations in the gene lead to the poor metaboliser (PM) phenotype. As a proof of principle the known location of the gene was predicted by CHROMSCAN using genotype information from surrounding SNPs. The predicted location was within 2Kb of the actual location and shows the power of this method in a candidate region analysis (Maniatis, Collins, and Morton 2007). The genome-wide properties of this method still require investigation.

To test and develop association mapping methods it is possible to analyse simulated data that mimic some of the important features of real data. However, this study collaborates with a research group on a genome-wide association study with the understanding that the disease is not disclosed. In this way the many questions raised by advances in association mapping on a genome-wide scale can be realistically addressed before the expected flood of data requiring such analysis and also allowing testing of the CHROMSCAN-cluster program.

6.2 Aims

The aims of this chapter are to make use of genome wide LDU maps in a genome-wide association study using anonymous data. The parallel CHROMSCAN-cluster program will be tested and the properties and problems of genome wide studies will be investigated. The most significant single SNP chi-square (msSNP) in each region will be determined and compared and combined with evidence from the composite likelihood results of

CHROMSCAN-cluster, with a view to selecting promising regions from a stage 1 analysis for further analysis in stage 2.

6.3 Methods

6.3.1 Data

The data consist of 239,146 SNPs genotyped across 798 individuals (403 cases, 395 controls). 17 SNPs removed due to lack of a Kb location, some of which mapped to multiple locations in the USCS May04 sequence. Only the autosomes were considered, leaving 233,686 genotypes. The data were then filtered to remove SNPs which deviated from Hardy-Weinberg Equilibrium with a χ_1^2 of 10 or more, leaving a total of 230,400 SNPs for analysis. Genome-wide cosmopolitan LDU maps updated with the latest (Phase II) HapMap data were used for this analysis, since no information was available on the population sample.

6.3.2 CHROMSCAN-cluster

For analysis by CHROMSCAN-cluster, the SNP data were split into non-overlapping regions which cover at least 10 linkage disequilibrium units (LDU) and contain a minimum of 30 SNPs without breaking blocks of linkage disequilibrium (LD). This gives 5,387 regions across the genome. Multiple SNPs across a region are simultaneously tested for the presence of a causal locus by using composite likelihood to model association which reduces the number of tests required. CHROMSCAN-cluster is based on the Malecot equation, describing the exponential decline of association between a SNP and an unknown causal location.

$$\hat{Z} = (1 - L)Me^{-\varepsilon\Delta(\hat{s}-s)} + L$$

A new parameter (s) is included in the model to estimate this location. Two sub-hypotheses of the Malecot model are used to test for a causal polymorphism within each region. Model A, which assumes no association between affection status and SNPs, is taken as the null hypothesis and compared with Model D which estimates disease location ($\hat{S}$), the intercept (M), and residual association (L). The test statistic, X, is determined by the difference in the sum of squares between these two models

$$X = \Lambda_A - \Lambda_D$$

Where

$$\Lambda = \sum_i K_{zi}(\hat{z}_i - z_i)^2$$

In order to account for autocorrelation between SNPs as a result of LD, the significance of this test statistic (X) is determined empirically by a rank-based permutation test. The case-control status is randomly shuffled and the test repeated to give a distribution of X_i under the null hypothesis (H_o). The replicates are sorted and assigned p values by rank/n. The corresponding χ^2_3 and variance are calculated from this. Values of variance assigned to X values surrounding the X under H1 are used to assign a variance to X under H1, from which a χ^2_3 and p value can be calculated. The chance of encountering a very significant association by chance increases as the number of replicates increases. For example, it is expected to see one p value of approximately 0.001 when there are 1000 replicates i.e. 1/1000. Thus to determine accurate levels of significance on this distribution, the number of permutation replicates must approach 10/P so that interpolation of the variance under H_1 is reliable.

6.3.3 Investigating the number of replicates

CHROMSCAN-cluster uses a rank-based permutation method to determine an empirical significance based on a null distribution. The number of replicates required to produce this distribution determines the computational time of the program. It is therefore important to determine an optimal number to have confidence in the results and still run the program efficiently. Initially the program was run with 100 replicates, however due to the increased speed of CHROMSCAN-cluster this was increased this to 1000 without incurring an unreasonable computational load. These results were compared to investigate the effect replicate number has on the p values. To ensure that the p values obtained were accurate the most significant regions ($p<0.01$) were then repeated using 50,000 replicates. The speed of CHROMSCAN-cluster allows 50,000 replicates to be run for a subset of regions without difficulty; this number will be unnecessary in most cases but allows accurate p values to be determined approaching 0.0002.

6.3.4 Correcting the msSNP p values

In order to compare evidence from CHROMSCAN and single SNPs, we identify the most significant single SNP (msSNP) from each region. However, selecting the msSNP from a large number of SNPs (30 or more) biases the nominal χ_1^2 and conventional P value computed on the null hypothesis. To determine the effects of the region definition on the p values for msSNPs and composite likelihood, a stepwise regression model was tested. The dependent variable was the p value, and the independent variables were SNP number and LDU length. The only significant result showed the bias in msSNP p values caused by the number of SNPs in the region.

Using the principle that msSNP P values should correspond to $\chi_2^2 = -2\ln P$ (Fisher, 1950), the variance of this nominal χ_2^2 among regions in a genome scan with the same number of SNPs under H_0 should be $V = 4$ and the mean $\mu = 2$. If selection of msSNPs were unbiased, adjustment of V would give an estimate of μ near 2, whereas adjustment of μ is less sensitive to small values of P, and therefore would not provide a good estimate of V. The bias in μ must be reduced before adjusting V.

Since the regions defined by CHROMSCAN-cluster vary in the number of SNPs, subsets of regions with limited diversity, but including at least 100 regions, were selected in which to estimate the Bonferroni parameter R. R is the effective number of independent SNPs in a subset with S SNPs. For each subset the weighted mean number of SNPs is $S = \sum f_i m_i / \sum f_i$, where f_i is the number of regions with m_i SNPs. The Bonferroni model assumes a corrected P value of $P_{ci} = 1-(1-P_{ni})^R$ (Gibson et al. 2008).

To obtain a mean of $\chi_2^2 = 2$ when $\chi_2^2 = -2\ln P_{ci}$, the formula is rearranged to give the equation $\sum f_i + \sum \ln [1-(1-P_{ni})^R] = 0$.

As below,

$$\left[\frac{-2\sum \ln P_{ci}}{\sum f_i}\right] = \frac{-2\sum \ln\left[1-(1-P_{ni})^R\right]}{\sum f_i} = 2$$

This equation was then solved by *regula falsi* to find the Bonferroni R giving the desired mean χ_2^2 of 2. This method requires two estimates of R either side of the real value so that one gives a negative solution and the other gives a positive solution. These values of R are then incremented and iterated until a solution sufficiently close to zero, in this case to 5 decimal places, is obtained. The relationship between R and S was then determined by regression so that a value of R could be assigned to each region given S. Corrected P values for msSNPs are then given by $P_{ci} = 1-(1-P_{ni})^R$. This corrected P value is then converted to χ_2^2 by $\chi_2^2 = -2\ln P$.

To set the variance of χ_2^2 to 4 requires dividing both χ_2^2 and μ by $\beta = \sqrt{\frac{\sum(\chi_2^2-2)^2}{4(\sum f_i - 1)}}$ to give the desired variance with mean $2/\beta$, which is acceptable only if $\beta \sim 1$.

Analysis of composite likelihood is simpler, since only a correction of variance is required, after conversion to χ_2^2 by $\chi_2^2 = -2\ln P$.

6.3.5 Combination of evidence from msSNPs and composite likelihood

The relationship between the corrected msSNP χ_2^2 and the composite likelihood χ_2^2 was determined using a correlation analysis. A principal component analysis based on this correlation matrix gives a first principal component (PC1) which brings together the 2 variables and a second principal component (PC2) which shows the differences. PC1 was used to order and rank all the regions, this rank was converted to a P value by rank/n which was then converted to a χ_2^2 by -2lnP. This gives a χ_2^2 for each region based on the combination of evidence.

6.3.6 Investigating discrepancies

In order to investigate the properties involved in cases where there are discrepancies between the results of the msSNP and composite likelihood analyses, PC2 was investigated. Regions which had extreme values of PC2 (≥4 or ≤4) were investigated further. To check the role of the LDU map, the genotype data for each region were used to create data-specific 'local' LDU maps, which were then used by CHROMSCAN-cluster. The results and the quality of these local LDU maps were compared with the HapMap cosmopolitan LDU maps. SNP density and coverage was investigated by the number of SNPs in the region and by determining the inter-marker distances between the SNPs flanking the msSNP. The extent of LD around the msSNP was also determined to examine whether the msSNP was located within a 'step' or a 'hole', regions with high recombination, which may explain the lack of other disease associated SNPs nearby. The arbitrary location of the region when the msSNP falls at the beginning or end was also investigated.

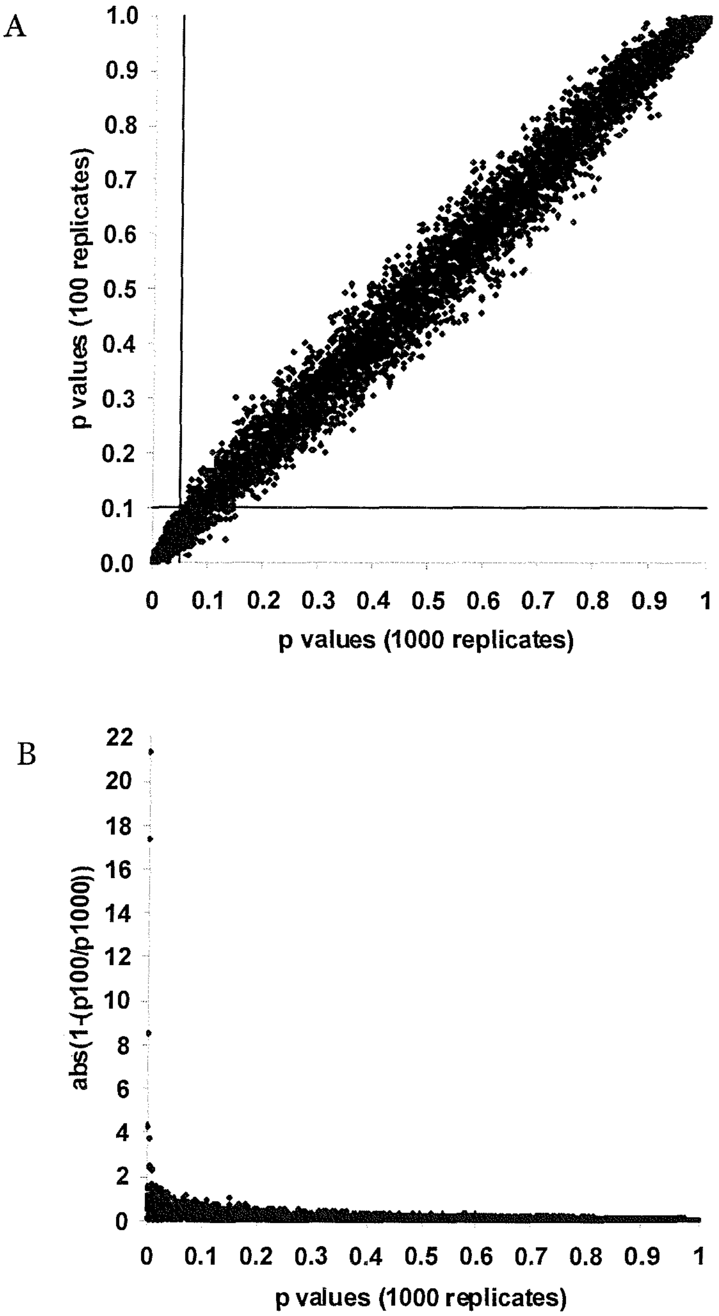
6.4 Results

6.4.1 Investigating the number of replicates

Running CHROMSCAN-cluster with 100 replicates, gave 16 regions with $p < 0.001$ and 61 regions with $p < 0.01$. With 1000 replicates, there were 8 regions with $p < 0.001$ and 53 regions with $p < 0.01$. P values for all 5387 regions, with 100 and 1000 replicates, across the whole genome are shown in figure 6.1. Figure 6.1A shows the p values are well correlated and have a linear relationship, although there seems to be more variation in the middle than at the extremes. The black lines indicate that a cut off of 0.1 at 100 replicates is required to capture all the regions with p values of < 0.05 at 1000 replicates.

However, the values actually differ more dramatically at the extreme low end, but the numbers are so small that they are not visible in figure 6.1A but are shown by a graph of the ratio of the two P values in figure 6.1B. This shows that the number of replicates is of more importance at the lower end of the scale, thus the 53 regions with a P value < 0.01 at 1000 replicates were repeated using 50,000 replicates.

Figure 6.1 A) p values with 100 and 1000 replicates. B) Ratio of p values for 100 and 1000 replicates.



6.4.2 msSNP p value bias

The msSNP p values are artificially exaggerated by the number of SNPs in the region (≥30) and are always below 0.3, whereas the p values for composite likelihood range 0-1 (figure 6.3A). This bias was illustrated by regressing p values on LDU length and SNP number. For the composite likelihood p values, neither variable was significant. For the msSNP p values only the SNP number was significant ($p<0.0001$) with an $R^2 = 0.0295$, showing, as expected, the higher the number of SNPs in a region the more chance of finding a highly significant p value.

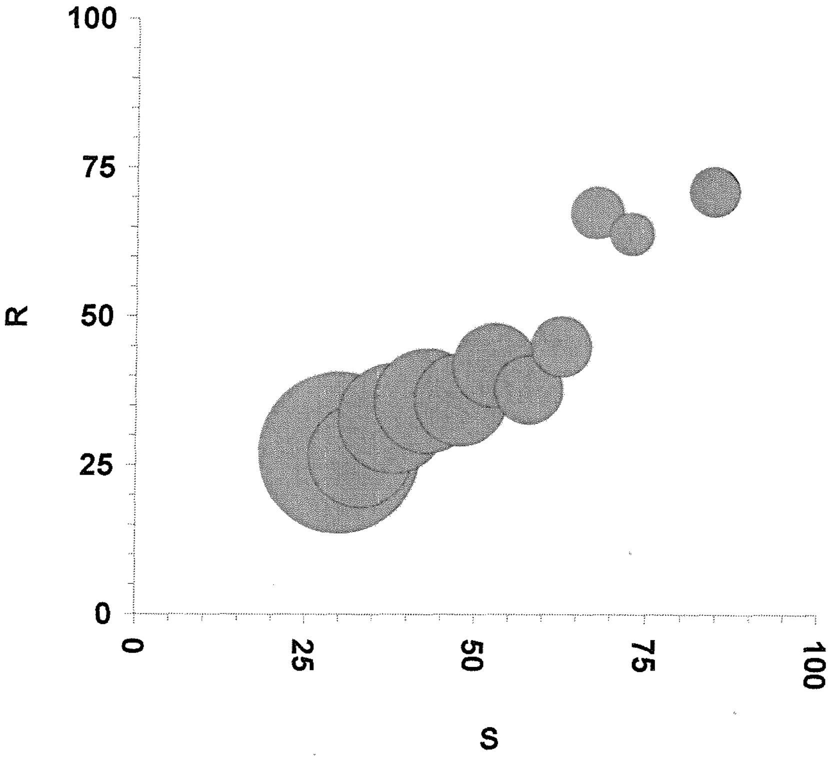
All regions were assigned to 11 subsets on the basis of the number of SNPs in the region. S was calculated as the weighted mean of the number of SNPs in the subset, R was calculated as the effective number of independent SNPs in the subset by *regula falsi,* and m is the number of regions in each subset. These values are given in table 6.1.

Table 6.1 Subsets of regions

SNP range	m	R	S
30	1536	27.10	30.00
31-35	638	26.55	33.12
36-40	712	32.86	37.96
41-45	650	35.76	42.87
46-50	508	35.99	47.89
51-55	421	41.73	52.93
56-60	280	37.66	57.91
61-65	221	44.89	62.77
66-70	163	67.41	67.82
71-75	109	63.89	72.83
76+	149	71.06	84.63
Total	5387		

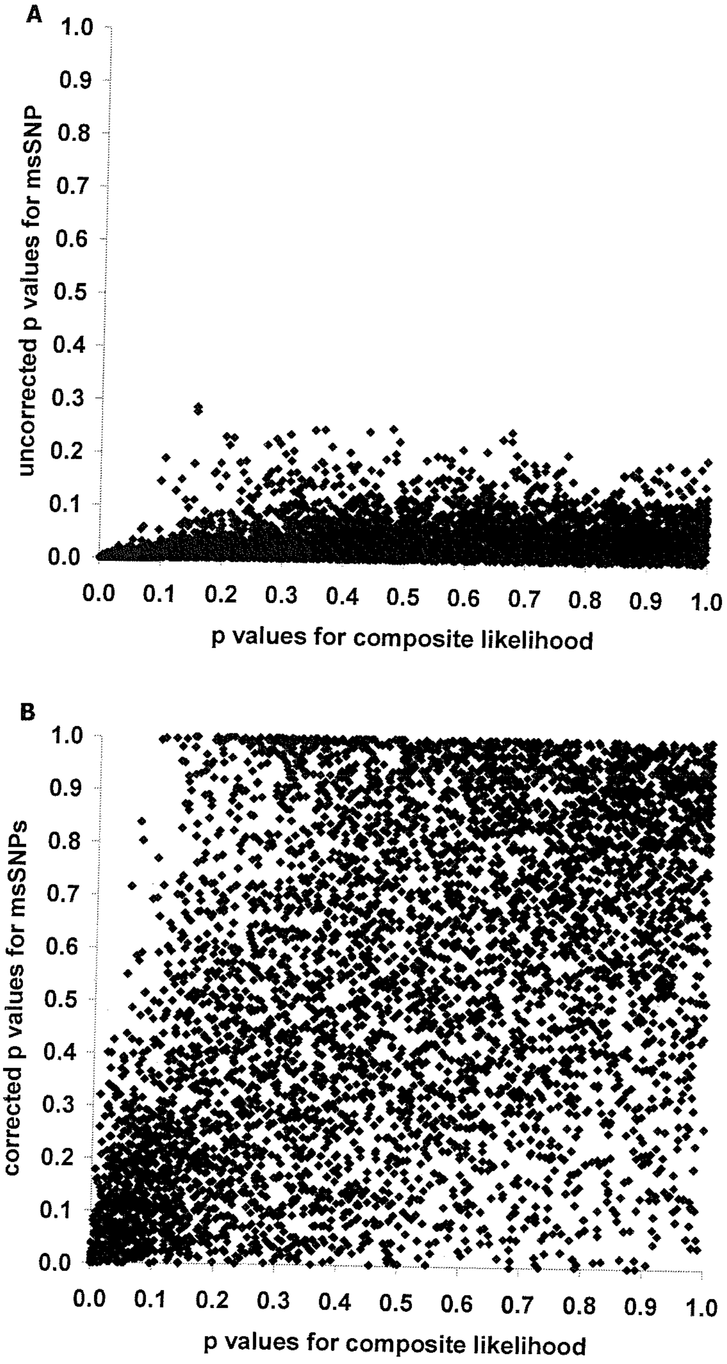
The relationship between R and S was shown to be linear by regression analysis and is illustrated in figure 6.2. This relationship allows R to be calculated for any S value.

Figure 6.2 The linear relationship between R and S, the area of each datapoint shows the number of regions per subset (m).



For each region the p value was corrected using the appropriate R value. This correction greatly reduces the significance of msSNPs, but conserves the order of the nominal p values. Converting these corrected p values to χ_2^2 gives a mean (μ) and variance (V) over all regions of $\mu = 2.0$ and $V = 5.2$. When V is constrained to its expected value of 4 under H_0, the estimate of μ becomes 1.8, corresponding to $\beta = 1.1$. The composite likelihood variance was 4.2 and after adjustment became, $V = 4$ and $\mu = 2$. The corrected msSNP p values are shown in figure 6.3 B.

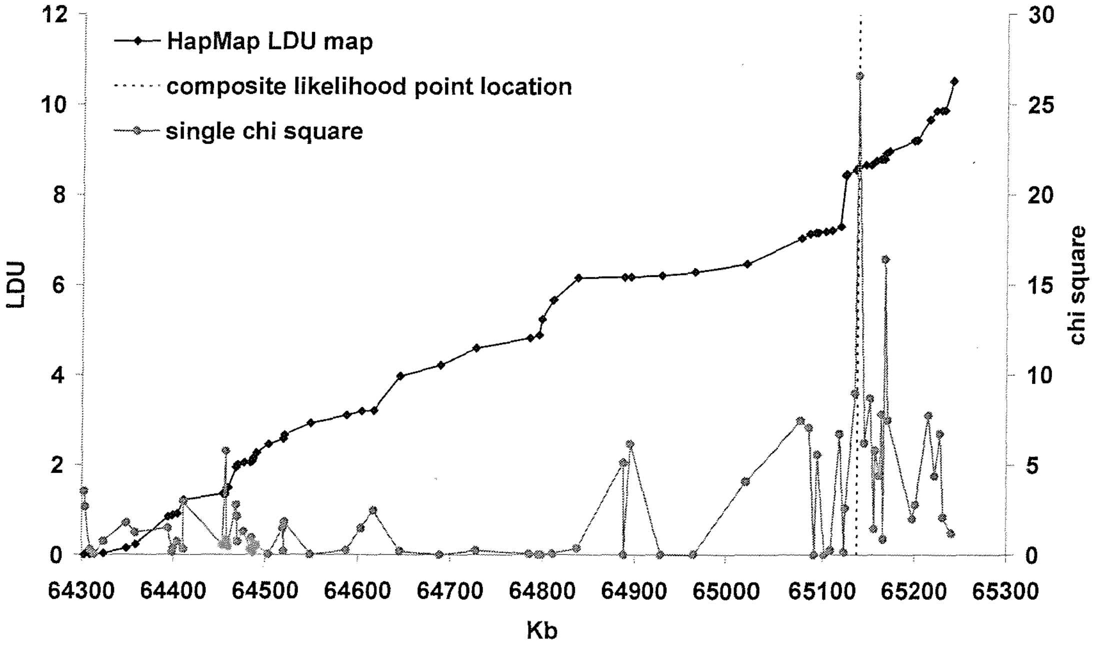
Figure 6.3 A) msSNP p values before correction against composite likelihood p values B) msSNP p values after correction against composite likelihood p values.



6.4.3 Combining evidence

A principal component analysis was applied to all regions for the adjusted χ_2^2 values for composite likelihood and msSNPs. The first principal component was converted to a rank, which was then transformed to p and χ_2^2 as for composite likelihood (Ewens 2003). The largest combined χ_2^2 is 17.2 and the top 50 are all greater than 9.3. A Bonferroni correction would require a critical significance level of .05/N, when N is the number of regions analysed. This corresponds to χ_2^2 of 23.17, which no region met, however this is considered a conservative correction. The region with the highest combined χ_2^2 is shown in figure 6.4.

Figure 6.4 Region ranked 1 by the combined metric (composite likelihood =1, msSNP=2)



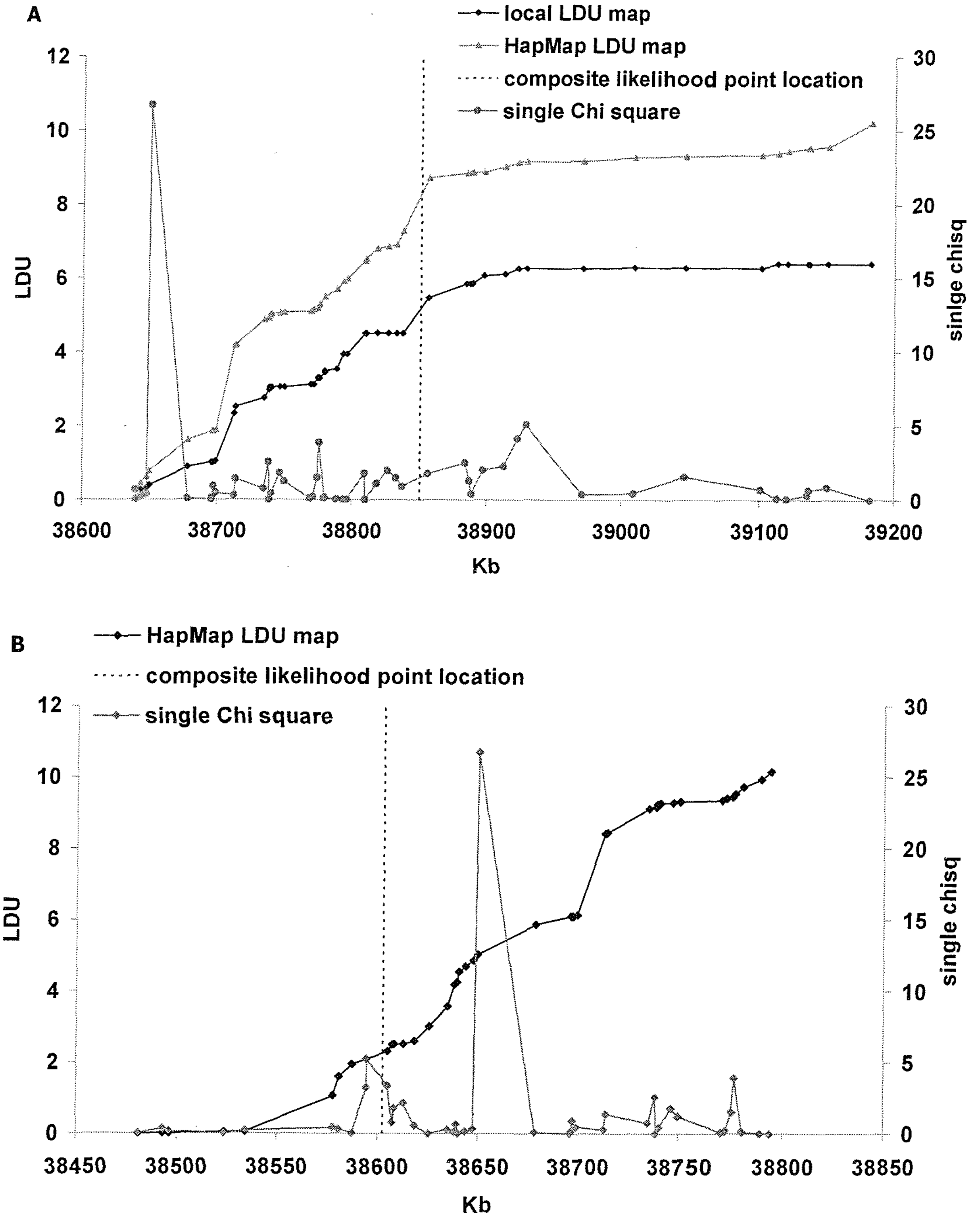
Among the 50 most significant regions ordered by the combined χ_2^2 values, 3 had a second principal component with a value greater than 4, indicating substantially greater significance for the msSNP than for composite likelihood. In no instance was the opposite observed (second principal component < -4). Local LDU maps for these outlier regions, constructed from control data, and the results of CHROMSCAN-cluster with the local maps were compared with initial results using the cosmopolitan HapMap LDU map. There was little difference between the 2 LDU maps in terms of the fit to the data or the structure and length of the maps. The composite likelihood results were also very similar and ε values of close to 1 for the regions validate the use of 1 as the epsilon parameter in CHROMSCAN-cluster. The ranks of the regions on the 3 different scales show the extent of the discrepancy (table 6.2). The number of SNPs and holes in each region, the distance in Kb between the SNPs flanking the msSNP and the LDU/Mb between these flanking SNPs are shown in table 6.2.

Table 6.2 Outliers favouring evidence from msSNPs

Rank			No. SNPs	No. holes (LDU>2.5)	flanking SNPs		msSNP	HapMap map	local map	
msSNP	composite likelihood	combined			Kb	LDU/Mb	χ^2_2	χ^2_3	χ^2_3	ε
1	3695	19	51	0	30.573	34.41	20.27	1.49	0.96	1.07
3	4731	29	30	0	3.956	57.38	18.60	0.68	0.45	1.12
4	569	16	30	2	3.970	19.14	16.88	6.33	5.20	1.11

When the regions were centred on the msSNP, maximising evidence from markers on both sides, the χ_3^2 remained low. For the region ranked 1 by msSNP, the composite likelihood χ_3^2 was 1.49, and in the msSNP centred region it was 1.55 (figure 6.5 A&B).

Figure 6.5 A) Region ranked 1 by msSNP (composite likelihood=3695, combined=19) B) Region centred on the msSNP.



The χ_2^2 values for the top 50 regions by combined rank are given in figure 6.6 and table 6.3. The 3 regions with PC2 >4 are shown in figure 6.6 as combined ranks 16, 19 and 29, and are shaded in table 6.3. In the top 50, the msSNP result is more significant than the composite likelihood result 24 times. Discrepancies occur when a very significant SNP is isolated with no other evidence from surrounding SNPs, composite likelihood gives more weight to regions with clusters of highly significant SNPs. This can be seen in figure 6.4. The 3 cases with the largest discrepancy (PC2 >4) were also the 3 with the largest difference in χ_1^2 between the msSNP and the next most significant SNP in the region (table 6.3).

Figure 6.6 The top 50 regions by the combined metric.

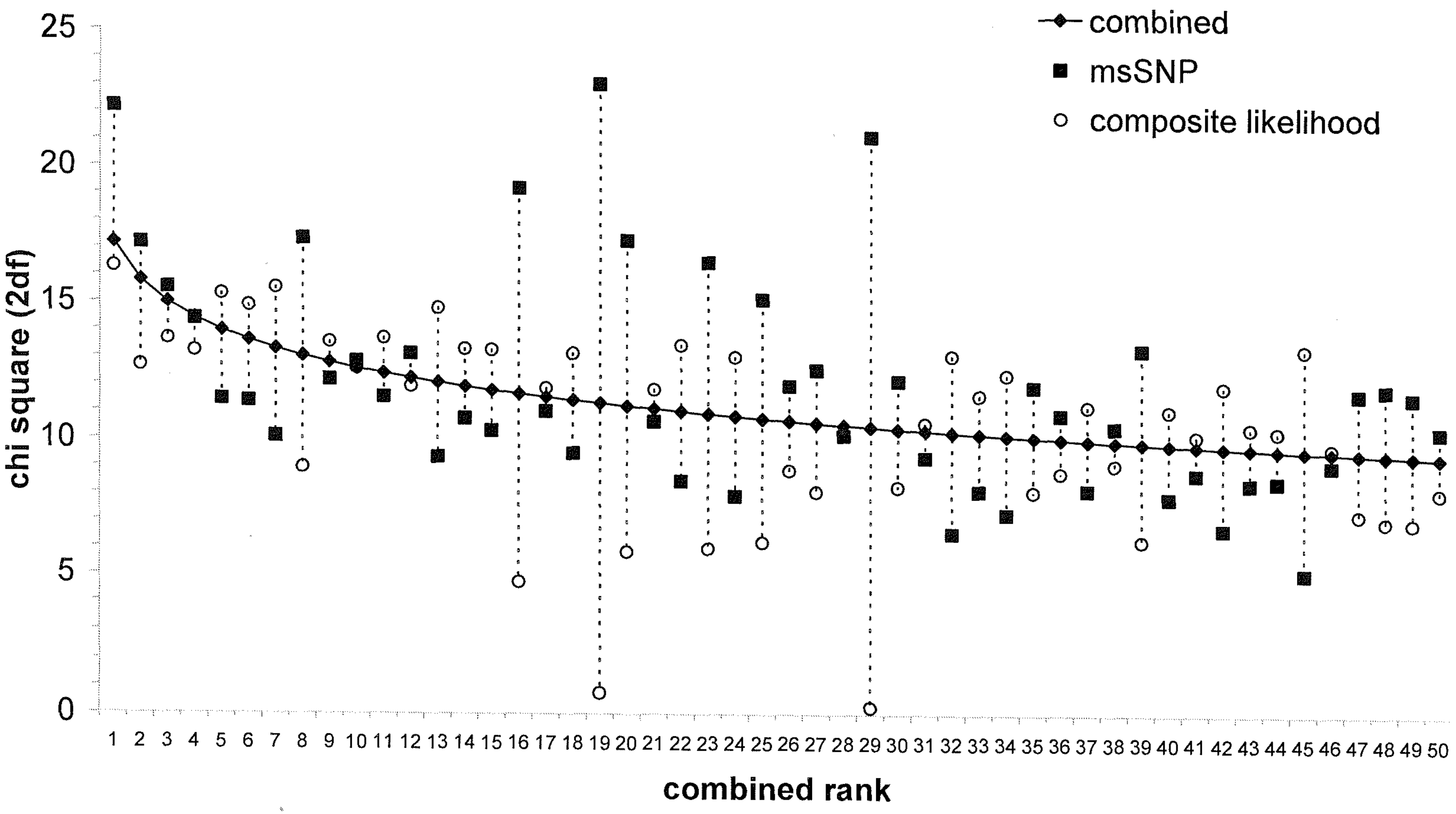


Table 6.3 The top 50 regions by the combined metric.

PC1	PC2	Rank			χ^2_2			No. SNPs	LDU/Mb	LDU difference	χ^2_1 difference
		combined	composite likelihood	msSNP	combined	composite likelihood	msSNP				
11.173	1.367	1	1	2	17.18	15.86	19.51	75	11.214	0.000	10.19
8.368	1.056	2	17	7	15.80	12.33	15.10	60	7.485	0.035	5.75
8.190	0.210	3	7	9	14.99	13.27	13.65	37	21.384	0.093	8.42
7.679	0.002	4	13	12	14.41	12.85	12.64	55	14.053	0.362	0.97
7.491	-1.634	5	3	37	13.96	14.89	10.06	45	33.129	0.159	2.02
7.316	-1.511	6	4	38	13.60	14.47	9.98	44	19.654	0.228	4.74
7.148	-2.132	7	2	54	13.29	15.11	8.87	30	80.039	0.000	1.00
7.133	2.393	8	64	5	13.02	8.69	15.25	37	41.994	0.000	9.21
7.109	-0.803	9	8	23	12.79	13.18	10.69	32	42.365	0.515	7.63
6.987	-0.266	10	18	19	12.58	12.24	11.28	31	38.276	0.000	4.23
6.954	-1.050	11	6	35	12.39	13.31	10.12	38	36.385	0.241	6.73
6.848	0.053	12	21	18	12.21	11.60	11.53	42	22.649	0.614	1.54
6.655	-2.110	13	5	75	12.05	14.38	8.20	34	29.741	0.826	1.55
6.575	-1.159	14	11	43	11.91	12.93	9.43	48	19.849	0.488	1.21
6.429	-1.285	15	12	50	11.77	12.90	9.05	30	60.868	0.658	1.49
6.247	4.437	16	569	4	11.64	4.55	16.88	30	61.102	0.000	18.16

6.165	-0.582	17	22	41	11.52	11.53	9.67	30	97.337	0.438	2.77
6.124	-1.500	18	14	72	11.40	12.77	8.32	85	9.528	0.364	0.28
6.098	6.984	19	3695	1	11.29	0.74	20.27	51	18.822	7.506	21.64
6.043	3.450	20	325	6	11.19	5.66	15.20	55	14.345	0.000	11.99
6.035	-0.678	21	23	44	11.09	11.48	9.35	51	19.360	0.008	0.03
5.915	-1.920	22	9	115	11.00	13.07	7.43	55	13.453	0.110	0.62
5.836	3.156	23	307	8	10.91	5.78	14.49	30	20.883	0.000	12.53
5.601	-1.942	24	16	144	10.83	12.66	6.95	30	93.065	0.000	0.06
5.498	2.651	25	279	11	10.75	6.02	13.30	64	11.492	1.559	12.18
5.424	0.749	26	67	27	10.67	8.60	10.50	30	30.086	2.052	0.98
5.341	1.200	27	103	20	10.59	7.85	11.03	40	21.783	0.000	3.44
5.340	-0.288	28	33	53	10.52	9.95	8.92	30	47.445	0.184	1.18
5.334	6.562	29	4731	3	10.45	0.25	18.60	30	102.507	1.326	17.61
5.280	1.018	30	95	24	10.38	8.02	10.68	37	28.991	0.000	3.36
5.206	-0.664	31	28	76	10.32	10.29	8.20	56	20.805	0.000	0.56
5.194	-2.389	32	15	274	10.25	12.71	5.74	40	46.479	2.343	1.46
5.177	-1.403	33	24	132	10.19	11.29	7.11	38	51.511	1.409	2.37
5.172	-1.930	34	19	201	10.13	12.03	6.36	37	26.273	0.000	1.50
5.156	1.011	35	100	28	10.07	7.85	10.50	39	38.868	0.000	6.01
5.071	0.453	36	71	42	10.02	8.52	9.59	45	17.275	5.518	5.08
5.055	-1.256	37	25	130	9.96	10.91	7.15	61	20.055	0.670	2.14

5.032	0.205	38	58	46	9.91	8.82	9.18	62	16.412	3.288	5.01
4.953	2.064	39	270	16	9.86	6.08	11.70	30	17.309	0.000	12.55
4.915	-1.287	40	26	148	9.81	10.76	6.91	30	64.687	0.379	1.70
4.864	-0.701	41	36	98	9.76	9.86	7.66	55	15.468	0.156	1.85
4.859	-1.952	42	20	259	9.71	11.62	5.89	30	41.413	0.103	0.96
4.857	-0.913	43	30	119	9.66	10.15	7.35	48	48.821	0.068	0.12
4.828	-0.843	44	31	116	9.62	10.01	7.41	35	30.483	0.075	0.26
4.809	-2.940	45	10	520	9.57	12.95	4.42	30	27.879	1.236	0.57
4.807	-0.442	46	41	85	9.53	9.41	7.95	48	46.781	0.053	0.12
4.795	1.219	47	152	32	9.48	7.05	10.28	46	20.270	0.007	4.74
4.757	1.357	48	174	30	9.44	6.80	10.42	30	27.112	4.215	5.51
4.651	1.287	49	188	34	9.40	6.75	10.17	30	40.621	0.303	10.59
4.649	0.511	50	104	49	9.36	7.84	9.07	30	33.717	0.030	3.67

χ_1^2 difference = difference in χ_1^2 between the most significant SNP (msSNP) and the next most significant SNP in a region, LDU difference = difference in LDU between the msSNP and the point location given by composite likelihood, LDU/Mb = LDU per megabase over the region, PC1/PC2 = 1st and 2nd principal components

6.5 Discussion

The data consist of an unknown phenotype coded 1, 0 for cases and controls, and are dealt with anonymously to gain some insight into the nature of problems encountered by association mapping, and the use of CHROMSCAN-cluster for genome-wide analysis. The genome is analysed in non-overlapping regions each containing 10 LDU (or at least 30 SNPs). The CHROMSCAN-cluster program handled the genome-wide data without difficulty, since the file is split into batches of regions which are run in parallel. The efficiency of CHROMSCAN-cluster allowed the whole genome to be analysed with both 100 and 1000 replicates. Comparison of these results showed that the p values are stable except at the most extreme lower end of the distribution, we therefore chose to rerun regions with a p value <0.01 with 50000 replicates. It is likely that a replicate number of 10/p would be adequate. Regression showed that a Bonferroni correction was required for the msSNP p values, which was complicated by the diversity of their regional lengths and SNP densities. However the *regula falsi* approach allowed an adjustment to be made based on the effective number of independent SNPs in each region. The corrected p values were in the range 0-1 and were no longer biased by the number of SNPs in the region. Figure 6.3 shows the msSNP p values before and after correction. The lower right of the graph in figure 6.3 (B) shows that there is a cluster of cases where the p value for the msSNP result is significant and the composite likelihood result is not. The lack of data points in the top left of the graph show that the opposite situation is rare.

Three χ_2^2 metrics were considered for each region; the most significant SNP in the region (msSNP), the composite likelihood result, and a metric combining both results by principal component analysis. Figure 6.4 shows the region ranked 1 by the combined metric. The msSNP and composite likelihood point-location implicate the same SNP which has the highest χ_1^2 in the region. However, some of the surrounding SNPs have very low χ_1^2 despite being in close proximity to the msSNP on the Kb and LDU scale.

All regions were ranked by the 3 metrics, and 3 regions with large discrepancies between results (PC2>4) were examined. These regions have a high msSNP rank and a low composite likelihood rank. To explain the lack of evidence from composite likelihood several options were investigated. A local LDU map showed little difference to the cosmopolitan HapMap LDU map used in this case, the maps were very similar in length and structure (figure 6.5 A). The composite likelihood results obtained using the local LDU map were very similar though slightly weaker than those obtained using the HapMap LDU map, possibly due to the lower SNP density in the local map (table 6.3). This evidence suggests the discrepancy is not due to the use of an inappropriate LDU map.

Low SNP density should not be a problem because CHROMSCAN-cluster accounts for SNP number in the region definition (10 LDU and ≥30 SNPs), the 3 regions have 51, 30 and 30 SNPs (table 6.3). However, a localised lack of SNPs could explain the lack of evidence from composite likelihood. The inter-marker distance between the SNPs flanking the msSNP are 3.96, 3.97 and 30.57Kb (table 6.3) which are not excessive and figure 6.4 shows that SNP coverage is generally even and the msSNP is not isolated on the Kb scale. This suggests that neither low SNP density nor insufficient SNP coverage explain the discrepancy. However, figure 6.3 shows that, for the region with a combined ranked of 1, there are several markers close to the msSNP with very low χ_1^2. It is possible that in these 3 cases increasing SNP density would find other significant SNPs nearby, and the isolated high χ_1^2 is due to a random lack of typing associated markers.

It is also possible that the msSNP is located in a step in the LDU map, a region of particularly high recombination, or a hole where an arbitrary LDU distance is applied to neighbouring SNPs when no LD (background only) is detected between SNPs. However, the msSNPs were not isolated on the LDU scale and the LDU/Mb between flanking SNPs was 19.14, 34.41 and 57.38 (table 6.3). The genome average LDU/Mb is 20.2 for the HapMap CEU sample and 28.4 for the YRI sample, therefore 57.38 indicates only a small step. The recent publication describing analysis of the Phase II HapMap data, show the presence of SNPs in regions of very high recombination (hotspots), which are described as 'untaggable' SNPs (Frazer et al. 2007). For association analysis in regions of high recombination these SNPs would need to be directly tested since surrounding SNPs have insufficient LD to provide any information. Regions of very high recombination are described as 'holes' in the LDU map. Although not directly close to the msSNP one region has 2 holes (≥2.5 LDU between SNPs), this may have the effect of reducing information and the number of SNPs in LD with the msSNP in the region. There were no holes in the 2 other regions.

Composite likelihood gives more weight to regions where there is a cluster of SNPs with high or moderate χ_1^2 results. These 3 regions all contain an msSNP with no other high χ_1^2 in the region (table 6.3). In 2 of the discrepant regions, the msSNP is at the beginning of the region (but not the first SNP). This would reduce information from the left of the msSNP. However, re-aligning the regions with the msSNP in the centre, to maximise the evidence from either side, failed to change the composite likelihood results (figure 6.5 B).

It may be the case that these 3 signals represent type 1 errors (false positives). However, only larger sample sizes and higher density genotyping will be able to answer this question. A known phenotype would also allow information from previous studies and functional considerations which may help determine the likelihood of type 1 error. At present there can be no objective recognition of the more reliable test and since it would be undesirable to miss a possible association, a combined metric was devised to help choose regions for follow up in stage 2. The 50 most significant regions representing approximately 1% of the genome seem a reasonable sample to investigate in stage 2. Combining evidence allows the very significant msSNP results to be included in the top 50 and is the best way to combine evidence for stage 2 to avoid losing potential candidate regions. This is a preliminary analysis, and although no signals met a Bonferroni adjustment for the number of regions, there were interesting findings. Further to this project a strategy for stage 2 would involve increasing SNP density in these 50 regions, which may resolve the discrepancies between msSNP and composite likelihood results, regions can then be prioritised based on candidate genes. Meta-analysis including previously published findings may help give more weight to results and could help to narrow down a region of interest.

There have been several examples of the successes of association mapping such as the recent publication of findings from the Wellcome Trust Case Control Consortium (WTCCC 2007). The release of this and other genome-wide data to the scientific community for further analysis, will allow comparison of methods, meta-analysis, and cohorts of controls for use in the study of other diseases (WTCCC 2007; Genetic Association Information Network (GAIN) 2007; Cancer Genetic Markers of Susceptibility (CGEMS) 2007). Databases of association analysis results will also be of great use in meta-analysis (database of Genotype and Phenotype (dbGaP) 2007).

The WTCCC managed to validate many of their findings with independent samples or previously published results. However, in general, association analyses suffer from a lack of convincing replication and publication bias towards positive results. As well as false positives there are many other explanations for non replication; samples from differing populations, differences in phenotype classification or assessment between studies, or an insufficiently sized replication sample. A larger sample is required for replication due to the increased chance of finding an association when it is over represented in the initial sample. Larger samples and combining samples by meta-analysis are approaches to increase power to find genes and validate results. However, the problems of sample consistency and population stratification become even more important.

Another possibility to increase power is to make use of LD patterns. One way is to impute genotypes based on LD patterns. The theory is that if a causal marker is not typed, an observed marker in LD can be used to detect the association, but power is reduced relative to any departure from perfect LD between the 2 markers. Using patterns of LD from the HapMap data to predict the genotypes at un-typed SNPs may regain some lost power, though only if the imputation is accurate. It is estimated that accuracy levels of >98% can be achieved, however, this is highly dependent of the local LD patterns, and accuracy would be severely compromised in regions of low LD or recombination hotspots. Also, although broad LD patterns are very similar across populations, fine scale differences mean that imputation would be inaccurate if the population sample investigated was not closely related to one of the HapMap study samples.

Statistically inferring genotypes in this way, which can then be analysed by any method, may provide an additional source of power for association studies, but must be used with caution (Marchini et al. 2007; Clark and Li 2007). The composite likelihood approach implemented in CHROMSCAN-cluster has the benefit of analysing multiple SNPs in a region reducing the need for multiple testing adjustments, and uses LD patterns described by LDU maps to increase power. Mapping on the LDU map rather than the Kb map is more powerful and was also shown to increase accuracy of the point location in the test case of CYP2D6 and the poor-metaboliser phenotype (Maniatis, Collins, and Morton 2007). Simulations were also carried out using SNPs from chromosome 4 of the Age-Related Macular Degeneration data described by Klein *et al.* (2005). These results show a 47% increase in accuracy of location estimate and a 5% increase in power when using the LDU map compared to the Kb map (Collins and Lau 2007).

Future studies are likely to involve analysis of more genetic variation, higher density genotyping, whole genome sequencing, as well as copy number variations (CNVs). Structural variants which have frequencies of >1% are considered genuine heritable polymorphisms, the structural Variation Database (Human Genome Structural Variation Project 2007) describes around 4,000 CNV loci. CNVs have also been identified in the 270 HapMap individuals and new SNP array technologies are being developed to score them, though SNP associations may not be sufficient to detect all CNVs. Knowledge of the extent of CNV contribution to phenotype is incomplete, gene dosage effects by duplication or deletion of a genes as well as regulatory influences by CNVs located outside of genes are thought to be involved. An example of a CNV with a phenotypic affect in complex disease is the CCL3L1 variant known to influence susceptibility to HIV-1 and rheumatoid arthritis (McKinney and Merriman 2007; Clark and Li 2007; Komura et al. 2006).

There are already examples of novel therapies and clinical interventions arising from association results, for example, clinical trials of Abatacept (CTLA4Ig) have shown evidence of its efficacy in rheumatoid arthritis (Ruderman and Pope 2005), though translating genetic risk into clinical relevance can be challenging. The best analyses will still miss rare moderate risk variants and small risk variants due to lack of power. The chances of transforming low risk or very rare variants directly into clinical interventions are not high, but may still give insights into disease pathways. There are still many statistical challenges to overcome when analysing genome-wide SNP data and the optimum approach has not yet been defined. The challenge will be to determine how best to exploit the massive accumulation of genomic data soon to be released.

6.6 Conclusions

This work has allowed testing of CHROMSCAN-cluster with genome-wide association mapping data, showing that the program is able to cope with high density data without difficulty. The results of this stage 1 analysis showed several regions with evidence for association, though none were significant after Bonferroni correction. msSNPs (most significant single SNPs) were also defined for each region analysed by CHROMSCAN-cluster. In three cases the evidence for association did not agree between the msSNP and CHROMSCAN-cluster results. The reasons for this discrepancy are not clear; however it is likely that higher density data and larger samples will resolve the issue. However, for this first scan of the data, a metric was devised to allow selection of regions for follow up, based on the combined evidence.

Chapter 7 – Summary and discussion

Linkage Disequilibrium (LD) describes the tendency for alleles to be inherited together more often than would be expected under random segregation. There has been increased interest, over the last few years, in a complete description of the structure and intensity of LD in the human genome across different population samples. The first descriptions of LD patterns and their relationship to recombination were published in 2001. Jeffreys *et al.* studied the MHC region on chromosome 6 and concluded that recombination was not evenly distributed across the genome but limited to small regions of 1-2Kb, referred to as recombination hotspots, which were separated by regions of high LD. This work was carried out by observing meioses in sperm (Jeffreys, Kauppi, and Neumann 2001). Daly *et al.* investigated a region of chromosome 5 using haplotypes to show that there are regions or 'blocks' of low haplotype diversity (high LD), separated by recombination sites (Daly et al. 2001). These findings led to a view of LD in the human genome that could be described by blocks of high LD interspersed with recombination hotspots.

An LDU map describes these patterns of LD in the form of an additive map. The LDMAP (and LDMAP+) program produces a description of LD using genotype (diplotype) data and the Malecot model which is used to model the decline of LD over distance. The map determines a location in Linkage Disequilibrium Units (LDU) for each SNP marker (Maniatis et al. 2002). The LDMAP program has proven capable of reproducing the block structure (LDU map plateaus) shown with the Daly *et al.* data and the hotspots (LDU map steps) described in the Jeffreys *et al.* data. This allowed validation of the LDMAP method since it was able to recover information about recombination hotspots from genotype data, which were originally detected by direct observation in sperm data (Zhang et al. 2002).

With the increasing availability of genotype data for this type of analysis, individuals with European ancestry were analysed for the first LDU map of a whole chromosome (22) with marker density ranging 1 SNP every 15-23Kb, showing the structure of LD and the high correspondence between LDU and linkage maps (Tapper et al. 2003). High density data (1 SNP every 2Kb) was produced on a 10Mb region of chromosome 20, enabling a more fine scale description of LD and investigation into the effects of SNP density, this showed the robustness of the LDU map (Ke et al. 2004). These data were then used to investigate LDU maps of different populations and the feasibility of a cosmopolitan LDU map (chapter 2). These data consisted of genotypes on 3 continental populations, with East Asian, African and European descent. Based on a previous study with smaller samples and regions (Lonjou et al. 2003) this work showed the high similarity in LDU patterns between populations, the differences in LD extent and the usefulness of a standard cosmopolitan LDU map that can be scaled to be applied to various population samples (Gibson et al. 2005). The similarity of LD patterns across populations described here, has also been shown for 3 chromosomes across 4 populations (De La Vega et al. 2005) and across chromosome 22 in 11 population isolates and one outbred European sample. The 'younger' isolates were shown to have more extensive LD than the outbred sample (Service et al. 2006). The major limitation of the work described in chapter 2 is that it is based on a region of a single chromosome and the results are interpreted to apply to the whole genome. Extension of this work was only possible when genome-wide data became available (International Hapmap Group 2005) and modifications to the LDMAP program (LDMAP+) allowed such large scale analysis.

Interest in creating a description of haplotype structure and LD across the Human Genome led to the initiation of the International HapMap project in 2002. Advances in genotyping technology enabled increasingly high density SNP genotyping in 270 individuals in 4 populations and the data were publicly released periodically via the HapMap website (International Hapmap Group 2005). The first release (11) to contain high density genotype data across all chromosomes (CEU sample only) was used to create the first genome-wide LDU map. This work allowed an estimate of effective bottleneck time for the CEU population sample based on whole genome data (Tapper et al. 2005). A preliminary analysis was carried out to compare all 4 populations as soon as data were available (4 chromosomes only). These results confirmed previous work on the similarities and differences between LDU maps in different populations (chapter 2), and allowed a first estimate of effective bottleneck times across the 4 populations. The HapMap Phase I data was the first release to contain high density whole genome data on all 4 populations. Genome-wide LDU maps were made for each population (chapter 3). This allowed extension of the analysis of LD patterns across different populations, to genome-wide data, and also allowed comparison with the linkage map and estimation of population age by calculating effective bottleneck time (t). The genome-wide LDU maps of the 4 HapMap population samples showed the same trends as the work on chromosome 20 (chapter 2). The major difference in LDU maps was between the African and non-African samples, with the least extensive LD in the African sample. The patterns of LD were again very similar, on the broad scale, across all chromosomes in all population comparisons ($r^2 \geq 0.99$). Such a high similarity suggests recombination hotspots are co-localised in all populations, since recombination is the major force determining LD. Comparison of the LDU and linkage maps over the whole genome showed a remarkable correspondence (97-99%) confirming this. The estimated age of 29,440-36,800 years for the CEU population falls short of the estimated 100,000 years since the 'Out of Africa' bottleneck. However, the effective bottleneck time is influenced by subsequent smaller bottlenecks which have the effect of increasing LD by restricting the haplotypes in the population. As well as subsequent bottlenecks, it is possible that the estimate is influenced by the small sample size (60 individuals) and the specific population sample (Utah residents with northern and western European ancestry).

Whole genome historical recombination maps of the HapMap data have been created using a coalescent method implemented in the LDHAT program (McVean et al. 2004). The LDHOT program which analyses the historical recombination maps provided by LDHAT, has been used on a publicly available genome-wide dataset (Hinds et al. 2005) produced by the genotyping company Perlegen Sciences and also the HapMap Phase I data (Myers et al. 2005) to predict over 25,000 recombination hotspots across the genome, the results are provided in the UCSC genome browser and through the HapMap genome browser (UCSC Genome Browser 2007; International Hapmap Group 2005). In a comparison of these maps and LDU maps of the genome, the LDU maps, which are based on a much simpler theory, show marginally higher levels of similarity than the historical recombination maps to the only genome-wide recombination information available, the linkage map (Kong et al. 2004; Tapper et al. 2008). However, overall the historical recombination and LDU maps are very similar even though the LDU map allows the inclusion of the effects of other stochastic processes such as selection, whereas the LDHAT program models only recombination. The inclusion of the effects of other processes in the LDU map allows the evidence of selection, for example, to remain and be investigated.

An accurate description of 'normal' variation is valuable in any analysis of disease causing variation. The whole genome LDU maps and the genotype data produced by the HapMap project provide the opportunity to carry out analyses of genome-wide variation such as regions of extended homozygosity. Longer than expected tracts of homozygosity have been shown in CEPH individuals with European ancestry using microsatellites (Broman and Weber 1999), but this was in part due to an identifiable relationship between some pedigrees. Long regions of homozygosity tend to occur in families where there is a certain degree of consanguinity regardless of levels of LD in these regions. To a lesser extent long regions of homozygosity occur in isolated populations and generally show a lack of haplotype diversity, which can also be shown by patterns of LD (Service et al. 2006). Chapter 4 investigates the extent of long homozygous tracts in the outbred populations represented in the HapMap project, and shows that even in outbred populations extended tracts of homozygosity are present and have a strong relationship with patterns of LD as shown by the LDU map. Three individuals from the HapMap data were identified as having longer and more numerous tracts than other individuals from the same population sample. This suggests that these individuals were from families where there has been some consanguinity in the past few generations, thus reducing the haplotype diversity to less than would be expected in the general population. In this way high density SNP data were used to evaluate the levels of inbreeding in an individual's history (Gibson, Morton, and Collins 2006). Two of these individuals were also identified as showing cryptic relatedness, i.e. relatedness in the ancestors of the sampled individual, by the HapMap analysis group in their publication on the Phase I HapMap data, although a direct analysis of homozygous tracts was not carried out.

Further to this work several other studies have analysed extended homozygosity. Li *et al.* describe long contiguous stretches of homozygosity in Han Chinese, Taiwan aborigines, Caucasians and African-Americans. 17 of 20 homozygous tracts determined in the HapMap CHB sample (chapter 4) were also present in the Han Chinese sample studied. The possible alternative explanation for extended homozygosity, the presence of a deletion when hemizygotes are miscalled as homozygotes, was ruled out using DNA copy number determination by hybridization intensity analysis and real-time quantitative PCR (Li et al. 2006). Simon-Sanchez *et al.* analysed 276 DNA samples (from lymphoblast cell lines) from Caucasian subjects. They found 26 samples with contiguous tracts of homozygosity >5Mb, they also repeated analysis in a proportion of subjects with DNA extracted directly from blood samples. They were able to show that the process of creating lymphoblastic cell lines did not create long regions of homozygosity that were not present in the original sample. They did not directly rule out the possibility of segmental uniparental disomy as a cause, but concluded that it was unlikely since many of the subjects with one long region of homozygosity also had several other regions (Simon-Sanchez et al. 2007). Another study was able to determine that long regions of homozygosity are not due to uniparental disomy (Curtis 2007). This paper analysed genotype data on 10 CEPH individuals and their parents to determine the presence of mendelian errors that would indicate uniparental disomy, for example, mother=AA, father=BB and child=AA or BB. It was determined that although these type of errors appeared within long homozygous regions they did so less than would be expected by chance and did not occur contiguously as might be expected if segmental uniparental disomy was the cause of the homozygosity (Curtis 2007). The latest release of the HapMap data (Phase II) has been published and an analysis of homozygosity was included (Frazer et al. 2007). The analysis detected the 3 individuals with unusually high levels of homozygosity, highlighted in the work described in chapter 4, and 'identified 79 regions over 3 Mb in 51 individuals, with many segments extending over 10 Mb' (Frazer et al. 2007).

Homozygosity usually occurs in inbred samples, and is particularly common in consanguineous pedigrees where a child is likely to have inherited the same haplotypes from both parents because they are related. Autozygosity mapping exploits this to determine the location of genetic variants causing autosomal recessive conditions. The use of high density SNP array data is becoming a popular way of determining regions of autozygosity that could be causal (Gutierrez-Roelens et al. 2006; Melin et al. 2007; Puffenberger et al. 2007). Following on from the work to determine the extent of homozygosity in outbred individuals (chapter 4), a study was undertaken to determine a candidate region for autosomal recessive congenital nephrotic syndrome in individuals from a large consanguineous pedigree (chapter 5). The analysis was designed to make use of high density SNP array technology, avoiding a traditional linkage approach which would rely on uncertain pedigree information and few individuals (4 affected). The high density SNP data allows determination of regions of homozygosity (presumed autozygosity) and increased power was obtained from use of LDU maps, since the correlation of LD patterns and homozygosity has been shown (chapter 4). Two regions of interest were determined, one of which contained a strong candidate gene. Further laboratory work is currently underway, preliminary results have determined a 4 base-pair deletion in exon 3 of the PLCE1 gene, present in all the affected individuals which generates a premature translational termination codon. However, this mutation also seems to be present and homozygous in one of the parents of an affected child, and further investigation is required.

The data for autozygosity mapping was provided with 4 individuals genotyped on 2 high density platforms, the Affymetrix 500K chip array and the Illumina humanhap550 bead array. This allowed comparison of the genotype calls and results show that low call rates and low call scores for an individual correlate with more discrepant genotypes where the platform with the poorer sample calls a heterozygote. Overall the Affymetrix platform had more discrepant heterozygous calls than the Illumina platform. Although this does not necessarily indicate that Affymetrix has more errors, an excess of inaccurate heterozygote calls would break up otherwise long homozygous regions which is critical for this type of analysis. Increasing the quality score threshold used to define a 'NoCall' genotype, decreased the number of discrepancies between the 2 platforms. Therefore, an increased quality score threshold was used to reduce the number of potential errors in the data prior to analysis. These results should direct the choice of genotyping platform and quality score threshold appropriate for future autozygosity mapping projects.

LD offers gene mapping at a much increased resolution than linkage mapping, although investigation is ongoing to determine the power, accuracy and sensitivity of these methods (Maniatis et al. 2004; Maniatis et al. 2005; Zaykin, Meng, and Ehm 2006; Morris et al. 2003). However, many studies use a simple single SNP χ^2, at least as a stage 1 scan (WTCCC 2007) and patterns of LD are also important for choosing the most appropriate SNPs for analysis, avoiding redundancy.

After using LDU maps to increase power for a disease gene search in a consanguineous pedigree (chapter 5), the maps were then used to search for a disease gene in an unrelated case-control sample by genome-wide association analysis (chapter 6). The cosmopolitan LDU maps created from Phase I HapMap data were used for this study. This was a stage 1 analysis, using an unknown disease, to determine regions for follow up in stage 2. CHROMSCAN-cluster is based on the Malecot model, like LDMAP, but determines association between markers and a disease. Results revealed several regions with evidence of association, however, none met a strict Bonferonni correction. The msSNP (most significant SNP) in each region analysed by CHROMSCAN-cluster was also determined. Three regions showed discrepancies, where the evidence from the msSNP analysis did not agree with CHROMSCAN-cluster results. Evidence from both sets of results were combined to determine regions for follow-up in a stage 2 analysis to avoid missing potentially important regions. This was a preliminary analysis with no information about the disease or, therefore, any candidate genes. However, the aim of determining regions for follow up was accomplished (Gibson et al. 2008).

There are several aspects to this project which could be investigated further in the future. Genome-wide association analysis using CHROMSCAN-cluster and msSNPs in higher density data and larger samples may resolve discrepancies between the 2 sets of results. Several large datasets have recently been released which would offer an ideal opportunity for more investigation of these methods (WTCCC 2007) (Cancer Genetic Markers of Susceptibility (CGEMS) 2007)

(Genetic Association Information Network (GAIN) 2007). This project also included some initial investigation into the evidence for selection detectable using high density SNP genotype data. Fine-scale differences in LD patterns between populations in particular regions and extended regions of homozygosity are both possible indicators of a selective sweep. However, LD and homozygosity are highly correlated and analysis must take both into account when determining putative regions under selection (Wang et al. 2006). Several studies have carried out genome scans for evidence of selection (Carlson et al. 2005; Zhang et al. 2006; Voight et al. 2006; Tang, Thornton, and Stoneking 2007; Sabeti et al. 2007). Analysing extended homozygosity with reference to LDU maps has the potential to give an advantage over these methods, and an analysis which makes use of the most recent high density data and the forthcoming HapMap data on 7 new population samples (Frazer et al. 2007), should provide interesting results.

It would also be valuable to build on the success of the autozygosity mapping work (chapter 5), with analysis of new datasets, this would provide validation of the method used. It is also possible to modify the method, for homozygosity mapping of recessive disease in outbred populations (Simon-Sanchez et al. 2007; Miyazawa et al. 2007). One such method was able to determine highly penetrant recessive loci in schizophrenia using long stretches of homozygosity (Lencz et al. 2007). Using homozygosity to search for selection and disease variants is an exciting and current field of research, however careful consideration of LD patterns and interpretation of results is required. LDU maps which provide a high resolution metric map of the amount and structure of LD in the genome will be of great value.

Appendices

Appendix 1. Chapter 3 - Properties of LDU maps, for each chromosome.

chr	pop	pairs	loci	E_kb	L_kb	M_kb	Swept radius	v_kb	E_LDU	L_LDU	M_LDU	v_LDU	holes	LDU length	kb length
1	CEU	2998419	59118	0.009	0.156	0.883	117.581	2.135	1.042	0.156	0.971	0.827	189	4182.493	245448
2	CEU	3147075	61911	0.008	0.150	0.893	127.747	2.374	1.033	0.150	0.975	0.916	164	4100.860	243341
3	CEU	2579693	50715	0.008	0.152	0.877	129.370	2.434	1.036	0.152	0.977	0.890	205	3629.121	199130
4	CEU	2203783	43357	0.008	0.151	0.912	117.742	2.209	1.038	0.151	0.975	0.892	192	3421.510	191682
5	CEU	2221207	43687	0.008	0.152	0.879	128.768	2.417	1.032	0.152	0.977	0.902	125	3197.480	180747
6	CEU	2417066	47504	0.008	0.154	0.897	121.525	2.341	1.040	0.154	0.975	0.930	140	3074.203	170676
7	CEU	1842213	36308	0.009	0.154	0.892	111.968	2.202	1.048	0.154	0.968	0.895	153	2906.975	158412
8	CEU	2782669	54740	0.009	0.145	0.911	115.267	2.565	1.032	0.145	0.985	0.881	102	2678.105	146141
9	CEU	2210255	43495	0.011	0.150	0.885	92.086	2.485	1.039	0.150	0.973	0.948	109	2597.887	136218
10	CEU	1748957	34496	0.009	0.153	0.870	111.698	2.335	1.045	0.153	0.969	0.887	124	2649.018	134989
11	CEU	1644502	32326	0.008	0.153	0.872	118.104	2.368	1.040	0.153	0.975	0.872	137	2543.549	134292
12	CEU	1760581	34834	0.009	0.154	0.868	114.708	2.343	1.043	0.154	0.970	0.891	149	2692.728	131958
13	CEU	1292984	25441	0.009	0.155	0.898	107.098	2.212	1.037	0.155	0.974	0.892	109	1987.097	96193
14	CEU	1085385	21344	0.009	0.154	0.863	111.748	2.284	1.038	0.154	0.965	0.902	91	1818.273	87057
15	CEU	624023	12440	0.007	0.152	0.802	144.904	2.578	1.041	0.152	0.966	0.884	143	1898.181	81862
16	CEU	871772	17360	0.012	0.155	0.842	84.909	2.070	1.059	0.155	0.957	0.838	128	1931.467	89882
17	CEU	896052	17724	0.009	0.152	0.839	114.478	2.501	1.053	0.152	0.977	0.804	111	1924.621	81652
18	CEU	1483222	29136	0.011	0.148	0.937	89.223	2.380	1.028	0.148	0.985	0.913	108	1894.649	76111

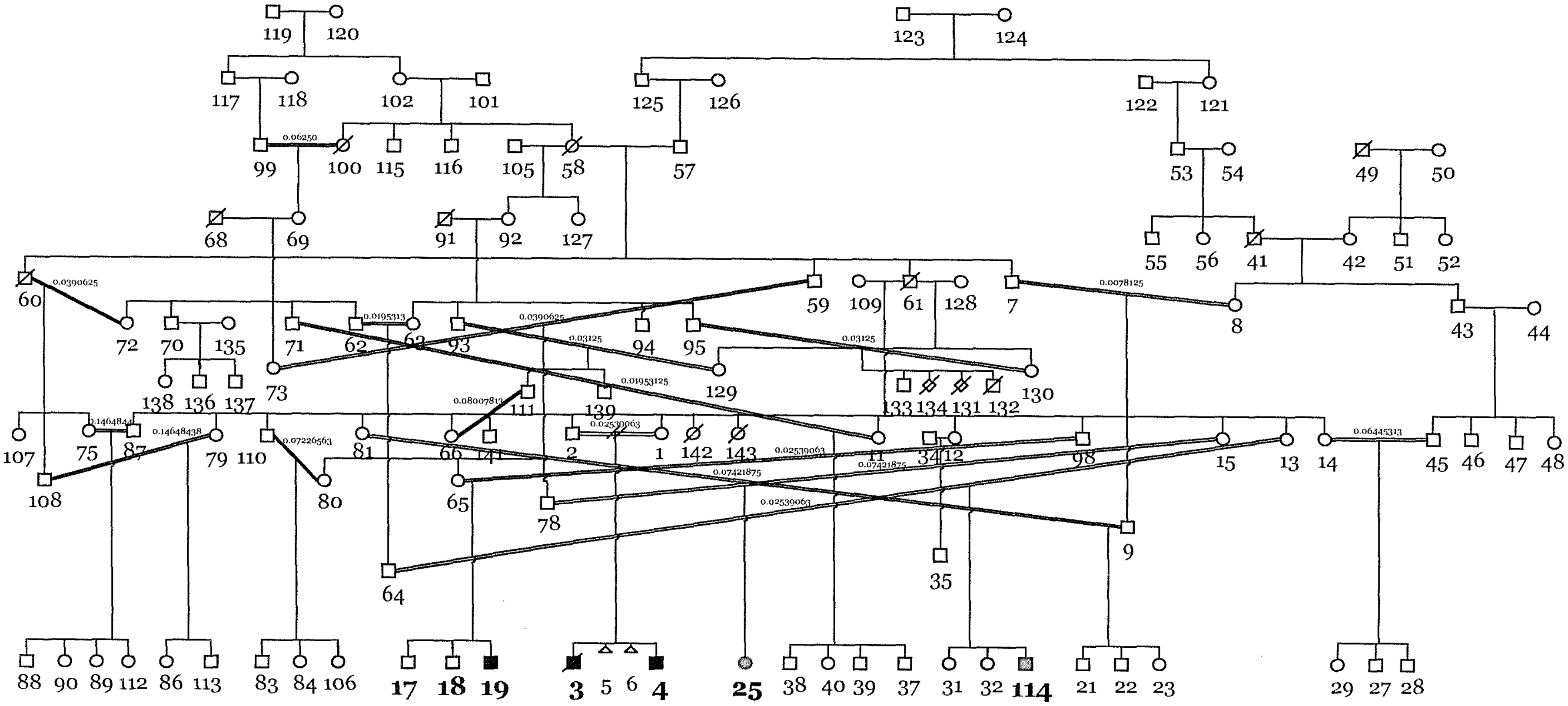
19	CEU	626105	12339	0.013	0.157	0.845	79.852	1.948	1.052	0.157	0.969	0.786	115	1725.310	63742
20	CEU	777285	15327	0.011	0.156	0.855	88.119	2.130	1.054	0.156	0.967	0.822	109	1633.093	63585
21	CEU	757237	14889	0.015	0.153	0.934	67.524	2.262	1.032	0.153	0.977	0.891	51	990.788	37027
22	CEU	709324	13959	0.012	0.151	0.808	84.744	2.676	1.043	0.151	0.977	0.860	38	1023.686	34923
23	CEU	1660095	33615	0.004	0.197	0.876	245.897	1.395	1.025	0.197	0.990	0.471	119	1749.021	150761
1	HCB	2551176	50373	0.009	0.173	0.888	116.352	1.831	1.029	0.173	0.978	0.664	356	4809.644	245227
2	HCB	2762531	54430	0.008	0.172	0.906	122.924	1.744	1.023	0.172	0.979	0.670	301	4471.459	243341
3	HCB	2164361	42615	0.008	0.173	0.892	124.999	1.793	1.027	0.173	0.981	0.650	295	3921.675	199131
4	HCB	1886361	37171	0.009	0.174	0.923	116.124	1.591	1.028	0.174	0.979	0.669	274	3653.231	191652
5	HCB	1901227	37456	0.008	0.172	0.889	125.445	1.779	1.022	0.172	0.983	0.650	253	3514.323	180747
6	HCB	2210501	43418	0.008	0.174	0.884	118.121	1.811	1.038	0.174	0.954	0.799	212	3356.298	170669
7	HCB	1499404	29638	0.009	0.178	0.910	110.852	1.564	1.032	0.178	0.972	0.656	238	3111.463	158406
8	HCB	2604974	51218	0.009	0.164	0.924	113.686	1.969	1.020	0.164	0.993	0.618	186	2965.858	146141
9	HCB	2058288	40471	0.011	0.170	0.895	89.856	1.854	1.030	0.170	0.976	0.677	220	2985.126	136216
10	HCB	1596965	31540	0.010	0.174	0.882	104.161	1.738	1.033	0.174	0.970	0.673	227	3017.485	134944
11	HCB	1440908	28366	0.008	0.174	0.878	118.605	1.798	1.028	0.174	0.975	0.676	237	2859.712	134292
12	HCB	1534157	30209	0.010	0.178	0.855	101.073	1.659	1.034	0.178	0.944	0.780	246	3061.044	131980
13	HCB	1182325	23232	0.009	0.174	0.904	106.387	1.698	1.019	0.174	0.978	0.661	178	2239.873	96206
14	HCB	946760	18618	0.009	0.174	0.875	111.134	1.704	1.032	0.174	0.973	0.657	141	1953.690	87047
15	HCB	832855	16519	0.009	0.175	0.820	113.945	1.827	1.041	0.175	0.973	0.641	187	2099.593	81777
16	HCB	750898	15043	0.013	0.177	0.859	78.619	1.510	1.056	0.177	0.968	0.631	206	2208.123	89882
17	HCB	767925	15263	0.009	0.173	0.840	108.729	1.825	1.037	0.173	0.983	0.615	192	2221.497	81626
18	HCB	1333632	26218	0.012	0.168	0.938	86.636	1.803	1.022	0.168	0.988	0.644	174	2156.418	76111
19	HCB	557644	11009	0.013	0.179	0.877	74.484	1.423	1.041	0.179	0.977	0.601	218	1982.320	63584

20	HCB	647048	12785	0.012	0.177	0.873	83.026	1.598	1.035	0.177	0.969	0.636	180	1869.510	63585
21	HCB	761137	14956	0.015	0.172	0.940	65.677	1.773	1.016	0.172	0.983	0.662	93	1116.590	37027
22	HCB	686607	13512	0.013	0.169	0.867	75.016	1.988	1.030	0.169	0.982	0.642	64	1121.687	34760
23	HCB	1428435	29172	0.005	0.223	0.882	217.866	1.098	1.012	0.223	0.994	0.370	201	1989.874	150761
1	JPT	2530333	49977	0.008	0.175	0.884	124.400	1.771	1.035	0.175	0.978	0.674	269	4316.699	245219
2	JPT	2744504	54072	0.008	0.173	0.899	131.438	1.736	1.031	0.173	0.977	0.703	202	4019.888	243341
3	JPT	2144533	42238	0.007	0.175	0.886	135.452	1.785	1.034	0.175	0.979	0.683	232	3519.691	199124
4	JPT	1866053	36756	0.008	0.176	0.926	120.593	1.563	1.034	0.176	0.978	0.679	224	3392.487	191652
5	JPT	1888568	37215	0.008	0.175	0.892	130.342	1.737	1.027	0.175	0.982	0.697	191	3198.336	180747
6	JPT	2194520	43105	0.008	0.175	0.882	125.219	1.816	1.035	0.175	0.952	0.837	174	3108.028	170669
7	JPT	1483986	29330	0.008	0.180	0.897	119.641	1.581	1.039	0.180	0.968	0.676	195	2848.551	158406
8	JPT	2590769	50945	0.008	0.166	0.925	121.012	1.959	1.028	0.166	0.993	0.642	158	2759.806	146141
9	JPT	2053174	40372	0.011	0.172	0.893	93.685	1.843	1.036	0.172	0.976	0.712	160	2685.329	136216
10	JPT	1591294	31403	0.009	0.175	0.874	114.228	1.760	1.038	0.175	0.971	0.698	192	2810.696	134944
11	JPT	1436562	28279	0.008	0.176	0.876	127.292	1.800	1.032	0.176	0.976	0.711	155	2493.952	134292
12	JPT	1527551	30085	0.009	0.179	0.850	106.310	1.657	1.039	0.179	0.939	0.812	185	2740.319	131958
13	JPT	1176240	23112	0.009	0.176	0.892	115.510	1.730	1.032	0.176	0.974	0.697	139	2016.941	96206
14	JPT	937082	18439	0.008	0.175	0.868	120.132	1.733	1.037	0.175	0.975	0.670	130	1873.782	87047
15	JPT	823239	16336	0.008	0.177	0.820	118.712	1.779	1.044	0.177	0.969	0.664	146	1868.918	81777
16	JPT	750704	15029	0.012	0.179	0.863	81.416	1.518	1.059	0.179	0.967	0.646	159	2013.295	89882
17	JPT	760932	15136	0.009	0.175	0.835	114.202	1.795	1.044	0.175	0.983	0.619	130	1921.419	81626
18	JPT	1319489	25938	0.011	0.169	0.934	91.234	1.821	1.024	0.169	0.988	0.653	130	1933.176	76111
19	JPT	538975	10668	0.013	0.179	0.869	78.795	1.436	1.044	0.179	0.972	0.632	159	1742.227	63584
20	JPT	642269	12685	0.011	0.177	0.870	90.328	1.611	1.053	0.177	0.971	0.670	120	1590.950	63585

21	JPT	756081	14857	0.014	0.174	0.944	69.282	1.708	1.028	0.174	0.985	0.703	56	967.773	37027
22	JPT	685070	13483	0.013	0.171	0.868	78.595	1.992	1.030	0.171	0.979	0.666	86	1139.183	34760
23	JPT	1362571	27910	0.005	0.228	0.885	214.724	1.023	1.024	0.228	0.988	0.387	139	1694.528	150761
1	YRI	2959312	58317	0.013	0.163	0.734	76.863	1.790	1.107	0.163	0.881	1.066	229	6092.728	245265
2	YRI	2998624	58512	0.014	0.161	0.754	82.000	1.810	1.101	0.161	0.890	1.010	150	5875.767	243402
3	YRI	2582077	50716	0.012	0.162	0.748	86.068	1.905	1.106	0.162	0.919	1.006	178	4986.423	199151
4	YRI	2253989	44318	0.013	0.162	0.773	75.802	1.642	1.110	0.162	0.902	1.000	155	4605.885	191649
5	YRI	2121833	41737	0.012	0.163	0.755	81.346	1.721	1.099	0.163	0.909	0.966	159	4528.654	180751
6	YRI	2462162	48364	0.013	0.163	0.780	77.628	1.823	1.096	0.163	0.914	1.034	109	4299.258	170675
7	YRI	1742456	34373	0.012	0.164	0.761	82.645	1.701	1.125	0.164	0.908	0.971	163	3930.962	158406
8	YRI	2953573	58035	0.015	0.158	0.812	64.734	2.148	1.096	0.158	0.941	1.092	75	3943.892	146141
9	YRI	2307503	45386	0.017	0.158	0.767	59.785	2.080	1.104	0.158	0.915	1.062	122	3714.281	136290
10	YRI	1870403	36878	0.014	0.162	0.742	70.749	1.823	1.109	0.162	0.907	0.985	113	3827.724	134890
11	YRI	1628897	32032	0.012	0.163	0.746	81.149	1.789	1.111	0.163	0.913	0.976	125	3472.884	134291
12	YRI	1813663	35789	0.014	0.162	0.752	71.450	1.773	1.100	0.162	0.900	0.997	135	3748.051	131993
13	YRI	1425104	28015	0.014	0.163	0.759	72.722	1.796	1.103	0.163	0.914	1.004	78	2750.522	96190
14	YRI	1065775	20958	0.012	0.163	0.726	83.443	1.767	1.115	0.163	0.908	0.977	92	2481.317	87057
15	YRI	955882	18954	0.013	0.164	0.682	74.614	1.720	1.119	0.164	0.890	0.936	111	2490.796	81777
16	YRI	892103	17712	0.013	0.162	0.658	78.833	1.808	1.164	0.162	0.902	0.902	138	2582.023	89882
17	YRI	870795	17225	0.013	0.163	0.690	76.442	1.697	1.129	0.163	0.905	0.902	150	2679.498	81626
18	YRI	1618944	31809	0.019	0.159	0.828	53.209	1.989	1.080	0.159	0.938	1.052	137	2922.153	76111
19	YRI	615655	12127	0.014	0.163	0.706	71.530	1.612	1.136	0.163	0.913	0.864	120	2153.041	63580
20	YRI	746647	14725	0.016	0.164	0.654	61.738	1.533	1.129	0.164	0.818	1.011	164	2420.084	63585
21	YRI	809890	15914	0.025	0.161	0.818	39.359	1.765	1.098	0.161	0.918	1.000	29	1451.245	37027

22	YRI	736975	14511	0.020	0.160	0.710	49.112	2.064	1.114	0.160	0.908	1.056	55	1523.018	34897
23	YRI	1965649	39360	0.007	0.221	0.728	136.152	1.230	1.086	0.221	0.941	0.621	171	3019.187	150761

Appendix 2 Chapter 5 - Large consanguineous pedigree diagram, genotype data is available for the individuals in bold.



A black symbol indicates a confirmed affected individual and a grey symbol indicates a mildly affected individual (Diagram provided by Beverley Dell, Wessex Clinical Genetics Service).

Appendix 3 Chapter 5 - Genes associated with the kidney or kidney disease.

Gene symbol	Description	chr	Start kb	End kb	source	
ABCB1	ATP-binding cassette subfamily B member 1	7	86777.599	86987.215	2	
ABCC1	ATP-binding cassette, subfamily C, member 1	16	15950.935	16143.774	2	
ABCC6	ATP-binding cassette, subfamily C, member 6	16	16151.491	16224.815	2	
ACE	angiotensin I converting enzyme	17	58915.909	58952.935	2	Associated with Nephrotic syndrome
ACTN4	a-Actinin-4	19	43830.166	43913.010	3	Focal-segmental glomerulosclerosis
AGTR1	angiotensin II receptor, type 1	3	149898.363	149943.486	2	
APC	adenomatosis polyposis coli	5	112101.483	112209.834	2	
APOA1	apolipoprotein A-I precursor	11/22	random	random	2	
APOA2	apolipoprotein A-II precursor	1	158005.156	158006.491	2	
APOE	apolipoprotein E	19	50100.879	50104.489	2	Associated with Nephrotic syndrome
AR	Androgen receptor	X	66571.704	66727.140	1	
AREG	Amphiregulin (schwannoma-derived GF)	4	75675.888	75685.760	1	
BAX	BCL2-associated X protein	19	54149.929	54156.866	1	
BBS1	Bardet-Biedl sydrome 1	11	66034.695	66057.660	2	
BCL2	B-cell chronic lymphocytic leukemia/lymphoma 2	18	58941.559	59137.025	1	
BDKRB1	bradykinin receptor B1	14	95792.312	95800.851	2	
BF	complement factor B preproprotein	6	32021.761	32027.839	2	Associated with Nephrotic syndrome
BHD	folliculin	17	17056.254	17081.221	2	
BSND	barttin	1	55176.638	55486.485	2	
C3	complement component 3 precursor	19	6628.878	6671.660	2	Associated with Nephrotic syndrome
C4A	complement component 4A preproprotein	6	32090.550	32111.173	2	Associated with Nephrotic syndrome
CA9	carbonic anhydrase IX precursor	9	35663.915	35671.152	2	

CCL2	small inducible cytokine A2 precursor	17	29606.409	29608.331	2
CCND1	cyclin D1	11	69165.054	69178.422	2
CDH1	Cadherin 1, type1, E-cadherin (epithelial)	16	67328.756	67424.940	1
CDKN1A	Cyclin-dependent kinase inhibitor 1A (p21, cip1)	6	36754.465	36763.086	1
CDKN2A	cyclin-dependent kinase inhibitor 2A	9	21957.758	21965.038	2
CDKN2B	cyclin-dependent kinase inhibitor 2B	9	21992.903	21999.312	2
CFTR	systic fibrosis transmembrane conductance regulator ATP-binding cassette subfanily C member 7	7	116713.968	116902.666	2
CLCN5	cholride channel 5	X	49537.192	49560.557	2
COL4A3	type IV alpha 3 collagen	2	227854.786	228002.091	2
COL4A4	alpha 4 type IV collagen precursor	2	227692.935	227852.780	2
COL4A5	type IV alpha-5 collagen	X	107489.299	107746.920	2
COL4A6	type IV alpha-6 collagen	X	107204.991	107487.805	2
CSF1	Colony-stimulating factor 1 (macrophage)	1	110165.499	110184.397	1
CSTA	Cystatin A (stefin A)	3	123526.701	123543.503	1
CTGF	Connective tissue growth factor	6	132311.018	132314.147	1
CTNNB1	catenin (cadherin-associated protein) beta1 88kDa	3	41216.016	41256.938	2
CUL2	cullin 2	10	35338.814	35419.300	2
CYLD	cylindromatosis (turban tumor syndrome)	16	49333.530	49393.347	2
CYP1A1	cytochrome P450, family 1, subfamily A, polypeptide 1	15	72798.943	72804.930	2
CYP2E1	cytochrome P450, family 2, subfamily E,	10	135229.748	135241.501	2

	polypeptide 1					
EGFR	epidermal growth factor receptor	7	54860.934	55049.239	2	
EGR1	Early growth response 1	5	137829.080	137832.903	1	
ERBB2	v-erb-b2 erythroblastic leukemia viral oncogene homologue 2	17	35109.780	35438.441	2	
ESR1	estrogen receptor 1	6	152220.800	152516.520	2	
EYA1	eyes absent 1	8	72272.222	72437.021	2	
F5	coagulation factor V precursor	1	166215.067	166287.379	2	Associated with Nephrotic syndrome
FGF1	Fibroblast growth factor 1 (acidic)	5	141953.307	142045.812	1	
FH	fumerate hydratase precursor	1	237986.947	238009.095	2	
FHIT	fragile histidine triad gene	3	59710.078	61212.164	2	
FKBP6	FK506-binding protein 6	7	72186.951	72217.292	2	
FOXD1	Forkhead box D1	5	72777.843	72780.108	1	
FRAS1	Fraser syndrome 1	4	79336.275	79822.602	2	
GATA3	GATA binding protein 3	10	8136.673	8157.170	2	
GLA	galactosidase alpha	X	100458.942	100469.096	2	
GLI3	GLI-Kruppel family member GLI3	7	41776.920	42036.135	2	
GNAI2	G protein, a inhibiting activity polypeptide 2	3	50248.651	50271.790	1	
GSTP1	glutathione transferase	11	67107.862	67110.699	2	
H19	H19	11	1972.984	1975.280	2	
HPRT1	hypoxanthine phosphoribosyltransferase 1	X	133319.777	133360.216	2	
HRAS	v-Ha-ras Harvey rat sarcoma viral oncogene homologue	11	522.243	525.550	2	
HSD11B2	Hydroxysteroid (11-beta) dehydrogenase 2	16	66022.537	66028.953	2	

HSPA1A	Heat shock 70-kDa protein 1A	6	31891.316	31893.698	1	
IFNG	interferon, gamma	12	66834.817	66839.788	2	Associated with Nephrotic syndrome
IGF1R	Insulin-like growth factor 1	15	97010.288	97319.034	1	
IGF2	Insulin-like growth factor 2 (somatomedin A)	11	2110.364	2116.780	1	
IL11	Interleukin-11	19	60567.569	60573.626	1	
IL1RN	interleukin 1 receptor antagonist	2	113591.701	113607.823	2	Associated with Nephrotic syndrome
INHA	Inhibin a	2	220262.459	220265.932	1	
INSR	insulin receptor	19	7067.049	7245.011	1	
KAL1	Kallmann syndrome 1 protein	X	8306.651	8509.963	2	
KCNJ1	potassium inwardly-recifying channel J1	11	128213.125	128242.478	2	
KRAS2	c-K-ras2 protein	12	25249.447	25295.121	2	
LAMB2	Laminin b2 chain	3	49133.551	49145.603	3	Pierson's syndrome
LMX1B	LIM homeobox transrciption factor 1 beta	9	126456.354	126538.284	3	Nial-patella syndrome
LTA	lymphotoxin alpha precursor	6	31648.072	31650.077	2	
LYZ	lysozyme precursor	12	68028.431	68034.280	2	
MET	met proto-oncogene precursor	7	115906.410	116032.390	2	
MMP1	matrix metalloproteinase 1	11	102165.861	102174.104	2	
MSH2	mutS homologue 2	2	47541.914	47622.011	2	
MTHFR	Methylenetetrahydrofolate reductase	1	11780.945	11800.248	2	
MYB	v-myb avian myeloblastosis virus (AMV) oncogene homologue	6	135544.146	135582.002	1	
MYC	v-myc AMV oncogene homologue	8	128817.686	128822.853	1	
MYCN	v-myc myelocytomatosis viral related oncogene, neuroblastoma derived	2	16031.281	16037.726	2	

MYH9	myosin heavy polypeptide 9 non-muscle	22	35001.827	35108.481	2	
NME1	nucleoside-diphosphate kinase 1	17	46585.919	46594.449	2	
NOS3	nitric oxide synthase 3 (endothelial cell)	7	150125.795	150149.323	2	Associated with Nephrotic syndrome
NOV	Nephroblastoma overexpressed gene	8	120497.822	120505.776	1	
NPHP1	nephrocystin	2	110237.281	110319.969	2	
NPHP4	nephroretinin	1	5857.136	5986.797	2	
NPHS1	Nephrin	19	41008.696	41034.579	3	Congenital nephrotic syndrome of the Finnish type
NPHS2	nephrosis 2, idiopathic, steroid-resistant (podocin)	1	176251.333	176276.725	3	corticosteroid-resistant nephrotic syndrome
NPY1R	neuropeptide Y receptor Y1	4	164602.722	164611.353	2	Associated with Nephrotic syndrome
NR0B1	Nuclear receptor subfamily 0, group b, member 1	X	30082.243	30087.149	1	
NR3C2	nuclear receptor subfamily 3 group C member 2	4	149357.525	149721.128	2	
OCRL	phosphatidylinositol polyphosphate 5-phosphatase	X	128399.787	128452.063	2	
ODC1	Ornithine decarboxylase 1	2	10531.106	10539.051	1	
PAX2	Paired box gene 2	10	102495.322	102579.687	1	
PGDFA	Platelet-derived growth factor a polypeptide	7	98637.240	98650.943	1	
PIGR	polymeric immunoglobulin receptor	1	203490.267	203508.202	2	Associated with Nephrotic syndrome
PKD1	polycystin 1 precursor	16	2078.712	2125.900	2	
PKD2	polycystin 2	4	89285.999	89356.107	2	
PKHD1	polycystic kidney and hepatic disease 1	6	51588.104	52060.382	2	
PLA2G7	phospholipase A2, group VII (platelet-activating	6	46780.238	46811.389	2	Associated with Nephrotic syndrome

	factor acetylhydrolase, plasma)					
**PLCE1*	phospholiase C,epsilon 1	10	95780.559	96078.136	4	Nephrotic Syndrome with diffuse mesangial sclerosis (DMS)
PNN	Pinin, desmosome associated protein	14	38714.151	38721.178	2	
PON1	paraoxonase 1	7	94571.639	94598.495	2	
PPARG	peroxisome proliferative activated receptor gamma	3	12367.959	12450.840	2	Associated with Nephrotic syndrome
PTEN	phosphatase and tensin homologue	10	89613.175	89716.382	2	
PTHLH	parathyroid hormone-like hormone	12	28006.521	28016.183	2	
RARA	Retinoic acid receptor a	17	35740.896	35767.420	1	
RASSF1	Ras association domain family 1	3	50342.221	50353.371	2	
RB1	retinoblastoma 1	13	47775.912	47954.023	2	
RNF139	ring finger protein 139	8	125556.189	125570.040	2	
SAH	SA hypertension-associated homologue	16	20682.813	21715.979	2	
SALL1	sal-like 1	16	49727.830	49742.653	2	
SALL2	Sal (Drosophila)-like 2	14	21059.074	21075.177	1	
SCGB1A1	secretoglobin, family 1A, member 1 (uteroglobin)	11	61943.099	61947.242	2	Associated with Nephrotic syndrome
SDC1	Syndecan 1	2	20322.188	20346.822	1	
SERPINE1	plasminogen activator inhibitor-1	7	100363.887	100375.741	2	
SLC12A1	sodium potassium chloride cotransporter 2 family 12 member 1	15	46285.790	46383.568	2	
SLC2A2	solute carrier family 2 (facilitated glucose transporter) member 2	3	172196.839	172227.470	2	
SLC34A1	solute carrier family 34 (sodium phosphate)	5	176744.061	176758.454	2	

	member 1					
SLC4A1	solute carrier family 4 anion exchanger member 1	17	39682.566	39700.993	2	
SLC4A4	solute carrier family sodiumbicarbonate cotransporter member 4	4	72569.852	72802.834	2	
SLC7A7	solute carrier family 7 (cationic amino acid transporter y+ system) member 7	14	22312.274	22354.852	2	
SLIT2	slit homologue 2	4	19931.504	20297.057	2	
SMARCB1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin subfamily b, member 1	22	22453.704	22501.258	2	
SOD1	Superoxide dismutase 1	21	31953.806	31963.112	1	
TAP1/ABCB2	transporter 1, ATP-binding cassette, subfamily B	6	32920.965	32929.726	2	Associated with Nephrotic syndrome
TCF2	transcription factor 2	17	33162.729	33179.182	2	
TGFB1	Tranforming growth factor B1	19	46528.491	46551.656	1	
THBS1	Thrombospondin 1	15	37660.572	37676.959	1	
THRA	thyroid hormone receptor alpha	17	35472.686	35503.611	2	
THRB	thyroid hormone receptor beta	3	24139.236	24511.317	2	
TIAM1	T-cell lymphoma invasion and metastasis 1	21	31238.438	31853.161	2	
TIMP3	tissue inhibitor of metalloproteinase 3	22	31521.362	31583.581	2	
TNF	tumour necrosis factor alpha	6	31651.329	31654.091	2	
TP53	tumour protein 53	17	7512.464	7531.642	2	
TRAa	T-cell antigen receptor, alpha polypeptide	14	21961.312	22090.938	2	Associated with Nephrotic syndrome
TRPC6	TRPC6	11	100827.582	100959.869	3	Focal-segmental glomerulosclerosis

TSC1	tuberous sclerosis 1	9	132796.290	132849.574	2	
TSC2	tuberous sclerosis 2	16	2038.600	2078.713	2	
UMOD	uromodulin	16	20251.875	20271.538	2	Associated with Nephrotic syndrome
VDR	Vitamin D (1,25-dihydroxyvitamin D3) receptor	12	46521.589	46585.081	1	
VHL	von Hippel-Lindau tumour suppressor	3	10158.319	10168.744	2	
WT1	Wilms tumor 1	11	32365.897	32413.643	3	Denys-drash syndrome

Source; 1. (Renshaw et al. 2004) 2. (Human kidney Gene DataBase 2004) 3. (Tryggvason, Patrakka, and Wartiovaara 2006) 4. (Hinkes et al. 2006)

Appendix 4. Chapter 5 - Affymetrix 50K and 500K array, top 10 regions common to ID3, ID4, ID19 and ID25, ordered by genetic length on the LDU scale.

		Location (Kb)					No. SNPs
					LDU	No.	in following
	Chr	Start	End	Kb length	length	SNP	region
Affy50K	10	132872.4	135126.6	2254.18	70.5	5	0
	1	836.73	3127.56	2290.83	49.22	8	0
	16	83445.94	84394.93	948.98	45.04	13	0
	6	41551.67	42430.49	878.83	42.03	5	0
	19	61906.99	62826.08	919.1	37.27	6	0
	10	125356.9	126428	1071.14	36.78	5	0
	22	46949.92	47337.59	387.67	30.31	8	0
	22	43945.37	44400.71	455.34	28.84	6	0
	13	26159.37	26711.95	552.58	28.33	9	1
	4	25204.63	25687.53	482.9	27.13	5	0
Affy500K	3	192033.4	192151.3	117.92	17.29	9	0
	19	56116.92	56177.01	60.09	17.20	7	0
	19	58781.76	58823.45	41.69	16.09	8	0
	4	33115.56	33882.9	767.33	14.72	52	0
	2	157803	158054.7	251.67	13.42	37	0
	8	62165.45	62175.6	10.15	12.35	5	2
	5	94765.45	94804.7	39.25	12.30	10	9
	19	18708.61	18966.45	257.84	11.87	9	1
	6	151239.2	151295.1	55.88	11.50	15	1
	18	10132.32	10188.74	56.42	11.38	8	3

Appendix 5. Chapter 5 - Affymetrix 50K and 500K array, top 10 regions common to ID3, ID4 and ID19, ordered by genetic length on the LDU scale.

		Location (Kb)					No. SNPs
					LDU	No.	in following
	Chr	Start	End	Kb length	length	SNP	region
Affy50K	13	22793.60	26016.28	3222.68	110.73	88	14
	10	132872.40	135126.60	2254.18	70.50	5	0
	10*	95364.46	99548.30	4183.84	64.75	78	0
	22	46627.38	47337.59	710.21	50.24	9	0
	1	836.73	3127.56	2290.83	49.22	8	0
	6	41551.67	42598.50	1046.83	48.85	6	0
	19	42606.99	45229.64	2622.65	47.25	5	0
	16	83445.94	84394.93	948.98	45.04	14	0
	4	8171.57	9668.55	1496.98	44.22	13	0
	19	60069.02	61064.52	995.49	43.21	8	3
Affy500K	13	23906.63	25639.08	1732.45	42.64	322	177
	13	25651.25	26431.64	780.39	35.38	177	81
	13	23458.33	23868.71	410.38	31.83	136	0
	10	97017.08	98526.09	1509.02	28.75	310	0
	10*	95282.73	96261.83	979.1	19.00	165	66
	4	32756.28	33882.9	1126.61	18.36	103	0
	3	192033.4	192151.3	117.92	17.29	9	0
	19	56116.92	56177.01	60.09	17.20	7	0
	19	58781.76	58823.45	41.69	16.09	8	0
	5	150898.3	151007.2	108.89	15.29	23	0

* Regions containing PLCE1 gene.

References

1. Abecasis GR and Cookson WO (2000) GOLD--graphical overview of linkage disequilibrium. *Bioinformatics.* 16 (2):182-183.

2. Adaimy L, Chouery E, Megarbane H, Mroueh S, Delague V, Nicolas E, Belguith H, de MP, and Megarbane A (2007) Mutation in WNT10A is associated with an autosomal recessive ectodermal dysplasia: the odonto-onycho-dermal dysplasia. *Am.J.Hum.Genet* 81 (4):821-828.

3. Aligianis IA, Johnson CA, Gissen P, Chen D, Hampshire D, Hoffmann K, Maina EN, Morgan NV, Tee L, Morton J, Ainsworth JR, Horn D, Rosser E, Cole TR, Stolte-Dijkstra I, Fieggen K, Clayton-Smith J, Megarbane A, Shield JP, Newbury-Ecob R, Dobyns WB, Graham JM, Jr., Kjaer KW, Warburg M, Bond J, Trembath RC, Harris LW, Takai Y, Mundlos S, Tannahill D, Woods CG, and Maher ER (2005) Mutations of the catalytic subunit of RAB3GAP cause Warburg Micro syndrome. *Nat Genet* 37 (3):221-223.

4. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, and Donnelly P (2005) A haplotype map of the human genome. *Nature* 437 (7063):1299-1320.

5. Altug-Teber O, Dufke A, Poths S, Mau-Holzmann UA, Bastepe M, Colleaux L, Cormier-Daire V, Eggermann T, Gillessen-Kaesbach G, Bonin M, and Riess O (2005) A rapid microarray based whole genome analysis for detection of uniparental disomy. *Hum.Mutat.* 26 (2):153-159.

6. Barrett JC, Fry B, Maller J, and Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* 21 (2):263-265.

7. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, and Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am.J.Hum.Genet* 74 (6):1111-1120.

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

9. Broman KW, Rowe LB, Churchill GA, and Paigen K (2002) Crossover interference in the mouse. *Genetics* 160 (3):1123-1131.

10. Broman KW and Weber JL (1999) Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am.J.Hum.Genet* 65 (6):1493-1500.

11. Bruce S, Leinonen R, Lindgren CM, Kivinen K, hlman-Wright K, Lipsanen-Nyman M, Hannula-Jouppi K, and Kere J (2005) Global analysis of uniparental disomy using high density genotyping arrays. *J.Med.Genet* 42 (11):847-851.

12. Cancer Genetic Markers of Susceptibility (CGEMS) (2007) http://cgems.cancer.gov/index.asp. *accessed 09/12/07.*

13. Carlson CS, Eberle MA, Kruglyak L, and Nickerson DA (2004) Mapping complex disease loci in whole-genome association studies. *Nature* 429 (6990):446-452.

14. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, and Nickerson DA (2005) Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res.* 15 (11):1553-1565.

15. Carr IM, Sheridan E, and Bonthron DT (2007) Intuitive tools for pedigree-free detection of autozygous regions in SNP data. *in preparation.*

16. Carr IM, Flintoff KJ, Taylor GR, Markham AF, and Bonthron DT (2006) Interactive visual analysis of SNP data for rapid autozygosity mapping in consanguineous families. *Hum.Mutat.* 27 (10):1041-1046.

17. Celedon JC (2005) http://innateimmunity.net/files/CANDGENES/siframes.html. *accessed 25/04/05.*

18. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, Brooks LD, Cardon LR, Daly M, Donnelly P, Fraumeni JF, Jr., Freimer NB, Gerhard DS, Gunter C, Guttmacher AE, Guyer MS, Harris EL, Hoh J, Hoover R, Kong CA, Merikangas KR, Morton CC, Palmer LJ, Phimister EG, Rice JP, Roberts J, Rotimi C, Tucker MA, Vogan KJ, Wacholder S, Wijsman EM, Winn DM, and Collins FS (2007) Replicating genotype-phenotype associations. *Nature* 447 (7145):655-660.

19. Chiang AP, Beck JS, Yen HJ, Tayeh MK, Scheetz TE, Swiderski RE, Nishimura DY, Braun TA, Kim KY, Huang J, Elbedour K, Carmi R, Slusarski DC, Casavant TL, Stone EM, and Sheffield VC (2006) Homozygosity mapping with SNP arrays identifies TRIM32, an E3 ubiquitin ligase, as a Bardet-Biedl syndrome gene (BBS11). *Proc.Natl.Acad.Sci.U.S.A* 103 (16):6287-6292.

20. Clark AG (2003) Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Curr.Opin.Genet.Dev.* 13 (3):296-302.

21. Clark AG and Li J (2007) Conjuring SNPs to detect associations. *Nat Genet* 39 (7):815-816.

22. Clayton D (2002) http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt. *accessed 25/11/07.*

23. Collins A and Lau W (2007) CHROMSCAN: genome-wide association using a linkage disequilibrium map. *J.Hum.Genet.*

24. Collins A, Lau W, and De La Vega F (2004) Mapping genes for common diseases: the case for genetic (LD) maps. *Hum.Hered.* 58 (1):2-9.

25. Collins A, Lonjou C, and Morton NE (1999) Genetic epidemiology of single-nucleotide polymorphisms. *Proc.Natl.Acad.Sci.U.S.A* 96 (26):15173-15177.

26. Collins A and Morton NE (1998) Mapping a disease locus by allelic association. *Proc.Natl.Acad.Sci.U.S.A* 95 (4):1741-1745.

27. Collins FS, Green ED, Guttmacher AE, and Guyer MS (2003) A vision for the future of genomics research. *Nature* 422 (6934):835-847.

28. Collins FS, Morgan M, and Patrinos A (2003) The Human Genome Project: lessons from large-scale biology. *Science* 300 (5617):286-290.

29. Cottingham RW, Jr., Idury RM, and Schaffer AA (1993) Faster sequential genetic linkage computations. *Am.J.Hum.Genet* 53 (1):252-263.

30. Curtis D (2007) Extended homozygosity is not usually due to cytogenetic abnormality. *BMC.Genet* 8:67.

31. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, and Lander ES (2001) High-resolution haplotype structure in the human genome. *Nat Genet* 29 (2):229-232.

32. database of Genotype and Phenotype (dbGaP) (2007) http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap. *accessed 09/12/07.*

33. Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SCL, Jenkins SC, Palmer SM, Balfour KM, Rowe BR, Farrall M, Barnett AH, Bain SC, and Todd JA (1994) A Genome-Wide Search for Human Type-1 Diabetes Susceptibility Genes. *Nature* 371 (6493):130-136.

34. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, and Altshuler D (2005) Efficiency and power in genetic association studies. *Nat Genet* 37 (11):1217-1223.

35. De La Vega FM, Isaac H, Collins A, Scafe CR, Halldorsson BV, Su X, Lippert RA, Wang Y, Laig-Webster M, Koehler RT, Ziegle JS, Wogan LT, Stevens JF, Leinen KM, Olson SJ, Guegler KJ, You X, Xu LH, Hemken HG, Kalush F, Itakura M, Zheng Y, de TG, O'Brien SJ, Clark AG, Istrail S, Hunkapiller MW, Spier EG, and Gilbert DA (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 (4):454-462.

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

37. Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, and Weissenbach J (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380 (6570):152-154.

38. Ewens WJ (2003) On estimating P values by the Monte Carlo method. *Am.J.Hum.Genet* 72 (2):496-498.

39. Farrall M and Morris AP (2005) Gearing up for genome-wide gene-association studies. *Hum.Mol.Genet* 14 Spec No. 2:R157-R162.

40. Field LL, Tobias R, Robinson WP, Paisey R, and Bain S (1998) Maternal uniparental disomy of chromosome 1 with no apparent phenotypic effects. *Am.J.Hum.Genet* 63 (4):1216-1220.

41. Forshew T and Johnson CA (2004) SCAMP: a spreadsheet to collate autozygosity mapping projects. *J.Med.Genet* 41 (12):e125.

42. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW, Shields B, Harries LW, Barrett JC, Ellard S, Groves CJ, Knight B, Patch AM, Ness AR, Ebrahim S, Lawlor DA, Ring SM, Ben-Shlomo Y, Jarvelin MR, Sovio U, Bennett AJ, Melzer D, Ferrucci L, Loos RJ, Barroso I, Wareham NJ, Karpe F, Owen KR, Cardon LR, Walker M, Hitman GA, Palmer CN, Doney AS, Morris AD, Smith GD, Hattersley AT, and McCarthy MI (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316 (5826):889-894.

43. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT,

Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, and Stewart J (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449 (7164):851-861.

44. Genetic Association Information Network (GAIN) (2007) http://www.fnih.org/GAIN2/home_new.shtml. *accessed 09/12/07.*

45. Genetic Epidemiology and Bioinformatics group (2008) http://cedar.genetics.soton.ac.uk/public_html/LDB2000/release.html. *last accessed 09/12/07.*

46. Gibbs JR and Singleton A (2006) Application of genome-wide single nucleotide polymorphism typing: simple association and beyond. *PLoS.Genet* 2 (10):e150.

47. Gibson J, Morton NE, and Collins A (2006) Extended tracts of homozygosity in outbred human populations. *Hum.Mol.Genet* 15 (5):789-795.

48. Gibson J, Tapper W, Cox D, Zhang W, Pfeufer A, Gieger C, Wichmann HE, Kaab S, Collins AR, Meitinger T, and Morton N (2008) A multimetric approach to analysis of genome-wide association by single

markers and composite likelihood. *Proc.Natl.Acad.Sci.U.S.A* 105 (7):2592-2597.

49. Gibson J, Tapper W, Zhang W, Morton N, and Collins A (2005) Cosmopolitan linkage disequilibrium maps. *Hum.Genomics* 2 (1):20-27.

50. Gutierrez-Roelens I, Sluysmans T, Jorissen M, Amyere M, and Vikkula M (2006) Localization of candidate regions for a novel gene for Kartagener syndrome. *Eur.J.Hum.Genet* 14 (7):809-815.

51. Helgason A, Yngvadottir B, Hrafnkelsson B, Gulcher J, and Stefansson K (2005) An Icelandic example of the impact of population structure on association studies. *Nat.Genet.* 37 (1):90-95.

52. Herbert A, Gerry NP, McQueen MB, Heid IM, Pfeufer A, Illig T, Wichmann HE, Meitinger T, Hunter D, Hu FB, Colditz G, Hinney A, Hebebrand J, Koberwitz K, Zhu X, Cooper R, Ardlie K, Lyon H, Hirschhorn JN, Laird NM, Lenburg ME, Lange C, and Christman MF (2006) A common genetic variant is associated with adult and childhood obesity. *Science* 312 (5771):279-283.

53. Hill WG (1974) Estimation of linkage disequilibrium in randomly mating populations. *Heredity* 33 (2):229-239.

54. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, and Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* 307 (5712):1072-1079.

55. Hinkes B, Wiggins RC, Gbadegesin R, Vlangos CN, Seelow D, Nurnberg G, Garg P, Verma R, Chaib H, Hoskins BE, Ashraf S, Becker C, Hennies HC, Goyal M, Wharram BL, Schachter AD, Mudumana S, Drummond I, Kerjaschki D, Waldherr R, Dietrich A, Ozaltin F, Bakkaloglu A, Cleper R, Basel-Vanagaite L, Pohl M, Griebel M, Tsygin AN, Soylu A, Muller D, Sorli CS, Bunney TD, Katan M, Liu J, Attanasio M, O'toole JF, Hasselbacher K, Mucha B, Otto EA, Airik R, Kispert A, Kelley GG, Smrcka AV, Gudermann T, Holzman LB, Nurnberg P, and Hildebrandt F (2006) Positional cloning uncovers mutations in PLCE1 responsible for a nephrotic syndrome variant that may be reversible. *Nat Genet* 38 (12):1397-1405.

56. Huie ML, nyane-Yeboa K, Guzman E, and Hirschhorn R (2002) Homozygosity for multiple contiguous single-nucleotide polymorphisms as an indicator of large heterozygous deletions: identification of a novel heterozygous 8-kb intragenic deletion (IVS7-19 to IVS15-17) in a patient with glycogen storage disease type II. *Am.J.Hum.Genet* 70 (4):1054-1057.

57. Human Genome Structural Variation Project (2007) http://humanparalogy.gs.washington.edu/structuralvariation/. *accessed 09/12/07.*

58. Human kidney Gene DataBase (2004) http://www.urogene.org/kgdb/index.htm. *accessed 10/11/07.*

59. International Hapmap Group (2005) http://www.hapmap.org. *accessed 25/04/05.*

60. Jeffreys AJ, Kauppi L, and Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat.Genet.* 29 (2):217-222.

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

62. Jorgenson E, Tang H, Gadde M, Province M, Leppert M, Kardia S, Schork N, Cooper R, Rao DC, Boerwinkle E, and Risch N (2005) Ethnicity and human genetic linkage maps. *Am.J.Hum.Genet.* 76 (2):276-290.

63. Kaback DB (1996) Chromosome-size dependent control of meiotic recombination in humans. *Nat.Genet.* 13 (1):20-21.

64. Ke X, Hunt S, Tapper W, Lawrence R, Stavrides G, Ghori J, Whittaker P, Collins A, Morris AP, Bentley D, Cardon LR, and Deloukas P (2004) The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum.Mol.Genet* 13 (6):577-588.

65. Kidd KK, Pakstis AJ, Speed WC, and Kidd JR (2004) Understanding human DNA sequence variation. *J.Hered.* 95 (5):406-420.

66. Klein C, Grudzien M, Appaswamy G, Germeshausen M, Sandrock I, Schaffer AA, Rathinam C, Boztug K, Schwinzer B, Rezaei N, Bohn G, Melin M, Carlsson G, Fadeel B, Dahl N, Palmblad J, Henter JI, Zeidler C, Grimbacher B, and Welte K (2007) HAX1 deficiency causes autosomal recessive severe congenital neutropenia (Kostmann disease). *Nat Genet* 39 (1):86-92.

67. 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, and Hoh J (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308 (5720):385-389.

68. Komura D, Shen F, Ishikawa S, Fitch KR, Chen W, Zhang J, Liu G, Ihara S, Nakamura H, Hurles ME, Lee C, Scherer SW, Jones KW, Shapero MH, Huang J, and Aburatani H (2006) Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.* 16 (12):1575-1584.

69. Kong X, Murphy K, Raj T, He C, White PS, and Matise TC (2004) A combined linkage-physical map of the human genome. *Am.J.Hum.Genet.* 75 (6):1143-1148.

70. Kruglyak L, Daly MJ, and Lander ES (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am.J.Hum.Genet* 56 (2):519-527.

71. Kruglyak L and Nickerson DA (2001) Variation is the spice of life. *Nat.Genet.* 27 (3):234-236.

72. Kuo TY, Lau W, and Collins AR (2007) LDMAP: the construction of high-resolution linkage disequilibrium maps of the human genome. *Methods Mol.Biol.* 376:47-57.

73. Kuruvilla, F, Green, T, Altshuler, D, Daly, M, and Gabriel, S. An evaluation of the Bayesian Robust Linear Modeling using Mahalanobis Distance (BRLMM) Genotyping Algorithm. Broad Institute of MIT and Harvard . 2006. 10-6-2006.
Ref Type: Electronic Citation

74. Lander ES and Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* 236 (4808):1567-1570.

75. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP,

Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la BM, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de JP, Catanese JJ, Osoegawa K, Shizuya H, Choi S, and Chen YJ (2001) Initial sequencing and analysis of the human genome. *Nature* 409 (6822):860-921.

76. Lau W, Kuo TY, Tapper W, Cox S, and Collins A (2007) Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. *Bioinformatics.* 23 (4):517-519.

77. Lencz T, Lambert C, Derosse P, Burdick KE, Morgan TV, Kane JM, Kucherlapati R, and Malhotra AK (2007) Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proc.Natl.Acad.Sci.U.S.A* 104 (50):19942-19947.

78. Lewontin RC (1964) The Interaction of selection and linkage. II. Optimum Models. *Genetics* 50:757-782.

79. Li LH, Ho SF, Chen CH, Wei CY, Wong WC, Li LY, Hung SI, Chung WH, Pan WH, Lee MT, Tsai FJ, Chang CF, Wu JY, and Chen YT (2006) Long contiguous stretches of homozygosity in the human genome. *Hum.Mutat.* 27 (11):1115-1121.

80. Lindner TH and Hoffmann K (2005) easyLINKAGE: a PERL script for easy and automated two-/multi-point linkage analyses. *Bioinformatics.* 21 (3):405-407.

81. Lonjou C, Zhang W, Collins A, Tapper WJ, Elahi E, Maniatis N, and Morton NE (2003) Linkage disequilibrium in human populations. *Proc.Natl.Acad.Sci.U.S.A* 100 (10):6069-6074.

82. Maniatis N, Collins A, Gibson J, Zhang W, Tapper W, and Morton NE (2004) Positional cloning by linkage disequilibrium. *Am.J.Hum.Genet.* 74 (5):846-855.

83. Maniatis N, Collins A, and Morton NE (2007) Effects of single SNPs, haplotypes, and whole-genome LD maps on accuracy of association mapping. *Genet Epidemiol.* 31 (3):179-188.

84. Maniatis N, Collins A, Xu CF, McCarthy LC, Hewett DR, Tapper W, Ennis S, Ke X, and 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 (4):2228-2233.

85. Maniatis N, Morton NE, Gibson J, Xu CF, Hosking LK, and Collins A (2005) The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. *Hum.Mol.Genet.* 14 (1):145-153.

86. Maraganore DM, de AM, Lesnick TG, Strain KJ, Farrer MJ, Rocca WA, Pant PV, Frazer KA, Cox DR, and Ballinger DG (2005) High-resolution whole-genome association study of Parkinson disease. *Am.J.Hum.Genet* 77 (5):685-693.

87. Marchini J, Howie B, Myers S, McVean G, and Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39 (7):906-913.

88. Marjoram P and Tavare S (2006) Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet* 7 (10):759-770.

89. McKeigue PM (2005) Prospects for admixture mapping of complex traits. *Am.J.Hum.Genet.* 76 (1):1-7.

90. McKinney C and Merriman TR (2007) The human genome and understanding of common disease: present and future technologies. *Cell Mol.Life Sci.* 64 (7-8):961-978.

91. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, and Donnelly P (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* 304 (5670):581-584.

92. MedlinePlus (2007) http://www.nlm.nih.gov/medlineplus/ency/article/001576.htm. *accessed 10/11/07.*

93. Melin M, Entesarian M, Carlsson G, Garwicz D, Klein C, Fadeel B, Nordenskjold M, Palmblad J, Henter JI, and Dahl N (2007) Assignment of the gene locus for severe congenital neutropenia to chromosome 1q22 in the original Kostmann family from Northern Sweden. *Biochem.Biophys.Res.Commun.* 353 (3):571-575.

94. Miyazawa H, Kato M, Awata T, Kohda M, Iwasa H, Koyama N, Tanaka T, Huqun, Kyo S, Okazaki Y, and Hagiwara K (2007) Homozygosity

haplotype allows a genomewide search for the autosomal segments shared among patients. *Am.J.Hum.Genet* 80 (6):1090-1102.

95. Morgan NV, Gissen P, Sharif SM, Baumber L, Sutherland J, Kelly DA, Aminu K, Bennett CP, Woods CG, Mueller RF, Trembath RC, Maher ER, and Johnson CA (2002) A novel locus for Meckel-Gruber syndrome, MKS3, maps to chromosome 8q24. *Hum.Genet* 111 (4-5):456-461.

96. Morgan NV, Pasha S, Johnson CA, Ainsworth JR, Eady RA, Dawood B, McKeown C, Trembath RC, Wilde J, Watson SP, and Maher ER (2006) A germline mutation in BLOC1S3/reduced pigmentation causes a novel variant of Hermansky-Pudlak syndrome (HPS8). *Am.J.Hum.Genet* 78 (1):160-166.

97. Morris AP (2006) A flexible Bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. *Am.J.Hum.Genet* 79 (4):679-694.

98. Morris AP, Whittaker JC, and Balding DJ (2000) Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Am.J.Hum.Genet* 67 (1):155-169.

99. Morris AP, Whittaker JC, Xu CF, Hosking LK, and Balding DJ (2003) Multipoint linkage-disequilibrium mapping narrows location interval and identifies mutation heterogeneity. *Proc.Natl.Acad.Sci.U.S.A* 100 (23):13442-13446.

100. Morton N, Maniatis N, Zhang W, Ennis S, and Collins A (2007) Genome scanning by composite likelihood. *Am.J.Hum.Genet* 80 (1):19-28.

101. Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok PY, and Collins A (2001) The optimal measure of allelic association. *Proc.Natl.Acad.Sci.U.S.A* 98 (9):5217-5221.

102. Myers S, Bottolo L, Freeman C, McVean G, and Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science* 310 (5746):321-324.

103. National Institutes of Health and National Human Genome Research institute. (2002) International Consortium Launches Genetic Variation Mapping Project. http://www.genome.gov/10005336. *accessed 25/04/05.*

104. National Institutes of Health and National Human Genome Research institute. (2005) International HapMap Consortium Expands Mapping Effort. http://www.genome.gov/13014173. *accessed 25/04/05.*

105. NCBI (2005) http://www.ncbi.nlm.nih.gov/projects/SNP/. *accessed 25/04/05.*

106. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Sato H, Hori M, Nakamura Y, and Tanaka T (2002) Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet* 32 (4):650-654.

107. Perneger TV (1998) What's wrong with Bonferroni adjustments. *BMJ* 316 (7139):1236-1238.

108. Puffenberger EG, Strauss KA, Ramsey KE, Craig DW, Stephan DA, Robinson DL, Hendrickson CL, Gottlieb S, Ramsay DA, Siu VM, Heuer GG, Crino PB, and Morton DH (2007) Polyhydramnios, megalencephaly and symptomatic epilepsy caused by a homozygous 7-kilobase deletion in LYK5. *Brain* 130 (Pt 7):1929-1941.

109. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, and Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am.J.Hum.Genet* 81 (3):559-575.

110. Rabbee N and Speed TP (2006) A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics.* 22 (1):7-12.

111. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, and Lander ES (2001) Linkage disequilibrium in the human genome. *Nature* 411 (6834):199-204.

112. Renshaw J, Orr RM, Walton MI, Te PR, Williams RD, Wancewicz EV, Monia BP, Workman P, and Pritchard-Jones K (2004) Disruption of WT1 gene expression and exon 5 splicing following cytotoxic drug treatment: antisense down-regulation of exon 5 alters target gene expression and inhibits cell survival. *Mol.Cancer Ther.* 3 (11):1467-1484.

113. Ruderman EM and Pope RM (2005) The evolving clinical profile of abatacept (CTLA4-Ig): a novel co-stimulatory modulator for the treatment of rheumatoid arthritis. *Arthritis Res.Ther.* 7 Suppl 2:S21-S25.

114. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallee C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD,

Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Johnson TA, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archeveque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, and Peterson JL (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* 449 (7164):913-918.

115. Sanger F and Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J.Mol.Biol.* 94 (3):441-448.

116. Sanger F, Nicklen S, and Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc.Natl.Acad.Sci.U.S.A* 74 (12):5463-5467.

117. Schaffer AA, Gupta SK, Shriram K, and Cottingham RW, Jr. (1994) Avoiding recomputation in linkage analysis. *Hum.Hered.* 44 (4):225-237.

118. Service S, DeYoung J, Karayiorgou M, Roos JL, Pretorious H, Bedoya G, Ospina J, Ruiz-Linares A, Macedo A, Palha JA, Heutink P, Aulchenko Y, Oostra B, van DC, Jarvelin MR, Varilo T, Peddle L, Rahman P, Piras G, Monne M, Murray S, Galver L, Peltonen L, Sabatti C, Collins A, and Freimer N (2006) Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nat Genet* 38 (5):556-560.

119. Shifman S, Kuypers J, Kokoris M, Yakir B, and Darvasi A (2003) Linkage disequilibrium patterns of the human genome across populations. *Hum.Mol.Genet.* 12 (7):771-776.

120. Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, Gibbs JR, Britton A, de Vrieze FW, Peckham E, Gwinn-Hardy K, Crawley A, Keen JC, Nash J, Borgaonkar D, Hardy J, and Singleton A (2007) Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Hum.Mol.Genet* 16 (1):1-14.

121. Smith UM, Consugar M, Tee LJ, McKee BM, Maina EN, Whelan S, Morgan NV, Goranson E, Gissen P, Lilliquist S, Aligianis IA, Ward CJ, Pasha S, Punyashthiti R, Malik SS, Batman PA, Bennett CP, Woods CG, McKeown C, Bucourt M, Miller CA, Cox P, Algazali L, Trembath RC, Torres VE, ttie-Bitach T, Kelly DA, Maher ER, Gattone VH, Harris PC, and Johnson CA (2006) The transmembrane protein meckelin (MKS3) is mutated in Meckel-Gruber syndrome and the wpk rat. *Nat Genet* 38 (2):191-196.

122. Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, Baker A, Jonasdottir A, Ingason A, Gudnadottir VG, Desnica N, Hicks A, Gylfason A, Gudbjartsson DF, Jonsdottir GM, Sainz J, Agnarsson K, Birgisdottir B, Ghosh S, Olafsdottir A, Cazier JB, Kristjansson K, Frigge ML, Thorgeirsson TE, Gulcher JR, Kong A, and Stefansson K (2005) A common inversion under selection in Europeans. *Nat.Genet.* 37 (2):129-137.

123. Stephens M and Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am.J.Hum.Genet* 73 (5):1162-1169.

124. Storey JD and Tibshirani R (2003) Statistical significance for genomewide studies. *Proc.Natl.Acad.Sci.U.S.A* 100 (16):9440-9445.

125. Sturtevant AH (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology* 14:43-59.

126. Tang K, Thornton KR, and Stoneking M (2007) A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. *PLoS.Biol.* 5 (7):e171.

127. Tapper W, Collins A, Gibson J, Maniatis N, Ennis S, and Morton NE (2005) A map of the human genome in linkage disequilibrium units. *Proc.Natl.Acad.Sci.U.S.A* 102 (33):11835-11839.

128. Tapper W et al.A comparison of methods to detect recombination hotspots. Hum.Hered. (in press)

129. Tapper WJ, Maniatis N, Morton NE, and Collins A (2003) A metric linkage disequilibrium map of a human chromosome. *Ann.Hum.Genet.* 67 (Pt 6):487-494.

130. Tapper WJ, Morton NE, Dunham I, Ke X, and Collins A (2001) A sequence-based integrated map of chromosome 22. *Genome Res.* 11 (7):1290-1295.

131. The International HapMap Consortium (2003) The International HapMap Project. *Nature* 426 (6968):789-796.

132. The SNP Consortium (2005) http://snp.cshl.org/. *accessed 25/04/05.*

133. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, Lowe CE, Szeszko JS, Hafler JP, Zeitels L, Yang JH, Vella A, Nutland S, Stevens HE, Schuilenburg H, Coleman G, Maisuria M, Meadows W, Smink LJ, Healy B, Burren OS, Lam AA, Ovington NR, Allen J, Adlem E, Leung HT, Wallace C, Howson JM, Guja C, Ionescu-Tirgoviste C, Simmonds MJ, Heward JM, Gough SC, Dunger DB, Wicker LS, and Clayton DG (2007) Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* 39 (7):857-864.

134. Tryggvason K, Patrakka J, and Wartiovaara J (2006) Hereditary proteinuria syndromes and mechanisms of proteinuria. *N.Engl.J.Med.* 354 (13):1387-1401.

135. U.S.National Library of Medicine. What is DNA? - Genetics Home Reference. http://ghr.nlm.nih.gov/handbook/basics/dna . 16-7-2006. Ref Type: Electronic Citation

136. UCSC Genome Browser (2007) http://genome.ucsc.edu/. *accessed 25/11/07.*

137. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, bu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di F, V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz

S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, and Nodell M (2001) The sequence of the human genome. *Science* 291 (5507):1304-1351.

138. Voight BF, Kudaravalli S, Wen X, and Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS.Biol.* 4 (3):e72.

139. Wang H, Lin CH, Service S, Chen Y, Freimer N, and Sabatti C (2006) Linkage disequilibrium and haplotype homozygosity in population samples genotyped at a high marker density. *Hum.Hered.* 62 (4):175-189.

140. Watson JD and Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171 (4356):737-738.

141. Weber S, Mir S, Schlingmann KP, Nurnberg G, Becker C, Kara PE, Ozkayin N, Konrad M, Nurnberg P, and Schaefer F (2005) Gene locus ambiguity in posterior urethral valves/prune-belly syndrome. *Pediatr.Nephrol.* 20 (8):1036-1042.

142. Woods CG, Valente EM, Bond J, and Roberts E (2004) A new method for autozygosity mapping using single nucleotide polymorphisms (SNPs) and EXCLUDEAR. *J.Med.Genet* 41 (8):e101.

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

144. Zaykin DV, Meng Z, and Ehm MG (2006) Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. *Am.J.Hum.Genet* 78 (5):737-746.

145. Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR,

Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, and Hattersley AT (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316 (5829):1336-1341.

146. Zhang C, Bailey DK, Awad T, Liu G, Xing G, Cao M, Valmeekam V, Retief J, Matsuzaki H, Taub M, Seielstad M, and Kennedy GC (2006) A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. *Bioinformatics.* 22 (17):2122-2128.

147. Zhang K, Qin Z, Chen T, Liu JS, Waterman MS, and Sun F (2005) HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics.* 21 (1):131-134.

148. Zhang W, Collins A, Gibson J, Tapper WJ, Hunt S, Deloukas P, Bentley DR, and Morton NE (2004) Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *Proc.Natl.Acad.Sci.U.S.A* 101 (52):18075-18080.

149. Zhang W, Collins A, Maniatis N, Tapper W, and Morton NE (2002) Properties of linkage disequilibrium (LD) maps. *Proc.Natl.Acad.Sci.U.S.A* 99 (26):17004-17007.

150. Zhu X, Luke A, Cooper RS, Quertermous T, Hanis C, Mosley T, Gu CC, Tang H, Rao DC, Risch N, and Weder A (2005) Admixture mapping for hypertension loci with genome-scan markers. *Nat.Genet.* 37 (2):177-181.

Publications

1. Tapper W, **Gibson J**, Morton NE, Collins A.
A comparison of methods to detect recombination hotspots.
Human Heredity. 2008 In press.

2. **Gibson J**, Tapper W, Cox D, Zhang W, Pfeufer A, Gieger C, Wichmann HE, Kääb S, Collins AR, Meitinger T, Morton N.
A multimetric approach to analysis of genome-wide association by single markers and composite likelihood.
Proc Natl Acad Sci U S A. 2008 Feb 19;105(7):2592-7. Epub 2008 Feb 11.

3. **Gibson J**, Morton NE, Collins A.
Extended tracts of homozygosity in outbred human populations.
Hum Mol Genet. 2006 Mar 1;15(5):789-95. Epub 2006 Jan 25.

4. Tapper W, Collins A, **Gibson J**, Maniatis N, Ennis S, Morton NE.
A map of the human genome in linkage disequilibrium units.
Proc Natl Acad Sci U S A. 2005 Aug 16;102(33):11835-9. Epub 2005 Aug 9.

5. **Gibson J**, Tapper W, Zhang W, Morton N, Collins A.
Cosmopolitan linkage disequilibrium maps.
Hum Genomics. 2005 Mar;2(1):20-7.

6. Zhang W, Collins A, **Gibson J**, Tapper WJ, Hunt S, Deloukas P, Bentley DR, Morton NE.
Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps.
Proc Natl Acad Sci U S A. 2004 Dec 28;101(52):18075-80. Epub 2004 Dec 16.

7. Maniatis N, Morton NE, **Gibson J**, Xu CF, Hosking LK, Collins A.
The optimal measure of linkage disequilibrium reduces error in association mapping of affection status.
Hum Mol Genet. 2005 Jan 1;14(1):145-53. Epub 2004 Nov 17.

8. Maniatis N, Collins A, **Gibson J**, Zhang W, Tapper W, Morton NE.
Positional cloning by linkage disequilibrium.
Am J Hum Genet. 2004 May;74(5):846-55. Epub 2004 Mar 26.

The following published papers were included in the bound thesis, but are not made available due to copyright restrictions. Digital object identifiers (DOI) to the published papers are provided. Pages 16 –105 are removed from the digitised PDF.

Gibson, J. et al (2007) A multimeric approach to analysis of genome-wide association by single markers and composite likelihood. *PNAS* **105** (7) 2592-2597
doi:10.1073/pnas.0711903105

Gibson, J et al (2006) Extended tracts of homozygosity in outbred human populations. *Human Molecular Genetics* **15** (5) 789-795
doi:10.1093/hmg/ddi493

Tapper, W et al (2005)A map of the human genome in linkage disequilibrium units. *PNAS* **102** (33) 11835-11839
doi:10.1073/pnas.0505262102

Gibson, J et al (2004) Cosmopolitan linkage disequilibrium maps. *Human Genomics* **2** (1) 1479-7364
http://eprints.soton.ac.uk/24708/

Zhang, W et al (2004) Impact of population structure, effective bottleneck time, and allele frequency on linkage disequilibrium maps. *PNAS* **101** (52) 18075-18080
doi:10.1073/pnas.0408251102

Maniatis, N et al (2005) The optimal measure of linkage disequilibrium reduces error in association mapping of affection status. *Human Molecular Genetics* **14** (1) 145-153
doi:10.1093/hmg/ddi019

Maniatis, N et al (2004) Positional cloning by linkage disequilibrium. *Human Genetics* **73** (5) 846-855
doi:10.1086/383589