# Chapter V

# Sequence variation

# 5.1 Introduction

## *5.1.1 Human variation*

Inherited differences in DNA sequence contribute to phenotypic variation, influencing an individual's anthropometric characteristics, risk of disease and response to the environment (ISNPMWG, 2001). SNPs are the most common type of DNA variation in the human genome and large-scale discovery projects, aiming to catalogue them, are under way (section 1.7.1).

Although common human genetic variation is limited compared to other species, it remains impractical to discover and test all SNPs for a role in each disease. One attractive proposal suggests the testing of only functional SNPs (Collins *et al*., 1997). Functional variants are likely to either introduce an amino acid change or alter the base composition of regulatory sequences. With a partial gene list in hand and poor knowledge of the location of regulatory elements, this approach is currently not fully applicable.

The availability of an alternative strategy was revealed by empirical studies, which demonstrated that nearby SNPs often display strong correlation in the population: Inheritance of one SNP allele is tightly linked to the state of other, closely linked sites. These correlations exist because the unit of human inheritance is not the individual SNP, but rather the ancestral segment that has undergone minimal historical recombination, and thus has been handed down from generation to generation with little modification. A biological basis for defining these ancestral segments (haplotypes) is to examine the

genomic patterns of recombination. This can be achieved using linkage disequilibrium analysis.

### *5.1.2 Theoretical aspects of linkage disequilibrium*

Linkage Disequilibrium (LD) refers to the non-random association of alleles at linked loci. Such associations underlie all forms of genetic mapping. However, linkage analysis is based upon associations in well-characterised pedigrees, whereas LD refers to the associations within populations of "unrelated" individuals. Nonetheless, there is a close relationship between the two approaches, because the "unrelated" individuals in a population are unrelated only in a relative and approximate sense (Nordborg and Tavare, 2002). In other words, the "unrelated" individuals will have a common ancestor at some point in the distant past. This makes LD particularly suitable to fine-scale mapping because it allows a lot more opportunities for recombination to take place.

LD is quantified using statistics of association between the allelic states at pairs of loci. D is one of the earliest measures of LD proposed (Lewontin, 1964), and quantifies disequilibrium as the difference between the observed frequency of a two-locus haplotype and the expected frequency if the alleles were segregating at random (i.e. if they were in linkage equilibrium). Consider two loci A and B with alleles $A_1/A_2$ and $B_1/B_2$. The proportion of chromosomes on which alleles $A_1$ and $B_1$ co-occur in the population is the observed frequency, denoted by $P_{11}$. The expected frequency under linkage equilibrium is the product of the allele frequencies in the population. Thus,

$$D = P_{11} - p_1 q_1 \qquad (1)$$

where the allele frequencies are symbolised as follows: $p_1$ = f($A_1$); $p2$ = 1 - $p_1$ = f($A_2$); $q_1$ = f($B_1$); $q_2$= 1- $q_1$ = f($B_2$).

If D differs significantly from zero, LD is said to exist. The degree of LD between two loci is dependent on both the recombination fraction, θ, and time in generations, *t*. Thus, D will tend to be smaller when two loci are located further apart, and D will decrease through time as a result of recombination (Jorde, 2000).

Although D captures the intuitive concept of LD, its numerical value is of little use for measuring the strength of and comparing levels of LD. This is due to the dependence of D on allele frequencies in the population: its maximum value is given by $D_{max}$ = min($p_1q_2$, $p_2q_1$), whereas its minimum value is given by $D_{min}$ = max(-$p_1q_1$, -$p_2q_2$) As a result, several alternative measures, based on D, have been devised (reviewed in Devlin and Risch, 1995; note that although they are all based on Lewontin's D, they have different properties and measure different things (Ardlie *et al*., 2002)). The most common measures are the absolute value of D' and $r^2$.

The absolute value of D' is determined by dividing D by $D_{max}$ (Lewontin, 1964).

$$D' = \frac{D}{D_{max}} \qquad (2)$$

D' = 1 (complete LD) if, and only if, two SNPs have not been separated by recombination (or recurrent mutation or gene conversion) during the history of the sample. In this case, at most three of the four possible two-locus haplotypes are observed

in the sample (Figure 5.1). Values of D' < 1 indicate that the complete ancestral LD has been disrupted, but the relative magnitude of values of D' < 1 has no clear interpretation. Estimates of D' are inflated in small samples, especially for SNPs with rare alleles. In addition, samples are difficult to compare because the magnitude of D' depends strongly on sample size. Therefore, statistically significant values of D' that are near 1 provide a useful indication of minimal historical recombination, but intermediate values should not be used for comparisons of the strength of LD between studies, or to measure the extent of LD (Ardlie *et al*., 2002).

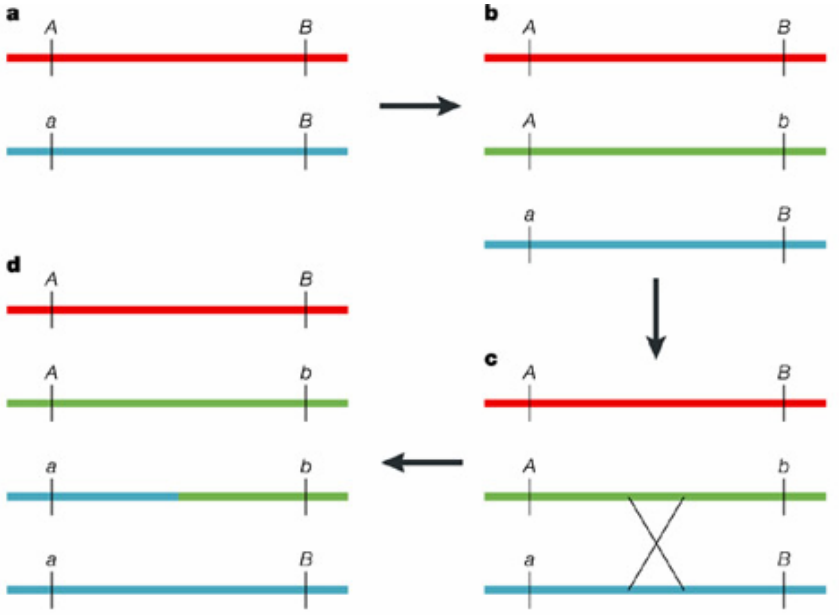



**Figure 5.1 (reproduced from Ardlie *et al*., 2002): The erosion of linkage disequilibrium by recombination. (a) At the outset, there is a polymorphic locus with alleles A and a. (b) When a mutation occurs at a nearby locus, changing an allele B to b, this occurs on a single chromosome bearing either allele A or a at the first locus (A in this example). So, early in the lifetime of the mutation, only three out of the four possible haplotypes will be observed in the population. The b allele will always be found on a chromosome with the A allele at the adjacent locus. (c) The association between alleles at the two loci will gradually be disrupted by recombination between the loci. (d) This will result in the creation of the fourth possible haplotype and an eventual decline in LD among the markers in the population as the recombinant chromosome (a, b) increases in frequency.**

The measure $r^2$ (Hill and Robertson, 1968; also labelled $R^2$ or $\Delta^2$) is in some ways complementary to D', and has recently emerged as the measure of choice for quantifying and comparing LD in the context of mapping (Pritchard and Przeworski, 2001; Weiss and Clark, 2002).

$$r^2 = \frac{D^2}{(p_1 p_2 q_1 q_2)} \quad (3)$$

$r^2 = 1$ (perfect LD) if, and only if, the markers have not been separated by recombination and have the same allele frequency. In this case, exactly two out of the four possible two-locus haplotypes are observed in the sample and observations at one marker (locus) provide complete information about the other marker (locus). Note that the value of $r^2$ is related to the amount of information provided by one locus about the other and that this property correctly takes into account differences in allele frequencies at the two loci. However, it also means that two markers that are immediately adjacent might show different $r^2$ values with a third marker, and that a low pairwise $r^2$ is not necessarily indicative of high ancestral recombination in the region. The sample size required to detect statistically significant LD is inversely proportional to $r^2$. $r^2$ also shows much less inflation in small samples than does D'. Typically, $r^2$ is lower than D' for any chromosomal distance (Ardlie *et al*., 2002; Weiss and Clark, 2002).

Mutation and recombination might have the most evident impact on LD, but there are additional contributors to the extent and distribution of disequilibrium. Most of these involve demographic aspects of a population, and tend to sever the relationship between LD strength and the physical distance between loci. Examples include population growth,

admixture/migration, population structure, natural selection, variable recombination rates, variable mutation rates and gene conversion.

### *5.1.3 Allelic associations and common disease*

Many studies have examined LD across small regions, like genes. These studies have generally concluded that the extent of disequilibrium is highly variable both across and between regions, and also differs between populations (reviewed in Pritchard and Przeworski, 2001; Boehnke, 2000; Jorde, 2000).

Studies across large contiguous genomic regions show a variable pattern of LD with regions of nearly complete LD interspersed with regions that show little, or no LD. Furthermore, they increasingly suggest that the human genome can be parsed objectively into haplotype blocks: sizeable regions over which there is little evidence for historical recombination, and within which only a few common haplotypes are observed (Olivier *et al*., 2001; Reich *et al*., 2001; Patil *et al*., 2001; Dawson *et al*., 2002; Gabriel *et al*., 2002).

With most human genetic variation being attributable to a limited set of common haplotypes, scanning the genome for regions of association to disease becomes feasible by testing the subset of variants able to report the common variants. However, for the study of complex, common diseases, a key question is whether the causative variants that confer disease susceptibility are likely to be common or rare. The answer cannot be known with certainty, of course, until these variants are identified and characterised; but there is a growing list of examples of common variants that predispose to common disease (examples include the SNPs in the apolipoprotein E gene (Davignon *et al*., 1988) and the factor V Leiden mutation (Bertina *et al*., 1994)).

To systematically test this hypothesis in a given population, we need a map of haplotype blocks that captures most of the genome. Although a dense SNP map is already available, empirical studies suggest that many more SNPs may be needed to achieve coverage of >90% of the genome in haplotype blocks (Jeffreys *et al*., 2001); such a project will require testing well above one million SNPs in multiple populations. This can only be achieved through the development of high throughput and cost effective genotyping platforms (reviewed in section 1.7.2). Although no one technique can be designated as the method of choice, the application of mass spectrometry in SNP genotyping emerges as a serious contender.

## *5.1.4 Mass spectrometry*

### 5.1.4.1 Background

Recent technological innovations have made nucleic acids accessible to mass spectrometric analysis. Due to its inherently high specificity, accuracy and throughput, mass spectrometry is an attractive detection method for SNP genotyping (Jackson *et al*., 2000; Kwok, 1998; Leushner, 2001).

Mass spectrometry (MS) is used to measure the mass-to-charge ratio (m/z) of ions, which can be used to infer their molecular weight. The recent advent of electrospray ionisation (ESI, Fenn *et al.*, 1989) and matrix-assisted laser desorption-ionisation (MALDI, Karas and Hillenkamp, 1988) techniques enable the routine mass spectroscopic analysis of various biomolecules, including peptides/proteins, lipids and carbohydrates (reviewed in Griffiths *et al*., 2001; Harvey, 2001), as well as nucleic acids.

Before the advent of ESI and MALDI, it was not possible to acquire the mass spectra of non-volatile, thermally labile, intact molecules with molecular weights greater than ~1-2 KDa. ESI and MALDI allow the production of gas-phase ions from solution and solid phases respectively. Different ionisation methods can be used with different mass analysers but most commonly ESI is coupled to either a quadrupole or ion-trap analyser whereas MALDI is coupled to a time-of-flight (TOF) analyser (Jackson *et al*., 2000). Of these, only MALDI-TOF will be discussed in more detail.

In MALDI-TOF the compound to be analysed (the analyte) is co-crystallised with excess light-absorbing matrix. Under high vacuum, the sample crystal is irradiated with an ultraviolet laser pulse that vaporises and ionises both analyte and matrix at the same time. Since the matrix absorbs most of the laser energy, sample fragmentation does not usually occur for smaller molecules. In the TOF analyser, the molecular ions are accelerated and passed over a flight tube, during which the ions are separated according to their m/z ratios. The smaller the ion, the faster it reaches the end of the tube. By measuring the ions at the end of the tube over a short time, a mass spectrum is generated. Under optimal conditions, MALDI produces singly charged ions. This simplifies the calculation of molecular masses that can be determined with high accuracy (0.01% to 0.1%). The entire process, including data acquisition, can be completed in approximately 3 seconds (Leushner and Chiu, 2000). A schematic diagram of MALDI-TOF is shown in Figure 5.2.

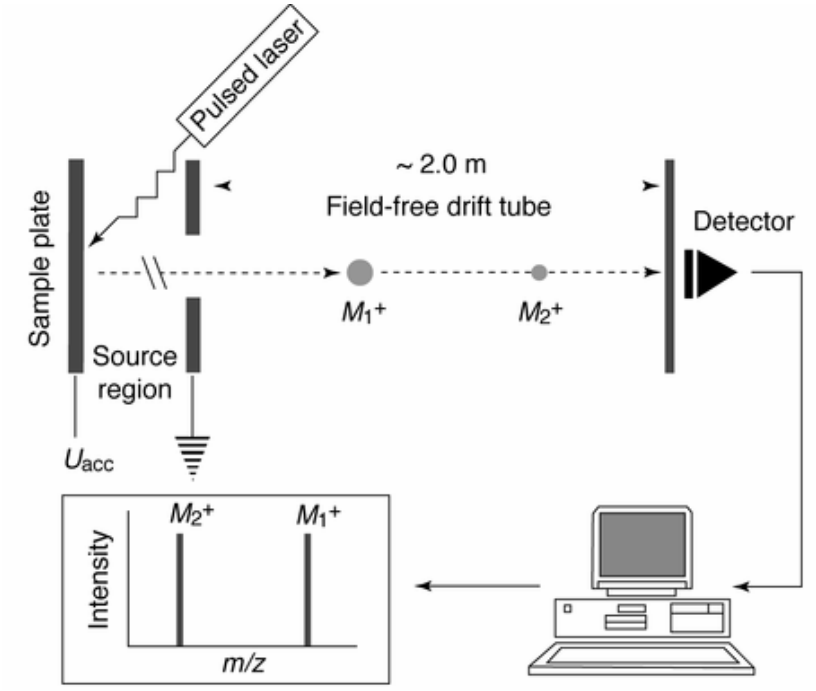



**Figure 5.2: MALDI-TOF MS (reproduced from Griffin and Smith, 2000, 2002). Matrix and analyte ions are desorbed and ionised upon irradiance with a laser pulse in the source region; a potential ($U_{acc}$) applied to the sample accelerates the ions into the field-free drift tube. The time-of-flight of each ion is measured and converted into m/z. The diagram shows the separation and detection of two positive, single-charged ions with different masses, $M_1$ and $M_2$.**

### 5.1.4.2 Genotyping methods using mass spectrometry

The short oligonucleotide mass analysis (SOMA) is one of the few techniques reported that employs ESI rather than MALDI for SNP analysis (Laken *et al*., 1998). The genomic region to be analysed is PCR amplified with primers containing a sequence for a type IIS restriction enzyme. Enzymatic digestion of the PCR products yields fragments as small as 7 bp. HPLC is used for improved purification of the samples, which are then analysed by ESI MS.

One of the MALDI methods developed for genotyping uses allele specific, mass-labelled, peptide nucleic acid (PNA) hybridisation probes (Griffin *et al*., 1997). PNA probes are structural analogs that have increased hybridisation stability and specificity over conventional DNA probes. The biotinylated target DNA (e.g. a PCR amplicon) is immobilised by binding to streptavidin-coated magnetic beads. The non-biotinylated strand is removed, followed by PNA probe hybridisation. Stringent washing conditions are used to remove the unbound probe and achieve proper discrimination. MALDI is then used to determine the mass of the bound PNA probes.

MALDI-TOF MS can also determine genotypes by characterising the products of primer extension (mini-sequencing) reactions. Application of a mini-sequencing based approach involves PCR amplification of the sequence of interest, followed by the annealing of a primer extension probe hybridising immediately upstream of the variable site. The probe is extended and the mass of the extended products is then used to determine the composition of the variable site. In the PROBE (primer oligonucleotide base extension) assay (Braun *et al*., 1997a, 1997b), the annealed probe is extended through the SNP site in the presence of three dNTPs and one ddNTP (incorporation of a ddNTP terminates the extension reaction). In the PinPoint assay (Haff and Smirnov, 1997; Ross *et al*., 1998; Fei *et al*., 1998) the primer extension probe is extended by only one base, in the presence of four ddNTPs. Finally, in the very short extension assays (Sun *et al*., 2000), the primer extension probe is extended in the presence of one dNTP and three ddNTPs, which tends to produce very short extension products. Platforms that employ this approach include the MassEXTEND assay (Sequenom) and the GOOD assay (Sauer *et al*., 2000).

MALDI was also been used in a miniaturised, chip-based probe annealing, extension and termination approach (Tang *et al*., 1999). It was also employed to analyse the products of Invader assays (Griffin *et al*., 1999). Finally, genotyping could potentially be achieved through sequencing by mass spectrometry (Köster *et al*., 1996; Kirpekar *et al*., 1998; Fu *et al*., 1998)

### *5.1.5 This chapter*

This chapter describes:

i. The identification of exonic SNPs in 20q12-13.2 by comparing the finished reference sequence to EST and mRNA sequences. The annotation of the region was used to classify each SNP as either non-synonymous (amino acid change), or synonymous (no amino acid change), or in UTR, whereas a subset was experimentally verified using the homogeneous MassEXTEND assay (Sequenom).

ii. Selection, genotyping and analysis of a set of circa 2,200 SNPs distributed across 119 individuals from three populations: twelve unrelated individuals of Asian origin, twelve unrelated individuals of African American origin and 95 Caucasian individuals from twelve multigenerational pedigrees. The three panels (each of twelve individuals) were used to verify the SNPs and estimate their allele frequencies in each population. The genotype data from the twelve pedigrees was used to obtain haplotype data and investigate the extent of LD across the region.

# 5.2 Exonic SNP discovery in 20q12-13.2

## *5.2.1 Identification of exonic SNPs in silico*

BLAST searches (performed as part of the automated analysis, section 3.2) were used to identify human expressed sequences with high homology to the finished sequence. The graphical BLAST viewer Blixem (Sonnhammer and Durbin, 1994) was used to manually inspect the alignments of the homologous human EST, cDNA and in-house generated sequences (Figure 5.3).

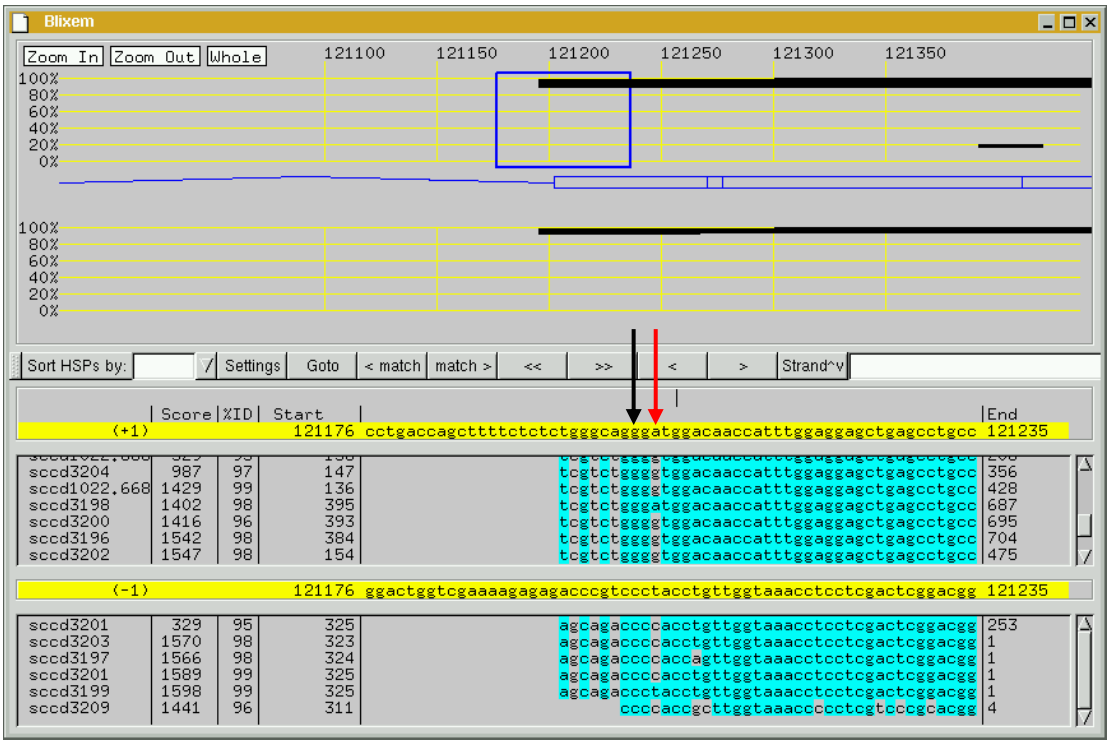



**Figure 5.3: Blixem view of homologous sequences (also see Figure 4.2). The region shown corresponds to the sequence from clone dJ453C12 that encodes part of the last exon of C20orf35. The exon starts at position 121,202 of the clone sequence (black arrow). Homologous expressed sequences corresponding to either the forward (+1) or reverse (-1) strand are shown below (sequence names at far left; not all sequences are shown). For each expressed sequence, nucleotides identical to the reference sequence are highlighted blue. Various vectorette sequences and other ESTs (e.g. BE908675 and AW37112; not shown) identify a candidate A□G variation at position 121,204 (red arrow).**

Each position, where two or more expressed sequences differed from the reference genomic sequence, was tagged as harbouring a putative exonic SNP. Expressed sequences that differ at multiple positions from the reference sequence (< 95% ID) were not used for SNP identification.

This approach yielded 124 putative SNPs. As shown in Figure 5.4 more than 60% of these SNPs are supported by at least four expressed sequences. Fifteen of the SNPs were known (previously identified by other SNP discovery projects). The 109 new SNPs were incorporated into 20ace (http://webace.sanger.ac.uk/cgi-bin/webace?db=acedb20) and submitted to dbSNP (http://www.ncbi.nlm.nih.gov/SNP/snp_search.cgi?searchType=byBatch&batch_id=4640).

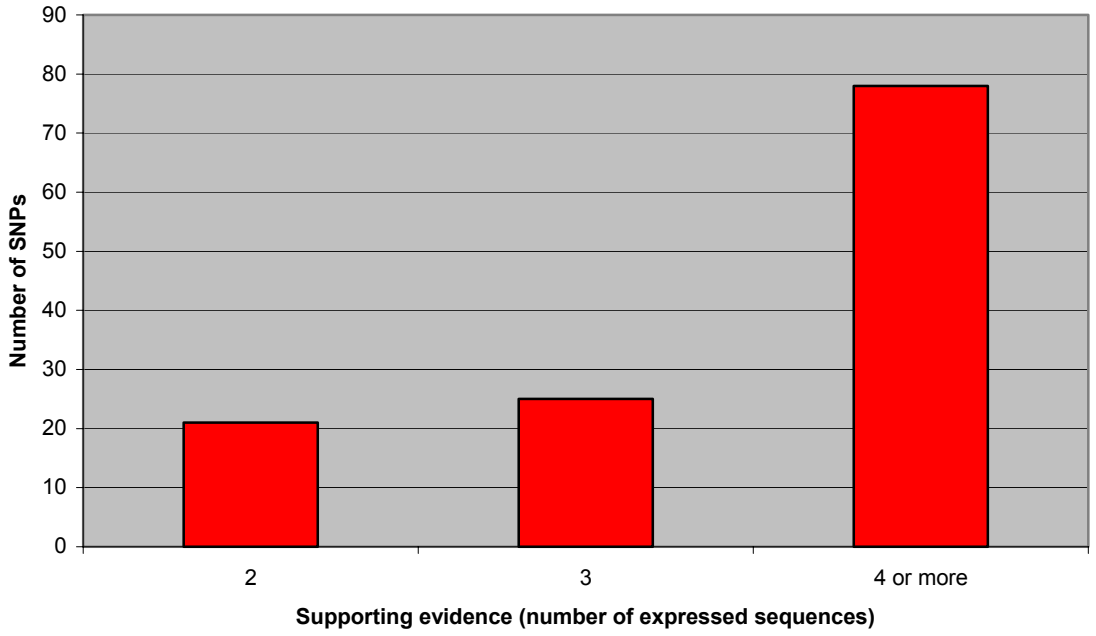



**Figure 5.4: Supporting evidence for exonic SNPs.**

### *5.2.2 Features of exonic SNPs*

Under random mutation, half as many transitions as transversions are expected to occur (Dawson *et al*., 2001; Table 5.1). Since the strand on which the changes occurred is not known, no distinction can be made between A↔G and C↔T and between C↔A and G↔T. In this data set (124 SNPs) transitions (66.9%) occur twice as often as transversions (33.1%) (Table 5.2), a pattern already reported in other SNP identification studies (Horton *et al*., 1998; Dawson *et al*., 2001; Deutsch *et al*., 2001). The most common change was C↔T (G↔A) (83/124), which probably reflects the deamination of 5-methylcytosine that occurs frequently at CpG dinucleotides. The other thing to note from Table 5.2 is that the occurrence of A↔T (T↔A) is less than half compared to any other variation (Dawson *et al*., 2001; Smink, 2000).

**Table 5.1: Expected distribution of transitions and transversions.**

| | A | T | C | G |
|---|---|---|---|---|
| A | - | | | |
| T | Transversion | - | | |
| C | Transversion | Transition | - | |
| G | Transition | Transversion | Transversion | - |

**Table 5.2: Distribution of transitions and transversions.**

| Variation | Number | Number |
|---|---|---|
| C↔T (G↔A) | 83 | 66.9% |
| **Transitions** | **83** | **66.9%** |
| C↔A (T↔G) | 28 | 22.6% |
| C↔G | 9 | 7.3% |
| A↔T | 4 | 3.2% |
| **Transversions** | **41** | **33.1%** |

On the basis of the 20q12-13.2 annotation, at least one exonic SNP was identified for 47 of the 99 coding genes in the region (Appendix 14). Most (92/124) SNPs are in the UTRs, whereas the remaining 32 (25.8%) map within the ORF of twenty genes (Table 5.3). Of the 21 variations that correspond to a 3' codon position change (Figure 5.5), only one results in an amino acid change. In contrast, five of the six first base changes and four of the five second base changes result in amino acid changes.

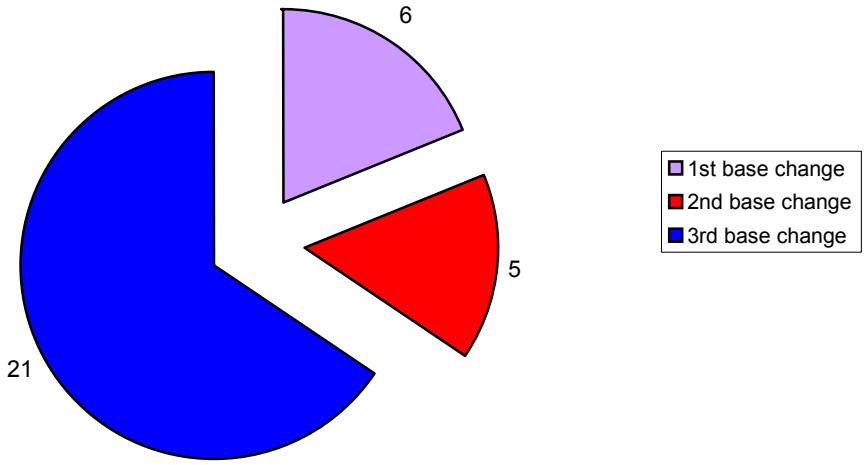



**Figure 5.5: Codon position changes for coding exonic SNPs.**

**Table 5.3: Coding changes for exonic SNPs.**

| Clone name | SNP position (clone sequence) | Gene | Allele change | Codon change | Encoded amino acid | Amino acid change |
|---|---|---|---|---|---|---|
| dJ511B24 | 64,001 | PLCG1 | T ↔ C | ATC ↔ ACC | ile ↔ thr | yes |
| dJ862K6 | 107,817 | SFRS6 | T ↔ C | CCT ↔ CCC | pro ↔ pro | |
| dJ138B7 | 73,871 | C20orf9 | G ↔ A | GTG ↔ GTA | val ↔ val | |
| dJ1028D15 | 93,173 | MYBL2 | T ↔ C | CCT ↔ CCC | pro ↔ pro | |
| dJ1028D15 | 96,061 | MYBL2 | C ↔ G | GCC ↔ GCG | ala ↔ ala | |
| dJ148E22 | 56,722 | YWHAB | A ↔ C | CGA ↔ CGC | arg ↔ arg | |
| dJ148E22 | 60,103 | YWHAB | T ↔ A | GTG ↔ GAG | val ↔ glu | yes |
| dJ1069P2 | 89,939 | TOMM34 | A ↔ G | GCA ↔ GCG | ala ↔ ala | |
| dJ1069P2 | 94,339 | C20orf119 | G ↔ A | CTC ↔ CTT | leu ↔ leu | |
| dJ1069P2 | 94,345 | C20orf119 | G ↔ A | CTG ↔ TTG | leu ↔ leu | |
| dJ1069P2 | 95,369 | C20orf119 | T ↔ C | TCA ↔ TCG | ser ↔ ser | |
| dJ453C12 | 121,204 | C20orf35 | A ↔ G | ATG ↔ GTG | met ↔ val | yes |
| dJ453C12 | 135,622 | PIGT | G ↔ A | ACG ↔ ACA | thr ↔ thr | |
| dJ453C12 | 135,781 | PIGT | C ↔ T | AGC ↔ AGT | ser ↔ ser | |
| dJ453C12 | 136,979 | PIGT | T ↔ C | TAT ↔ TAC | tyr ↔ tyr | |
| dJ461P17 | 119,487 | C20orf170 | T ↔ C | AAT ↔ AGT | asn ↔ ser | yes |
| dJ447F3 | 93,739 | TNNC2 | C ↔ A | ACG ↔ ACT | thr ↔ thr | |
| dJ337O18 | 14,252 | PTE1 | G ↔ A | GTC ↔ GTT | val ↔ val | |
| bA465L10 | 100,183 | MMP9 | G ↔ C | CGG ↔ CCG | arg ↔ pro | yes |
| bA394O2 | 49,352 | KIAA1834 | A ↔ G | GAT ↔ GAC | asp ↔ asp | |
| dJ28H20 | 2,734 | C20orf64 | G ↔ A | GCC ↔ GCT | ala ↔ ala | |
| dJ1049G16 | 55,543 | NCOA3 | G ↔ A | GGG ↔ GGA | gly ↔ gly | |
| dJ1049G16 | 67,054 | NCOA3 | A ↔ G | CAA ↔ CAG | gln ↔ gln | |
| dJ1049G16 | 67,057 | NCOA3 | G ↔ A | CAG ↔ CAA | gln ↔ gln | |
| dJ1049G16 | 67,084 | NCOA3 | A ↔ G | CAA ↔ CAG | gln ↔ gln | |
| bA269H4 | 64,588 | KIAA1415 | T ↔ C | AAG ↔ GAG | lys ↔ glu | yes |
| dJ998C11 | 2,266 | KIAA1415 | G ↔ C | CAC ↔ CAG | his ↔ gln | yes |
| dJ155G6 | 63,791 | CSE1L | T ↔ C | TCT ↔ CCT | ser ↔ pro | yes |
| dJ155G6 | 63,887 | CSE1L | A ↔ C | AAA ↔ CAA | lys ↔ gln | yes |
| dJ686N3 | 10,895 | DDX27 | C ↔ T | TTC ↔ TTT | phe ↔ phe | |
| dJ686N3 | 34,718 | KIAA1404 | A ↔ G | CAT ↔ CAC | his ↔ his | |
| dJ686N3 | 34,741 | KIAA1404 | G ↔ C | CTT ↔ GTT | leu ↔ val | yes |

# 5.3 Studying sequence variation across 20q12-13.2

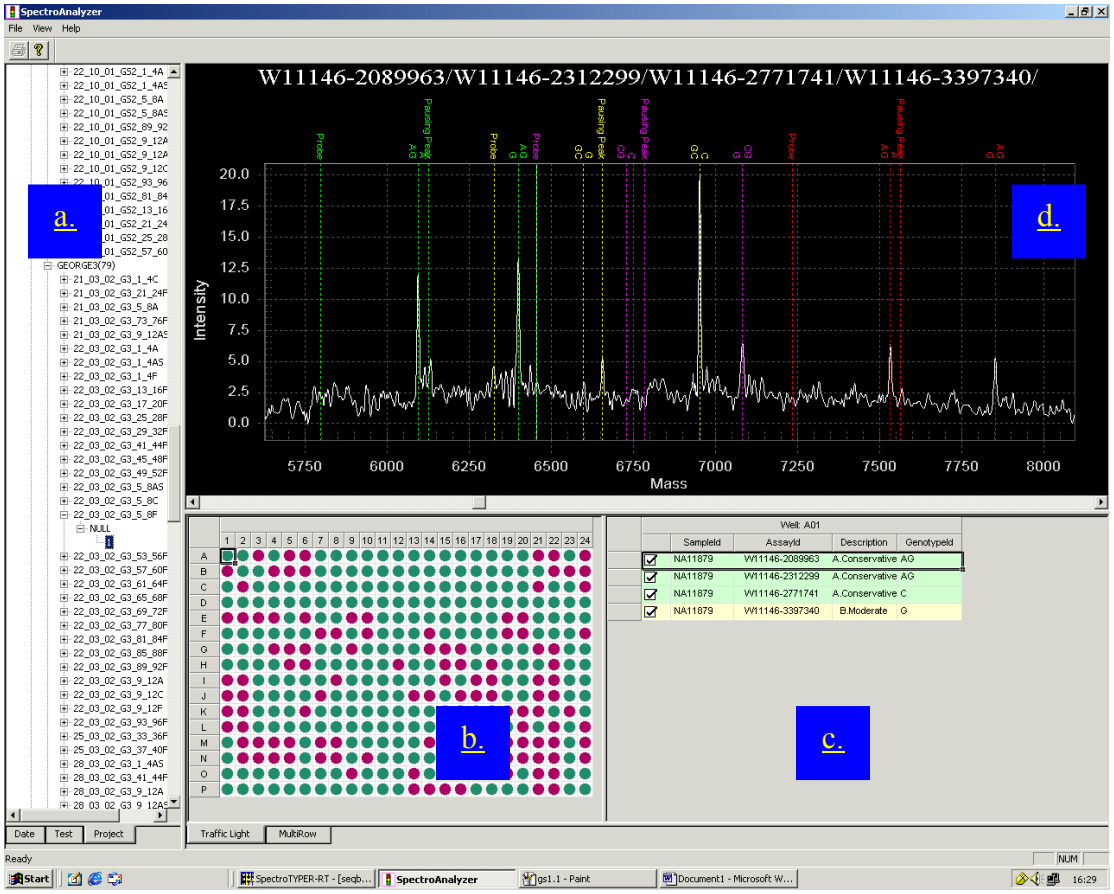
## *5.3.1 SNP selection and high-throughput genotyping*

Homogeneous MassEXTEND assays were designed for 2,208 SNPs mapping in 20q12-13.2 using the SpectroDesigner software. During the first round of SNP selection, SNPs mapping more than >15 Kb apart were selected. During subsequent SNP selection rounds, neighbouring SNPs were selected to replace failed or non-polymorphic ones; additional SNPs mapping between polymorphic SNPs were also selected to decrease the average SNP distance to ~5 Kb. In general, SNPs mapping <2 Kb apart were avoided. 106 of the SNPs selected were identified by this study (section 5.2), whilst the remaining 2,102 were imported from dbSNP or generated in-house by a parallel SNP discovery project on chromosome 20. In brief, chromosome 20 was flow-sorted from a Caucasian, an African American, an African pygmy and a Chinese cell line. Individual 2 Kb insert libraries were prepared and shotgun sequenced to a depth of 2x coverage. The SSAHA (Ning *et al*., 2001) software was used to align reads to the genome assembly and over 110,000 new SNPs were discovered (Deloukas, pers. communication). Note that this rich resource became available at a very late stage of this study and was only used to smooth the initially uneven distribution of the selected SNPs.

Assays were designed as quadruplex reactions and genotyping was performed as described in section 2.3. Genotyping was attempted in 119 individuals from three ethnic groups (twelve African Americans, twelve Asians and 95 Caucasians). DNA sample information is listed in Table 2.7 (Chapter II). Automated call analysis was performed

using the SpectroTyper RT package and data was stored in an Oracle database. An example of genotype calls is shown in Figure 5.6.

**A.**



**B.**

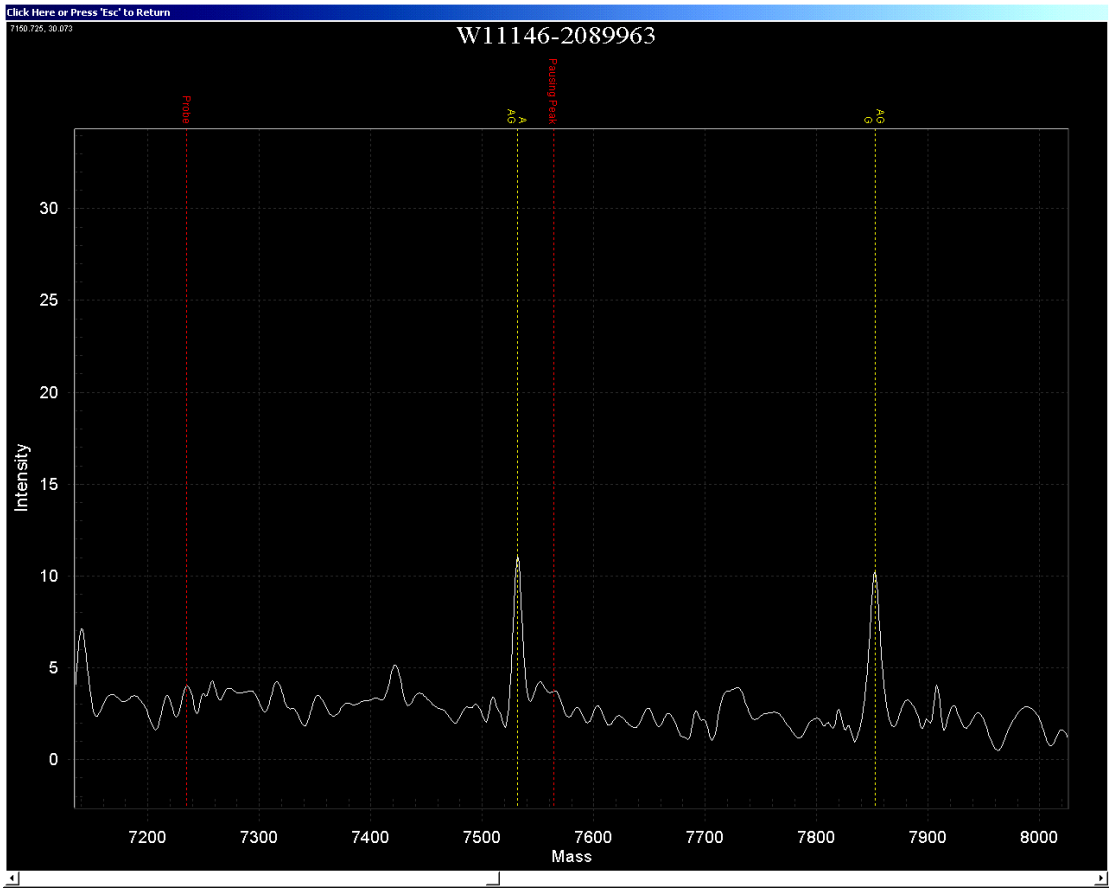



**Figure 5.6: Viewing genotyping results in SpectroAnalyser (SpectroTyper RT software package). (A) Sub window a lists genotyped plates; plate 22_03_02_G3_5_8F was selected for viewing. Sub window b shows a virtual 384-well plate, highlighting the type of genotyping results obtained for each well. Wells with four conservative/moderate calls are highlighted green, the rest are highlighted red. For a selected virtual well (A1 in this example) the individual assay information are listed in sub window c. Data listed include the DNA tested, well/SNP assay numbers, type of calls and the genotypes obtained. Sub window d shows the spectra obtained for these assays. Each colour corresponds to a particular assay. (B) Detailed view of the W11146-2089963 spectra. The "Probe" line (red) shows the expected peak position for un-incorporated (not extended) probe. Both "Probe" and "Pausing Peak" lines help monitor the completeness of the extension reaction. The two yellow lines show the peak positions for the two alleles obtained (A and G).**

### *5.3.2 Error checks and quality assessment of data*

All checks described in this section were performed by Sarah Hunt, using commercial software and customised in-house perl scripts.

The genotypes obtained from the twelve Caucasian families were used to test the SNP assays for Hardy-Weinberg equilibrium (Pedigree Statistics (c) 1999-2001, Gonçalo Abecasis). 48 assays violated H-W equilibrium ($\chi^2$>10) and they were excluded from further analyses.

Mendelian checks (PedCheck 1.1 (c) 1997-1999, Jeff O'Connell; MERLIN 0.8.8 (c) 2000-2001, Gonçalo Abecasis) were first performed in parents/offspring trios and then in whole families. 1,459 genotypes (from 202 SNPs) were involved in Mendelian errors and were excluded from further analyses.

Two independent tests were used to estimate error rates. Comparison of duplicate calls in the raw Sequenom data (all genotype calls) from this study suggests an error rate of 1.4% (983/69,618). Note that this percentage corresponds to the raw data (which includes all assays).

The results were also compared to data obtained by an independent study, which used the Illumina SNP assay platform. In total, 566 SNPs had conclusive data from both methods and 295 differences were detected in the combined set of 27,901 genotypes (1.1%). Of the total 295 discrepant calls, 215 (72.9%) are associated with only 21 SNP assays (3.7%). This and previous studies, using the Sequenom platform in our laboratory have shown that 2-3% of SNP assays (randomly selected from public resources) are sub-optimal. Although the exact reason is not fully understood, imbalanced allele

amplification appears to be the most likely cause. 490 SNPs (86.6%) showed no discrepant calls for any DNA. Less than half of the total differences (119/27,901; 0.43%) corresponded to high quality genotype calls (i.e. conservative calls for Sequenom and confidence of "5" for Illumina calls).

### *5.3.3 Estimation of allele frequencies in three populations*

Of the 2,208 SNPs genotyped in the three ethnic groups, circa 15% (331) failed, i.e. did not produce any genotype data, or failed during data checks (section 5.3.2). For a further 438 (20%) SNP assays, genotype data was incomplete for at least one ethnic group (genotype data was obtained for less than seven of the twelve individuals from each group). In total, 1,439/2,208 (65%) SNP assays gave genotype data for at least seven individuals in each group (Figure 5.7). This set of 1,439 "complete" SNPs was used for the analyses described below.

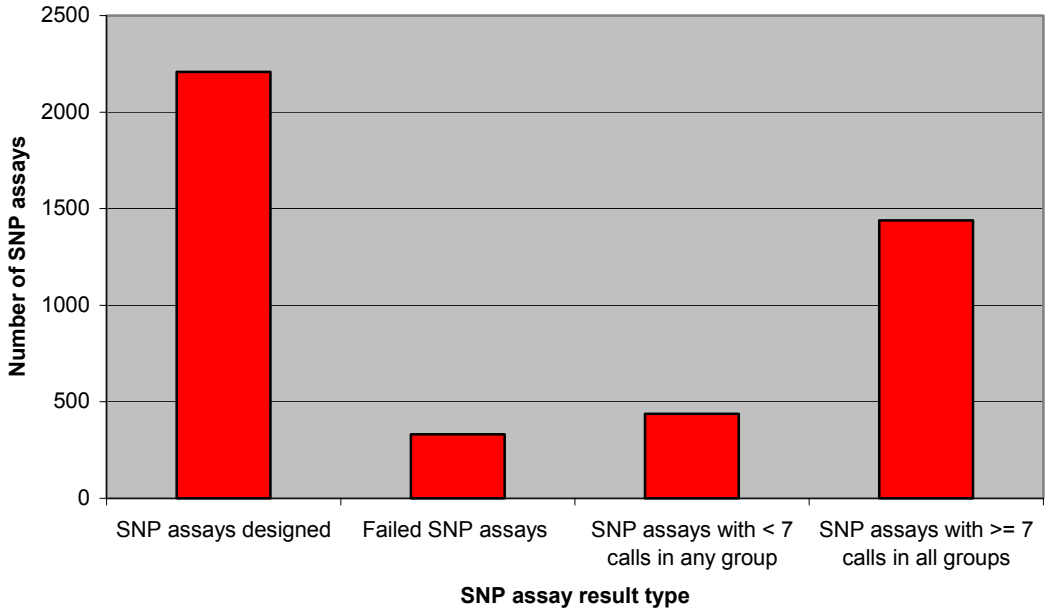



**Figure 5.7: Overall breakdown of SNP assay results.**

A total of 46,627 non-redundant genotypes were obtained for the 1,439 SNPs with "complete" assays (out of 51,804 possible genotypes). On average, 32.4 individuals were successfully genotyped for each SNP (32.4/36, 90% complete). 340 (23.5%) SNPs were found to be non-polymorphic in all ethnic groups (for the individuals tested), whilst the remaining 1,099 SNPs (76.5%) were found to be polymorphic in at least one population. The results are shown in Figure 5.8. The corresponding figures for the exonic SNPs identified *in silico* (section 5.2) were 48% non-polymorphic and 52% polymorphic in at least one population. This suggests that the *in silico* approach either yields more rare SNPs, or has a high false positive rate due to sequencing errors.

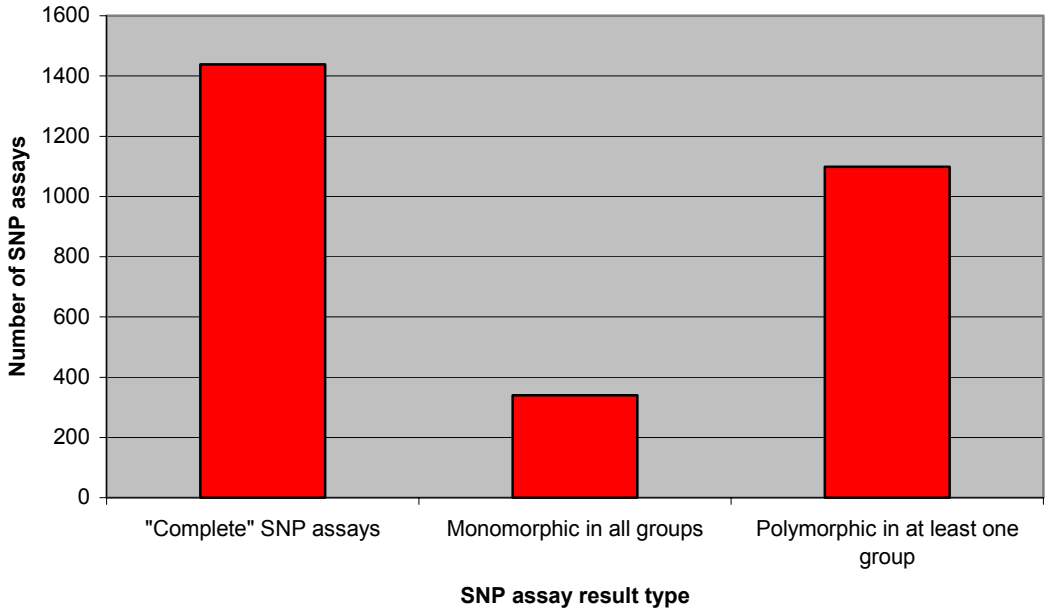



**Figure 5.8: Breakdown of "complete" SNP assay results.**

Figure 5.9 shows the distribution of polymorphic and non-polymorphic SNPs across the three populations. Compared to the other two populations, African Americans had the highest number of polymorphic SNPs (985/1,099). The corresponding numbers for Asians and Caucasians were 850 and 904, respectively. Approximately 66% (725/1,099) of SNPs were polymorphic in all groups.

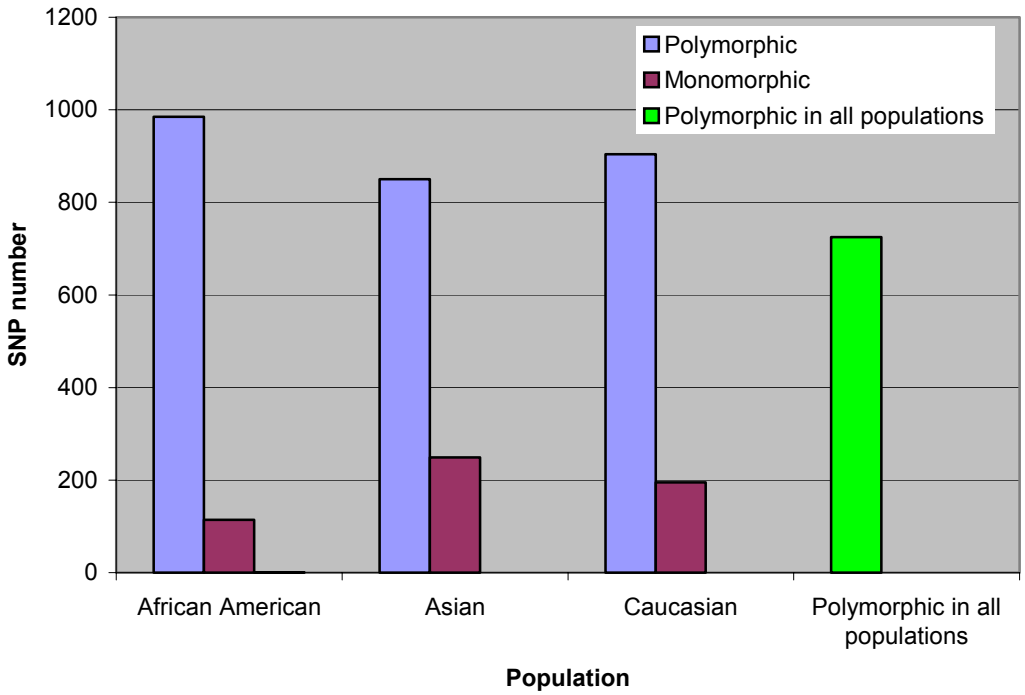



**Figure 5.9: Distribution of polymorphic and monomorphic SNPs across the three ethnic groups. Only polymorphic SNPs, in at least one group, were considered (1,099 SNPs).**

As illustrated in Figure 5.10, the African Americans showed more than two-fold more unique polymorphisms than either the Asians, or Caucasians. Of the SNPs found to be polymorphic in only two groups, the African American and Caucasian groups shared the most. Compared to Asians and Caucasians, the African Americans and Asians also shared more SNPs (polymorphic in only two groups).

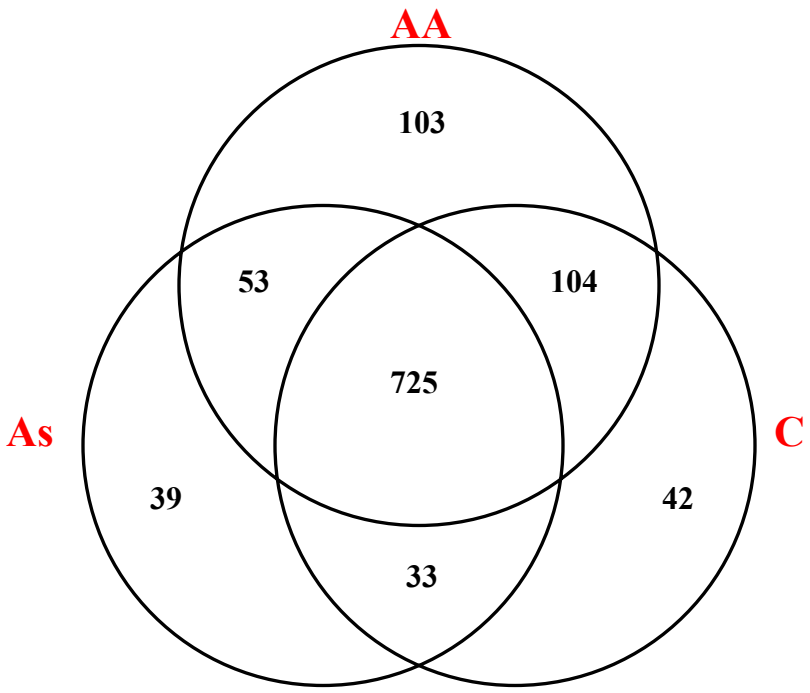



**Figure 5.10: Distribution of polymorphic SNPs in the three groups (African Americans, AA; Asians, As; Caucasians, C).**

Overall, the African Americans had the largest proportion of SNPs at the lower end of minor allele frequencies (40% of SNPs had a Minor Allele Frequency (MAF) of 4-20%). The corresponding numbers for Asians and Caucasians were 31.5% and 27.4%, respectively. The reverse was observed in the Caucasian population, for which 55% of SNPs had a MAF of 21-50%. The corresponding numbers for the Asian and African American populations were 45.8% and 49.5%, respectively (Figure 5.11).

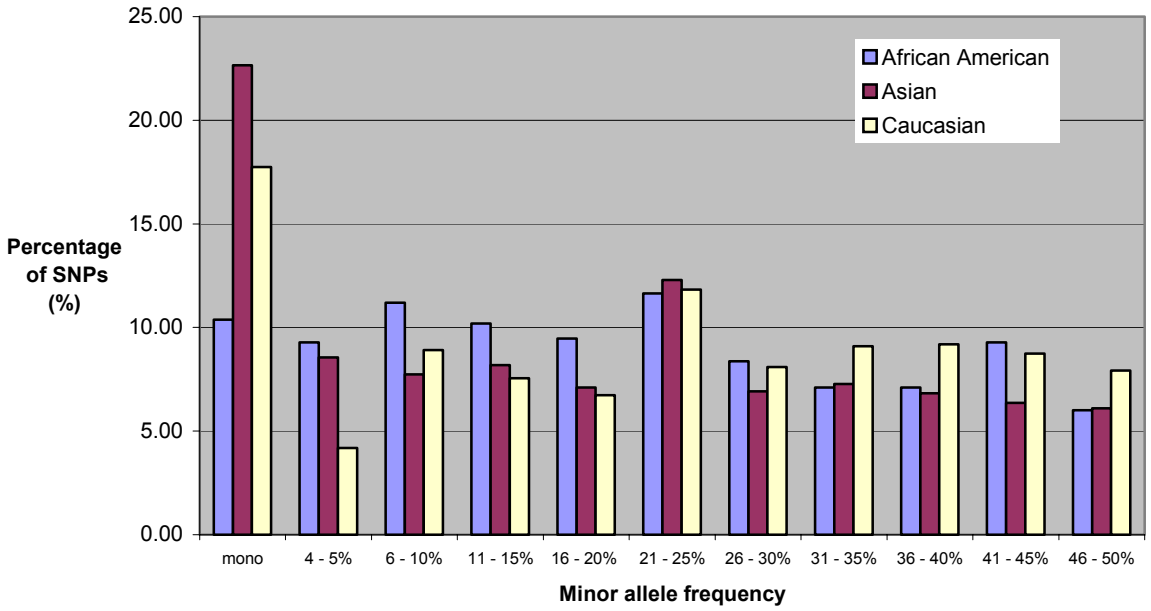
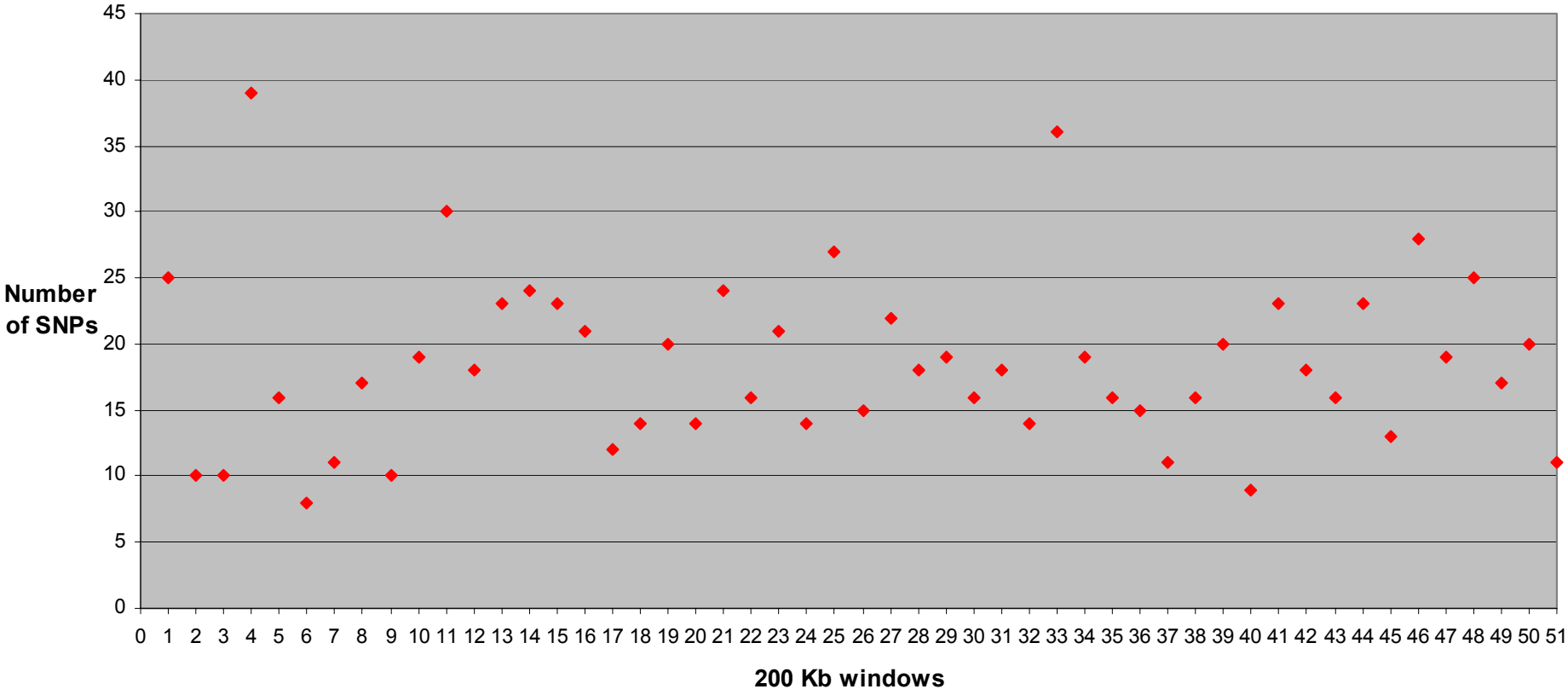



**Figure 5.11: Distribution of minor allele frequencies in ethnic populations.**

# 5.4 A first generation LD map of 20q12-13.2 in Caucasians

The 2,208 SNP assays were used to genotype twelve three-generation CEPH families with a total of 95 individuals (section 5.3.1). Allele frequencies were calculated using only the founder chromosomes (grandparents; 47 individuals). Following quality checks, a set of 943 SNPs, with MAF of ≥5% and with at least thirteen calls from founder chromosomes, was selected (Appendix 15). The distribution of these SNPs across 20q12-13.2 is shown in Figure 5.12.

**Figure 5.12: SNP distribution across the region of interest. The y-axis reports the total number of SNPs in non-overlapping windows of 200 Kb (x-axis).**



The average distance between SNPs is 10,709 bp (median 7,564 bp). As shown in Figure 5.13, 62.6% of neighbouring SNPs (590 of 942 pairs) are separated by less than 10 Kb, whilst 2.8% of SNP pairs (27) are separated by more than 40 Kb. The longest interval is 85,498 bp.

The sequence coverage per interval size (given in 5 Kb windows) is shown in Figure 5.14.

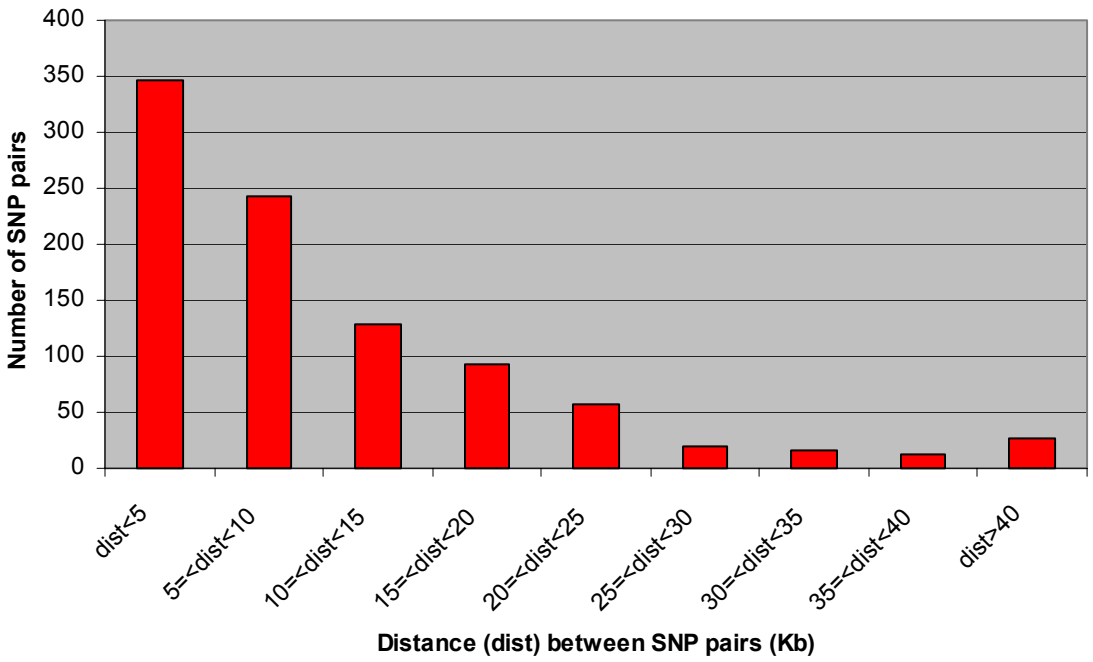



**Figure 5.13: Distance between neighbouring SNPs (SNP pairs).**

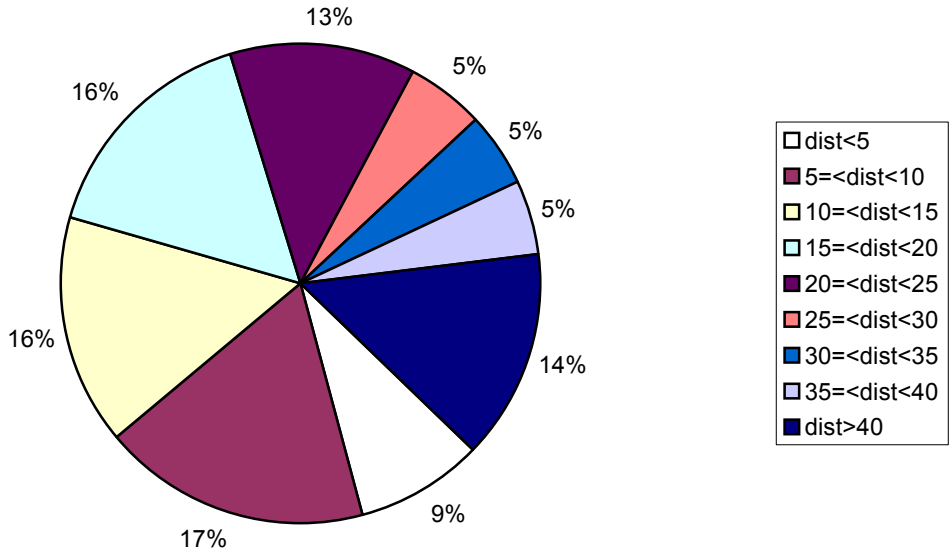
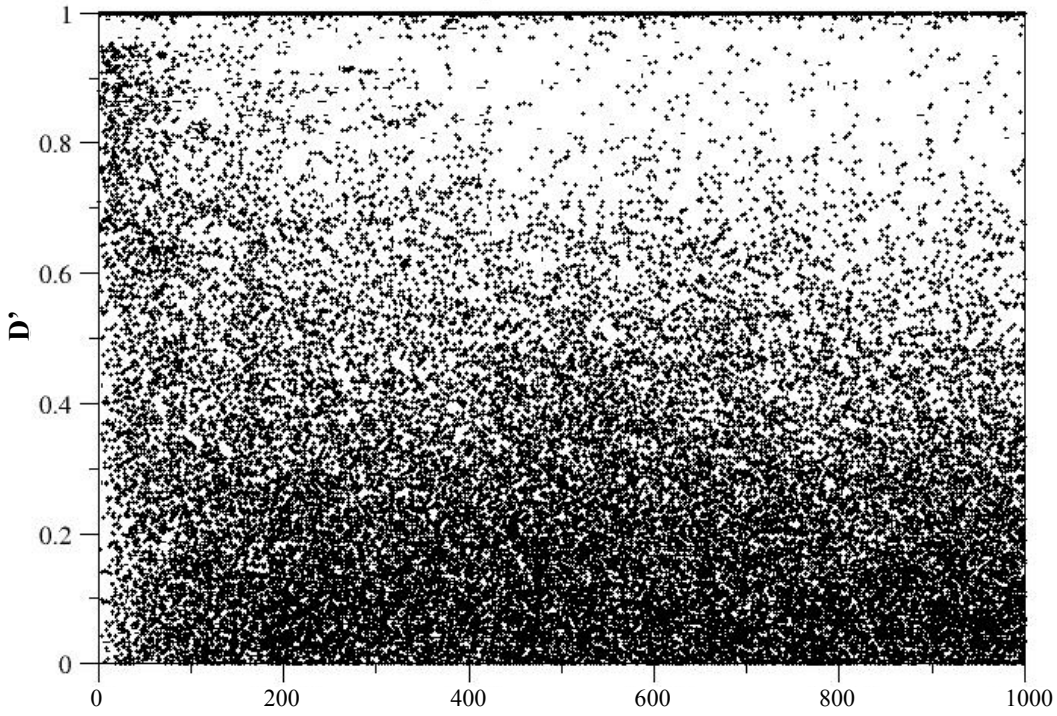



**Figure 5.14: Proportion of sequence occupied by the various types of SNP pairs.**

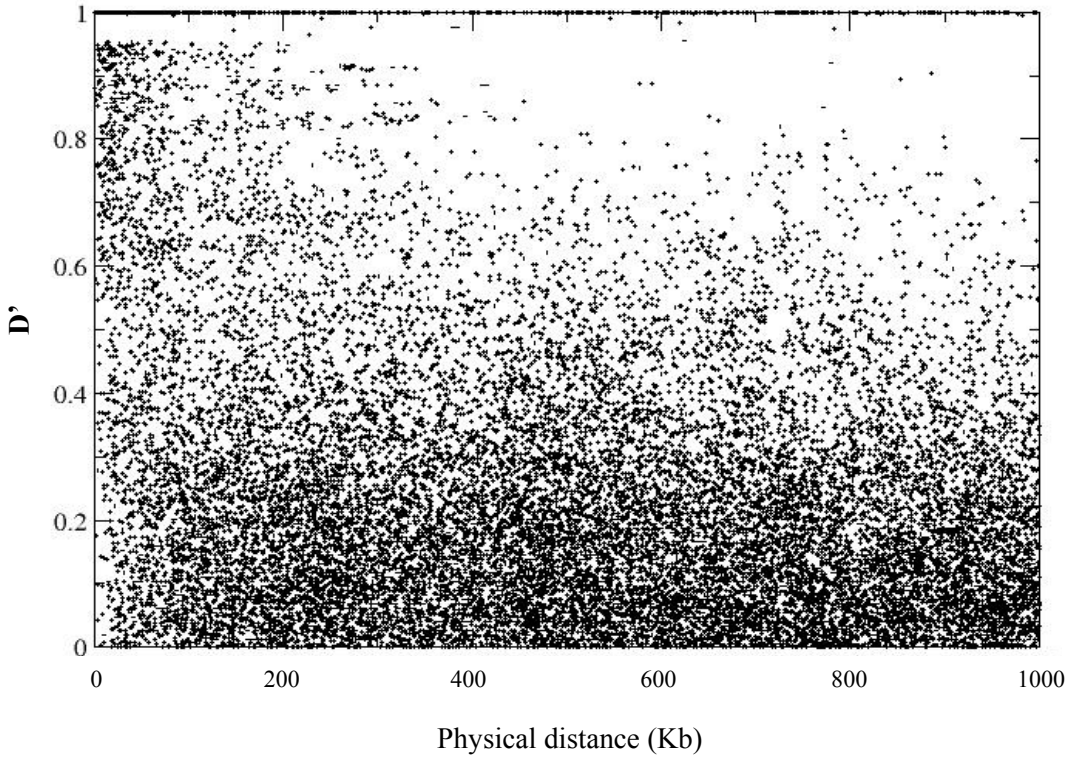
Of all polymorphic SNPs in Caucasians, those with >80% of all possible genotype calls were selected for further analysis. These 879 SNPs give an average spacing of 11.5 Kb. On average >93% of all genotyping calls were obtained for these SNPs. Statistical analyses were performed by Robert Lawrence at Oxford University.

LD between pairs of markers was calculated using D' and $r^2$. Calculations were performed for SNPs with MAF>10% (685), SNPs with MAF>20% (485), and all polymorphic SNPs (879). As shown in Figure 5.15, LD decays with increasing distance, but also shows extensive variability. Very high D' values (>0.9) extend up to distances over 250 Kb, contrasting with occurrences of very low D' values (<0.2) for SNPs less than 5 Kb apart (Figure 5.15 A, B). The corresponding $r^2$ values also show extensive variability (Figure 5.15 C, D). LD decay as a function of increasing physical distance was calculated using the average D' and $r^2$ in successive 10 Kb windows and the plots are shown in Figure 5.16. For common SNPs (MAF>20%), average D' declines from 0.89 for markers less than 10 Kb apart to ~0.22 for markers > 250 Kb apart (Figure 5.16 B), whereas the corresponding values for $r^2$ are 0.5 and 0.03 respectively (Figure 5.16 D). Although the two measures differ in scale, their decay profiles are similar. Note that when the average D' is calculated using the 685 SNPs with MAF>10% the average D' values decline from 0.89 to 0.31. The observed differences between the two D' values are expected, since the estimates of D' are known to be inflated by SNPs with rare alleles. (section 5.1, also see Appendix 16). Also note that when the less common SNPs are included, the maximal average $r^2$ decreases (Figure 5.16 C, D and Appendix 16).

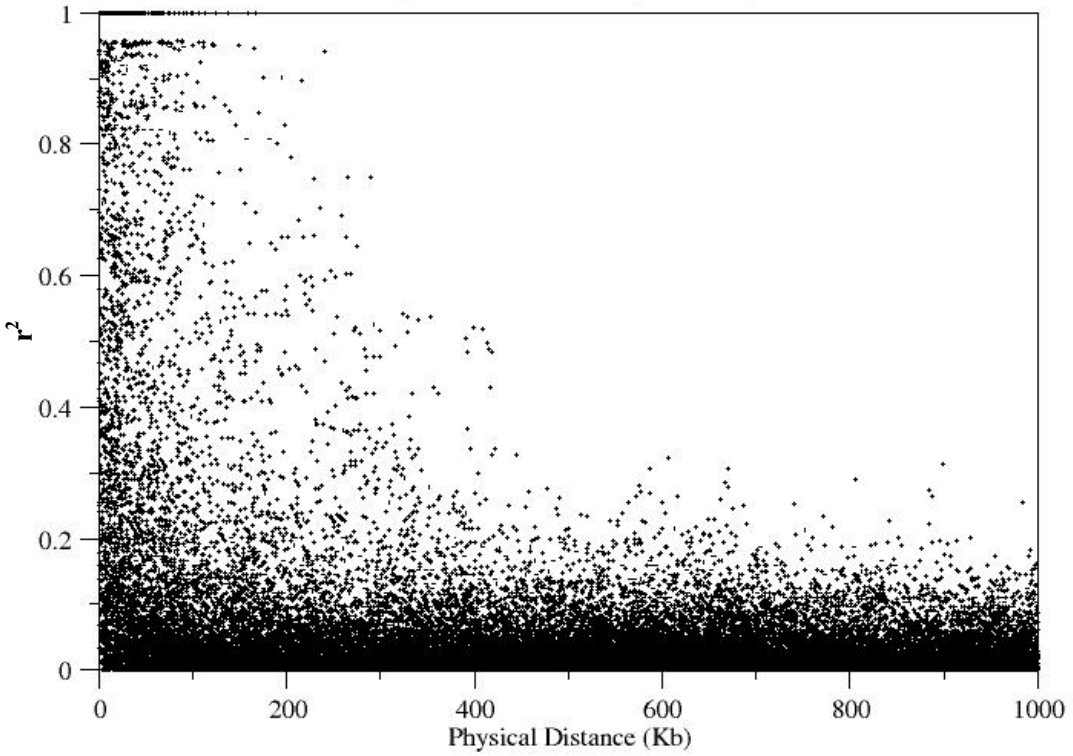
**A. D' (using SNPs with minor allele frequency >10%)**




**B. D' (using SNPs with minor allele frequency >20%)**



**C. $r^2$ (using SNPs with minor allele frequency >10%)**




**D. $r^2$ (using SNPs with minor allele frequency >20%)**

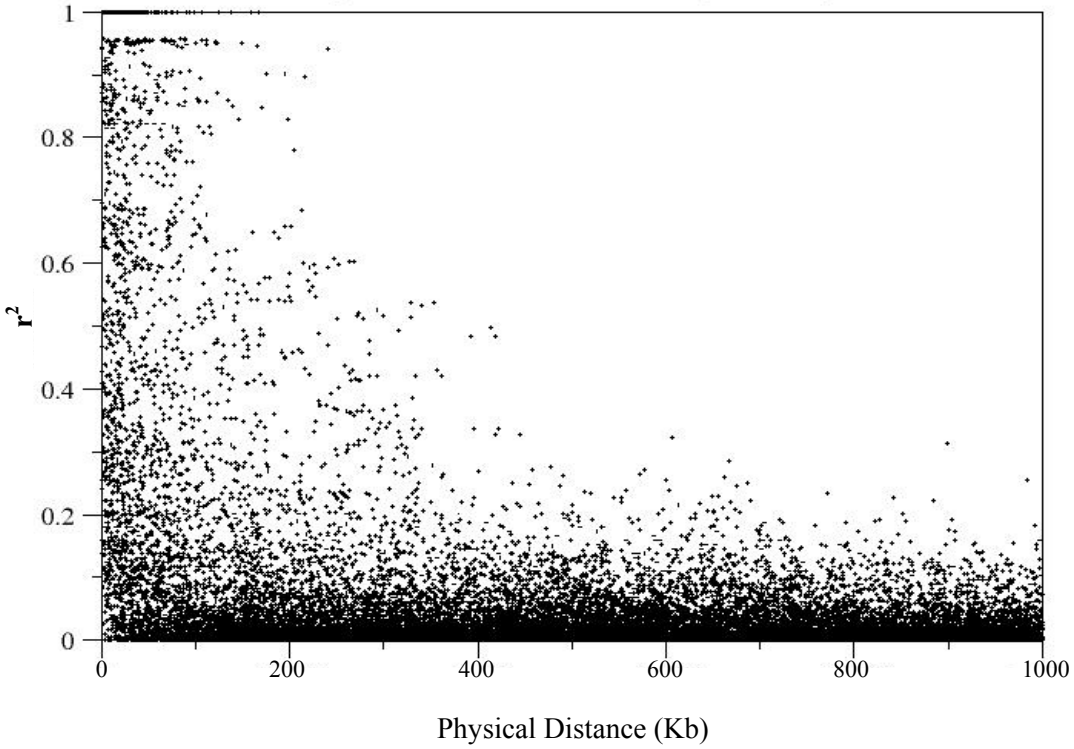
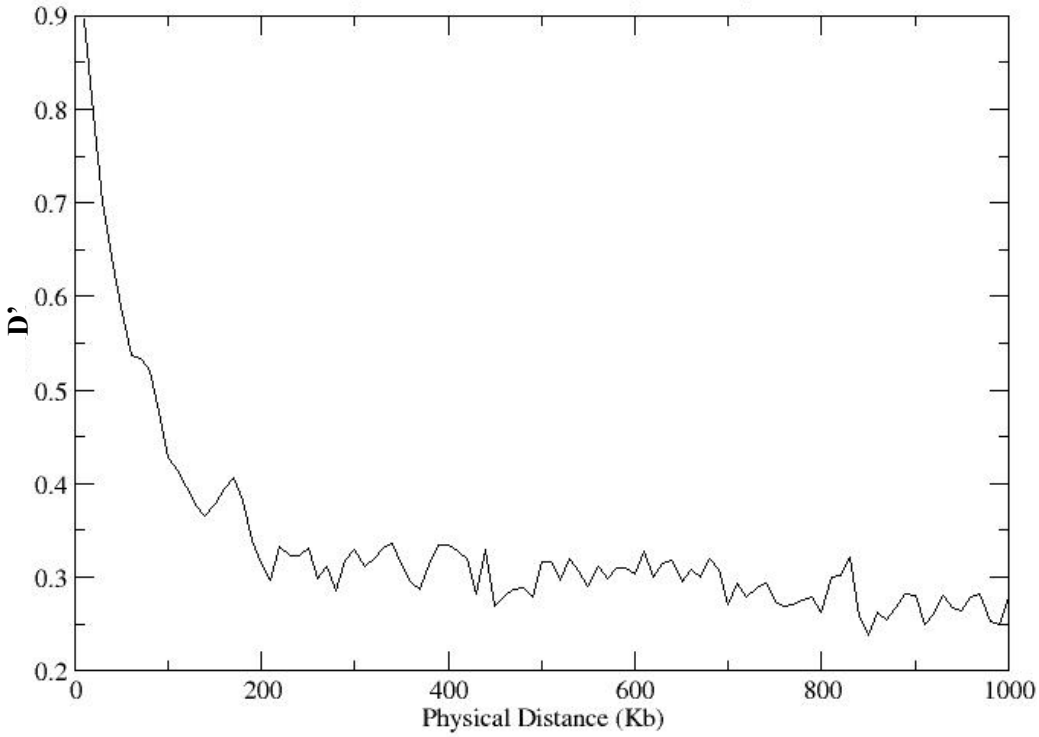
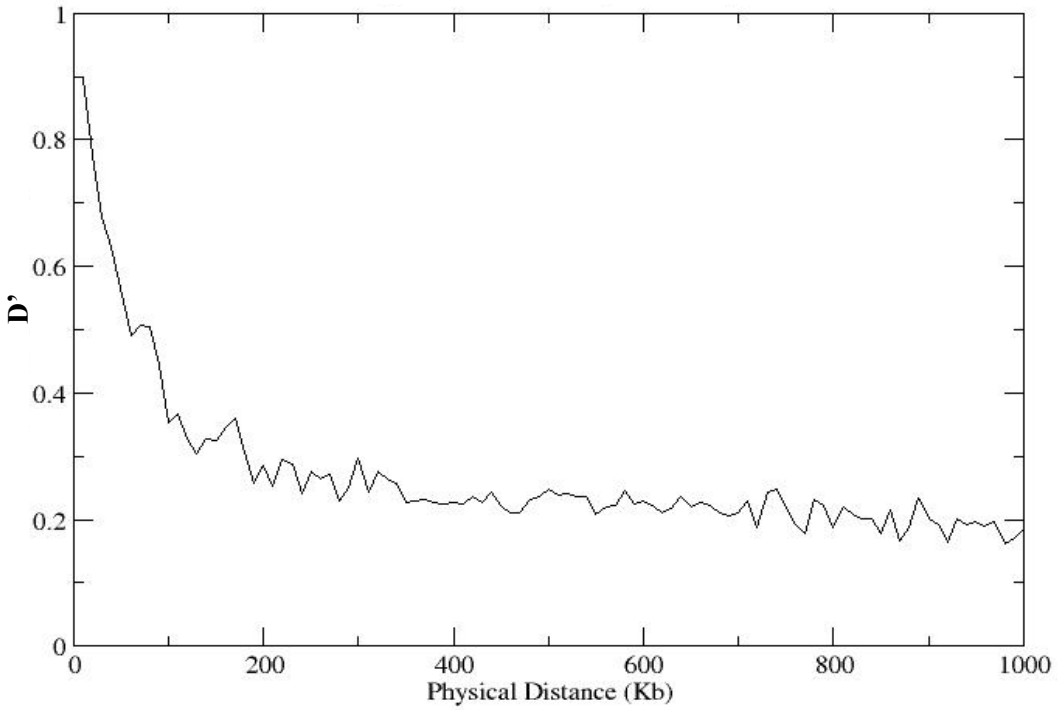



**Figure 5.15: Variability of D' (A, B) and $r^2$ (C, D) using the pairwise values for SNPs separated by □1 Mb. The corresponding D' and $r^2$ scatter plots using all polymorphic SNPs are shown in Appendix 16.**

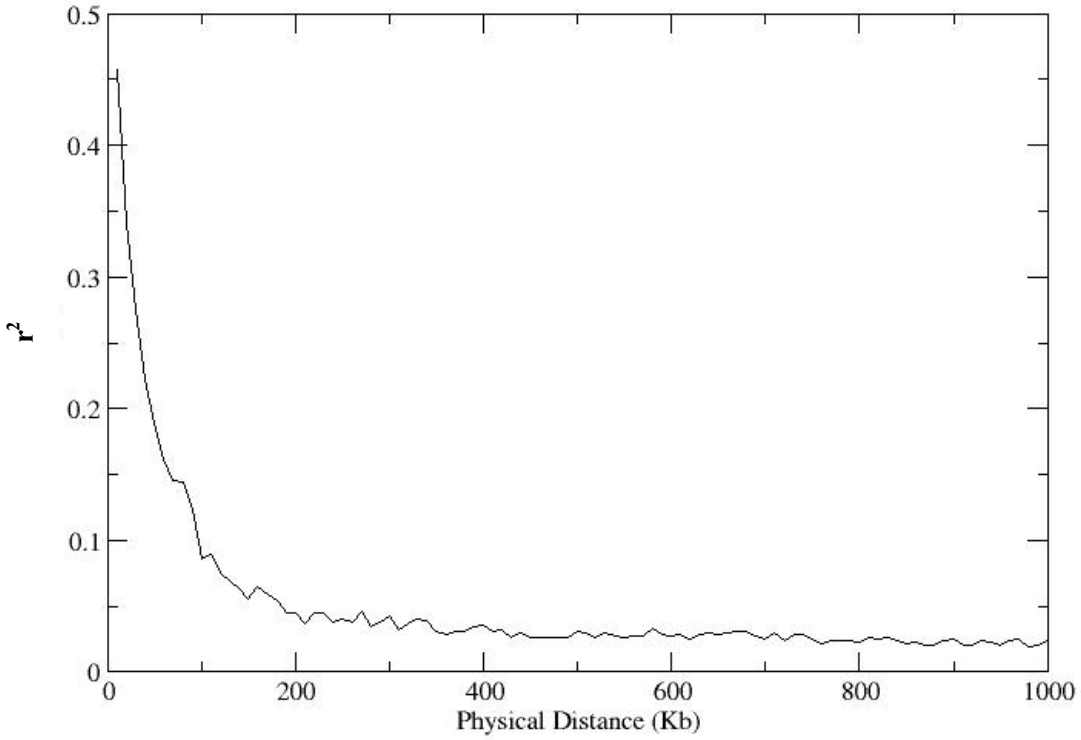
**A. D' (using SNPs with minor allele frequency >10%)**




**B. D' (using SNPs with minor allele frequency >20%)**



**C. $r^2$ (using SNPs with minor allele frequency >10%)**




**D. $r^2$ (using SNPs with minor allele frequency >20%)**

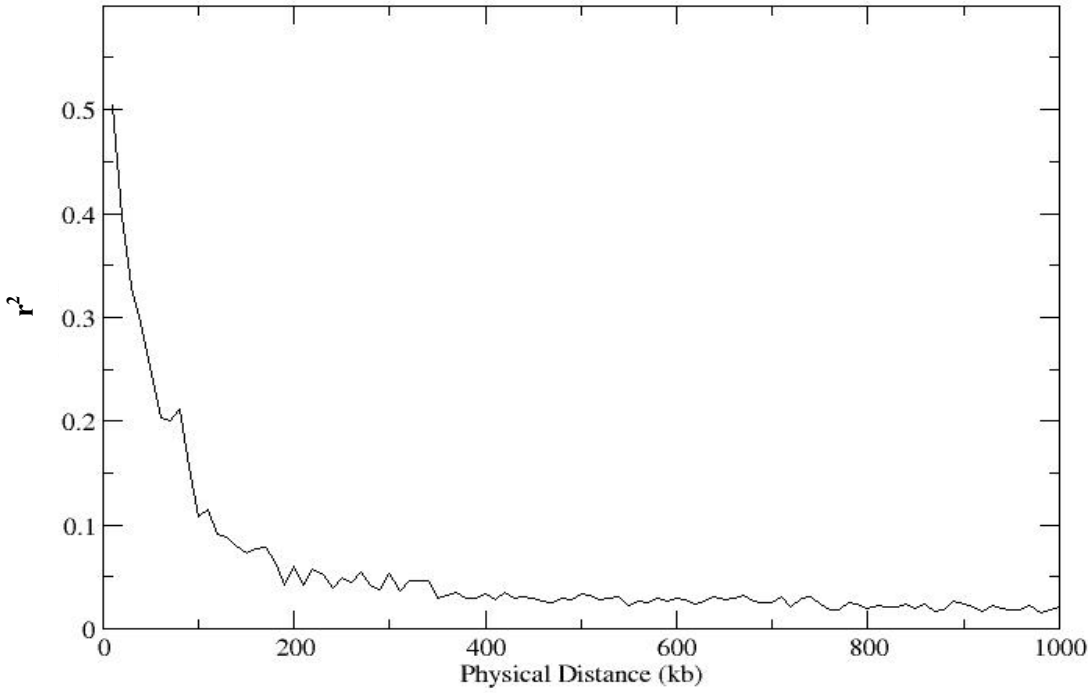



**Figure 5.16: Average D' (A, B) and $r^2$ (C, D) in 10 Kb, non-overlapping, physical distance bins (up to 1 Mb). The average D' and $r^2$ plots using all polymorphic SNPs are shown in Appendix 16.**

The pattern of LD along the 20q12-13.2 sequence was assessed by calculating the average D' and $r^2$ for polymorphic SNPs within continuous stretches of DNA (500 Kb windows, sliding by 250 Kb). As shown in Figure 5.17, the results highlight areas with high levels of LD, with peaks at 1-1.5 Mb, 3.5-4 Mb (detected only by D'), 5.5-6 Mb, 7.5-8 Mb (detected only by D') and 8.7-9.2 Mb.

There is considerable evidence that sites of recombination in humans are not randomly distributed, but are often localised into specific hotspots (Jeffreys et al., 2001). Current, low-resolution genetic maps can be used to model local recombination rates and provide additional predictors of LD beyond physical distance (Dawson *et al*., 2002). The region of 20q12-13.2 has an elevated degree of recombination, averaging 2.1 cM $Mb^{-1}$ in the Marshfield sex-averaged genetic map (Figure 5.18), compared to the genome average of approximately 1.3 cM $Mb^{-1}$. With the exception of the LD peak at ~5.5-6 Mb all other peaks are situated within regions of very low recombination (<1 cM $Mb^{-1}$). Note that in the 5.5-6 Mb area two markers are wrongly placed on the genetic map, which complicates the correlation of the two maps (Figure 5.18). In addition, areas of higher recombination frequency are associated to steep decreases in LD. For example, this is observed at 4-5 Mb and 8-8.6 Mb (recombination rate >3 cM $Mb^{-1}$). These data suggest a strong correlation between the rate of recombination and the extent of LD which is in agreement with the findings of the recent chromosome 22 study (Dawson *et al*., 2002).

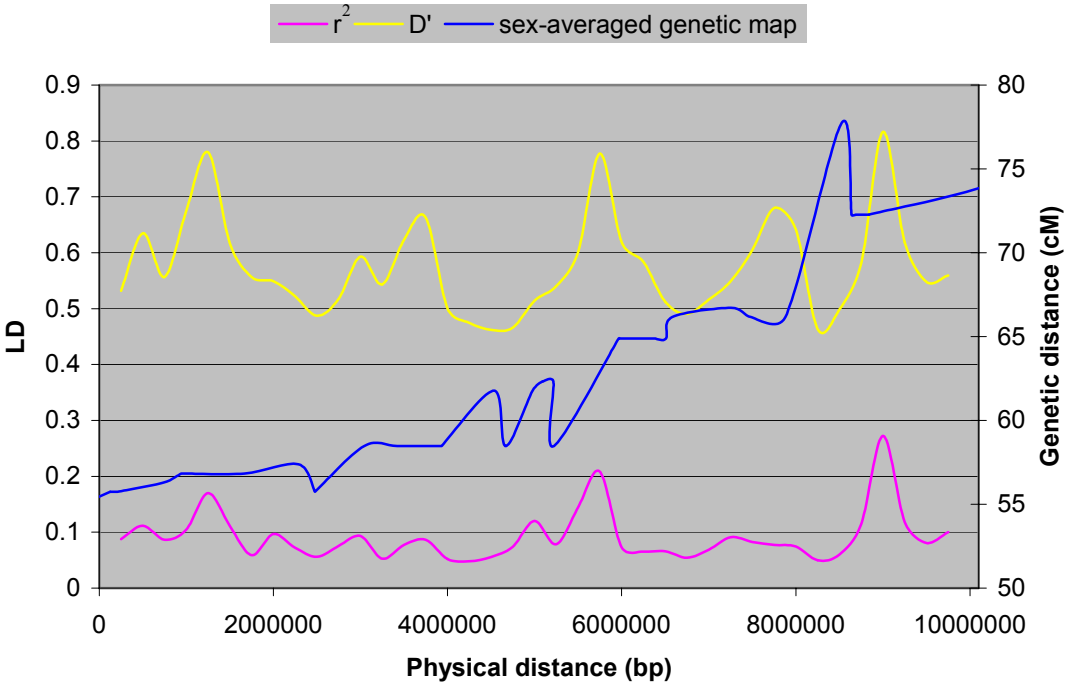



**Figure 5.17: Linkage disequilibrium across 20q12-13.2. The average D' and $r^2$ are plotted in 500 Kb windows (sliding by 250 Kb) containing all polymorphic SNPs. The sex averaged genetic map for the region is also shown (also see Figure 5.18).**

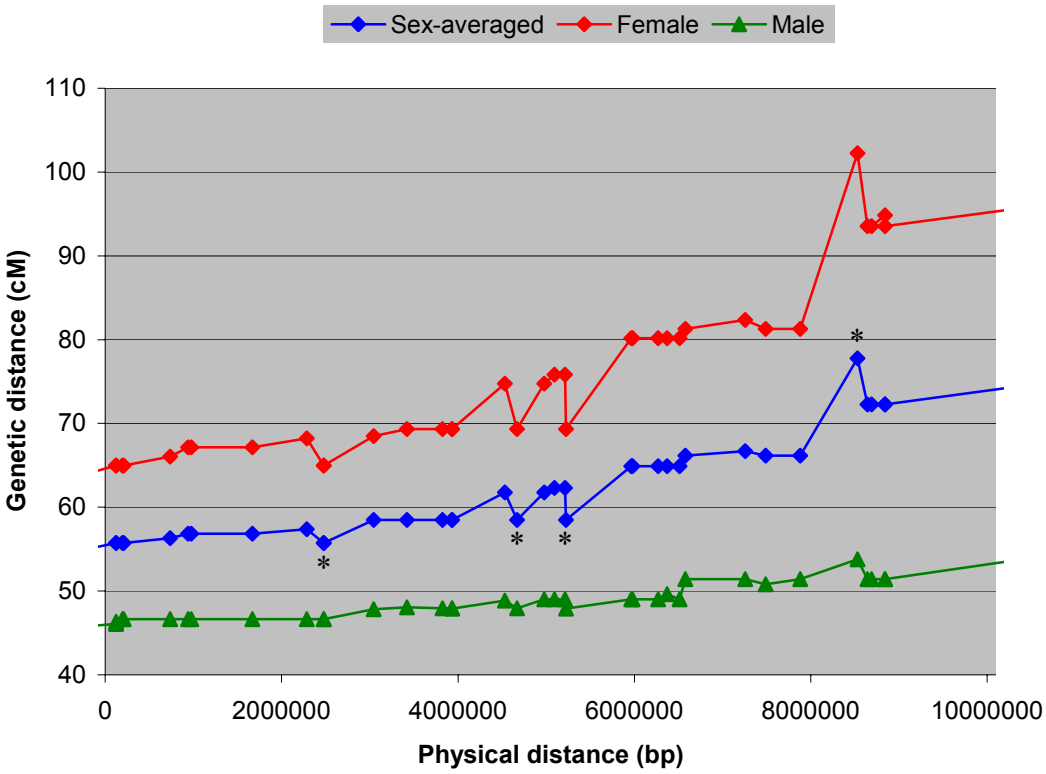
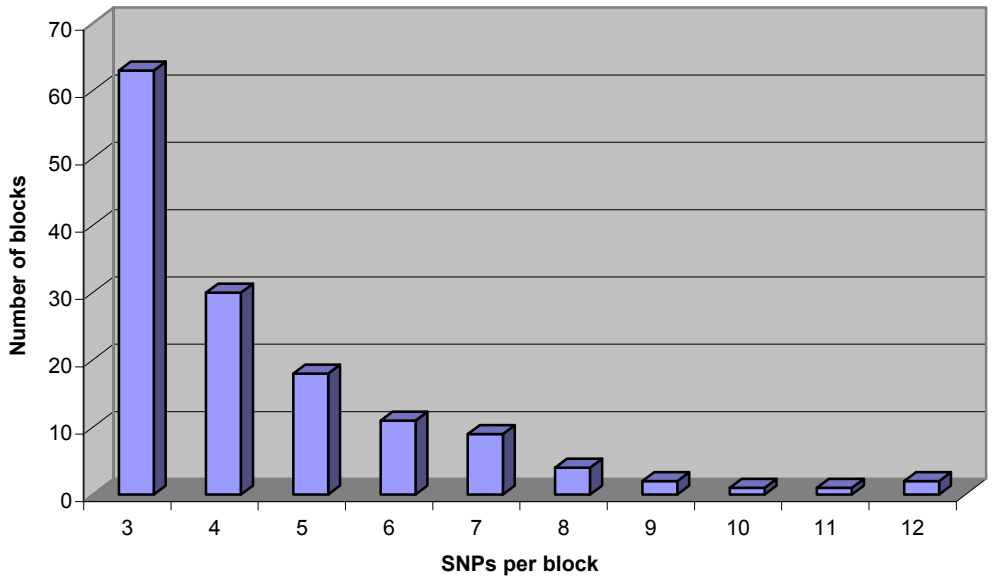



**Figure 5.18 (reproduced from Deloukas *et al*., 2001 (only the region of interest shown)): The integrated Marshfield male, female and sex-averaged genetic maps (Yu *et al*., 2001), aligned to the sequence map of the region. The four discrepant points, marked with an asterisk, correspond to markers D20S466, D20S454, D20S424 and D20S427 (Utah Center of Genome Research markers UT1688, UT275, UT654 and UT1521 respectively).**

Regions with three or more SNPs, for which all SNP pairs have D'>0.9, were identified as "LD blocks". A total of 141 such blocks were identified, which span 5.02 Mb (~50% of the region) and harbour 597 of the 879 analysed SNPs (68%). The mean block size is 35.6 Kb and the mean number of SNPs per block is 4.2 (corresponding median values 27.1 Kb and 4 respectively). The smallest and largest block identified span 2,332 and 142,291 bp respectively. As shown in Figure 5.19 A, most blocks (55%) contain four or more SNPs with two blocks having twelve SNPs each. Only 25% of the total sequence in blocks belongs to blocks with three markers (Figure 5.19 B).

**A.**




**B.**

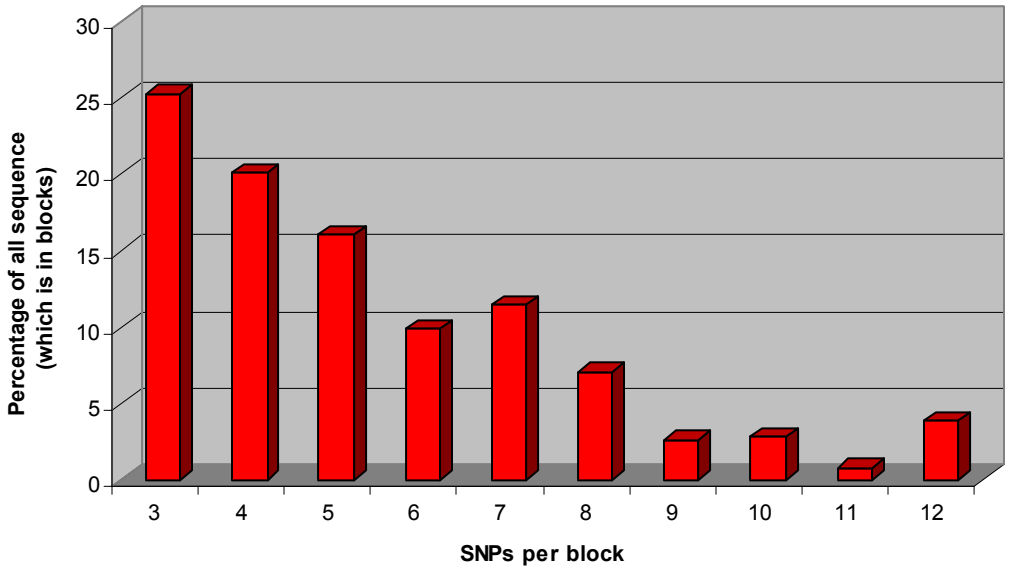
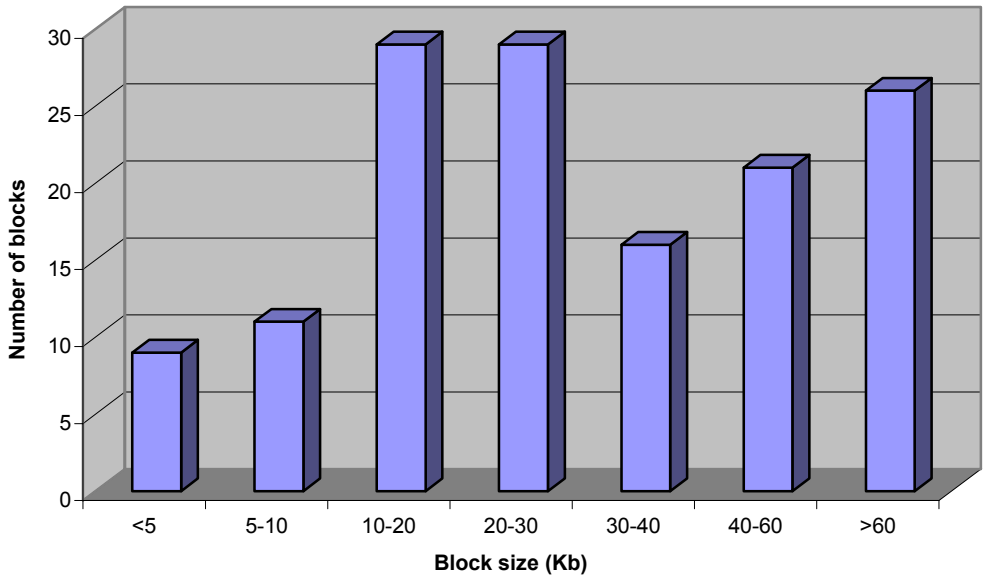



**Figure 5.19: Correlation of "LD blocks" and SNPs. (A) "LD blocks" binned according to the number of SNPs in each block. (B) Proportion of all genome sequence spanned by blocks, binned according to the number of SNPs in each block.**

Most of the LD blocks (65%) are more than 20 Kb long and 18% are more than 60 Kb long (Figure 5.20 A). Also, blocks over 40 Kb in size account for most (63%) sequence in blocks (Figure 5.20 B). The distribution of the various types of blocks is shown in

Figure 5.21 A, whereas the percentage of the total sequence in blocks across the region is shown in Figure 5.21 B.

**A.**




**B.**

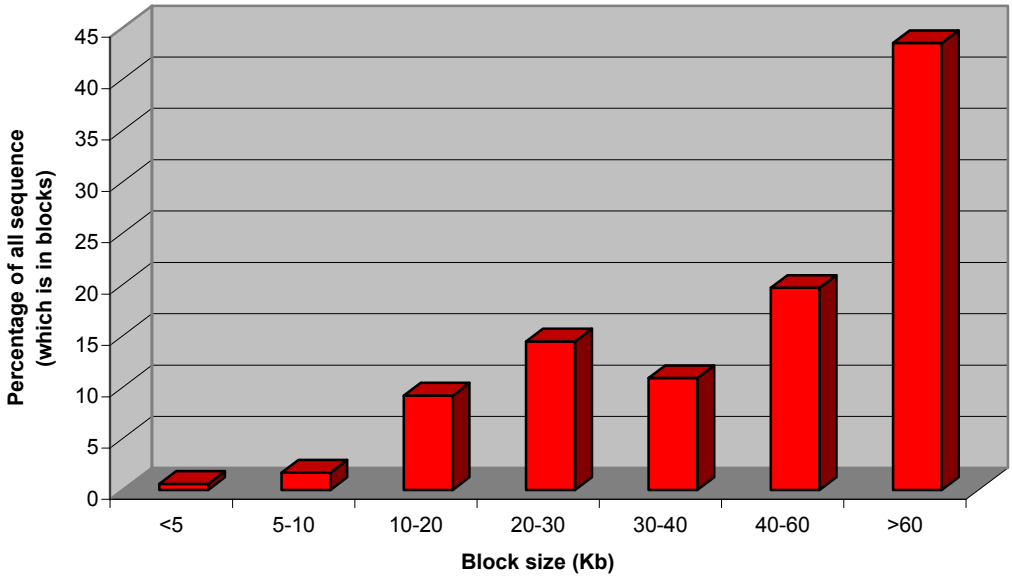
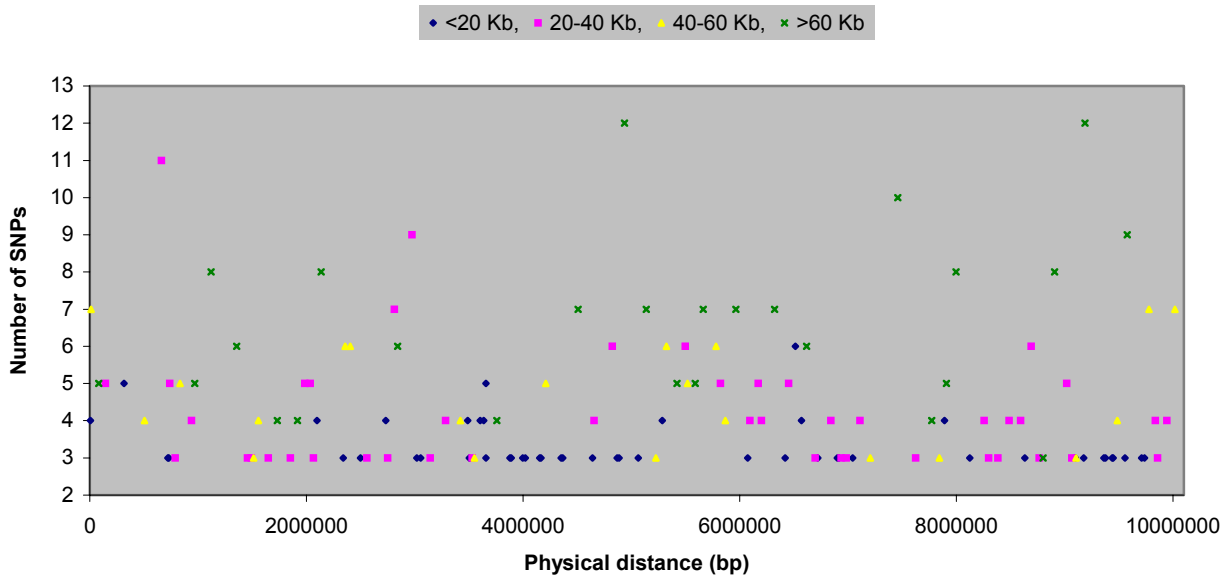



**Figure 5.20: Size correlation of "LD blocks". (A) Size distribution of "LD blocks" (B) Proportion of all sequence in blocks binned according to the size of each block.**

**A.**




**B.**

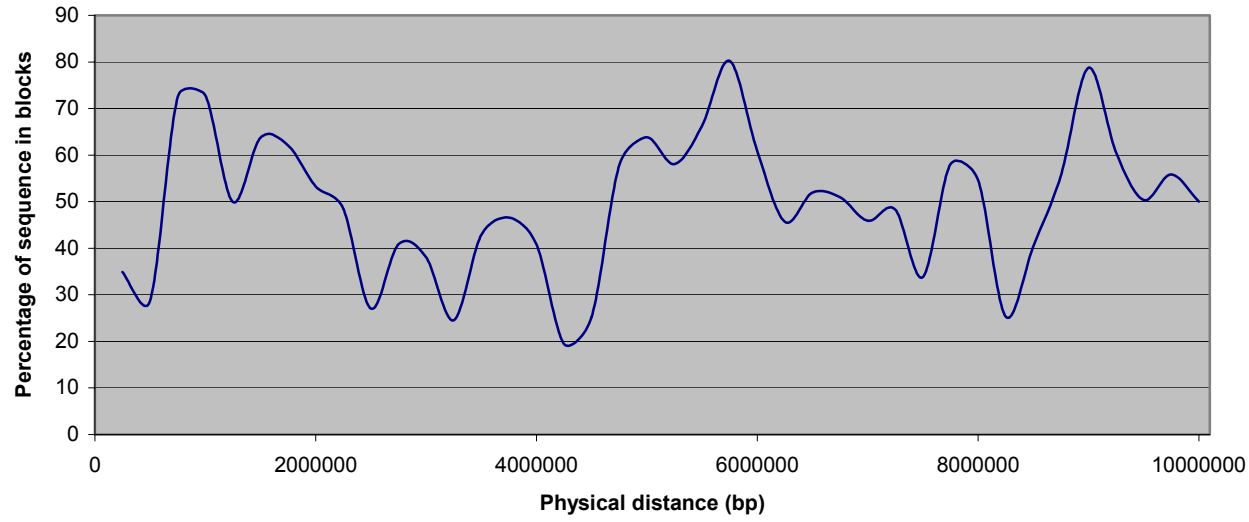



**Figure 5.21: The distribution of "LD blocks" across 20q12-13.2. (A) The position of blocks across the region. Blocks were binned in four size ranges, each represented by a different type of data point. The number of SNPs in each block is also shown. (B) The percentage of all genomic sequence in blocks across the region (calculated as 500 Kb sliding windows, overlapping by 250 Kb).**

## 5.5 Discussion

The available reference sequence of human chromosome 20 and expressed sequences were used to identify over 100 putative SNPs *in silico*, corresponding to 47 genes. This approach of SNP discovery is simple and inexpensive. In addition, it tackles exonic regions, which are more likely to harbour SNPs that are functionally important. The major disadvantage of this approach is that SNP discovery depends on the expression levels of genes: genes with lower expression levels will be represented by fewer ESTs, decreasing the chance of identifying SNPs. In addition, due to the 3' end bias of the EST databases, most identified SNPs reside within the 3' UTR of genes. Functionally important SNPs residing in non-transcribing genic regions (like promoters and introns) are also missed. Validation of the identified SNPs suggested that either there is a higher rate of false positives or many of these SNPs are rare (52% polymorphic in at least one population). Finally, the manual inspection of sequence reads and scoring of SNPs is a tedious process; thus for large-scale applications, an automated approach is required (as described for example in Deutsch *et al.*, 2001).

A set of 2,208 SNPs mapping across 20q1-13.2 were genotyped across 119 individuals from three populations, using the Sequenom MassEXTEND platform. In total, a non-redundant set of 188,307 genotype calls was obtained. Error checking using an independent platform (Illumina) suggests an error rate of 0.4%. However, 2-3% of assays are not robust and tend to inflate error rates in the raw data. Imbalanced allelic amplification is the most likely cause.

Approximately 50% of the SNPs with "complete" results were polymorphic in all three populations, whilst 12% were polymorphic in only one population. Overall, 76.5% of SNPs with "complete" results were polymorphic in at least one population.

The African American ethnic group represents a population that has very recently (in demographic terms) migrated from Africa. Whilst expected to be subject to the forces of admixture with the population in the new geographic region, studies found almost complete correlation between African and African American samples (Gabriel *et al*., 2002). For the purposes of this discussion, African Americans will be treated as Africans.

In agreement with previous studies (Frisse *et al*., 2001; Przeworski *et al*., 2000; Wall and Przeworski, 2000), this study has found higher levels of variation in African/African American populations relative to non-African populations and a skew in the frequency spectrum towards more common variants in non-African populations relative to African/African American populations. One possibility is that non-African populations experienced a phase of population size reduction, during which the rare variants were lost more quickly than the common ones. The deficit of less common variants outside Africa, combined with the fact that the non-African variation is a fraction of that found in Africa, is consistent with the "Out of Africa" model of modern human origins (Wall, 2001). This suggests that modern humans evolved in a small region in Africa 120,000-150,000 years ago, and from there they expanded and replaced existing hominid populations around the world (Stringer and Andrews, 1988).

The study shows that Caucasians and African Americans have more SNPs in common than the Asians with the African Americans. Assuming an "Out of Africa" model, this difference could imply that Asians and Caucasians may have arisen from separately

migrating populations. This is supported by the fact that the Asian and Caucasian groups share fewer polymorphisms with each other than either share with the African Americans.

Of course, even the most complex population history models that are currently available are likely to oversimplify the real history of human populations. Scenarios affecting the patterns of sequence variation could include population subdivision with a change in migration rates over time and admixture with archaic humans. Recent developments such as admixture, population growth and founder effects that have occurred in historical times are also likely to affect patterns of variation (Przeworski *et al*., 2000).

This study identified a set of 943 SNPs with minor allele frequencies of ≥5% in Caucasians. This set is already a useful resource for ongoing and future association studies for the common diseases linked to 20q12-13.2 (e.g. Graves, diabetes and obesity). Although the average distance between neighbouring SNPs is less than 11 Kb some larger gaps still remain despite repeated attempts to identify and verify polymorphic SNPs. This is due to the non-uniform distribution of public SNPs across the region. The new set of ~110,000 chromosome 20 SNPs identified recently at the Sanger Institute will allow the selection of SNPs in the remaining gaps. It is, however, clear that additional SNP discovery efforts are needed for tackling the whole genome.

The extent of LD across 20q12-13.2 was assessed in Caucasians using 879 polymorphic SNPs, for which at least 80% of the 95 possible genotypes were available (average 93%). Both the D' and $r^2$ measures were applied. The decay of average D' with distance is sensitive to the minor allele frequency cut off used. The inclusion of rare SNPs tends to

inflate D' estimates as opposed to the use of only common SNPs (MAF>20%). Little difference was seen between the MAF>10% only and MAF>20% only "half length of decay" (approximately 80 Kb in both cases). This is slightly elevated (by ~20 Kb), when compared to similar studies (Reich *et al*. (2001) and Dawson *et al*. (2002)).

The pattern of LD across the region (average D' and $r^2$ in 500 Kb sliding windows) shows clear fluctuation with areas of high and low LD. There is strong correlation between recombination rate and extent of LD; the same observation was made in the chromosome 22 study Dawson *et al*. (2002). It would be interesting to use a higher resolution genetic map like the one recently reported by deCODE (Kong *et al*., 2002) for a more detailed analysis.

DNA regions were defined as "LD blocks" if they harboured three or more polymorphic SNPs with D' values of more than 0.9 for all possible SNP pairs (stringent cut off). In total, 141 such blocks were identified, covering approximately 50% of the sequence. Sequence coverage is a minimum as blocks are likely to extend further with the addition of more SNPs. Block size varied enormously between 142.3 Kb (largest) and 2.3 Kb (smallest). As in the present study we did not include the number of common haplotypes per block (ongoing effort), we used the term "LD" instead of "haplotype" block. The long-range haplotype of each founder chromosome is available. The next steps beside the calculation of the number of common haplotypes per block are to increase the coverage of the region in blocks and extend the study to other populations.

This study contributes to the overall effort to understand the long-range organisation of LD across the human genome and provides a first generation LD map as a tool for association studies in 20q12-13.2.