THE DEVELOPMENT OF THE GENETIC MAP OF HUMAN CHROMOSOME 16 BY LINKAGE ANALYSIS.

A thesis submitted for the degree of Doctor of Philosophy
to the University of Adelaide.

by

Helen Kozman B. Sc (Hons.)

Department of Paediatrics, Women's and Children's Hospital, University of Adelaide, South Australia

August, 1994

Statement of Declaration

This thesis contains no material that has been accepted for the award of any other degree or diploma in any University and to the best of my knowledge and belief, the thesis contains no material previously published or written by another person, except where due reference is made in the text of this thesis. I consent to the thesis being made available for photocopying and loan (if applicable), if accepted for the award of the degree.

Helen Kozman

Table of Contents

List of Tables

List of Figures

List of Abbreviations

$(AC)_n$	-	an AC repeat sequence, repeated *n* times.
CEPH	-	Centre d'Étude du Polymorphisme Humain
CLN3	-	Batten disease, juvenile onset neuronal ceroid lipofuscinosis
cM	-	centimorgan
EM	-	Expectation and maximisation algorithm
ESTs	-	expressed sequence tags. These are short coding regions of the genome which can be amplified by a defined pair of primers.
FMF	-	familial Mediterranean fever
FRA16A	-	folate sensitive fragile site at 16p13.11
FRA16B	-	Distamycin A sensitive fragile site at 16q22.1
GALNS	-	N-acetylgalactosamine-6-sulfutase (mucopolysaccharidosis type IV A, Morquio A syndrome)
ISLM	-	The international system for human linkage maps
Kb	-	kilobase
Mb	-	megabase
MLE	-	Maximum Likelihood Estimate
PCR	-	Polymerase Chain Reaction
PFGE	-	pulsed field gel electrophoresis
PKD1	-	adult polycystic kidney disease
RFLP	-	Restriction Fragment Length Polymorphism
RTS	-	Rubinstein-Taybi syndrome
STRs	-	Simple Tandem Repeats. See STSs

STSs	-	Sequence tagged sites. These are short non-coding regions of the genome which can be amplified by a defined pair of primers. Coding tagged sequences are designated ESTs.
TSC2	-	gene for tuberous sclerosis on chromosome 16
UV	-	ultraviolet
VNTR	-	variable number tandem repeats
WCH	-	Women's and Children's Hospital, South Australia
YACs	-	Yeast artificial chromosome. These are vectors for cloning large DNA fragments.

Abstract

Genetic maps of entire chromosomes are constructed on the basis of recombination relationships between genes or anonymous DNA markers. Multipoint linkage analysis determines the order of loci, the likelihood support for that order, and the distance in centiMorgans between the loci. These maps enable chromosomal assignment and regional localisation of genes for familial disorders. This information is applied to genetic counselling of family members and represents the starting point for positional cloning of the disease gene.

Multipoint genetic linkage maps of human chromosome 16 were constructed in three stages. Firstly, an integrated genetic and physical map covering 99% of the chromosome was developed (Kozman *et al.*, 1992; NIH/CEPH collaborative mapping group, 1992). This was followed by the construction of a high resolution PCR based microsatellite map (Shen *et al.*, 1994). Finally the CEPH consortium map, recognised internationally as the ultimate map from the current CEPH database was constructed (Kozman *et al.*, 1994). These genetic maps provide a firm basis for the assignment and regional localisation of disease genes to this chromosome. Localised high resolution linkage maps were developed during the same period for specific regions of the chromosome immediately prior to more detailed physical mapping. Thes regions were targetted because they contained fragile sites (*FRA16B*, *FRA16A*) or disease genes (*CLN3*, *FMF*) (Kozman *et al.*, 1991; 1993; Mitchison *et al.*, 1993; 1994). *FRA16A* was subsequently isolated (Nancarrow *et al.*, 1994), and work continues on the isolation of *FRA16B*, *CLN3* and *FMF*.

A total of 191 polymorphic loci (79 PCR formatted markers, 102 classical RFLP loci and 10 VNTR markers) were incorporated within these maps of chromosome 16. Markers were genotyped on the panel of forty reference families made available by

CEPH (Centre d'Etude du Polymorphisme Humain). Multipoint linkage analysis was performed by the computer program packages LINKAGE and CRI-MAP.

The genetic and cytogenetic-based physical maps were constructed simultaneously as a part of the process of map validation and refinement of map distance.

The goal of the human genome project addressed in this thesis, was the construction of a genetic linkage map with a resolution of between 2-5 cM by the year 1995 (Collins and Galas, 1993). The CEPH consortium map, with framework loci (loci placed with odds of 1000:1 or greater) of resolution 2.8 cM, and resolution 1.6 cM with all loci included, has easily achieved this aim for chromosome 16.

Acknowledgement

I would like to thank the Department of Cytogenetics and Molecular Genetics at the WCH for providing me with the opportunity to study for the Degree of Doctor of Philosophy, during the years of 1989 to 1994. My time in the department has been very stimulating, both mentally and socially.

I would like to thank my supervisors Professor Don Roberton and Professor Grant Sutherland for their help and advice during the course of the project.

John Mulley, my day to day supervisor, has been very helpful and understanding during my entire time in the department, for which I am very grateful. Others in the laboratory who worked with me, Hilary Phillips, Agi Gedeon, Kathy Friend, Sinoula Appostolou, provided me with support, friendship and entertainment during all the trials and tribulations of the years of my study. They are wonderful friends, and I thank them. I also acknowledge others in the department for providing a good working environment.

Many of the colleagues in the department contributed to the information for the construction of this linkage map: Zhong Chen, Kathy Friend, Agi Gedeon, Kathy Holman, Sharon Lane, John Mulley, Julie Nancarrow, Hilary Phillips, Yang Shen, Jean Spence, Yu Sui, Graeme Suthers, Andrew Thompson, by isolating, characterising, physically mapping and genotyping many of the markers included in the map, as indicated in the text.

I was supported during most of this period of study by the US Department of Energy, contract number DEF0289ER6086.

Lastly, I would like to thank my family for their willing support over the years; my husband Mark Heddle for his constant love and understanding; my sister Alison Watkins, and her husband Peter, and my parents, Florence and Viktor Kozman.

Note to Publications

Most of the work performed during the course of this project has been published. The published manuscripts have been cited in the text, and the papers can be found in appendix A.

"Science can be a tool for liberation, but it can also become a tool for oppression; in order to avoid this deflection, those responsible for its progress must share their knowledge by expressing themselves in such a way as to be understood by all."

Albert Jacquard,

Philosopher

"Endangered by Science"

CHAPTER 1

LITERATURE REVIEW

Table of Contents

Associated Papers (See Appendix A):

Identification of a highly polymorphic marker within intron 7 of the *ALAS2* gene and suggestion of at least two loci for X-linked sideroblastic anemia. Cox, T.C., Kozman, H.M., Raskind, W.H., May, B.K., and Mulley, J.C. 1992. *Hum. Mol. Genet.* **8**:639-641.

A linkage map of microsatellite markers on the human X chromosome. Donnelly, A., Kozman, H.M., Gedeon, A.K., Webb, S., Lynch, M., Sutherland, G.R., Richards, R.I., and Mulley, J.C. 1994. *Genomics* **20**:363-370.

A (CA)n repeat polymorphism for the human skeletal muscle α-actinin gene ACTN2 and its localisation on the linkage map of chromosome 1. Beggs, A.H., Phillips, H.A., Kozman, H.M., Mulley, J.C., Wilton, S.D., Kunkel, L.M., and Laing, N.G. 1992. *Genomics* **13**:1314-1315.

1.1. Introduction

Like every organism, mankind owes its characteristics to the inherited blueprint, the genes, the medium through which living organisms transmit genetic information from one generation to the next. A genetic linkage map specifies the relative order of genes and marker loci along the chromosome and the distances between them, as determined by recombination. Linkage mapping of Mendelian diseases for which the gene product is unknown remains a pre-requisite for the identification of the gene responsible for the disease. Linkage to a specific chromosome is detected when a genetic marker of known location is shown to co-segregate with a disease phenotype. The location of the marker allows the position of the gene on the chromosome to be determined, leading to the eventual isolation of the gene by focusing on that region of the genome. Once the gene is isolated, the spectrum of genetic defects leading to disease pathology can be determined from the study of affected families. Thus, the evolving human genetic linkage map is an essential tool for the localisation and the eventual isolation of many human disease genes.

The linkage approach can be powerful when applied to large families; however, this strategy has been hampered in the past by the paucity and poor utility of available polymorphic loci. This limitation is being removed, as technology has progressed in the last two decades to uncover numerous highly polymorphic markers which are easily genotyped. When more than two loci are involved in any mapping problem, ordering of genetic loci belonging to a linkage group, and determination of distances between them, requires the analysis of joint segregation. This presents a computational challenge due to the mathematical complexity in modelling genetic recombination with the exponential increase in the number of possible multilocus genotypes and possible gene orders as additional multiallelic loci are studied jointly. Methods for efficient multilocus analysis were devised to meet this challenge.

The aim of this project was to develop the multipoint genetic linkage map of human chromosome 16. This linkage map, together with maps from all the other chromosomes, will represent the linkage map of the entire human genome.

In this literature review a brief history of the development of human gene mapping and linkage analysis will be presented. The evolution of genetic markers and the establishment of a panel of reference pedigrees will be reviewed and their utility to genetic mapping will be discussed. The role of human genetic linkage mapping in human genome science will be considered, with particular reference to the Human Genome Project and its contribution in furthering genetic linkage studies for localising genes of unknown function. The methodologies involved in multipoint linkage analysis will be reviewed and the status of the chromosome 16 linkage map, at the commencement of this project, will be described. The review will conclude with a description of the rationale for the research presented in the remainder of this thesis.

1.2. Historical Review

The foundations of genetic linkage were discovered by Gregor Mendel in his breeding experiments with peas, described in his famous paper in 1865. He proposed that hereditary characteristics are determined by elementary units transmitted between generations in a uniform, predictable fashion. Mendel's achievement lay dormant until the turn of the century when it was rediscovered by de Vries and Morgan, and Mendel's laws received the plaudits that they deserved. The laws formulated by Mendel are the fundamental principles of genetics. The first law states that alleles for different characters segregate in families (law of segregation), and the second law states that genes are inherited independently (law of independent assortment). The second law is true if genes are not linked; linkage was subsequently discovered and is now utilised for building genetic maps.

In 1911, Morgan suggested that genes are linked as a result of being carried on the same chromosome. The alleles of genes which were closer would be inherited together more frequently than those of genes more distant from one another. The chiasma frequency increases proportionately to the distance between loci. The first notion of a genetic map was documented when, from Morgan's postulations, Sturtevant (1913) proposed that it should be possible to map genes in linear order using the frequency of recombination as an index of relative position on the chromosome. He determined the relative position along a chromosome of five sex-linked genes in *Drosophila melanogaster*. This was the first multilocus map to be reported, and has since been proven to be remarkably accurate (Crowe, 1988). Soon after, a genome map of *D. melanogaster* was established, when Bridges and Morgan reported the second and third maps of this organism (Bridges and Morgan, 1919; 1923; Morgan *et al.*, 1925).

Progress in human gene mapping was slow, since humans cannot be experimentally manipulated, and specified genotypes cannot be constructed to test hypotheses. The scarcity of large families with known phase in the informative parent(s) was another major difficulty, as was the deficiency of informative genetic markers. These analytical problems were addressed by a number of earlier investigators, who proposed a variety of mathematical approaches to analyse the linkage relationship between two loci (a two point linkage analysis) in human pedigrees. More recently, the lack of informative markers was overcome by the discovery of highly polymorphic genetic markers.

Penrose introduced the sib-pair method of analysis in 1935. The principle of this method is that if linkage exists between two conditions, examination of large numbers of pairs of siblings with these conditions would show an excess of cases where the siblings are similar in both characters, over those cases in which the siblings are similar for one character and dissimilar for another. At the same

time, Fisher (1935), proposed the first likelihood approach in linkage analysis and developed a maximum likelihood scoring procedure in the form of *u statistics*.

Later Haldane and Smith (1947) applied Bayes Theorem to linkage analysis and developed the probability ratio test, where the ratio is the probability of a specified value of the recombination frequency, Θ given the data, divided by the probability of independent assortment ($\Theta = ½$).

In order to allow additivity of scores among families, the decimal logarithm of the probability ratio ($L(\Theta)$) was used; hence the term lod (log odds) score ($Z(\Theta)$), as introduced by Barnard in 1949.

$$\text{i.e.} \qquad Z(\Theta) = \log_{10} L(\Theta)/L(½).$$

The best estimate of Θ is when $Z(\Theta)$ is at its maximum. Because of a variety of complicating factors such as incomplete penetrance, missing data, and unequal male and female recombination values, the lod scores of modern genetics are rarely calculated manually, and computer programs have been designed to perform these arduous tasks (see Section 1.6.6).

In 1955, Morton proposed a method of two point linkage analysis involving a sequential test. This combined the advantages of Fisher's *u statistics* and the probability ratio test of Haldane and Smith, and Barnard. This test involved collecting data from separate families sequentially, until there was sufficient information to accept or reject the hypothesis of linkage. The end point of this sampling was defined in terms of the risk of falsely detecting linkage (Type I error) and the risk of falsely not detecting linkage (Type II error) (Morton, 1955; Conneally and Rivas, 1980). Morton suggested certain criteria for linkage, designed to help the researcher decide if the test of an hypothesis of linkage was significant of not. Lod scores (decimal log of the probability ratio as developed by Haldane and Smith (1947) and Barnard (1949)) were accumulated over families; if the total is 3 or greater, Morton concluded that the frequency of

recombination is significantly less than ½ (i.e. the two loci are linked); if the total is -2 or less, he concluded that the frequency of recombination is significantly greater than the value of Θ for which the lods were calculated, i.e. linkage can be ruled out at that value of Θ or less. If the total lod score lies between -2 and 3, no conclusion can be drawn, and more data needs to be accumulated.

Another method proposed for linkage analysis was the general approach of Elston and Stewart (1971) which allows the likelihood estimation for pedigrees of greater complexity, using maximum likelihood estimation and likelihood ratio test procedures. They presented a recursive method, the *Elston-Stewart algorithm*, for fast and exact calculation of pedigree likelihoods.

These above methods were applicable to two point analysis only, and linkage analysis was carried out for a disease gene versus a marker locus, as a sequence of two point analyses. The computer program LIPED was developed, implementing Morton's concepts and the algorithm of Elston and Stewart (Ott, 1974; 1991), to perform a two point linkage analysis on any number of families of arbitrary size and structure.

1.3. DNA Polymorphisms

Genetic mapping advanced dramatically when the hundred or so cell surface antigen and isozyme markers (Race and Sanger, 1968; Harris *et al.*, 1977) were supplemented by Restriction Fragment Length Polymorphisms (RFLPs), of which there are many (Botstein *et al.*, 1980). RFLPs are genetic loci defined by cloned DNA segments which can represent genes of known specificity, or represent anonymous non-coding sequences. These common polymorphisms in the DNA sequence provided sufficient amount of DNA sequence variation among individuals to allow the systematic study of linkage in humans. Botstein *et al.* proposed that using RFLPs, it would be feasible to construct a complete linkage map of the human genome. Linkage analysis was revolutionized, and human

genetics revitalised. They suggested that a genetic map of the genome could be developed with just several hundred polymorphic loci. Such a map could be used to position disease genes within the human genome.

With the discovery of RFLPs, a great interest developed to analyse the linkage relationships of three or more loci simultaneously (multipoint linkage analysis). Consequently, more efficient methods of analysis also had to be developed (Section 1.6) for estimating the linkage distances and gene orders of all markers simultaneously.

Within seven years of the discovery of these markers, sufficient RFLP loci had been identified to generate the first multipoint genetic linkage map, of the human genome, covering approximately 95% of the genome (Donis-Keller *et al.*, 1987). This increased the efficiency of the localisation of genetic loci causing diseases where there was no prior knowledge of the function of the responsible genes. Polymorphic loci must have high heterozygosity for efficient linkage mapping, and although systematic screening of the human genome has revealed that RFLP loci are abundant throughout the genome (Willard *et al.*, 1986), many of them show only moderate heterozygosity.

Another source of polymorphisms was discovered by Jeffreys *et al.*, (1985) and Nakamura *et al.*, (1987). These were regions containing tandem repeats of short DNA sequences called variable number tandem repeats (VNTRs), or mini satellites (Jeffreys *et al.*, 1985; Nakamura *et al.*, 1987). These were identified with GC-rich oligonucleotide probes derived from previously identified hypervariable regions in the genome found within the myoglobin, insulin and α and τ globin genes (Bowcock and Cavalli-Sforza, 1991). VNTR markers were good for linkage mapping since they were more polymorphic then their less heterozygous counterparts, the common RFLPs. However, it was soon discovered that they were non randomly distributed with greater concentration at the telomeric regions (Armour *et al.*, 1989; Royle *et al.*, 1988; Nakamura *et al.*, 1988). Using

a combination of RFLP and VNTR loci, the next genetic map for the entire genome was developed (NIH/CEPH Collaborative Mapping Group, 1992).

Further advances in molecular biology lead to the isolation of many microsatellite repeat elements known as simple tandem repeats (STRs). A microsatellite is a genomic sequence that consists of a mono-, di-, tri-, or tetra- nucleotide repeated in multiple tandem copies. Although microsatellites containing all nucleotide combinations have been identified, the class found most commonly in mammalian genomes contains a $(AC)_n.(TG)_n$ dinucleotide repeat, and is often referred to as an AC-repeat. The high density of AC-repeats (50,000 to 100,000 copies in the genome) and their random distribution throughout the genomes of a variety of higher eukaryotic species (they are now known to occur on average every 20-60 kb, (Bowcock and Cavalli-Sforza *et al.*, 1991; Stallings *et al.*, 1991)), was first demonstrated a decade ago (Hamada *et al.*, 1982). In 1989, Weber and May found variation in the AC-repeat number among alleles at 90% of microsatellite loci examined and developed a method for typing polymorphic microsatellites by PCR. They also proposed that in general, the longer the AC repeat, the greater the probability it is polymorphic, and the greater its heterozygosity is likely to be. Many tri- and tetra-nucleotide repeats have also been shown to be polymorphic (Economou *et al.*, 1990; Hoffman, 1994).

These highly polymorphic loci, abundant and ubiquitous throughout the genome, were thus exploited for genetic mapping by assay using PCR and DNA sequence gels (Weber and May, 1989; Litt and Luty, 1989). There is no need for cloning of probes, Southern blotting or hybridisation, and results can be obtained relatively rapidly compared with Southern analysis. Many of these PCR formatted markers now exist as primer sequences and have been incorporated into high resolution genetic maps (Wiessenbach *et al.*, 1992; Gyapay *et al.*, 1994; See Section 1.5). Not only are microsatellites useful as anonymous loci but the presence of these tags in regions adjacent to coding sequences allows the

incorporation of functional genes into genetic linkage maps (Weber, 1990). The density of AC-repeats is so great that any gene will have a microsatellite nearby that can be used as a tag or marker for that gene. For example, a dinucleotide repeat has been identified within intron 7 of the *ALAS2* gene on the X chromosome which has subsequently been incorporated into a genetic linkage map of this region (Cox *et al.*, 1992). Similarly, a polymorphic dinucleotide repeat has been isolated from within the *ACTN2* gene on chromosome 1, and was then incorporated in the linkage map of chromosome 1 (Beggs *et al.*, 1992).

Microsatellite loci have a higher frequency of mutation than RFLPs with estimates ranging from 10^{-3} to 10^{-5} per locus per gamete per generation (Kwiatkowski *et al.*, 1992). The mutation rate of the 814 microsatellite markers studied by Weissenbach *et al.* (1992) was 0.1%, a greater mutation rate than that for the mutation prone *DMD* gene (Hoffman, 1994). As a consequence, microsatellite markers have a higher level of variation than RFLPs. This mutability could generate problems in family linkage studies when mutation leads to an apparent departure from Mendelian inheritance, and particularly if an allele being followed is converted to a different allele compatible with Mendelian segregation in the family, causing misinterpretation of the linkage data.

Primer sequences used to detect STR polymorphisms represent STSs on high resolution genetic maps. These STSs can be easily placed on YACs and YAC contigs. This allows placement and orientation of YACs along the chromosome and contributes to the development of the contig map of the human genome, an essential foundation for determination of the complete DNA sequence of the human genome - the ultimate genome map.

1.4. CEPH

The availability of a collection of common reference pedigrees and a shared genotype repository has facilitated the development of genetic maps. The Centre

d'Etude de Polymorphisme Humain (CEPH), established in 1984 by Jean Dausset and Daniel Cohen to hasten construction of a primary genetic map of the human genome (Botstein et al., 1980), provides high quality genomic DNA produced from cultured lymphoblastoid cell lines derived from each member of a reference panel of forty large nuclear pedigrees (517 individuals) and a database contributed to and shared by the collaborating investigators. The object of CEPH was to maintain a genotype database that is distributed regularly, and sponsor the construction of consortium linkage maps (Dausset *et al.*, 1990). Since collaborators use the same reference panel of families, data generated by one group can be used by others to detect new linkages, and to construct reference maps. A key premise of the CEPH collaboration was that the human genetic map will be more efficiently achieved by *collaborative* research on DNA from the same set families.

The consortium maps, one for each chromosome, will be based on all genotypes determined from the CEPH families, and will represent the ultimate validation for the genotypic data in the CEPH database. Each consortium map will be prepared by a committee of collaborating investigators who have contributed genotypic data for markers localised to the particular chromosome. The consortium maps will provide the basis for localising genes that determine diseases and other genes of interest.

1.5. The Role of Human Genetic Linkage Mapping and the Human Genome Project.

Detection of linkage is the first step toward the identification of a gene for monogenic disorders if the gene product is unknown. Flanking markers for a disease gene on a genetic map can allow afflicted families to benefit from carrier detection, presymptomatic diagnosis, or prenatal diagnosis during the period between the detection of linkage and the identification of the gene. These flanking

markers can also be used for positional cloning of the gene, which is the starting point for understanding the biology of the disease and for designing therapies where this is possible. The major disease genes isolated by positional cloning include Duchenne muscular dystrophy (*DMD*) (Koenig *et al.*, 1987), cystic fibrosis (*CF*) (Dean *et al.*, 1989), fragile X syndrome (Verkerk *et al.*, 1991), myotonic dystrophy (*DM*) (Brook *et al.*, 1992), tuberous sclerosis (*TSC2*) (European Chromosome 16 Tuberous Sclerosis Consortium, 1993), Huntington's Disease (*HD*) (The Huntington's Disease Collaborative Research Group, 1993), and polycystic kidney disease (*PKD1*) (The European Polycystic Kidney Disease Consortium, 1994). Maps with greater resolution may even enable "polygenic" and "multifactorial diseases" to be mapped and resolved into genetic and environmental components in the future (Sefton and Goodfellow, 1992).

The progress of construction of human genome maps in the past two and a half decades has been spectacular. The first gene assigned to a human autosome was the Duffy blood group locus to chromosome 1 in 1968 (Donahue *et al.*, 1968). With the introduction of molecular probes exploiting DNA sequence variation, an exponential increase in the number of mapped loci and genes resulted, and by 1991, more than 10,000 loci were defined by DNA markers and 1524 cloned genes were identified (Williamson *et al.*, 1991), and genetic maps of the human genome were constructed (NIH/CEPH collaborative mapping group, 1992; Weissenbach *et al.*, 1992).

The medical, biological and technological benefits that will accrue from detailed maps of the human genome have prompted a political enterprise to "map" the human genome. The initiative has been named the "Human Genome Project". It is an international effort, designed to fulfil three main goals, the ultimate goal being to determine the DNA sequence of the human genome.

The goals are:

(i) The construction of high resolution linkage maps containing highly polymorphic loci (called simple tandem repeats, STRs; see section 1.2), with average spacing of 2.5 cM, within the first five years (from 1991 to 1995) (Collins and Galas, 1993).

(ii) The construction of a high resolution physical map of the genome containing large segments (2,000 Kb) of contiguous cloned DNA (Hoffman, 1994), and a complete STS map at a resolution of 100 Kb (Collins and Galas, 1993).

(iii) Utilisation of *(i)* and *(ii)* above for achieving the ultimate goal, the determination of the complete DNA sequence of the human genome (Collins and Galas, 1993). This will enable all genes to be located on the genome and be incorporated into the genetic and physical maps. One of the more important reasons to do this is to understand human diversity, both normal variation and that responsible for inherited diseases.

Since the Human Genome Project was initiated, international efforts have been launched to systematically develop a *high resolution genetic map* of the human genome. Twenty seven genetic linkage maps based on STRs markers have already been published (Table 1.1). Although the resolution of most of these maps has not reached the first five-year goal of the Human Genome Project, the highly polymorphic loci on these maps have already been useful for the localization, isolation and characterisation of disease genes. The maps of chromosomes 15, 16, 21, and the X chromosome have an average genetic distance of 3.5, 3.2 , 2.5 and 3.8 cM respectively. As the recently released Genethon markers (Gyapay *et al.*, 1994) are integrated into the previously published STR maps, the resolution of the other single chromosome maps will approach the 2.5 cM, 1995 goal. A prediction is that 1995 will see the publication of a map containing 5,000 - 10,000 highly polymorphic, easily characterised PCR-based markers (Lathrop, 1994), with a conservative estimated average interlocus resolution of between 0.7 cM and 0.3

Table 1.1: *Simple Tandem Repeat (STR) maps generated for specific chromosomes.*

Chromosome	Reference	Number of STRs
Chromosome 1	Engelstein *et al.*, 1993	38
Chromosome 3	Schmidt *et al.*, 1993; abst.	40
Chromosome 4	Mills *et al.*, 1992	16
Chromosome 5	Weber *et al.*, 1991	13
Chromosome 6	Wilkie *et al.*, 1993	11
Chromosome 8	Tomfohrde *et al.*, 1992	22
Chromosome 9	Wilkie *et al.*, 1992	10
	Attwood *et al.*, 1994	59
Chromosome 9q	Kwiatkowski *et al.*, 1992	14
Chromosome 9q34	Henske *et al.*, 1993	16
Chromosome 10	Decker *et al.*, 1992	10
Chromosome 11	Litt *et al.*, 1993	25
Chromosome 12	Dawson *et al.*, 1993	22
Chromosome 13	Petrukhin *et al.*, 1993	21
Chromosome 13q	Bowcock *et al.*, 1993a	12
Chromosome 14	Wang and Weber, 1992	9
Chromosome 15	Beckman *et al.*, 1993	55
Chromosome 16	Shen *et al.*, 1994	81
Chromosome 17	O`Connell *et al.*, 1993	4
Chromosome 18	Straub *et al.*, 1993	14
Chromosome 20	Hazan *et al.*, 1992	26
Chromosome 21	McInnis *et al.*, 1993	43
Chromosome 22	Buetow *et al.*, 1993	24
	Vallada *et al.*, 1994	14
X Chromosome	Donnelly *et al.*, 1994	62
Whole genome	Weissenbach *et al.*, 1992	813
	Gyapay *et al.*, 1994	2066

cM (assuming a genetic map length of the human genome of approximately 3,300 cM (White *et al.*, 1985)). This will be an important resource for the study of genetic diseases, and a significant advance for the human genome program. These high resolution genetic maps will enable disease genes of unknown function to be easily mapped to regions of the genome, initially for application to genetic counselling, and ultimately for gene identification by positional cloning.

1.6. Methods of Multipoint Linkage Analysis

1.6.1. General introduction: idea and theory

Over the last century, map making has been transformed from a relatively arcane pursuit (Sturtevant, 1913) to an intensive, highly collaborative process. This achievement has been enhanced by the assembly of the panel of CEPH reference pedigrees, the organisation of an international collaboration, and by the propagation of many new highly polymorphic STR markers. The development of statistical methods for linkage calculations, embodied in computer programs (Lathrop *et al.*, 1985; Lander and Green, 1987; Green, 1992; Lander *et al.*, 1987), and the availability of more sophisticated, high speed work-stations for computation has further enhanced the assembly of whole chromosome, high resolution genetic maps containing many markers (White *et al.*, 1990; Dracopoli *et al.*, 1991; Spurr *et al.*, 1992; Bowcock *et al.*, 1993a; 1993b; Attwood *et al.*, 1994; Table 1.1).

Linkage analysis originated for application to pairs of unassigned loci, as a sequence of pairwise (two-point) comparisons between a trait locus and each of a number of marker loci. For each comparison, the trait locus versus the i^{th}

marker, or marker versus marker, lod scores[1] are computed and combined over families. Dramatic increase in the availability of informative genetic markers stimulated the search for more efficient methods to construct a human genetic map. The method of multipoint linkage analysis was devised. Early contributions to the development of the technique of multipoint linkage analysis of three of more loci include those of Renwick and Bolling (1971), Cook *et al.*, (1974), and Meyers *et al.* (1975, 1976). In contrast to the methods of traditional two-point linkage analysis, multipoint linkage analysis involves the simultaneous estimation of linkage between three or more loci, making a considerable difference to the efficiency and precision of estimates of the recombination frequency, and to the estimates of support for gene order (White, 1985). The consideration of multiple loci concurrently is more efficient than the traditional two-point analysis (Lathrop *et al.*, 1984), or even three-point analysis, in estimating the recombination fraction, since different branches within families will be informative for different subsets of loci, allowing the accumulation of information on genetic locations from all the available data. As shown by Thompson (1984), "three point data provides more information, particularly with regard to the problem of ordering the loci"; where individuals informative for all three loci are observed, information is lost by summarising the data in pair wise-form (Lathrop, 1994). Lathrop *et al.* (1985) calculated the relative efficiency of 3-locus versus 2-locus estimates of recombination fractions and found, making no assumptions on interference, that three point data can be over five times more efficient, thus requiring less than 1/5 the number of observations than two point for the same precision. The

---

1. The most familiar lod is

$$Z(\Theta) = \log_{10} \frac{L(\Theta)}{L(\tfrac{1}{2})},$$

where $L(\Theta)$ denotes the probability of the data on a given pair of loci when the recombination value is Θ (Morton, 1955). Then the maximum value of $Z(\Theta)$, say $Z(\hat{\Theta})$ gives the maximum likelihood estimate $\hat{\Theta}$. Linkage is detected with statistical significance when $Z > 3$, $Z < -2$ suggests no linkage and a value between -2 and 3 is suggestive of linkage.

examination of multiple loci simultaneously also provides powerful linkage evidence for locus order by reason of the drastically reduced likelihood of close multiple exchanges providing good evidence for the elimination of most of the possible locus orders (White, 1985).

The process of multipoint linkage analysis involves three stages:

(i) the detection of linkage groups,

(ii) inferences regarding the order of the loci along the chromosome, and,

(iii) estimation of the genetic map distances, based on recombination frequency, between the loci.

If there are n loci there are potentially $n!/2$ distinct orders to be considered. However, the biologically natural choice among these orders is that order which satisfies the maxim that recombination between flanking markers is greater than or equal to that between adjacent loci. Thus, the choice is the ordering of the loci where the frequency of the higher order recombinants are all less than the frequency of the lower order recombinants, and so eliminating many of the $n!/2$ possible orders. It can be seen that there is a necessity to develop more sophisticated techniques to perform linkage analyses, and develop computer programs to automate the task.

The statistical property of *likelihood* is used to determine the locus order and the recombination fractions for a set of loci. This involves the methods of *likelihood ratios* and *maximum likelihood estimation* (Fisher, 1935; Haldane and Smith, 1947; Morton, 1955; Ott, 1991), and utilises mathematical *map functions* such as those proposed by Haldane (1919) and Kosambi (1944). Computer programs have been written to perform multipoint linkage analysis, combining relevant information from all families to determine the most likely order of loci and to evaluate the maximum likelihood estimate of the recombination fraction between the loci (Lathrop and Lalouel, 1988). There are other types of statistical methods

used for linkage analysis, for example sib-pair analysis, but these are not relevant to this thesis.

1.6.2. Likelihood, likelihood ratios, and maximum likelihood estimates

Hypotheses are verified on the basis of observations. The statistical quantity that seems most suited to serve as a measure for testing a particular hypothesis based on the observations is the *likelihood* (Fisher, 1970; Edwards, 1984). The likelihood, L, for a hypothesis H given a set of observations F is defined as the probability,

$$L(H) = P(F;H),$$

with which the observations have occurred. The odds in favour of one hypothesis (H_0) over another (H_1) are expressed as a *likelihood ratio*,

$$R = L(H_0)/L(H_1).$$

The two hypotheses being considered most commonly in a linkage analysis are an hypothesis of free recombination (H_0), and an hypothesis of linkage (H_1). They represent two distinct observations, defined through the value of the recombination fraction, free recombination corresponding to $\Theta = ½$ and linkage corresponding to $\Theta < ½$. That is, a set of loci are linked as compared to the alternative of not linked. Typically in a linkage analysis, one wants to determine either the order and recombination frequencies between a set of loci, or place an additional locus on a predefined map. Once the most likely order is determined using the likelihood ratio statistic, the recombination frequencies are estimated using the method of maximum likelihood estimation.

To determine the order of a set of loci and estimate the recombination fractions between them, the probability that a specific order would have given rise to the observed data, is computed. This probability is called the "*likelihood of the map*".

Consider a situation where there is a cluster of loci whose order is not known; that is, the hypotheses of interest may be different locus orders. The best order among the possible orders, given a set of loci, is determined by calculating the likelihood for each order, and by comparing the maximum likelihood order to that of alternative orders - that is, the likelihood ratio statistic, or odds ratio. Clearly, different locus orders generally lead to different recombination fraction estimates between loci. If L_1 and L_2 are two maximum likelihoods under two different orders, then the likelihood ratio, or the *odds*, of order 1 versus order 2 is given by;

$$R_1 = L_1/L_2,$$
$$\text{or } R_1{:}1.$$

Large odds (odds greater than 1000:1, as discussed in Section 1.6.3) is firm evidence for the rejection of a given alternative order, and increases our confidence in the inferred map. Thus, the "best" map is the map with the highest likelihood, and the ratio of the likelihoods between two maps provides a simple measure of how much "better" one fits the data than the other. Interpretation of the statistic, however, is hampered by the lack of distribution theory that could lead to the calculation of significance levels. The standard theory of the asymptotic distribution of the likelihood ratio statistic does not apply to the comparison of orders, because the parameter spaces are often multimodal, and may have the same dimension under the different alternatives (Lathrop *et al.*, 1987).

The hypotheses considered above (H_0 and H_1) contain parameters, the recombination frequencies, whose values are unknown and need to be investigated. The method of maximum likelihood estimation is a standard statistical procedure where the maximum likelihood estimate (MLE) of Θ, $\hat{\Theta}$, (the recombination fractions of the hypothesis, H_1), of the likelihood function $L(\Theta)$, are determined;

the values chosen for the recombination rates are those that maximise the maximum likelihood function, $\hat{L}(\Theta)$ (Lathrop, 1994). The method assumes a fixed order for the marker loci. Maximisation is performed by taking the first derivative, $dL/d\Theta$, of $L(\Theta)$, with respect to Θ, and setting it equal to zero. When the likelihood equation, $dL/d\Theta = 0$ is solved for Θ, it results in the MLE, $\hat{\Theta}$. As a general rule, in complex multipoint analyses, the likelihood cannot be maximised analytically, and will generally be solved numerically by varying the values of the parameters of interest until an approximate maximum is found.

Other situations are ofttern encounted in linkage problems (Keats *et al.*, 1989), and are outlined below. Suppose there exists a sequence of n loci of known order 1,2,3,...,n, where the distance between locus i and locus j is given by Θ_{ij}. The multipoint likelihood function for this map of loci is given by:

$$L(\Theta_{12}, \Theta_{23}, \Theta_{34},\ldots,\Theta_{(n-1)n}).$$

Supposing the mutual distance, the MLEs, between each of these loci is to be established, then

$$R_2 = \frac{L(\Theta_{12}, \Theta_{23}, \Theta_{34},\ldots,\Theta_{(n-1)n})}{L(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2},\ldots, \tfrac{1}{2})}$$

is the odds ratio, providing a measure of support for linkage (H_1) at a distance of Θ_{ij}, as opposed to no linkage (H_0) ($\Theta=\frac{1}{2}$).

Alternatively, it may be necessary to determine the position of a locus, relative to a map of loci, and estimate its distance, x, in that position. This is performed by placing the additional locus in various positions on the map formed by the other loci, estimating the MLE, $\hat{\Theta}$, for each position, and comparing the likelihood for each. That is:

$$R_3 = \frac{L(x, \Theta_{23}, \Theta_{34}, .., \Theta_{(n-1)n}).}{L(\frac{1}{2}, \Theta_{23}, \Theta_{34}, , \Theta_{(n-1)n})}$$

In this instance, R_3 is the likelihood odds ratio of the locus being on the map with fixed distance x, against the locus not being on the map at all.

Taking the decimal logarithms of R_2 and R_3 leads to the *multipoint lod scores* corresponding to the above situations. That is;

$$Z_2 = \log_{10}(R_2), \qquad \text{and} \qquad Z_3 = \log_{10}(R_3).$$

Z_3 is analogous to the situation of ordinary 2-point lod scores, as one parameter (map distance) is required to be estimated (Cook, 1974; Keats *et al.*, 1989). The critical values of these multipoint lod scores for autosomal analyses are as proposed by Morton (1955), for two point analysis, where:

$$-2 < Z_i < 3.$$

If the localization of a test locus relative to a map of loci is considered, another transformation of the likelihood ratio, R_3, called the *location score* (Lathrop *et al.*, 1984), is often performed. That is:

$$S(x) = 2\ln(R_3),$$

or a simple multiple of the lod score above,

$$S(x) = 2\ln(10)Z_3 = 4.6Z_3.$$

The location score provides a measure of the relative likelihood that the test locus is located at a specified point on the genetic map, versus the relative likelihood of that locus not being located on the predefined map. If the value of the location score is at its maximum, it represents the most likely location of the test locus.

Asymptotically, the random variable, S(x), follows a chi-square distribution with 1 degree of freedom for large samples (Ott, 1991).

1.6.3. Criteria for linkage maps

In order to achieve the goal of the Human Genome Project (Section 1.5, *(i)* the construction of a linkage map using highly informative STR markers, with resolution of 2 - 5 cM), a set of criteria for genetic maps and markers were established. Markers that meet the specific criteria need to be found throughout the genome at regular intervals and incorporated onto the genetic and physical maps (Skolnick, 1991). Linkage maps are defined by locus order, interlocus distances and statistical support for marker location(s). The most important criterion for maps is that genetic markers be unambiguously ordered. Unambiguous order may be established by either linkage or physical methods. When a multipoint linkage map is reported, genetic distances between loci are generally given in terms of map distances (cM) and the recombination fraction estimates (Θ) are not given (Keats *et al.*, 1989).

The International System for Human Linkage Maps, (ISLM, 1990) (Keats *et al.*, 1991) formulated four measures of statistical support for marker locations, each defined by $\log_{10} (L_1/L_2)$, where L_1 and L_2 represent the maximum likelihood functions of each of the situations indicated below for each specified type of support described below, depending on the hypotheses considered. In each instance, significant support for an hypothesis is reached if:

$$\log_{10} (L_1/L_2) > 3.$$

That is, the support for various positions, x, of a single locus, relative to a map of loci can be expressed as a map-specific multipoint lod score. This corresponds to a likelihood odds ratio (for example R_1, R_2, and R_3 described above) of at least 1000:1 or greater.

1. Global Support is a measure of support for a locus belonging to a linkage group. In this instance, L_1 is taken as the maximum likelihood when the locus is present on the map, and L_2 is the maximum likelihood when the locus is not on the map. "Definite" linkage of the locus to the map of linkage loci is said to be established when global support is 3 or greater.

2. Interval support is a measure of support that a locus is in a specified interval on the map, where L_1 is the maximum likelihood with the locus in that interval and L_2 is the maximum likelihood with the locus in any other interval. If interval support is greater than or equal to 3, then the locus is more likely to be in that corresponding interval over any other interval on the map. Loci with interval support of 3 or greater are termed *framework loci* and a map of framework loci is a *framework map* (Keats *et al.*, 1991). Framework maps are of low resolution with high support for the order of framework loci, each uniquely placed with a map specific multipoint lod score support of 3 or greater (odds 1000:1 in favour of that specific location over next best localization). Conversely, *comprehensive maps* are of high resolution with low local support for the order of loci as they consist of all syntenic markers, the *comprehensive loci*. These are placed in their most likely location regardless of the level of support. Desirable framework loci are those that are evenly spaced, highly informative, easily assayed, and with well established physical locations. A map of highly polymorphic, well spaced framework markers provides the basis for the assignment of genes to chromosomes whereas dense maps of comprehensive markers are useful for precise regional localization of additional loci, especially disease genes.

3. Support for order. In this situation, L_1 is the maximum likelihood for the best supported order and L_2 is maximum likelihood for any other order. As for the above, an order is said to be significantly favoured over alternative orders if the support for order is 3 or greater (odds of 1000:1 in favour of one order compared with alternative orders).

4. Generalized Lod Score. A fourth form of multilocus support is Generalized support or generalized lod score, i.e.

$$\log (L_1/L_0) = \log \frac{L(\Theta_{12}, \ldots,\Theta_{n-1,n})}{L(\frac{1}{2}, \ldots\frac{1}{2})},$$

where Θ_{ij} is the recombination fraction between loci *i* and *j*. Generalized support provides evidence that a set of loci form a linkage group, where L_1 is obtained as the maximum likelihood under the best supported order and L_2 is the likelihood obtained by assuming all loci are unlinked. This is the lod score, Z_2, from Section 1.6.2 above.

1.6.4. Interference and map functions

Multilocus linkage analysis involves the construction of a linear genetic map from recombination data, and requires assumptions about the mathematical relationships between map distance, expressed in Morgans (units of cross-over), and recombination frequency, thus defining a mapping function.

If multiple cross-overs between two loci can be excluded, that is, the model of complete interference[2] over small distances is assumed, then the relationship between the recombination fraction (Θ) and the map distance (x) is linear (also known as Morgan's map function; Morgan, 1928).

$$\Theta = x$$

This relationship also holds if the loci are close, and the simplest method of determining map distances among the loci is to estimate the recombination fraction in each interval of adjacent loci, and the map distance between two more distant

---

2. Interference is the frequent observation that recombination events do not occur independently of each other on chromosomes. Specifically, the occurrance of one chiasma reduces the chance that a second chiasma will form nearby, so that the second chiasma occurs at a lower frequency than could be attributable to chance alone.

loci can be obtained as the sum of map distances in the intervals between these loci (Sturtevant, 1913). However, as the distance between loci increases, the linear relationship between the map distance and recombination fraction can no longer be assumed, and a mapping function is required.

Two mapping functions will be considered in this thesis. The first mapping function was derived by Haldane in 1919, based on the assumption of no interference. He expressed the distance between two loci as:

$$x = -\tfrac{1}{2}\,[\ln(1-2\theta)], \text{ for } 0 < \theta < \tfrac{1}{2},$$

$$\text{or} \qquad \theta = \tfrac{1}{2}\,(1-e^{-2x})$$

In 1944, Kosambi derived a mapping function, the one generally used for human genetic studies (Conneally and Rivas, 1980), based on assumptions of marginal interference:

$$x = \tfrac{1}{4}\,\ln[(1+2\theta)/(1-2\theta)]$$

$$\text{or} \qquad \theta = \tfrac{1}{2}\,(e^{4x}-1)/(e^{4x}+1).$$

Interference is known to occur in humans (Weeks *et al.*, 1993), and there has been a continuing debate regarding what level of interference should be incorporated into multipoint linkage analyses. The assumption of no interference is made for multipoint linkage analysis because it greatly simplifies the development and application of multipoint analysis programs and simplifies the computation of the probability of transmitting a gamete from parent to child. The fact that the degree of interference can vary according to chromosomal location (White and Lalouel, 1986) represents a further complication. If the markers in the map are ordered with certainty, and the distances between adjacent loci are less than 10 cM, then the assumption of no interference will have little effect on the estimation of map distance, and omission of interference increases the efficiency of analysis.

However, if the markers in the map are not ordered, or not ordered uniquely, then ignoring interference may reduce the power to resolve the correct order (Weeks *et al.*, 1993).

When interference does exist between loci, and the recombination is estimated assuming the absence of interference, these estimates are biased. The bias is well within the error range of recombination estimates for the relatively small sample sizes currently used in human studies (Lathrop *et al.*, 1985; White and Lalouel, 1986). As more loci are mapped, the genetic distance between adjacent loci will get smaller, and the bias of not incorporating interference will be reduced, further supporting the validity of the assumption of excluding interference in multipoint linkage analyses.

The effects of ignoring interference on estimates of recombination fraction were studied in greater detail by Bishop and Thompson (1988). These authors showed that, for gametes from phase-known triple heterozygotes, no bias in the recombination fraction for adjacent loci results from the assumption of the lack of interference even if interference is present, although for samples from pairwise informative data, there is bias introduced. Consequently, when data are analysed jointly at multiple loci, the effect of disregarding interference in estimates of Θ can be expected to be negligible (Ott, 1991).

Morton *et al.* (1986) provide a contrary view that maps constructed on the basis of no interference may lead to the exaggeration of the effects of multiple recombination, resulting in the inflation of map distances. As a consequence, the error in genetic risk estimation when flanking markers are used in genetic counselling is exaggerated, the power to distinguish alternative orders is reduced, the likelihood of the multipoint data is falsified, and the genetic map is spuriously nonlinear (Morton *et al.*, 1986; Pascoe and Morton, 1987; Morton, 1988). Morton (1988) concluded that for efficient use of multipoint data to determine locus order and map distances, it would be necessary to consider the occurrence of

interference between chiasmata. When Haldane (1919) introduced the assumption of no interference, it was generally understood that the fit to genetic data was poor. However, no multipoint mapping program incorporating the phenomenon of interference had been devised, and in the absence of such a program these biased estimates would suffice for estimating genetic locations (using Haldane's map function). Morton (1988) states that current maps should be treated as provisional, with a strong likelihood of errors in recombination estimates and locus order. While many mistakes will be resolved by high resolution mapping, the effect of technical errors must increase with resolution. Under an interference model the impact of technical errors is greatest for classes with the smallest expectations and so increases with the number of loci in multipoint mapping.

Weeks *et al.* (1993) have developed a computer program that calculates multipoint likelihoods of the genotypes of the CEPH families while taking interference into account, in which interference is modelled by using a map function to convert genetic distances into recombination fractions. They use a modified form of the LINKAGE analysis package, but still view their results as tentative.

1.6.5. Effect of genotype errors on map construction

Another consideration important to multipoint linkage mapping is the undetected errors in marker typing. Genotypes obtained from marker typing are checked for consistency with Mendelian inheritance. While many typing errors, such as those due to sample mix-ups, clerical mistakes when the genotypes are transferred to the database, or mutations within the lymphoblastoid cell lines (Banchs *et al.*, 1994; Weber and Wong, 1993), will show up when scrutinised in this manner, there will be a proportion of errors undetected as they do not lead to inconsistencies.

Undetected genotype errors in the data used to construct genetic maps need to be eliminated; they are one of the principal difficulties in attempting to resolve locus order in a genetic map, as they have three consequences:

(i) reduction of the maximum lod score achievable, and thus reduction in the power to discriminate different orders,

(ii) over-estimation of the recombination frequency between loci, thus inflating the overall map length, and

(iii) support of erroneous orders over correct orders (Ott, 1977; Buetow, 1991; Morton, 1991a; Lasher *et al.*, 1991).

Genotype errors in the database become evident primarily from the detection of apparent double recombinants within short intervals. A small proportion of these may in fact be real recombination events, since at least one is required to be observed to determine the order of loci. Since this is biologically unlikely, these observations flag potential errors, which should be checked. A 1% error rate in a map of markers 1 cM apart can generate false map orders in 10% of simulated maps and result in a three-fold inflation of the total map distance (Skolnick, 1991), the effect being strongest when the markers are more tightly linked (Haines, 1992). Ott (1977) showed that the introduction of error at one locus at rate p resulted in an inflation of the true recombination estimate (Θ) by $p(1-2\Theta)$. On the basis of genotyping of classical genetic markers (and few microsatellite markers), Lathrop *et al.* (1983) estimated a rate of typing errors of approximately 1%, noting that the majority of typing errors do not lead to a genetic inconsistency. Consortium map reports using the CEPH reference data presented error rates of between 0.17% on chromosome 16 (Kozman *et al.*, 1994) to greater than 1% on chromosome 13 (Bowcock *et al.*, 1993b). It is important to note that the identification of multiple recombination events will identify genotype errors that mimic recombination events, but will not genotype errors that obscure recombinant events, thus producing potentially biased results.

High resolution linkage maps by definition have a low probability of recombination occurring in each interval (Buetow *et al.*, 1991). As the map gets denser and recombination frequency between markers approaches the error rate, a

significant proportion of all observed crossovers will be spurious (Lincoln and Lander, 1992), and a portion (approximately 1% or less) of the recombination events will be real (Buetow, 1991). If it is not possible to distinguish these true recombination events from data errors, a large fraction of misclassified recombinant gametes with statistical weight equal to that of true recombinant gametes will be introduced into the analysis. Thus, power to resolve the true locus order will be greatly reduced. As it is necessary to observe at least one recombination event between any two loci to determine their relative order within a collection of loci, maps with increasing resolution require data sets of sufficient size to ensure a reasonable probability of observing the necessary number of recombinant events. Recombinant genotypes need to be checked on the original data. If they can not be excluded on the basis of incorrect typing, then the genotype should not be altered for the analysis. Validating the order of loci on the multipoint genetic map against the order of loci independently derived on the physical map, will enable verification of the order for the linkage map. Buetow (1991) performed some simulation experiments, and concluded that the extended CEPH panel, of 65 families, would provide sufficient meioses to derive the order of a high resolution map of highly informative markers, although the increase in the number of individuals and the increase in the number of loci available would lead to an increase in the opportunities to generate genotype errors.

Microsatellite markers mutate more frequently than the classical RFLP, and in some situations an apparent double recombinant may be due to a mutation which has given rise to a new allele already segregating within the family. It is not possible to distinguish between the mutation and real recombination event, and so any double recombinant in a microsatellite marker should remain in the analysis, unless it is known to lead to incorrect locus order, as independently established by other mapping techniques.

Null alleles can arise by non PCR amplification of a dinucleotide repeat sequence, producing apparent non-mendelian inheritance. This has been reported by Phillips *et al.* (1991b), Weber *et al.* (1991), Callen *et al.* (1993), and Koorey *et al.* (1993). Null alleles can have important implications for linkage studies by feigning non-inheritance, possibly incorrectly attributed to non paternity. If undetected, a null allele will not corrupt the linkage analysis, as the individual with a segregating null allele will merely result in that individual being coded as a homozygote, resulting in loss of informativeness.

Statistical techniques, such as that in the CRI-MAP (option **chrompic**) (Lander and Green, 1987) or the program **chromlook** (Haines, 1992), can be used to detect these errors and other data anomalies (mutations or null alleles), by comparing the haplotype phase of each child's chromosomes to that of the parents, as deduced from the grandparents. In both these programs, chromosomal haplotypes inconsistent with that of the parents are indicated. The procedure involves the detection of families that contain several apparent recombinants between closely linked markers. These clusters are identified and the data verified, as a single genotype error in a parental or grandparental generation may lead to inference of an incorrect parental phase and the erroneous scoring of nonrecombinant meioses as recombinant in the offspring generation. If the physical map is constructed simultaneously with the genetic map, this can help to confirm the results of the linkage analysis.

1.6.6. Programs to construct multipoint linkage maps

Within the last decade, several multipoint linkage computer programs have been developed. The LINKAGE ANALYSIS PACKAGE (LINKAGE) is an integrated system of programs designed to perform linkage analysis to generate linkage maps and genetic risk calculation for an arbitrary number of loci. The package can be used to analyse genotype data obtained from the three generation CEPH pedigrees

and codominant marker loci, or for use with general disease pedigrees with markers and disease loci. Elston and Stewart (1971) provided the first general algorithm for computing the likelihood, as implemented in the LINKAGE ANALYSIS PACKAGE (Lathrop and Lalouel, 1984); thus, by determining the likelihood of many possible orders of multiple loci, one searches for the map with the maximum likelihood. This program package is well suited to the analysis of linkage relationships for general pedigrees and arbitrary traits and is the work horse for twopoint linkage analyses and risk analyses.

The Elston-Stewart algorithm, however, is not well suited to the sort of multilocus analysis involving a large number of loci required for the construction of high resolution genetic linkage maps of the human genome: the computation time needed to calculate likelihoods grows exponentially with the number of loci (Wilson, 1988). Consequently, multilocus linkage analysis incorporating many loci is "prohibitively time consuming, even on a super computer" (Morton *et al.*, 1986), and that "some other shorter, easier method is urgently required" (Smith, 1986).

Several modifications to multipoint analysis have been implemented to reduce the computation times. Lathrop *et al.* (1986) described a modification of the Elston-Stewart algorithm which incorporates the fact that the likelihood calculation can sometimes be factored into two or more parts - such as phase known and phase unknown. This approach has been incorporated in the LINKAGE PACKAGE (Version 5.1) for three generation CEPH pedigrees (Lathrop *et al.*, 1984). The programs for the CEPH pedigrees are optimised for rapid calculations with codominant data, and either sex-linked or autosomal loci can be analysed.

Using this modified algorithm, the LINKAGE programs determine the order of loci by optimising the likelihood over all orders, given a specific set of genotypic data. The maximum likelihood estimates of the recombination fraction between loci in this order are obtained by numerical optimisation using the GEMINI

program (Lalouel, 1979). The program, using Haldane's mapping function, makes assumptions about the mathematical relationships between map distance and recombination frequency.

Following is a quick guide to the programs available in the LINKAGE package for some common applications:

(i) Multipoint estimation of recombination rates and calculation of maximum lod scores: CILINK for the three-generation reference pedigrees; ILINK for general pedigrees, segregating with disease genes.

(ii) Twopoint lod score tables and multipoint risk analysis: MLINK (for general disease pedigrees).

(iii) Location Scores: CMAP for the three-generation CEPH reference pedigrees; LINKMAP for general disease pedigrees.

Lander and Green (1987) also described an algorithm for computing likelihoods, embodied in the computer program CRI-MAP. This program is made up of a series of options within a main program, and implements the **EM** (expectation and maximisation) algorithm (Lander and Green, 1987) and a novel and efficient method for calculating likelihoods and genetic distances (Weaver *et al.*, 1992) in its analysis routine. It is used for constructing multilocus linkage maps based on the CEPH data, and for detecting errors in the data. The advantages of using the EM algorithm, are that the likelihood increases monotonically, converging to a point (maximum Θ) where the first derivative of the likelihood expression is zero. EM searches tend to converge quickly to the vicinity of the maximum, even when started at a distant point. The favourable scaling properties of the algorithm make it practical to study large numbers of loci simultaneously.

In summary, the theoretical advantages of the EM algorithm are:

(*i*) less computation time per iteration;

(*ii*) increased likelihood on each iteration;

(*iii*) good initial convergence properties;

(*iv*) exact expressions for the first and second derivatives of the likelihood;

(*v*) ease of generalisation, inferring that sex-specific estimates, Θ_{male} and Θ_{female}, can be found easily.

Taking advantage of the EM algorithm, CRI-MAP can efficiently analyse many loci concurrently to determine the framework map of a set of loci (using the **build** option). The remaining loci can then be easily added to the map in their most likely positions using the option **all**. As stated previously, the **chrompic** option of CRI-MAP is very useful for determining the location of potential data errors.

Both of LINKAGE and CRI-MAP have strengths and foibles. For example, CRI-MAP is more efficient when all parental genotypes are available, but LINKAGE is more proficient at dealing with missing genotypes. CRI-MAP is more adept for the construction of reference maps with many loci; the LINKAGE package, whilst useful in the analysis of disease pedigrees, is not as practical at determining multilocus maps, on desktop computers.

Other programs have been written for multipoint linkage analysis. These include MAPMAKER, MENDEL, and MAP, and are described bellow.

1. MAPMAKER (Lander et al., 1987)

This is a program designed specifically for the construction of primary genetic linkage maps from the CEPH database, using the Lander-Green EM algorithm described above, to calculate the `best' map for any given order of loci. This program has been designed to maximise likelihood recursively over loci rather than over individuals in a family. This allows the analysis of a large number of loci, but limits the size of the pedigrees. It can be slow for phase unknown families, and fails to extract information from the data (White *et al.*, 1990).

2. MENDEL (Lange et al., 1988)

This is a program designed for analysis of general pedigrees. It provides routines for special tasks such as risk calculation and estimating parameters for homogeneity tests. It is similar to the LINKAGE package, but has the

disadvantage of requiring more memory. The advantage of using MENDEL is that it does not require the breakage of inbreeding loops, which is required using LINKAGE.

3. MAP (Morton and Andrews, 1989)

This program is based on multiple pairwise lods scores and uses the Rao mapping function (Rao *et al.*, 1977) to incorporate interference into the analysis. It essentially combines two point lod scores into a multipoint framework; the resulting multipoint likelihood may be approximate only (Ott, 1991). From the discussion under section 1.6.4, a multiple pairwise analysis incorporating interference, such as is possible using MAP, may have advantages over multipoint analysis programs that neglect interference (Morton, 1988).

In addition to the main linkage analyses packages, several other utility packages have been written, designed to support and complement the above packages and CEPH database software (Weaver *et al.*, 1992). These programs, PIC/HET, FAMINFO, PREINPUT, INHERIT, TWOTABLE, UNMERGE, GENLINK and LINKGEN are written in the c programming language (Kernighan and Ritchie, 1978). They have been designed to simplify input into the linkage analysis, and to interpret the results.

A more comprehensive description of LINKAGE and CRI-MAP and the above utility programs can be found in Chapter 2 (Sections 2.2.3. to 2.2.6.).

Although existing computational and database tools have been adequate for constructing linkage maps, they have been strained by the rapid growth of data. The LINKAGE package of programs has recently been rewritten, implementing new algorithms to perform rapid calculations of the likelihoods, more efficiently, in general pedigrees (Cottingham *et al.*, 1993). More recently, Matise *et al.* (1994) introduced a new software package, **MultiMap**, for automated map construction, which allows the rapid construction of genetic maps and verification of genotype data with little human intervention. Buetow *et al.* (1994) have

recently established a Cooperative Human Linkage Centre (CHLC) which incorporates a similar automated map construction procedure to MultiMap, where maps are constructed using a rigorous, semi-automated map construction algorithm that enables map validation during the construction process. These packages have extended the statistical methods and software which were initially applied to the construction of earlier RFLP-based maps. Both these packages use the CRI-MAP program as the analytical engine for their calculations.

1.7. Human Chromosome 16

Chromosome 16 contains approximately 98 Mb of DNA, thus representing about 3% of the human genome (Morton, 1991b). If the 50,000 to 100,000 genes that the human genome contains are evenly distributed throughout the genome, then it is expected that chromosome 16 would contain between 1,500 to 3,000 genes. This chromosome possesses two well characterised rare fragile sites, the folate-sensitive *FRA16A* at 16p13.11 and the distamycin A-inducible *FRA16B* at the interface between 16q21 and 16q22.1 (Fig. 1.1). In addition to these two fragile sites others have also been described for this chromosome. For example, *FRA16E* at 16p12.1 and *FRA16D* at 16q23.1 (Sutherland, 1993).

The known genes on chromosome 16 have been reviewed (Callen *et al.*, 1994); to date 38 cloned genes and 15 cDNA sequences have been physically mapped to chromosome 16 (Whitmore *et al.*, 1993). The genes associated with disease are shown in Fig. 1.2. These include the genes for autosomal polycystic kidney disease (*PKD1*) (Reeders *et al.*, 1988; Somlo *et al.*, 1992; The European Polycystic Kidney Disease Consortium, 1994), familial Mediterranean fever (*FMF*) (Shohat *et al.*, 1992; Aksentijevich *et al.*, 1993), tuberous sclerosis (*TSC2*) (Kandt *et al.*, 1992), Rubenstein-Taybi syndrome (*RTS*) (Breuning *et al.*, 1993), Batten disease (*CLN3*) (Mitchison *et al.*, 1993; 1994; Yan *et al.*, 1993) and Morquio A syndrome (Baker *et al.*, 1993).

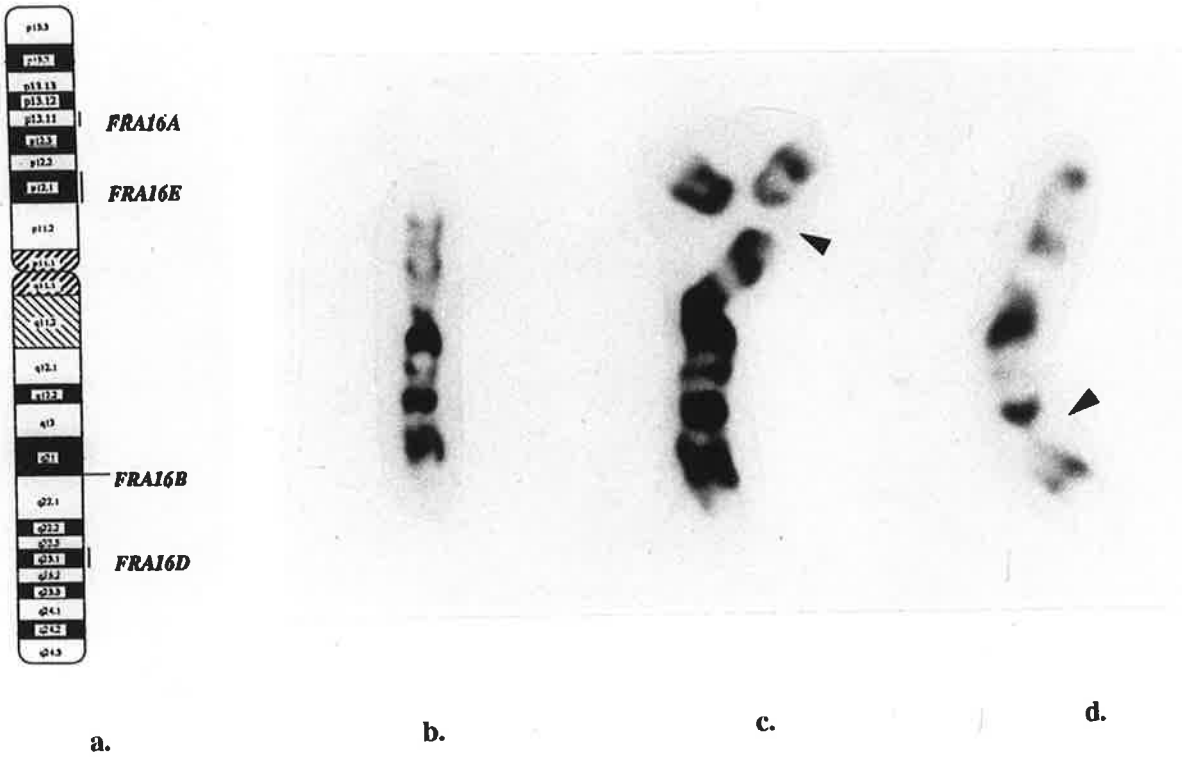

Figure 1.1: **a.** *Ideogram of Human Chromosome 16 demonstrating the cytogenetic location of four fragile sites on this chromosome.* **b.** *Normal G-Banded chromosome 16;* **c.** *G-banded chromosome 16 showing FRA16A;* **d.** *G-banded chromosome 16 showing FRA16B.*

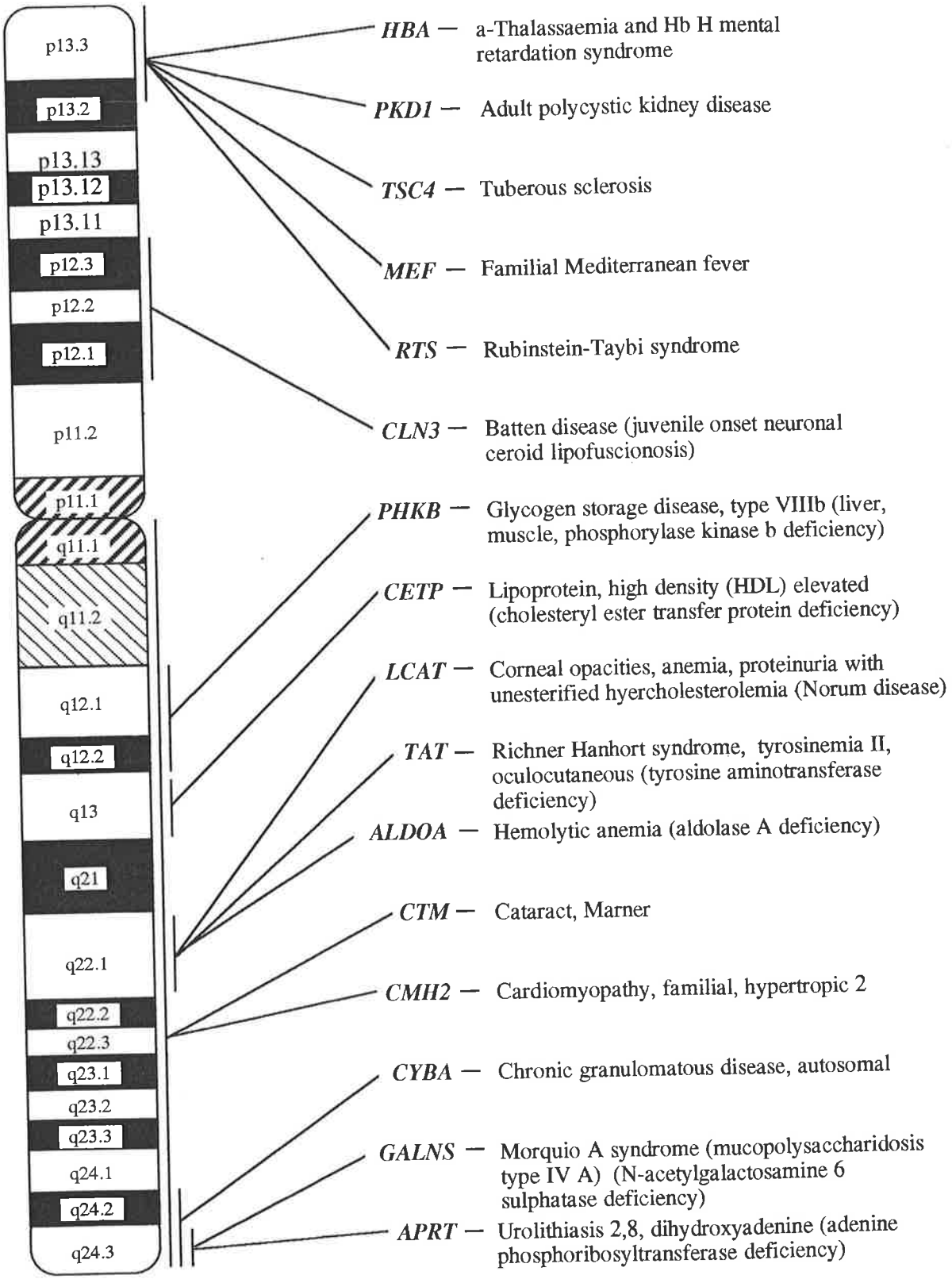

Figure 1.2: Localisation of Disease Genes on Human Chromosome 16

Various methods have been used to map these genes. Linkage analysis followed by homozygosity mapping and disequilibrium mapping (Lander and Botstein, 1987) have been exploited in refining the localisations of Batten disease (Mitchison *et al.*, 1993), and familial Mediterranean fever (Aksentijevich *et al.*, 1993); the technique of deletion mapping has been used for Rubenstein-Tabi syndrome (Breuning *et al.*, 1993), and linkage analysis, followed by high resolution physical mapping, have been used for the refinement the location of polycystic kidney disease (Germino *et al.*, 1992). The gene for polycystic kidney disease has since been isolated (The European Polycystic Kidney Disease Consortium, 1994).

Certain regions of the chromosome are known to be involved in rearrangements that cause diseases, such as deletions of specific regions of the chromosome (Callen *et al.*, 1993b), and also inversions of regions of the chromosome. In particular, a deletion in the region 16q22.1 has been suggested as critical for manifestation 16q- syndrome (Fujiwara *et al.*, 1992; Callen *et al.*, 1993). One region of the distal q arm, q24.2-q24.3, having never been reported to have been involved in a deletion suggests that it may be an essential region for fetal development. Rubenstein-Tabi syndrome is known to be associated with a deletion within a specific region of 16p (Breuning *et al.*, 1993). Aneuploidy for the entire length of chromosome 16 is lethal. Trisomy 16 is probably the most common chromosome anomaly to occur in humans and always results in early spontaneous abortion.

This chromosome has been targeted for extensive genetic and physical mapping over the last decade. These maps are the basis for the isolation of known disease genes on this chromosome, some of which have already been cloned and characterised. For example, the genes for *PKD1* (The European Polycystic Kidney Disease Consortium, 1994) and *TSC2* (European Chromosome 16 Tuberous Sclerosis Consortium, 1993).

1.7.1. Development of the linkage map of chromosome 16

The first multipoint linkage map presented for chromosome 16 was part of the genetic map for the entire human genome (Donis-Keller *et al.*, 1987) and was extended further by Julier *et al.* (1990) and Keith *et al.* (1990). These maps were based on the CEPH database, and are constructed entirely of RFLPs and VNTR markers. The first multipoint map of 42 loci (Donis-Keller *et al., 1987*) was 195 cM in length on the sex-average map. It covered almost the entire chromosome (from *HBA* at 16p13.3 to *D16S44* at 16qter), with a resolution of 4.6 cM. Only two of the loci were physically mapped. The map reported by Keith *et al.* (1990) incorporates only 4 additional loci to the map generated by Donis-Keller *et al.* in 1987. The sex-average length of this map was 149 cM, thus having a resolution of 3.2 cM. The order of loci between these two maps is slightly different, which may account for the great difference in the maps lengths of the two maps (195 cM for Donis-Keller *et al.*, 1987 and 149 cM for Keith *et al.*, 1990). The map generated by Julier *et al.* (1990) contained 24 loci, of which 20 had not been previously reported. This map extends from the distal region of the p arm (*HBA/HBZP* locus) to the distal region of the q arm (*D16S7*), although in the absence of established telomeric markers for this map, it is difficult to determine the full coverage of the chromosome. No sex-average length was reported for this map, however, the length of the male and female maps were 186 cM and 226 cM, respectively.

The apogee of genetic linkage mapping, as established as part of the human genome initiative, is to construct whole chromosome maps, with an average resolution of 2 cM, using highly polymorphic markers. Using Morton's predictions (1991b) and an estimated sex-average length of 149 cM (Keith *et al.*, 1990), 75 evenly spaced markers are required, leaving at least twenty-nine more markers to be added to the genetic map of chromosome 16 (assuming the additional markers fill the gaps on the existing maps).

Detailed linkage maps need to be developed for specific regions of chromosome 16 containing disease genes or fragile sites. The regions targeted were the regions containing the two rare fragile sites on this chromosome, *FRA16A*, and *FRA16B*, and the disease gene, *CLN3* on this chromosome. The placement of a disease gene or fragile site on the genetic map is the initial step in positional cloning since genetic distances between flanking markers verify the proximity of the markers to the target locus. This determines when studies should switch from linkage mapping to high resolution physical mapping.

More recently, whole chromosome maps have been constructed for chromosome 16 (Buetow *et al.*, 1994; Gyapay *et al.*, 1994; Matise *et al.*, 1994); however, none of these maps incorporate all the data currently available for this chromosome, nor do they adequately cover the chromosome, and few of the loci are physically mapped.

In order to efficiently localise disease genes on this chromosome, and to determine the nature of fragile sites and the molecular basis of the disease genes on this chromosome, a high-resolution linkage map, incorporating all available data and integrated with the physical map, was required. This was the aim of the present project.

1.7.2. Other methods of mapping multiple loci on chromosome 16, and integration of maps

Physical and linkage maps are colinear although the relationship between physical distances and genetic distances varies by location on the chromosome and by sex of the parent. Ultimately one map will emerge - the sequence of the human genome (Skolnick, 1991), depicting all functional elements, the genes.

Apart from genetic mapping, physical mapping strategies have been devised. These strategies have various levels of resolution. The crudest physical map is the cytogenetic based physical map, achievable by in situ hybridisation and the use of

a human-rodent somatic cell hybrid panel. Cytogenetic studies are the basis for physical maps of the human genome, by providing the visible reference point for positioning other maps. Chromosome 16 has now been divided into at least 69 intervals (Callen *et al.*, 1994), based on 77 human/rodent somatic cell hybrids constructed from rearrangements of chromosome 16, the fragile sites, and the centromere. Each interval averages approximately 1.2 Mb in length. 235 chromosome 16 markers have been used to verify the order of the hybrid breakpoints and fragile sites. This map is being constantly refined with the use of PCR based markers and new hybrids. Many of the markers on this cytogenetic-based physical map are polymorphic and have been genotyped on the CEPH panel of families, which has allowed the integration of the cytogenetic and genetic maps (Kozman *et al.*, 1993) as the initial step toward a correlated genetic and physical map.

Increased resolution is achievable by radiation hybrid mapping, in which the frequency of chromosome breakage between DNA markers is analysed statistically (based on the probability of cosegregation of markers in irradiated somatic cell hybrids), using a method analogous to genetic linkage mapping, to determine both the order and the distance between markers along the chromosome, in the Mb range (Cox *et al.*, 1990; Lawrence *et al.*, 1991). Cox *et al.*, (1990) have developed a method for mapping loci using data generated from radiation hybrids involving the estimation of a parameter comparable to a recombination frequency. A distinct advantage of radiation hybrid mapping over genetic linkage mapping is that non polymorphic markers, as well those that are polymorphic can be included in the map. Using this method of radiation hybrid mapping, it has been possible to order 38 fragments of human chromosome 16 independently of any other information and to estimate distances between them (Ceccherini *et al.*, 1992). Multiple pairwise analysis on the markers enabled estimation of distances between pairs of adjacent loci in addition to determining their most likely order. The order

and the distances of the loci obtained from this study were mostly consistent with the available data on genetic and physical mapping of these markers.

The construction of contigs of overlapping DNA fragments cloned into cosmids or YACs (Burke *et al.*, 1987), assembled by repetitive sequence fingerprinting (Stallings *et al.*, 1992), provides greater resolution on the physical map. The method of cosmid contig mapping exploits the high density of interspersed repetitive sequences, which provide highly informative "tags" to fingerprint cloned genomic DNAs found in complex genomes.. Each cosmid clone is fingerprinted using restriction fragment fingerprinting, combined with a "multiplex" hybridization strategy with selected classes of repetitive sequences (Stallings *et al.*, 1992). Using this "repeat fingerprinting" technique, 4000 cosmid clones have been arranged into 550 contigs on chromosome 16, covering approximately 60 % of the chromosome. This allows small regions of overlap (10-20%) to be unambiguously detected, and large contigs (>100 Kb) have been rapidly generated to produce a contig map. Markers on the contig map that are in common with the genetic map will allow their rapid integration (Stallings *et al.*, 1990; 1992). Completion of the cosmid contig map will enable the genetic and physical maps to be completely correlated, by expressing the recombination frequency per unit of physical distance for males and females.

The development of a method for cloning large DNA fragments, of 100 Kb to 2000 Kb (Chumakov *et al.*, 1992), into yeast artificial chromosomes (YACs) is making a major contribution to the characterisation of complex genomes. The ability to generate "YAC-sized" contigs, in a similar manner to that of a cosmid contig, provides the means for more rapid cloning of a chromosome into contigs. In the case of chromosome 16 where cosmid contig assembly is already well advanced, YACs will link existing cosmid contigs. Overlapping yeast artificial chromosomes currently provide the fastest route to the development of physical maps of entire human chromosomes. During map assembly, YACs are overlapped

to form contigs by their common content of STSs. Contig elongation can then proceed by using the ends of existing YAC contigs to achieve long range coverage of a region (Green and Green, 1991). For chromosome 16, the YAC map comprises 427 CEPH mega YACs and 125 flow-sorted 16-specific YACs that are localised to and ordered within each of the 69 breakpoint intervals with 260 STSs (Doggett *et al.*, 1994). The YAC map currently covers 97% of the euchromatic arms of the chromosome in 18 contigs with an average contig size greater then 5 Mb.

The construction and integration of the different maps described above test the integrity and consistency of the data of the constituent maps. Descrepancies between the different maps can be identified and resolved. These discrepancies might arise, for example, where genotyping errors lead to incorrect order of loci on genetic maps of where chimaerism and deletions, common within human fragments cloned into YACs, lead to errors in physical mapping. Detailed correlations between genetic and physical maps will enable gaps in the genetic and contig maps to be identified while under construction. They may be filled by employing a targeted approach; new genetic markers may be isolated from clones mapping into gaps on the genetic map, and YACs or cosmids may be identified by using polymorphic STS's which map into gaps on the physical map. Once chromosome 16 is completely cloned into a contig, any DNA sequence will be easily mapped to a specific clone without the need for tedious cytogenetic and genetic mapping which is presently the case. This will finally lead to the ultimate goal of the human genome project; the complete DNA sequence of the human genome, where the position of all coding elements are identified within the human genome.

1.8. The Aim and Strategy of the Project

With the Human Genome Initiative established, DNA technology has rapidly advanced and many new highly informative markers have become available. Ordering and mapping multiple highly informative genetic loci simultaneously to construct a high resolution linkage map is becoming considerably more feasible.

This project aims to develop the genetic linkage map of human chromosome 16, based on genotype data generated from the CEPH reference pedigrees. The denouement is a map covering the entire chromosome incorporating all available markers, the CEPH consortium map. The map that is constructed must be suitable to enable the efficient and accurate localisation of disease genes on this chromosome.

To generate a complete map of this chromosome, suitable loci need to be isolated, characterised, physically mapped and genotyped on the CEPH reference families. Then, basing inferences on data derived from observing possible recombinants from these genotypes, a genetic linkage map can be constructed.

To achieve this aim, methods of multipoint linkage analysis, embodied in the computer programs LINKAGE and CRI-MAP, will be employed with the following strategies:

(i) multipoint linkage maps in regions of interest on the chromosome, such as fragile sites or disease genes will be initially developed.

(ii) as more sophisticated techniques are developed, and more markers isolated, maps of the entire chromosome will be constructed, and integrated with the co-developing cytogenetic-based physical map, as a natural progression of map validation.

(iii) highly polymorphic PCR formatted loci will be incorporated into a high resolution linkage map for the entire length of the chromosome.

(iv) all the genetic markers for chromosome 16 (RFLPs, VNTRs, and STSs) will be incorporated into one map covering more than 99% of this

chromosome. This will be the CEPH consortium map for chromosome 16. This final map will also be integrated with the cytogenetic-based physical map, to test the integrity and consistency of the data.

1.9. Conclusion

Linkage mapping plays an important role in the chromosomal assignment and regional localisation of disease genes segregating in families, to precise locations within the human genome. An international genome initiative has been established, the human genome project, and multipoint linkage maps are being constructed for each chromosome.

The first human genetic linkage maps were constructed using protein polymorphisms. However, these markers were scarce, and so progress lagged. RFLPs and VNTRs were then discovered, and have enabled significant advances in genetic mapping. Using these markers, a rudimentary linkage map spanning much of the genome was constructed (Donis-Keller *et al.*, 1987).

With the development of laboratory techniques to generate information useful for linkage studies, so too has the technology for performing linkage analysis become more sophisticated. A variety of computer programs have been developed to perform the calculations involved in generating a linkage map. Two of the main programs involved in the analysis of chromosome 16 are CRI-MAP (Lander and Green, 1987) and LINKAGE (Lathrop *et al.*, 1985), both of which involve recursive algorithms to calculate the likelihood of the distance and order between loci. These methods of multipoint analysis have allowed the linear order and distance between genetic markers to be determined with greater accuracy than was previously possible (Chakravarti, 1991). Genetic map construction has greatly benefited from genotyping markers in common reference pedigrees (Dausset *et al.*, 1990) allowing different investigators to study a few DNA markers each, yet

enabling map construction on all markers simultaneously from a common data base.

The map of chromosome 16 began to develop in regions of interest with these genetic polymorphisms. Preliminary maps were generated in regions surrounding the two fragile sites, *FRA16B* (Chapter 3) and *FRA16A* (Chapter 4). The LINKAGE ANALYSIS PACKAGE was used to construct the map in the region of *FRA16B*, while CRI-MAP was used for the construction of the map in the vicinity of *FRA16A*.

Researchers then discovered microsatellites; abundant (with an estimated 50,000 within the genome), highly polymorphic, simple repeat sequences that are well distributed across the genome (Weber and May, 1989; Litt and Luty, 1989; Stallings *et al.*, 1991). Microsatellite markers can be assayed by PCR, and thus more quickly scored. These new markers were then incorporated into the genetic map of chromosome 16, and a high resolution map encompassing *CLN3* was developed (Chapter 5). For this map, and all subsequent map construction, the program CRI-MAP was used.

As more markers became available for this chromosome, maps were extended to cover the entire chromosome (Chapters 6, 7, and 8), and integrated with the co-developing physical map. The final map constructed for this chromosome, the CEPH consortium map incorporates the entire CEPH database (Chapter 8).

CHAPTER 2

MATERIALS AND METHODS

Table of Contents

Overview

The Materials and Methods section is divided into two areas. The first section describes methods used by the candidate in the laboratory, early in the project. The second section details the methods and computer programs used to determine the linkage map of chromosome 16. The technologies and strategies involved in linkage mapping will be described in detail in this section as they provide the framework for this thesis. These technologies and strategies were implemented by the candidate during the course of this project as the software was being developed and as hardware evolved, to enable linkage studies to be routinely used at the WCH. Initially linkage analyses were performed on a personal computer, an NEC APCIV. These were transferred to a Sun workstation IPC, when the need arose for greater computing power.

Part I: Laboratory Materials and Methods

2.1.1. Introduction

Methods employed for handling DNA during this project were previously established and used routinely in the Department of Cytogenetics and Molecular Genetics at the WCH, and are described briefly. The DNA samples used in this project were obtained on a collaborative basis from CEPH. The DNA markers used during the course of this project were isolated, characterised and physically mapped by colleagues and genotyped by the candidate, or by colleagues, as specified in Table 2.1. The human/rodent hybrid cell line panel used for physically mapping the genetic markers was generated at the WCH, and maintained by Sharon Lane under the supervision of Dr David Callen (WCH).

2.1.2. DNA Samples

DNA samples were provided on a collaborative basis by the Centre d'Etude Polymorphisme Humain (CEPH) (Dausset *et al.*, 1990). CEPH provides collaborating investigators with high quality cellular DNA produced from cultured lymphoblastoid cell lines derived from each member of a reference panel of large nuclear pedigrees. The initial panel consisted of 40 pedigrees and has now been extended to 65 three-generation families. The typical CEPH reference family consists of four grandparents, two parents and an average of eight offspring. These three generation families are ideal for genetic mapping. The presence of both sets of grandparents frequently allows determination of linkage phase in both parents, allowing unambiguous distinction of recombinants from non recombinants. Linkage phase cannot be determined where both set of grandparents are identical heterozygotes. The large size of the sibships provides substantial mapping data from each family, even when phase cannot be determined.

Collaborating investigators determine the genotypes of markers of the individuals from the panel of families. These data are processed into a uniform format by a set of programs developed by Jean-Marc Lalouel, at CEPH (see Section 2.2.2). The genotypes are then returned to CEPH for preparation of a consortium database which is returned to the collaborators for linkage analysis and map construction (called the CEPH consortium map).

The majority of the loci used in the analyses described in this thesis have been typed on the initial forty families (except where indicated in Table 2.1). Included in the 40 pedigrees distributed by CEPH are 9 Utah pedigrees (CEPH pedigrees 13291, 13292, 13293, 13294, 1331, 1333, 1340, 1341, and 1345). The DNA from these individuals was also maintained as cell lines in the laboratory by S Lane under the supervision of DF Callen (WCH). The candidate extracted the DNA from the cell lines, using either of the two extraction methods described below (Section 2.1.2.1).

A pedigree segregating for the rare folate sensitive fragile site, *FRA16A* (Simmers *et al.*, 1987), was used to localised the fragile site on to the background map in that region. This pedigree is described in more detail in Chapter 4.

During the course of this project, several inconsistent genotypes were discovered for the CEPH pedigrees. The genotypes for the CEPH family 104 were coded as unknown for the markers typed in our laboratory as the DNA was found to be misidentified. Later in the project, new samples were sent from CEPH to replace the misidentified ones. The individuals 1347-03 and 1416-10 were also found to be misidentified, and all genotypes generated at the WCH for these individuals were coded as unknown.

2.1.3. DNA Probe/Enzyme Systems

The 191 DNA markers used during the course of the project are described in Table 2.1. The information provided includes the location, locus, probe name and

Table 2.1: *Markers used for the development of the linkage map of human chromosome 16.*

Location	Gene/Locus	Probe/Enzyme	Type	Het.	No inf meioses	CEPH Lab.	Reference
16q24.3	*APRT*	M13-APRT/*Taq*I	RFLP	0.28	192	8	Williamson *et al.*, 1991
		M13-APRT/*Taq*I	RFLP	0.27	150	8, 20	Williamson *et al.*, 1991
		Aprt/*Bgl*II	RFLP	-	27	20	Williamson *et al.*, 1991
16q12.1-q13	*CETP*	pCETP.11A/*Taq*I	RFLP	0.47	98	20	Williamson *et al.*, 1991
16q23.1-q24.2	*CTRB1*	2pEKXp3B/*Pvu*II	RFLP	0.38	311	20	Williamson *et al.*, 1991
16p13.3	*HBZP1*	HBZP1/*Bgl*II	RFLP	0.47	130	01	Williamson *et al.*, 1991
		HBZP1/*Hinc*II	RFLP	0.80	246	01	Williamson *et al.*, 1991
16q22.1-q23.1	*HP*	hp2-alpha/*Msp*I	RFLP	0.50	336	01	Williamson *et al.*, 1991
16q12.1-q13	*MT2A*	MT2AA/*Taq*I	RFLP	0.25	118	43	Hyland *et al.*, 1989a
		MT2AB/*Taq*I	RFLP	0.20	-	43	Williamson *et al.*, 1991
16p12.1-p11.2	*SPN*	SPN/PCR	STR	0.96	463	43	Rogaev *et al.*, 1992
16q22.1	*D16S4*	ACH207/*Taq*I	RFLP	0.36	361	43	Hyland *et al.*, 1989b
		ACH207/*Msp*I	RFLP	0.44	-	43	Hyland *et al.*, 1989b
		ACH207/*Msp*I**	RFLP	0.47	-	43	Hyland *et al.*, 1989b
16q23.1-q24.2	*D16S5*	ACH224/*Rsa*I	RFLP	0.44	61	43	Hyland *et al.*, 1989b
16q24.3	*D16S7*	p79-2-23/*Taq*I	VNTR	0.64	201	01	Julier *et al.*, 1990
		p79-2-23/*Rsa*I	VNTR	0.82	677	20	Julier *et al.*, 1990
16p13.2-p13.11	*D16S8*	ACHF1/*Pvu*II	RFLP	0.50	354	43	Fratini *et al.*, 1988
16q21	*D16S10*	ACHF3/*Taq*I	RFLP	0.47	348	43	Fratini *et al.*, 1988
		ACHF3/*Rsa*I	RFLP	0.41	-	43	Fratini *et al.*, 1988
		ACHF3/*Msp*I	RFLP	0.18	-	43	Fratini *et al.*, 1988
16q23.1-q24.2	*D16S20*	phi8-9/*Bgl*II	RFLP	0.47	406	20	Julier *et al.*, 1990
16q22.1	*D16S38*	CRI-O2/*BamH*I	RFLP	0.35	73	01	Keith *et al.*, 1990
16q12.1-q13	*D16S39*	CRI-O3/*Pst*I	RFLP	0.49	173	01	Keith *et al.*, 1990
16q23.1-q24.2	*D16S40*	CRI-O15/*Taq*I	RFLP	0.50	172	01	Keith *et al.*, 1990
16q24.2-q24.3	*D16S41*	CRI-O43/*Taq*I	RFLP	0.36	85	01	Keith *et al.*, 1990
16p13.11-p12.1	*D16S42*	CRI-O66/*Pst*I	RFLP	0.36	103	01	Keith *et al.*, 1990
16q23.1-q24.2	*D16S43*	CRI-O84/*Hind*III	RFLP	0.33	111	01	Keith *et al.*, 1990
		CRI-O84/*Pst*I	RFLP	0.69	237	01	Keith *et al.*, 1990

Table 2.1 (cont'd): *Markers used for the development of the linkage map of human chromosome 16.*

Location	Gene/Locus	Probe/Enzyme	Type	Het.	No inf meioses	CEPH Lab.	Reference
16q24.3	*D16S44*	CRI-O89/*Bgl*II	RFLP	0.50	206	01	Keith *et al.*, 1990
16p13.3	*D16S45*	CRI-O90/*Eco*RI	RFLP	0.49	304	01	Keith *et al.*, 1990
		CRI-O90/*Dra*I		0.14	52	01	Keith et al., 1990
16q22.1	*D16S46*	CRI-O91/*Msp*I	RFLP	0.36	137	01	Keith *et al.*, 1990
		CRI-O91/*Taq*I	RFLP	0.38	104	01	Keith *et al.*, 1990
16q21-q23.1	*D16S47*	CRI-O95/*Eco*RI	RFLP	0.49	158	01	Keith *et al.*, 1990
16p12.1-p11.2	*D16S48*	CRI-O101/*Hind*III	RFLP	0.47	144	01	Keith *et al.*, 1990
16p13.3-p13.11	*D16S49*	CRI-O114AC/PCR	STR	0.66	113	43	Shen *et al.*, 1993c
		CRI-O114/*Eco*RI	RFLP	0.39	169	01	Keith *et al.*, 1990
16q23.1-q24.2	*D16S50*	CRI-O119/*Taq*I	RFLP	0.46	162	01	Keith *et al.*, 1990
16p13.3-p13.11	*D16S51*	CRI-O120/*Bam*HI	RFLP	0.60	220	01	Keith *et al.*, 1990
16q12.1-q13	*D16S52*	CRI-O123/*Hind*III	RFLP	0.41	92	01	Keith *et al.*, 1990
16p*	*D16S53*	CRI-O125/*Pst*I	RFLP	0.49	175	01	Keith *et al.*, 1990
16p*	*D16S54*	CRI-O126/*Hinc*II	RFLP	0.20	53	01	Keith *et al.*, 1990
		CRI-O126/*Eco*RI	RFLP	0.42	161	01	Keith *et al.*, 1990
16p13.3	*D16S55*	CRI-O128/*Hinc*II	RFLP	0.30	72	01	Keith *et al.*, 1990
16p13.3	*D16S56*	CRI-O129/*Eco*RI	RFLP	0.38	235	01	Keith *et al.*, 1990
		CRI-O129/*Bgl*II	RFLP	0.20	75	01	Keith *et al.*, 1990
16p11.2-p11.1	*D16S57*	CRI-O131/*Bam*HI	RFLP	0.40	79	01	Keith *et al.*, 1990
16p13.3	*D16S58*	CRI-O133/*Hind*III	RFLP	0.32	248	01	Keith *et al.*, 1990
16q*	*D16S59*	CRI-O134/*Msp*I	RFLP	0.33	122	01	Keith *et al.*, 1990
16p13.3-p13.11	*D16S60*	CRI-O136/*Hinc*II	RFLP	0.83	444	01	Keith *et al.*, 1990
16p*	*D16S61*	CRI-O144/*Eco*RI	RFLP	0.46	198	01	Keith *et al.*, 1990
16q24.2-q24.2	*D16S62*	CRI-O149/*Taq*I	RFLP	0.45	106	01	Keith *et al.*, 1990
16p13.3	*D16S63*	CRI-O327/*Hind*III	RFLP	0.38	276	01	Keith *et al.*, 1990
16p13.11-p12.1	*D16S64*	CRI-O373/*Hind*III	RFLP	0.46	195	01	Keith *et al.*, 1990
16q12.1-q13	*D16S65*	CRI-O377/*Bgl*II	RFLP	0.41	148	01	Keith *et al.*, 1990
		CRI-O377/*Hinc*II	RFLP	0.50	84	01	Keith *et al.*, 1990
		CRI-O377/*Taq*I	RFLP	0.24	56	01	Keith *et al.*, 1990

Table 2.1 (cont'd): *Markers used for the development of the linkage map of human chromosome 16.*

Location	Gene/Locus	Probe/Enzyme	Type	Het.	No inf meioses	CEPH Lab.	Reference
16p*	*D16S66*	CRI-O383/*Pst*I	RFLP	0.47	189	01	Keith *et al.*, 1990
16p13.11-p12.1	*D16S67*	CRI-O391AC/PCR	STR	0.77	395	43	Callen *et al.*, 1993
		CRI-O391/*Hinc*II	RFLP	0.46	128	01	Keith *et al.*, 1990
		CRI-O391/*Taq*I	RFLP	0.50	199	01	Keith *et al.*, 1990
16p12-qter*	*D16S68*	CRI-O393/*Bgl*II	RFLP	0.47	165	01	Keith *et al.*, 1990
		CRI-O393/*Pst*I	RFLP	0.28	100	01	Keith *et al.*, 1990
16q22.1-q24.3	*D16S69*	CRI-P84/*Bam*HI	RFLP	0.38	132	01	Keith *et al.*, 1990
16p*	*D16S70*	CRI-P130/*Msp*I	RFLP	0.48	195	01	Keith *et al.*, 1990
16q*	*D16S71*	CRI-P400/*Bam*HI	RFLP	0.48	139	01	Keith *et al.*, 1990
16p*	*D16S72*	CRI-P403/*Msp*I	RFLP	0.46	178	01	Keith *et al.*, 1990
16p*	*D16S73*	CRI-P477/*Taq*I	RFLP	0.38	128	01	Keith *et al.*, 1990
16p*	*D16S74*	CRI-P85/*Bgl*II	RFLP	0.49	196	01	Keith *et al.*, 1990
16p13.11-p12.1	*D16S75*	CRI-R99/*Hind*III	RFLP	0.51	162	01	Keith *et al.*, 1990
		CRI-R99/*Msp*I	RFLP	0.31	88	01	Keith *et al.*, 1990
16p*	*D16S76*	CRI-L223/*Taq*I	RFLP	0.46	160	01	Keith *et al.*, 1990
16q*	*D16S77*	CRI-L922/*Hinc*II	RFLP	0.36	112	01	Keith *et al.*, 1990
		CRI-L922/*Hind*III	RFLP	0.21	72	01	Keith *et al.*, 1990
16p13.3-p13.11	*D16S79A*	16AC66F3/PCR	STR	0.81	398	43	Phillips *et al.*, 1993
		36-1A/*Taq*I	RFLP	0.34	245	43	Breuning *et al.*, 1989
16p13.11-p12.1	*D16S79B*	36-1B/*Taq*I	RFLP	0.15	83	43	Breuning *et al.*, 1989
16p13.3	*D16S80*	24-1/*Taq*I**	RFLP	0.34	256	55	Callen *et al.*, 1989
16p13.3*	*D16S83*	pEKMDA2.1/*Hinf*I	VNTR	0.87	752	20	Julier *et al.*, 1990
16p13.3*	*D16S84*	pCMM65/*Eco*RI	VNTR	0.39	357	20	Julier *et al.*, 1990
16p13.3	*D16S85*	3'HVR/*Msp*I	VNTR	0.93	573	1	Keith *et al.*, 1990
		3'HVR/*Rsa*I/*Msp*I	VNTR	-	425	1	Keith *et al.*, 1990
		5'HVR/*Rsa*I	VNTR	0.79	257	1	Keith *et al.*, 1990
16q22.1	*D16S91*	LE12/*Rsa*I	RFLP	0.11	46	43	Davidson *et al.*, 1987
16p13.3	*D16S94*	16ACVK5/PCR	STR	0.51	383	43	Aksentijevich *et al.*, 1993
		VK5B/*Msp*I	RFLP	0.49	324	43	Hyland *et al.*, 1990

Table 2.1 (cont'd): *Markers used for the development of the linkage map of human chromosome 16.*

Location	Gene/Locus	Probe/Enzyme	Type	Het.	No inf meioses	CEPH Lab.	Reference
16p13.11-p12.1	*D16S96*	VK20A/*Taq*I	RFLP	0.49	347	43	Hyland *et al.*, 1989a
		VK20A/*Msp*I	RFLP	0.36	-	43	Hyland *et al.*, 1989a
		VK20B/*Msp*I	RFLP	0.59	-	43	Hyland *et al.*, 1989a
16p13.11-p12.1	*D16S131*	VK45C6/*Taq*I	RFLP	0.46	240	43	Hyland *et al.*, 1989c
16p*	*D16S137*	pKKA22/*Pst*I	RFLP	0.60	163	20	Julier *et al.*, 1990
16p13.11-p12.1	*D16S148*	CJ52.95M1/*Msp*I	RFLP	0.39	326	20	Julier *et al.*, 1990
16q12.1-q13	*D16S150*	1CJ52.161/*Taq*I	RFLP	0.38	281	20	Julier *et al.*, 1990
		2CJ52.161/*Taq*I	RFLP	0.04	31	20	Julier *et al.*, 1990
16q21	*D16S151*	1CJ52.209M/*Msp*I	RFLP	0.32	266	20	Julier *et al.*, 1990
		2CJ52.209M/*Msp*I	RFLP	0.25	178	20	Julier *et al.*, 1990
16q22.1	*D16S152*	1CJ52.1/*Taq*I	RFLP	0.31	158	20	Julier *et al.*, 1990
		2CJ52.1/*Taq*I	RFLP	0.50	236	20	Julier *et al.*, 1990
16q22.2-q23.1	*D16S153*	CJ52.10T2/*Taq*I	RFLP	0.46	466	20	Julier *et al.*, 1990
16q24.2-q24.3	*D16S154*	CJ52.105/*Taq*I	RFLP	0.69	466	20	Julier *et al.*, 1990
16q22-qter*	*D16S155*	CJ52.199/*Msp*I	RFLP	0.43	344	20	Julier *et al.*, 1990
16q22-qter*	*D16S156*	1CJ52.197/*Taq*I	RFLP	0.27	244	20	Julier *et al.*, 1990
		2CJ52.197/*Taq*I	RFLP	0.23	141	20	Julier *et al.*, 1990
16q23.1-24.3	*D16S157*	CJ52.96/*Taq*I	RFLP	0.25	148	20	Julier *et al.*, 1990
16p13.2-q21	*D16S158*	CJ52.112/*Taq*I	RFLP	0.15	113	20	Julier *et al.*, 1990
16p13.11-p12.1	*D16S159*	CJ52.94T1/*Taq*I	RFLP	0.50	481	20	Julier *et al.*, 1990
16q*	*D16S160*	CJ52.196M1/*Msp*I	RFLP	0.50	278	20	Julier *et al.*, 1990
16q21	*D16S164*	16PHAC-15/PCR	STR	0.38	226	43	Phillips *et al.*, 1991a
16q21	*D16S186*	16PHAC-101/PCR	STR	0.57	413	43	Phillips *et al.*, 1991a
16p13.3*	*D16S246*	218EP/*Pvu*II**	RFLP	0.46	265	55	Williamson *et al.*, 1991
16p13.3*	*D16S252*	pCMM103/*Pvu*II	VNTR	0.51	218	20	Williamson *et al.*, 1991
16q12.1-q13	*D16S261*	MFD24/PCR	STR	0.71	449	43	Weber *et al.*, 1990
16q21	*D16S265*	MFD23/PCR	STR	0.77	505	43	Weber *et al.*, 1990
16p13.3*	*D16S282*	14C19/*Hae*III**	RFLP	0.73	421	46	Lauthier *et al.*, 1991
16p13.3	*D16S283*	SM7/PCR	STR	0.64	407	55	Harris *et al.*, 1991

Table 2.1 (cont'd): *Markers used for the development of the linkage map of human chromosome 16.*

Location	Gene/Locus	Probe/Enzyme	Type	Het.	No inf meioses	CEPH Lab.	Reference
16q12.1-q13	*D16S285*	AluGT16/PCR	STR	0.83	496	10	Konradi *et al.*, 1991
16p13.11-p12.1	*D16S287*	16XE81/PCR	STR	0.78	502	43	Phillips *et al.*, 1991a
16p12.1-p11.2	*D16S288*	16AC7.1/PCR	STR	0.73	449	43	Shen *et al.*, 1991
16q23.1-q24.2	*D16S289*	16AC7.46/PCR	STR	0.77	396	43	Shen *et al.*, 1992
16p13.3	*D16S291*	16AC2.5/PCR	STR	0.79	519	43	Thompson *et al.*, 1992
16p13.2-p13.11	*D16S292*	16AC2.3/PCR	STR	0.74	450	43	Thompson *et al.*, 1992
16p13.11-p12.1	*D16S294*	16AC1/PCR	STR	0.49	235	43	Thompson *et al.*, 1992
16p13.11-p12.1	*D16S295*	16AC62F3/PCR	STR	0.66	430	43	Callen *et al.*, 1993
16p13.11-p12.1	*D16S296*	16AC62B4/PCR	STR	0.75	502	43	Callen *et al.*, 1993
16p13.11-p12.1	*D16S297*	16AC15H1H/PCR	STR	0.73	412	43	Callen *et al.*, 1993
		16AC15H1S/PCR	STR	0.69	124	43	Callen *et al.*, 1993
16p12.1-p11.2	*D16S298*	16AC3.12/PCR	STR	0.80	455	43	Thompson *et al.*, 1992
16p12.1-p11.2	*D16S299*	16AC6.17/PCR	STR	0.72	425	43	Thompson *et al.*, 1992
16p11.2-p11.1	*D16S300*	16AC1.1/PCR	STR	0.61	436	43	Thompson *et al.*, 1992
16q22.1	*D16S301*	16AC6.21/PCR	STR	0.64	409	43	Thompson *et al.*, 1992
16q24.3	*D16S303*	16AC6.26/PCR	STR	0.43	230	43	Thompson *et al.*, 1992
16q12.1-q13	*D16S304*	16AC1.14/PCR	STR	0.60	410	43	Thompson *et al.*, 1992
16q24.3	*D16S305*	16AC1.15/PCR	STR	0.82	494	43	Thompson *et al.*, 1992
16p*	*D16S307*	pMS637/*Hae*III**	RFLP	0.56	401	46	Williamson *et al.*, 1991
16q12.1-q13	*D16S308#*	16AC1.18/PCR	STR	0.77	113	43	Thompson *et al.*, 1992
16p13.3*	*D16S309*	pMS205/*Mbo*I**	RFLP	0.97	516	31	Royle *et al.*, 1992
16q21	*D16S310*	MIT-MH20/PCR	STR	0.67	472	43	Hudson *et al.*, 1992
16p13.11-p12.1	*D16S313*	MIT-MS79/PCR	STR	0.57	498	43	Hudson *et al.*, 1992
16q22.1	*D16S318#*	16AC8.20/PCR	STR	0.54	143	43	Shen *et al.*, 1992
16p13.11-p12.1	*D16S319*	16AC7.14/PCR	STR	0.52	315	43	Shen *et al.*, 1992
16q12.1-q13	*D16S320*	16AC8.52/PCR	STR	0.86	559	43	Shen *et al.*, 1992
16q23.2-q24.3	*D16S332*	16AC305D6/PCR	STR	0.53	324	43	Shen *et al.*, 1993a
16q22.1	*D16S347#*	16AC12F8/PCR	STR	0.76	147	43	Shen *et al.*, 1993b

Table 2.1 (cont'd): *Markers used for the development of the linkage map of human chromosome 16.*

Location	Gene/Locus	Probe/Enzyme	Type	Het.	No inf meioses	CEPH Lab.	Reference
16q12.1-q13	*D16S359*	16AC26E3B/PCR	STR	0.42	321	43	Shen *et al.*, 1994a
16q23.2-q24.3	*D16S363#*	16AC51G1/PCR	STR	0.78	164	43	Shen *et al.*, 1993a
16p12.1-p11.2	*D16S383*	16AC80B3/PCR	STR	0.45	235	43	Shen *et al.*, 1993d
16q21	*D16S389#*	16AC10B3/PCR	STR	0.77	170	43	Shen *et al.*, 1993b
16q12.1-q13	*D16S390*	16AC10F5/PCR	STR	0.80	511	43	Shen *et al.*, 1993b
16q23.2-q24.3	*D16S392#*	16AC305E9/PCR	STR	0.78	161	43	Shen *et al.*, 1993a
16q23.2-q24.3	*D16S393#*	16AC323H4/PCR	STR	0.87	159	43	Shen *et al.*, 1993a
16q23.2-q24.2	*D16S395*	16AC33G11/PCR	STR	0.69	467	43	Shen *et al.*, 1993b
16q22.1	*D16S397#*	MFD98/PCR	STR	0.70	168	43	Shen *et al.*, 1994b
16q22.1	*D16S398*	MFD168/PCR	STR	0.90	559	43	Shen *et al.*, 1994b
16p13.11-p12.1	*D16S401#*	AFM025tg9/PCR	STR	0.74	80	42	Weissenbach *et al.*, 1992
16q23.1-q24.3	*D16S402#*	AFM031xa5/PCR	STR	0.87	140	42	Weissenbach *et al.*, 1992
16p13.11-p12.1	*D16S403#*	AFM049xd2/PCR	STR	0.86	171	42	Weissenbach *et al.*, 1992
16p13.2-13.11	*D16S404#*	AFM056yf6/PCR	STR	0.82	158	42	Weissenbach *et al.*, 1992
16p13.2-13.11	*D16S405#*	AFM070ya1/PCR	STR	0.78	143	42	Weissenbach *et al.*, 1992
16p13.3-13.11	*D16S406#*	AFM079yh3/PCR	STR	0.82	152	42	Weissenbach *et al.*, 1992
16p13.2-13.11	*D16S407#*	AFM113xe3/PCR	STR	0.86	160	42	Weissenbach *et al.*, 1992
16q12.1-q13	*D16S408#*	AFM137xf8/PCR	STR	0.69	120	42	Weissenbach *et al.*, 1992
16p13.11-p12.1	*D16S410#*	AFM165yb6/PCR	STR	0.57	125	42	Weissenbach *et al.*, 1992
16q12.1-q13	*D16S411#*	AFM186xa3/PCR	STR	0.79	130	42	Weissenbach *et al.*, 1992
16p13.11-p12.1	*D16S412#*	AFM191wb10/PCR	STR	0.76	101	42	Weissenbach *et al.*, 1992
16q24.3	*D16S413#*	AFM196xg1/PCR	STR	0.85	154	42	Weissenbach *et al.*, 1992
16p13.2-13.11	*D16S414#*	AFM205za11/PCR	STR	0.61	137	42	Weissenbach *et al.*, 1992
16q12.1-q13	*D16S415#*	AFM205ze5/PCR	STR	0.74	121	42	Weissenbach *et al.*, 1992
16q12.1-q13	*D16S416#*	AFM210yg3/PCR	STR	0.43	53	42	Weissenbach *et al.*, 1992
16p13.11-p12.1	*D16S417#*	AFM220xb10/PCR	STR	0.73	155	42	Weissenbach *et al.*, 1992
16p13.3-13.11	*D16S418#*	AFM225xd2/PCR	STR	0.83	157	42	Weissenbach *et al.*, 1992

Table 2.1 (cont'd): *Markers used for the development of the linkage map of human chromosome 16.*

Location	Gene/Locus	Probe/Enzyme	Type	Het.	No inf meioses	CEPH Lab.	Reference
16q12.1-q13	*D16S419#*	AFM225zf2/PCR	STR	0.77	149	42	Weissenbach *et al.*, 1992
16p13.11-p12.1	*D16S420#*	AFM238xb2/PCR	STR	0.82	149	42	Weissenbach *et al.*, 1992
16q22.1	*D16S421#*	AFM240yh6/PCR	STR	0.57	140	42	Weissenbach *et al.*, 1992
16q23.1-q24.2	*D16S422#*	AFM249xc5/PCR	STR	0.80	166	42	Weissenbach *et al.*, 1992
16p13.3	*D16S423#*	AFM249yc5/PCR	STR	0.75	136	42	Weissenbach *et al.*, 1992
16q23.2-q24.3	*D16S449#*	16AC51A4/PCR	STR	0.85	134	43	Shen *et al.*, 1993a
16q22.2-q23.1	*D16S450*	16AC80H3/PCR	STR	0.52	232	43	Shen *et al.*, 1993b
16q21	*D16S451#*	16AC69F12/PCR	STR	0.84	145	43	Shen *et al.*, 1993b
16p13.11-p12.1	*D16S452#*	16AC33A4/PCR	STR	0.68	143	43	Shen *et al.*, 1993d
16p13.3	*D16S453*	16AC301G12/PCR	STR	0.45	335	43	Shen *et al.*, 1993c
16p13.2-p13.11	*D16S454#*	16AC45G5/PCR	STR	0.75	134	43	Shen *et al.*, 1993c
16q22.2-q23.1	*D16S522*	16AC8.21/PCR	STR	0.69	449	43	Shen *et al.*, 1994a
16p13.3	*D16S523#*	16AC13H1/PCR	STR	0.68	149	43	Shen *et al.*, 1993d
16p13.11-p12.1	*D16S524#*	16AC40A7/PCR	STR	0.76	157	43	Shen *et al.*, 1993d
16p13.3	*D16S525#*	16AC308G7/PCR	STR	0.91	180	43	Shen *et al.*, 1994a
16q12.1-q13	*D16S531#*	16AC8.15/PCR	STR	0.86	147	43	Shen *et al.*, 1994a
16p*	**	Hag1/*Hinf*I	RFLP	-	241	20	Williamson *et al.*, 1991
16p*	**	69A4/*Pst*I	RFLP	-	313	20	Williamson *et al.*, 1991
16q*	**	76BC6/*Msp*I	RFLP	-	490	20	Williamson *et al.*, 1991
16q*	**	LC9/*Msp*I	RFLP	-	243	20	Williamson *et al.*, 1991

\# - Markers typed on only 8 of the CEPH families.
* - Physical location not refined.
** - Markers not previously incorporated into a published genetic linkage map of chromosome 16.

enzyme, type, and reference for each marker. All of the markers used were extracted from the CEPH database. However, the majority of markers were isolated, characterised, physically mapped, and genotyped at the WCH by colleagues and subsequently contributed to the CEPH database (as indicated in Table 2.1). The probes used by the candidate to genotype the CEPH pedigrees are indicated on Table 2.1. These were *D16S4*, *D16S10*, *D16S91*, *D16S96*, and *MT*. One hundred and fifty-eight of the 191 DNA probes described in this thesis have been physically mapped to a human/rodent hybrid panel, containing cytologically defined portions of human chromosome 16 (Callen *et al.*, 1992a). One marker, *D16S265*, was physically mapped by the candidate; the methods used to physically map this marker are described in Chapter 6.

2.1.4. Methods of Extracting DNA

The two methods included in this section were used to extract DNA from preparations of lymphoblastoid cell lines (LCL) or fibroblast lines. Care was taken when extracting genomic DNA to ensure that the DNA was not sheared. Gloves were worn when performing the phenol-chloroform extractions, and also when staining with Ethidium Bromide (EtBr), which is mutagenic.

2.1.4.1. High salt extraction of cell line DNA (Miller et al, 1988)

Frozen pellets of fibroblast cell lines were thawed and washed twice in 30 ml cell lysis buffer (0.32 M sucrose, 10 mM Tris HCl, 5 mM $MgCl_2$, 1% Triton X-100), and incubated in the buffer on ice for thirty minutes; the tube was then centrifuged in a Jouan CR3000 centrifuge (4° C, 3,500 rpm for fifteen minutes) to pellet the lysed cells and the supernatant discarder. The pellet was resuspended in three ml of nuclei lysis buffer (10 mM Tris HCl, 400 mM NaCl, 2 mM EDTA, pH 8.0) was added and the pellet resuspended; 0.5 ml 10% SDS and 0.2 ml 1% Proteinase K (Boehringer Mannheim) were added for the digestion of unwanted proteins. The tube was sealed with parafilm and incubated overnight at 37° C on a rotating

wheel (10 rpm). One ml of saturated NaCl (6 M) was added and the tube shaken vigorously for fifteen seconds; the tube was centrifuged (4° C, 1000 *g* for fifteen minutes), and the supernatant transferred to a 50 ml tube. Two volumes of 100% ethanol (room temperature) were added, and the solution was gently mixed to precipitate the nucleic acids. The precipitated DNA was transferred to a new tube and washed with 70% ethanol (room temperature) and centrifuged to pellet the nucleic acids, which were then desiccated and resuspended in 2 ml nuclei lysis buffer, and incubated at 4° C overnight.

The DNA pellet was treated with 0.15 μg/ml RNase A (Boehringer Mannheim) at 37° C for four hours, followed by a further Proteinase K digestion (0.07 μg/ml, 37° C overnight) on a rotating wheel. The DNA was precipitated by the addition of 1/3 volume saturated NaCl and 1/5 volume of 50% PEG_{6000}, gently mixed, and incubated at 4° C overnight. The DNA was recovered by centrifugation (4° C, 1000 *g* for fifteen minutes), washed in 70% ethanol, desiccated, and resuspended in TE (10 mM Tris HCl, 1 mM EDTA).

2.1.4.2. Phenol - Chloroform extraction of cell line DNA (a modification of Blin and Stafford, 1976)

Frozen pellets of fibroblast cell lines were thawed and washed twice in 30 ml cell lysis buffer (0.32 M sucrose, 10 mM Tris HCl, 5 mM $MgCl_2$, 1% Triton X-100) and incubated in the buffer on ice for thirty minutes; the tube was then centrifuged in a Jouan CR3000 centrifuge (4° C, 3,500 rpm for fifteen minutes) to pellet the lysed cells and the supernatant was removed. Three ml of nuclei lysis buffer (10 mM Tris HCl, 400 mM NaCl, 2 mM EDTA, pH 8.0) was added and the pellet resuspended completely; 0.5 ml 10% SDS and 0.2 ml 1% Proteinase K (Boehringer Mannheim) was added for the digestion of unwanted proteins. The tube was sealed with parafilm and incubated overnight at 37° C on a rotating wheel (10 rpm).

The nucleic acids were extracted once each with:

(i) an equal volume of phenol (saturated with Tris pH 8.0 plus 8-hydroxyquinoline),

(ii) an equal volume of phenol : chloroform : isoamyl alcohol (25:24:1), and

(iii) an equal volume of chloroform : isoamyl alcohol (24:1).

To precipitate the DNA, 1/10 volume of 3 M sodium acetate (pH 4.6) and 2 volumes of ice cold 100% ethanol were added, the DNA pelleted by centrifugation, washed in 70% ethanol, desiccated and resuspended in an appropriate volume of 10 mM Tris HCl, 0.1 mM EDTA. If it was necessary, an RNase and Proteinase treatment (Section 2.1.3.1), followed by a further phenol - chloroform extraction, was performed.

2.1.5. Isolation of Plasmid DNA

2.1.5.1. Large scale isolation of plasmid DNA (Birnboim and Doly, 1979; Maniatis et al., 1982)

Five ml Luria Bertani (LB) medium containing ampicillin (50 μg/ml) was inoculated with a single colony. The culture was incubated overnight at 37° C with vigorous shaking. 250 ml of LB in a one litre conical flask containing ampicillin (50 μg/ml) was inoculated with 0.75 ml of the overnight culture, and incubated at 37° C with vigorous shaking until the OD_{600} read between 0.4 - 0.6 (two to three hours). Chloramphenicol (170 μg/ml) was added to the flask, which was again incubated overnight at 37° C with vigorous shaking. The 250 ml culture was placed in a 250 ml centrifuge bucket and centrifuged on a Beckman J2-21M/E centrifuge, in a JA 20 rotor (4° C, 6000 *g* for fifteen minutes). The pelleted cells were resuspended in 50 mM glucose, 25 mM Tris HCl pH 8.0, 10 mM EDTA, and lysozyme (13.3 mg/ml) (Boehringer Mannheim). The cell suspension was left at room temperature for five minutes, followed by the addition of 0.2 M NaOH, 1% SDS, gentle mixing, and incubation on ice for ten minutes.

½ volume of 5 M ice cold potassium acetate (pH 4.8) was mixed with the cell suspension by gentle inversion and incubated on ice for a further ten minutes. The mixture was centrifuged on a Beckman J2-21M/E centrifuge, in a JA 20 rotor (4° C, 12 000 rpm, for thirty minutes). The supernatant was collected and two volumes of ice cold 100% ethanol, and left at room temperature for fifteen minutes. The DNA was pelleted by centrifugation in the Beckman centrifuge (room temperature, 12 000 rpm, for thirty minutes). The pellet was washed twice in 70% ethanol, desiccated and resuspended in TE. To ensure complete resuspension, the DNA preparation was left at room temperature for one hour. The DNA solution was treated with DNase-free RNase (0.15 μg/ml) (Boehringer Mannheim), to digest RNA, for thirty minutes at 37° C. To digest proteins, the DNA preparation was Proteinase K (0.07 μg/ml) incubated for one hour at 37° C. To purify the DNA solution, a phenol - chloroform extraction was performed as previously described (Section 2.1.3.2). The plasmid was purified on a CsCl gradient where necessary (Section 2.1.7.). The DNA was precipitated by the addition of 1/10 volume 3 M sodium acetate and 2 volumes 100% ethanol, mixed and left overnight at -20° C. The DNA was pelleted in an Eppendorf centrifuge by centrifugation at 12 000 rpm for thirty minutes, washed with 70% ethanol, desiccated, and resuspended in TE.

2.1.5.2. Small scale isolation of plasmid DNA (Birnboim and Doly, 1979)

1.5 ml Luria Bertani (LB) medium containing ampicillin (50 μg/ml) was inoculated with a single colony, and incubated overnight at 37° C with vigorous shaking. The 1.5 ml overnight culture was then centrifuged in an Eppendorf centrifuge (14 000 rpm for two minutes). The pellet was resuspended in 100 μl (50 mM glucose, 25 mM Tris HCl pH 8.0, 10 mM EDTA), with the addition of lysozyme (13.3 mg/ml) (Boehringer Mannheim). The cell suspension was left at room temperature for five minutes. 200 μl (0.2 M NaOH, 1% SDS) was added to

the solution, mixed gently, and incubated on ice for five minutes. Half volume of 5 M ice cold potassium acetate (pH 4.8) was added and mixed well gentle inversion and incubated on ice for a further five minutes. The mixture was centrifuged at 12 000 rpm for ten minutes. To purify the DNA solution, a phenol - chloroform extraction was performed as prevoiusly described (Section 2.1.4.2). The DNA was then collected as for the previous method.

2.1.6. Recovery of DNA From Agarose

2.1.6.1. Electroelution of DNA (McDonnell et al., 1977)

DNA was recovered from low melting point (LMP) agarose or the agarose gel (Pharmacia) by electroelution after ethidium bromide (EtBr) staining and visualisation with UV illumination. The dialysis tubing (Promega) used for electroelution was prepared by boiling for ten minutes in one litre of 2% sodium bicarbonate, 1 mM EDTA. After rinsing thoroughly in distilled water, the dialysis tubing was boiled for ten minutes in distilled water before use.

A slice of agarose gel containing the DNA band was excised and placed into the pretreated dialysis tubing containing 0.5 x TBE (1 x TBE: 89 mM Tris base, 89 mM boric acid, 10 mM EDTA). DNA was electrophoresed out of the gel slice at 100 V for two to three hours in 0.5 x TBE. The current direction was reversed for two minutes and the buffer containing the DNA was recovered from the dialysis tubing. The DNA was rescued by ethanol precipitation, as described in Section 2.1.3.2.

2.1.6.2. "Freeze / squeeze" isolation of insert DNA

Whole plasmids containing insert DNA were digested with the appropriate enzyme to release the genomic DNA fragment. The digest was stopped by the addition of EDTA (final concentration 10 mM), and the DNA gel electrophoresed (Section

2.1.8.2.) and visualised by staining with EtBr. The band containing the genomic DNA was cut from the agarose gel and the agarose slice was either:

(i) placed in an eppendorf tube, snap frozen in liquid nitrogen, and the supernatant containing the nucleic acids was collected by centrifugation, or

(ii) placed in a small bag, sealed, and then squeezed to extract as much liquid out of the agarose slice as possible. This squashed agarose could also be snap frozen and spun if desired.

In both the above cases, the EtBr in the agarose was removed by an n-butanol extraction. An equal volume of n-butanol is added to the DNA sample, mixed well, and centrifuged at 1600 *g* for 1 minute. This causes the phases to separate and the EtBr is eliminated in the organic n-butanol phase, leaving the DNA sample clear of EtBr. This process was repeated until the organic phase was clear of any colour. The sample was twice extracted with water saturated ether to remove any remaining n-butanol. After the final wash, the ether was removed by evaporation. The nucleic acids were collected by ethanol precipitation, as described previously (Section 2.1.4.1).

2.1.7. Caesium Chloride Gradient For The Purification of Plasmid DNA

(Birnboim and Doly, 1979)

1.6 ml of 63% w/v CsCl solution was placed in each 2.2 ml polyallomer Quick-Seal tube (Beckman). Nucleic acids (isolated from one of the above plasmid isolation methods) were dissolved in 360 μl of 10 mM Tris (pH 7.5), 1 mM EDTA in an eppendorf microfuge tube. 0.63 gm CsCl was added to the nucleic acid solution, and mixed to dissolve. Sixty μl of EtBr (10 mg/ml) was added. The nucleic acid solution was carefully underlaid to the lower density CsCl solution in the Quick-Seal tube. The tubes were then completely filled with the addition of extra 0.63 gm/ml CsCl solution, and balanced in pairs to 0.01 gm precision, and sealed.

The tubes were centrifuged in the TL-100 centrifuge (Beckman) for 2½ hours (100,000 rpm, 20° C). On completion of the run, the plasmid bands were visualised under UV light, and extracted from the tube with a 21 gauge needle, by piercing the side of the tube and "sucking out" the DNA. The DNA solution was then extracted with n-butanol as described above (Section 2.1.6.2), and then ethanol precipitated as described in Section 2.1.5.2.).

2.1.8. Enzyme Digestion, Electrophoresis, and Southern Analysis

2.1.8.1. Restriction endonuclease reactions *of genomic DNA and plasmid DNA.*

Samples of genomic DNA and plasmid DNA, were digested with a variety of restriction enzymes (New England Biolabs).

1. Genomic DNA: 10 μg of DNA were digested under the conditions specified by the manufacturer of the enzyme, using the specified buffer. Usually 4 units of enzyme was allowed for each μg of DNA. The digest proceeded overnight at the temperature recommended by the manufacturer, which differed depending on the enzyme used. The reaction was stopped by the addition of EDTA to a final concentration of 10 mM.

2. Plasmid DNA: Plasmid DNA was digested to release the genomic insert for quantitation, and to be used as a probe. For quantitation, 1 μl of plasmid DNA was digested with 4 units of the appropriate enzyme and quantitated on a Shimadzu spectrophotometer. Routinely, 1 μg of plasmid/insert DNA was digested with 4 units of enzyme, in the buffer recommended by the manufacturer. The reaction proceeded for two to sixteen hours, at the temperature specified by the manufacturer. Genomic insert DNA was separated from the plasmid DNA by electrophoresis (Section 2.1.8.2.) and recovered from the agarose as described in Section 2.1.6.

2.1.8.2. Gel electrophoresis (Maniatis et al., 1982)

DNA samples were electrophoresed in 0.8% - 1.2% agarose gels, in 1 x TBE. Gels were run overnight for genomic DNA or for several hours for plasmid DNA, depending on the expected size of fragments, and the relative size of fragments that needed to be separated. The voltage used varied with the length of time the gel was to be run. The standard DNA marker used to determine fragment sizes was *Eco*RI digested Spp1 phage (Bresatec). DNA samples were prepared by adding 0.1 volume of 10 x gel-loading buffer (0.1 M Tris HCl (pH 8.0), 0.2 M EDTA (pH 8.0), 2% Sarcosyl (w/v), 20% Ficoll 400 (w/v), 0.1% Bromophenol blue (w/v), 0.1% Xylene cyanol (w/v)). The DNA was staining with Ethidium Bromide (EtBr) (0.0025 mg/ml) and visualised under UV light. Complete digestion of genomic DNA showed as a smear on an agarose gel. Plasmid digestion gave bands of expected size on the EtBr stained gel. Gels were photographed to document the positions of the DNA size markers for later determination of DNA fragment sizes from autoradiographs, and for the determination of the plasmid and insert sizes prior to being extracted from the agarose. The size of the fragments, both genomic and plasmid, were estimated using the program DNASIZE (Suthers, 1992).

2.1.8.3. Southern analysis (Southern, 1975)

After electrophoresis, the genomic DNA was transferred from the agarose gel to nylon filters using modifications of the Southern Blot method described by Southern (1975). The conditions of transfer were dependent on the membrane used, and were those specified by the manufacturer of the specific membrane. The membranes used were a nylon nitrocellulose filter (Hybond N+, Amersham), and Genescreen Plus (Dupont). Transfer proceeded overnight for the best results.

2.1.9. ^{32}P-Labelling of DNA Fragments

^{32}P is a carcinogen, and great care was taken when handling the radioactive material. Double stranded DNA fragments were ^{32}P-labelled using the Multiprime DNA labelling system (Amersham Kit 1601) (Random primer extension, Feinberg and Vogelstein, 1983) or by Nick translation (Amersham Kit N5000) (Rigby *et al.*, 1977) to incorporate ^{32}P-dCTP (Amersham) (1 mCi/ml) according to the appropriate protocol described in the Amersham product handbooks. Between 20-30 *ng* of DNA was usually radioactively labelled using the random primer extension method, and approximately 250 *ng* of DNA was used for nick translation reactions.

To improve the efficiency of the labelled markers during the hybridisation process, in some instances unincorporated radioactive nucleotides were removed from the labelled DNA solution by running the sample through a Sephadex G-50 (Pharmacia) column. The first peak detected by the radiation monitor was collected since it contained the ^{32}P-labelled DNA.

For some labelling, the nucleotide incorporation was also tested, to ensure the reaction was working efficiently. One of two methods were used for this:

(i) Dried filter discs, spotted with some of the labelling reaction, were placed in 10 ml toluene scintillant, and monitored on a scintillation counter.

(ii) The second protocol used TLC plastic sheet cellulose-F precoated paper (Merck), and is described in Chapter 3.

2.1.10. Hybridisation Methods

^{32}P-labelled DNA probes were hybridised to filters of genomic DNA, using a modified method described by Sealey *et al.* (1985). Autoradiography was used to locate the position of any DNA complementary to the radioactive probe. This technique can be used to detect specific sequences in both genomic and cloned

DNA. Two methods were employed, their major differences being in the temperature (either 42°C or 65°C) at which hybridisation was performed.

2.1.10.1. Hybridisation at 42° C

Nylon membranes were prehybridised at 42° C for 2 hours in hybridisation solution (5 x SSPE, 50% deionised formamide, 1% SDS, 7.5% dextran sulphate, 0.1 mg/ml sonicated denatured salmon sperm DNA (Sigma)). The heat denatured radiolabelled probe was added to the hybridisation solution, and incubated at 42° C overnight with gentle agitation. The nylon filters were removed from the solution and washed sequentially in 2 x SSC, 0.5% SDS at 65° C for thirty minutes, and 0.1 x SSC, 0.1% SDS. The time and temperature of the second wash were adjusted according to the background "noise" level of the filters when tested by a radiation monitor after the first wash. After washing, the filters were exposed to XO-mat XK-1 film (Kodak) at -80° C in the presence of intensifying screens. The duration of the exposure depended upon the strength of signal registered by the radiation monitor after the second wash.

The radiolabelled probe was later "stripped" from the nylon membrane to enable future reprobing, by washing first in 0.4 M NaOH for thirty minutes at 42° C, and then in 0.2 M Tris HCl (pH 7.5), 0.1% SDS, 1 x SSC for thirty minutes at 42° C. The membranes were then dried and stored for reuse.

2.1.10.2. Hybridisation at 65° C

Nylon membranes were pre hybridised in 1 M Na_2HPO_4/NaH_2PO_4 (pH 7.0), 7% SDS at 65° C for at least fifteen minutes, with gentle agitation. The ^{32}P-labelled probe was boiled for ten minutes to denature and added to the prehybridising filter and solution. Hybridisation was performed at 65° C overnight. After hybridisation, the filters were washed, exposed to X0-mat XK-1 film (Kodak) and stripped as described in Section 2.1.10.1.

Part II: Linkage Analysis

2.2.1. Introduction

During the course of this project, two main linkage analysis packages were utilised. Firstly, the LINKAGE ANALYSIS PACKAGE of programs (Version 5.1) was utilised to develop a multipoint linkage map in the region of *FRA16B*. The hardware used for analysis was an IBM-compatible Personal computer This program package and its implementation will be described. The LINKAGE package is an analysis program for performing risk analysis and for the calculation of lod scores on disease pedigrees. To analyse a pedigree segregating for the fragile site, *FRA16A*, the LINKAGE programs (Version 5.1) were used on a Sun workstation, enabling a more powerful analysis than was possible using a PC.

As more genetic loci were incorporated into the analysis, the computer program CRI-MAP (Version 2.4) was implemented. This was used for subsequent multipoint linkage analyses of the region on chromosome 16 encompassing the CLN3 gene, and for construction of the maps of the entire length of chromosome 16. This program was designed to analyse as many as 100 loci concurrently, enabling the entire chromosome 16 database to be incorporated into the linkage map of human chromosome 16. This program will also be described.

The CEPH Database Management program was used for the maintenance of the probe/polymorphism data and the genotype data generated from the CEPH pedigrees. This package is briefly described. Other utility programs required for the transformation of the genotype data into linkage analysis format and some for use with the analysis packages are also described.

2.2.2. CEPH Database Management Package

CEPH provide a database management program that enables efficient storage of all the genotype information and the probe/enzyme records. Versions 4.0 with

upgrades to 7.1 of the database management programs were used during the course of this project. The package includes facilities to generate the pedigree files and data files required for the analysis programs, and enables the researcher to input new genotype information and edit existing information. It provides the service to extract new genotype information and return it to CEPH, where it is incorporated with other new data from collaborating investigators; they then return it to the investigators, for map construction. During the course of this project the database was upgraded through stages from Version 4.0 to Version 6.0.

All data included in the CEPH database are automatically checked and corrected as necessary, prior to the analysis for the following inconsistencies:

(i) parental alleles incompatible with those of children,

(ii) discordances between the genotypes in the grandparents and grandchildren in cases in which the parental data is missing, and

(iii) intralocus recombinants.

2.2.3. LINKAGE Analysis Program

The LINKAGE ANALYSIS PACKAGE (LINKAGE) embodies two groups of programs. The first group, LINKAGE ANALYSIS PROGRAMS, consist of a "core" of linkage analysis programs, designed to perform maximum likelihood estimation of recombination rates (using iterative algorithms), calculation of lod scores and lod score tables, and the analysis of genetic risk. Activity at the "core" level requires a comprehension of how these programs interact, and a knowledge of how to prepare and adapt the appropriate data files.

The second group, LINKAGE SUPPORT PROGRAMS (LSP), consists of a sequence of programs that form an interactive "shell" around the "core" programs, providing a friendly, easy-to-use interface to the "core" programs. In theory, the shell provides the facility for a user to use the analysis programs without specific

knowledge of the programs or data files that support them, and indeed of computing itself.

The LINKAGE ANALYSIS PROGRAMS can be further divided into two groups. The first group of programs are optimised for the analysis of the three generation CEPH pedigrees with codominant marker loci, and are primarily intended for the rapid construction of genetic maps from the reference family data. The second group are for use with general disease pedigrees with codominant markers and disease loci; both groups of programs allow the analysis with either sex-linked or autosomal loci. The input to the LINKAGE programs is divided into two files. The "PEDFILE", as it is internally known by the linkage program, contains information regarding pedigree structure and the genotype data. Locus description, recombination rates, and gene order are incorporated in the "DATAFILE", as it is internally known by the LINKAGE programs. The pedigree and genotypic data must be processed prior to analysis by a series of preparatory programs that accompany the analytical programs in the LINKAGE package.

The method of analysing linkage relationships using the LINKAGE ANALYSIS PACKAGE can be broken into three phases: the data preparation and management phase, the data analysis phase, and the linkage results report generation phase.

The data preparation and management phase involves the generation of genotype data (described in Part I of this chapter), and its management and storage, which is performed by the CEPH database programs for the CEPH pedigrees (as described in Section 2.2.2.). The information for the *FRA16A* pedigree was stored in the pedigree file constructed prior to the analysis.

The next phase is the data analysis phase, which involves the construction of command files, delineating the analysis procedure and then running these files. The analysis itself is controlled by the support program LINKAGE CONTROL

PROGRAM (LCP), and builds the command files to perform the linkage analysis. This phase is described in the Sections 2.2.3.3., 2.2.3.4., 2.2.3.5., and 2.2.3.6. The final phase of linkage analysis is generating a report and interpreting the data. The LINKAGE REPORT PROGRAM (LRP) converts the output from the analytical programs into an easily read format. This program is described in Section 2.2.3.4.

2.2.3.1. Computer hardware and operating systems

The LINKAGE programs (Version 5.1) (Lathrop *et al.*, 1984, 1985; Lathrop and Lalouel, 1988) are available for IBM-PC computers, Vax computers and for Unix-based systems, like the Sun workstation. The LINKAGE analysis programs designed to run on an IBM compatible computer are written and compiled in Turbo-Pascal Version 5 (Borland International Inc) and the LINKAGE control programs are written in C (Borland International Inc). The analysis programs designed to run on the Sun SPARC station IPC are written in Unix-based Pascal, and the LINKAGE control programs use the language C. The steps involved to establish a working copy of these programs are detailed below. The LINKAGE programs were supplied as executable files and as source code, for both types of computers.

LINKAGE Programs run on an IBM personal computer:

The computer used was an NEC Powermate IV operating under MS-DOS (Version 3.2), with 640 K of core memory, a 40 Mb hard disc, an 80287 numerical co-processor, and 512 K of extended memory.

LINKAGE Programs run on an Sun Workstation:

The workstation was a Sun SPARC station IPC with 12 Mb of core memory and a 207 Mb hard disc. The operating system was SunOS Operating System, release 4.1.1.

2.2.3.2. Compilation and system constants

As stated before the LINKAGE programs run on the Sun Workstation did not need to be compiled, thus only the compilation on a personal computer will be discussed. This means only the compilation of those LINKAGE programs associated with the analysis of the three generation CEPH pedigrees will be discussed, and not those associated with disease pedigree analysis.

The compiler used was Turbo-Pascal Version 5 (Borland International INC). This compiler was presented in two formats. One format provided an integrated environment using a windows based system. The second format was a command-line compiler with the advantage of using less memory. It was possible to compile larger programs with the command-line compiler where it was not possible to do so using the integrated environment.

To compile the three analytical programs used for CEPH pedigree analysis, CMAP, CILINK, and CLODSCORE, the following sequence of files is called (CMAP is used as an example):

```
SWTHG.pas
program cmap(output,tempdat,tempped,outfile,stream);
{use overlays}
THGC,
THG1,
THG2,
THG3,
CMP4;
```

The first file to be called is a file called SWTHG.pas, which contains the directives for the compilation process. Next are four units that are common to all the main analytical programs, THGC.pas, THG1.pas, THG2.pas, and THG3.pas, which contain the general information for performing linkage analysis. Following these four units, subprograms specific to each main analysis program are called, CMP4.pas (for CMAP), CLK4.pas (for CILINK), and CLD4.pas (for CLODSCORE), and executable files for CMAP, CILINK and CLODSCORE are constructed.

To compile the LINKAGE programs, the files SWTHG.PAS and THGC.pas need to be modified. The various options available in the SWTHG.pas file are detailed in Ott's notes to version 5.03. It is used to compile the main linkage analysis programs. For SWTHG.pas, the options were as follows:

```
{$DEFINE CILINK}          {program name; CILINK, CMAP, or
                          CLODSCORE}
{$DEFINE autosomal}
{ DEFINE sexlinked}
{$DEFINE extended}        {Precision; single or double or extended or turbo}
{$O-}                     {Overlays; + or -}
{$N+}                     {Use numeric coprocessor; + or -}
{$E-}                     {Emulate coprocessor; + or -}
{$R+}                     {Range checking; + or -; do not change without
                          good reason;
```

In Turbo-Pascal, brackets {} are used to define either a comment or a critical command for the compiler. { DEFINE ....} is a comment for the programmers benefit; {$DEFINE ....} is a compiler directive. The "$" activates the statement when the program is compiled, and if it is in the wrong place the analytical programs will not compile properly. The analytical program to be compiled must be identified in the first line above. The precision switch (the fourth line above) determines the number of bytes used for a real variable. Using "extended" precision uses more bytes and increases computing time, but reduces the chance of an underflow[1], whereas using "turbo" precision will produce much faster running code, but increases the chance of an underflow occurring. Since a numeric coprocessor was installed, the precision switch was set at DEFINE "extended" (or sometimes "double"). In the CILINK program, the precision also sets the variable *nbit* representing the mantissa length of real variables. The following options were available:

---

1. Underflows occur when a real number becomes smaller than a critical limit, for example, $10^{(-38)}$. In Turbo Pascal, when an underflow has occurred in a variable, its value will simply be set equal to zero and computation continues, which may lead to apparent errors and inconsistencies that are difficult to pin point. Underflows were largely avoided by choosing the "double" or "extended" type of real variable which has a low critical bound.

Switch setting (DEFINE....)	Variable length (bytes)	Mantissa length (bits)
single	4	23
turbo	6	39
double	8	52
extended	10	63

For less than 10 loci, extended precision was used. If more loci were to be analysed, turbo precision, or even single precision, was used.

If the program was to be compiled to overlay files (i.e. {$O+}) the overlay path in the corresponding program had to be altered, as detailed in Ott's notes for version 5.03. The overlay files were placed on a RAMDISC in the extended memory buffer as detailed in the DOS manual. For example, in CMAP.pas, the following line needed to be included:

```
ovrinit('E:\CMAP.OVR');
```

where E was the disc name for the RAMDISC.

The running constants for each analysis are contained in the file THGC.pas; following are the main constants that needed to be altered:

```
{ SOME USER DEFINED CONSTANTS }
      dostream = true;          {STREAM FILE OUTPUT}
      maxlocus = 9;             {MAXIMUM NUMBER OF LOCI IN
                                   MAPPING PROBLEM }
      maxsystem = 9;            {MAXIMUM NUMBER OF LOCI IN
                                   ONE FAMILY AFTER
                                   TRANSFORMATION}
      maxind = 1405;            {MAXIMUM NUMBER OF
                                   INDIVIDUALS}
      maxped = 40;              {MAXIMUM NUMBER OF
PEDIGREES}
      maxrectype = 110;         {MAXIMUM NUMBER OF DIFFERENT
                                   RECOMBINATION PATTERNS}
      window = 0;
{CONSTANTS DETERMINING ARRAY SIZES}
      segstore = 500;
      hapstore = 1300;
      boolstore = 950;
{LOCUS LIMITS}
      maxall = 8;               {MAX NUMBER OF ALLELES AT A
                                   SINGLE LOCUS;
```

It was necessary to set the constants *maxlocus*, *maxind*, *maxped*, and *maxrectype* at the beginning of the file for each analysis specifically, and compile the respective analytical programs each time. The analysis would crash if a range check error

occurred with an error message "Array subscript out of range" or similar, where one of the constants written at the beginning of the THGC.pas file was too small for the problem to be analysed; this constant would need to be altered and the program recompiled. If the analysis was not carefully designed, insufficient memory would also occasionally cause a LINKAGE analysis to crash while running, with the message "Heap/stack overflow[2]" or "Insufficient memory". This was due to limitations placed on the computer memory capacity, applied by the DOS environment. The memory limitations were overcome to a certain extent by setting each of the constants at their minimum. For example, the number of loci was set to exactly that used in the current analysis. The problem was also overcome by:

(i) ensuring that all unnecessary programs were removed from memory,

(ii) incorporating into the analysis only those pedigrees that were informative for at least two of the markers to be analysed, thus reducing the number of pedigrees being studied to a minimum,

(iii) reducing all real variables from "extended precision" to "double", "turbo" or even to "single precision" in the SWTHG.pas file, reducing the memory required to store real numbers. This also reduced the accuracy of the analysis to some extent, but not sufficiently to cause concern.

(iv) Finally, by utilising the extra 512 K of extended memory installed in the computer, the LINKAGE programs were compiled to overlay files. This had the effect of slowing the analysis, but it extended the memory capacity.

---

2. Heap overflow occurs if there is not enough free memory to hold all the data. There are four possible solutions: 1. Reduce all program constants to the smallest possible values, 2. Reduce the compiler switch `DEFINE double' to `DEFINE single', 3. Set the compiler switch R-, and do not check for "out of range" errors. This should be done as a last option, as if there are any errors, the computer will freeze if an array bound is exceeded.

Thus, an analysis program was compiled by checking the analytical program setting in the SWTHG.pas file first, and then verifying the settings for the constants in the THGC.pas file.

Turbo Pascal limits the amount of memory available to all arrays in a program; this is a Turbo Pascal limitation, and not a DOS problem, and hence cannot be resolved by the addition of extra memory. This limitation meant that often the program would not compile at the size of the constants set, and the size of the analysis needed to be reduced. As more loci became available to be included in the analysis, these memory limitations and size restrictions occurred more often, and could not be overcome. Thus, the change to the CRI-MAP analysis package was crucial to accommodate analyses of increasing complexity.

The LINKAGE analysis package used on the Sun workstation IPC was then only used to analyse the pedigree segregating for FRA16A, and not for any general CEPH pedigree analysis. These programs were also distributed as executable files and as source code, although, for the scope of this project it was never necessary to recompile these programs. However, the process is essentially the same, and is outlined in the postscript files that are distributed with the package.

2.2.3.3. Program description: CEPH pedigrees and general pedigrees

The LINKAGE package programs for analysis of both the CEPH three generation pedigrees and the general disease pedigrees are essentially the same, but the formal structure of the CEPH pedigrees enables a simpler analytical routine program. The structures of the input files (a "PEDFILE", comprising the pedigree data; and a "DATAFILE" including the description of the loci, locus order, recombination frequencies, and gene frequencies) are the same, and hence will be discussed together, as will the description of the analytical programs. The main analysis files for the CEPH pedigrees are CMAP, CILINK and CLODSCORE,

and for the disease pedigrees, the analytical files are LINKMAP, ILINK, LODSCORE and MLINK.

The first step in the LINKAGE analysis is to create the pedigree "PEDFILE" file. In the case of the CEPH pedigrees, the CEPH database management programs create a pedigree file and data file automatically as described in Section 2.2.2., with no further modification required. For general pedigrees, the pedigree file is constructed using a text editor. For both types of programs, the format of the pedigree file should be as follows, running in columns across the page:

- Pedigree number
- ID number of a given individual (needs to be sequential)
- ID number of that individual's father (0 if unknown)
- ID number of that individual's mother (0 if unknown)
- Sex of individual: 1= male, 2= female
- phenotype at locus 1
- phenotype at locus 2, etc

Each item has to be separated from the others by at least one space, and there must be no spaces between the rows, or at the end of the file; spaces would cause the analysis to abort, with no indication of what the problem was!

After the disease pedigree file has been constructed, the file needed to be converted into the format for the LINKAGE analysis programs. The program MAKEPED performed the appropriate transformations.

The data file structure for both groups of pedigrees is also the same. Internally this file is called DATAFILE, which includes a description of the loci, the locus order, and recombination rates. There are a number of ways of defining loci for linkage analysis. For the programs for analysis of the three generation families, the codominant markers are described in terms of "binary factors" or "numbered alleles", providing information about a individual's genotype at a locus. *Binary factors* are a series of 1's and 0's, indicating the presence or absence of a phenotype factor, and are entered into the data file as a binary string. For example, in a two allelic system the allelic codes are:

1 0 (allele 1)
0 1 (allele 2)

Numbered alleles are codominant alleles at a single locus, and the numbers run consecutively from 1 to the maximum number of alleles observed. Thus the phenotype is represented by two allele numbers corresponding to a genotype. For example, 2 5 (alleles 2 and 5 are present), 0 0 denotes unknown. Homozygotes are given two identical numbers. Either system may be used but it is essential to use the same system in both the PEDFILE and the DATAFILE.

Affection status is used only for disease pedigree analysis, or in the case of this project, to specify presence or absence of the fragile site, *FRA16A*. It indicates the presence or absence of a disease (or fragile site), and is described by a numbered code. For example, individuals that could be induced to display the fragile site, *FRA16A*, were coded with the affection status of 2, indicating that they were "affected", 1 represented unaffected, and 0 implied that the individual's phenotypic status was unknown, because the individual was not cytogenetically tested for the presence of the fragile site. Penetrance (probability of being affected or of carrying the fragile site, given a certain genotype) may be incomplete in some individuals, and so liability classes can also be used. There were two liability classes assumed for the autosomal folate sensitive fragile site, *FRA16A*. Females were specified as fully penetrant, and males were specified at only 50% penetrance (Sherman and Sutherland, 1986).

The loci described in the DATAFILE should correspond the loci in the equivalent PEDFILE. The file consists of general information concerning the number of loci, the risk locus, sex-linked or autosomal, mutation rate, and the order of loci. It also contains information on the recombination frequencies between the loci and program specific information. The DATAFILE for the CEPH pedigree loci is created by the CEPH database programs, and the data file for the general pedigrees is created by the program PREPLINK.

2.2.3.4. LINKAGE programs for the CEPH pedigrees

The LINKAGE package for the CEPH pedigrees consists of 4 programs for analysing the genotype information; CFACTOR, CMAP, CILINK, and CLODSCORE. The calling sequence for these programs is:

input (pedfile, data file)
-> CFACTOR ->
output (tempped, tempdat)
-> CILINK or CMAP

CFACTOR applies factorisation and transformation rules to the genotypes from the PEDFILE and DATAFILE and inputs the modified information to the files *tempped* and *tempdat* respectively, before CLINK or CMAP are evoked. Three constants are defined in CFACTOR that can be altered to make the analysis more efficient, called elimb, elimu and esame. These constants were incorporated into CFACTOR to simplify the calculations by disregarding data at loci in families in which either or both the parents genotypes are unknown and cannot be uniquely inferred.

elimb = true:	data for a locus are eliminated if the genotypes of one or both parents cannot be uniquely inferred.
elimu = true	data for a locus are eliminated if the genotypes of both parents cannot be uniquely inferred
esame = true	data for a locus are eliminated if all the children are identical homozygotes, and if one parent is typed, he or she is also a homozygote.

The default setting is all three constants are false, in which case, no data is eliminated.

CILINK is a program for maximum likelihood estimation of recombination fractions for an arbitrary number of markers loci. Given a specific order, CILINK determines the maximum likelihood and the recombination

fractions at that maximum. Sex specific differences in the recombination rates can be incorporated.

CMAP is a program for calculation of location scores of one locus against a fixed map of other loci.

CLODSCORE is a program for the efficient calculation of LOD scores in the three generation families. This program reads information directly from PEDFILE and DATAFILE, and so does not run through CFACTOR prior to the calculations.

There are three program constants for CILINK and CMAP: *segstore*, *hapstore*, and *boolstore*. These constants define limits of array sizes, are specified at the start of the THGC file, and need to be changed to fit the analysis and the program recompiled. The program halts if these array sizes are exceeded. However, these constants can not be determined prior to the analysis; if the program crashes due to one of then being out of array size, it specifies the size it should be, and the THGC.pas file can be appropriately altered. Making the constants as low as possible enables a program to be compiled with the maximum number of loci and alleles.

2.2.3.5. LINKAGE programs for general pedigrees

The LINKAGE package for analysing general pedigrees of arbitrary structure consists of 5 programs for analysing the genotype information; UNKNOWN, LINKMAP, ILINK, MLINK, and CLODSCORE. The calling sequence for these programs is:

input (pedfile, datafile)
-> UNKNOWN ->
output (tempped, tempdat)
-> LINKMAP, ILINK, or MLINK

UNKNOWN This program takes the PEDFILE and DATAFILE, checks for data inconsistencies, and infers unknown genotypes. This program replaces CFACTOR for the CEPH pedigree analysis.

LINKMAP is a program that estimates the most likely position of an unknown locus on a known genetic map, calculating the location score. This was the most appropriate program used for localising *FRA16A* on the multipoint map of the region encompassing this fragile site.

ILINK is a program that iteratively calculates the maximum likelihood of a given locus order. It is used to determine the probable order and recombination fractions for a number of loci simultaneously, similarly to CILINK.

MLINK is used to calculate the LOD score for a given recombination fraction between two loci (Eg for preparing a LOD score table) and for risk analysis where required.

LODSCORE Unlike the other analysis programs for the general pedigrees, LODSCORE is not preceded by UNKNOWN. It calculates two-point LOD scores between pairs of loci. LODSCORE is used to check the recombination fraction between all pairs of loci before moving on to a more complex analysis. The DATAFILE needs to be modified slightly compared to the format for input into the other programs.

2.2.3.6. LINKAGE support programs: LCP and LRP

As mentioned in the introduction, the LINKAGE ANALYSIS PACKAGE is made up of two groups of programs. The first group are those described above, the analysis programs. The second group, the LINKAGE SUPPORT PROGRAMS control all the LINKAGE analysis programs, by setting up command files at the user's direction. The two programs are **LINKAGE CONTROL PROGRAM** (LCP) and **LINKAGE REPORT PROGRAM** (LRP).

The **LINKAGE CONTROL PROGRAM** (LCP), is an interactive menu-driven program for building command files to perform linkage analysis. The command file may take several minutes to run the analysis, or many hours. LCP is very efficient in generating a complete linkage analysis involving many stages, and

different analytical programs. There is the option of selecting different loci to be analysed, and to set parameter values such as recombination fractions. The choices that are made for the analysis are saved together in the command file (by default this file is called pedin.bat).

The LINKAGE REPORT PROGRAM (LRP), is also a menu-driven program for displaying and formatting the stream output of the linkage programs. The results of LRP can be exported to a report file.

2.2.4. CRI-MAP Analysis Program

CRI-MAP was developed to allow rapid, largely automated construction of multilocus linkage maps and assess the support relative to alternative orders, generate LOD tables, and detect data errors.

The goal of multilocus linkage analysis is to find the locus order having the highest likelihood, and identify alternative orders with comparable likelihoods. Because it is impossible to consider all possible orders for a large set of loci, it is necessary to adopt a strategy which makes decisions on the basis of subsets of loci. This approach may result in some loss of information, but it should not result in any bias in estimating recombination fractions (Green *et al.*, 1989). CRI-MAP deduces missing genotypes where possible, and incorporates information from individuals of unknown genotype at a given locus only when their genotypes can be inferred with certainty at that locus, and then computes the likelihood (as defined by Ott, 1985) based on analysis of the known or deduced genotypes. Genotypes where only one allele is known or deducible, are also utilized in the calculations as far as possible. For missing data at a particular locus in an individual, where possible genotypes at that locus may include a homozygous genotype, then CRI-MAP treats all meioses in that individual as uninformative for that locus. A likelihood analysis by LINKAGE would consider each possible genotype at the locus, assign relative probabilities to each based on the genotypes

of ancestors and descendants, and compute a likelihood which is the weighted sum of the likelihood expressions for each particular choice of genotype. CRI-MAP thus ignores some of the information available, but the loss appears to be small (Green *et al.*, 1989), and allows the analysis to be performed more rapidly.

CRI-MAP does not make use of allele frequencies; if an original parents' genotype is missing, a full likelihood analysis would use the population allele frequencies to assign probabilities to various possible genotypes, thus influencing the probability, and the likelihood, that the allele in any child of the original parent is derived from that parent. CRI-MAP deduces the parental origin for any allele in a child of an original parent from other genotype information if possible, but otherwise assigns equal probability to the two possible parental origins for the allele. Little information is lost by this procedure, except when the allele is rare; using codominant marker loci, little information is lost. Since allele frequencies estimated in different populations vary dramatically between populations, an analyses which makes no assumptions concerning allele frequencies such as that given by CRI-MAP, may be preferable.

Although CRI-MAP was originally designed to handle co-dominant loci (eg. RFLPs) scored on pedigrees "without missing individuals" such as CEPH or nuclear families, the algorithms in CRI-MAP have been extended to work on general pedigrees with single dominant markers of disease systems. In the course of this project however, CRI-MAP was only used with the CEPH pedigrees. CRI-MAP and the LINKAGE package are linked only in that the same pedigree files are used, with some slight modifications.

CRI-MAP is based on a computationally efficient algorithm for computing the maximum likelihood genetic map under the assumption of no interference. This algorithm belongs to a special class of maximum likelihood algorithms called EM (Expectation and Maximization) algorithms. The EM algorithm offers a powerful

approach to obtaining maximum likelihood estimates from incomplete data (Dempster *et al.*, 1977; Wu, 1983).

The EM algorithm has an interesting mathematical property that the likelihood never decreases with successive iterations (Dempster *et al.*, 1977; Theorem 1), and it converges to a point which is guarantied to be a local (possibly global) maximum (Wu, 1983). In addition, the EM algorithm has the practical advantage that it is often easy to program because the likelihood function for the complete data specification typically has a simple form. Finally, it is frequently observed that the estimates usually converge to the maximum from a wide range of initial values (Lander and Green, 1987).

2.2.4.1. Computer hardware and operating systems

CRI-MAP is distributed as source code only, in the language C. The program needs to be compiled with a C compiler. The one used was a Unix-based compiler built in to the Sun OS 4.1.1 operating system. The program requires at least 1 Mb of memory to perform an analysis incorporating 10 loci. Thus, it is desirable to run the package on a computer with at least 1 Mb of memory, preferably much more. CRI-MAP analyses were routinely run on a Sun SPARC station IPC, which has 12 Mb of memory, and a 207 Mb hard disc. In the initial phase of the project, attempts were made to run CRI-MAP on a PC, with limited success, due to limited hard disk space and computer memory.

2.2.4.2. File Structure

CRI-MAP is made up of a series of options within a main program, all of which use a similar file structure, unlike LINKAGE, which is comprised of a series of programs.

To run CRI-MAP, there are four main files that are required for input into the analysis; the file structures of these input files will be discussed.

These four files are:

.gen - the raw genotype data file

.par - the parameter file

.dat - the processed data file

.ord - the orders file, or "orders database".

The prefix to these files is in the format *chrx.file*, where the *x* represents the number of the chromosome, and file represents the extension (in this case it would be *chr16.gen*, for example). Each of these files are in ASCII format, and can be edited with a text editor. To use CRI-MAP for linkage analysis, it is only necessary to understand the file structure for the *.gen* and *.par* files, which are discussed in detail below. The file structures for the *.dat* and *.ord* files need not be understood, and will not be discussed in great detail.

.gen file

The *.gen* file contains the raw pedigree information, and must be supplied, prior to running the option **prepare**. This file is created from the ped.out file generated by the CEPH database programs (Section 2.2.2.) by a program called ***mkcrigen*** (Section 2.2.5.), written by Santosh Mishra in St Louis (Personal communication). The data items in this file are as follows:

{# of families}
{# of loci (*n*)}
{locus name 1} *{each locus is named,*
{locus name 2} *the name consisting*
" *of at most 15*
" *characters}*
{locus name *n*}

For each family:
{Pedigree ID}
{# of members}
For each member in the family (in columns running across the page):
ID number of a given individual (needs to be sequential)
ID number of that individual's mother (0 if unknown)
ID number of that individual's father (0 if unknown)
Sex of individual {0=female, 1=male, 3=unknown}
phenotype at the *n* loci (starting on a new line):
{locus1 allele1} {locus1 allele2} {locus2 allele1} {locus2 allele2}
{etc}.

The pedigree structure must be completely specified, including the original parents, who's IDs are coded as 0. Alleles must be represented by integers, with missing alleles scored as 0. Each locus name may consist of up to 15 characters, with no embedded blanks. Each family ID may be any character strings not containing blanks.

.par file

The *.par* file is constructed from the *.gen* file by running the option **prepare**. This file must be constructed before running any of the other CRI-MAP options. Once the file is created, it may be edited using a text editor, as opposed to running **prepare** every time the analysis is to be altered. The *.par* file contains the running information for the analysis, and the format of this file is:

*dat.*file *chrx.dat* *
*gen.*file *chrx.gen* *
*ord.*file *chrx.ord* *
nb_our_alloc 3000000 *
SEX_EQ 1 *
TOL 0.01 *
PUK_NUM_ORDS_TOL 6 *
PK_NUM_ORDS_TOL 8 *
PUK_LIKE_TOL 3.0 *
PK_LIKE_TOL 3.0 *
use_ord_file 0 *
write_ord_file 1 *
ordered_loci {index # of 1st ordered locus} {index # of 2nd ordered locus}..{index # of *n*th ordered locus} *
inserted_loci {index # of 1st inserted locus} {index # of 2nd inserted locus}..{index # of *n*th inserted locus} *
hap_sys {index # of 1st locus in system} {index # of 2nd locus}... *
hap_sys0 {index # of 1st locus in system} {index # of 2nd locus}... *

The constant *nb_our_alloc 3000000* is the initial memory allocation (in bytes), and if additional memory is needed during the analysis, it is allocated automatically. The default value was sufficient for even the large number of loci to be analysed on chromosome 16.

Using CRI-MAP, it is possible to specify an analysis where the recombination between the sexes is equal and constant (*SEX_EQ 1*), or allow a variable recombination rate between the sexes (*SEX_EQ 0*).

The tolerance for determining convergence of the layered EM algorithm can be specified, such that when $\log_{10}$likelihoods from successive "phase unknown" iterations increase by less than this amount (*TOL 0.01*), iteration terminates. A more stringent analysis can be performed by setting this value to 0.001 instead; the linear nature of EM convergence guarantees that, if the estimates had not converged with TOL = 0.01, then a substantial improvement in likelihood should be apparent with the more stringent tolerance.

The next two constants in the *.par* file, *PUK_NUM_ORDS_TOL 6* and *PK_NUM_ORDS_TOL 8* refer to the maximum number of orders allowed in the phase unknown and phase known parts of a **build** analysis, respectively, in the current map.

PUK_LIKE_TOL 3.0 and *PK_LIKE_TOL 3.0* are the constants that set the tolerance for discarding locus orders, such that, if the $\log_{10}$likelihood of an order is less than the $\log_{10}$likelihood of some other order for the same loci by an amount exceeding these constants, that order is not reported. Used only by **build**, **all** and **flips***n*. With **twopoint**, LOD values are displayed only for locus pairs whose LOD exceeds *PUK_LIKE_TOL*. *PK_LIKE_TOL* applies only to analysis of the phase known data in the option **build**.

When the parameter *use_ord_file* is set to 1, the orders generated by the options **all** and **flips***n* are prescreened against the respective *.ord* file created from a previous **build** analysis, prior to computing likelihoods, thus effectively eliminating orders incompatible with the orders database. When the parameter is set at zero, the information in the orders database is not used.

write_ord_file applies only when the **build** option is being executed. If the parameter is set to 1, the results of the current **build** run are used to update the orders database, and when it is 0 the database is not be updated.

The next two lines in the *.par* file refer to the loci to be included in the analysis; the *ordered_loci* and the *inserted_loci*. In all options (except **twopoint**) the "ordered_loci" are assumed to be in their known, unique order; the remaining loci, the "interted_loci", are those loci to be placed in the framework defined by the *ordered_loci*.

hap_sys and *hap_sys0* represent a list of haplotyped systems which are to be grouped together in an analysis. When the parameter *use_haps* is 1, the primary locus (that locus first in the list) is used whenever ordered sets of loci are constructed, and the secondary loci in a system (the rest of the loci in the list), "tag along". However, if a locus is being inserted into the map, or a new map is being created by permuting a collection of loci, only the primary locus will be utilized, and once the order is constructed, the secondary loci are inserted immediately following the corresponding primary locus to "fill out" the order, prior to calculating likelihoods and map distances. If *hap_sys0* is specified, distances within the system are forced to 0. If the *use_haps* is set to zero, any haplotyped systems specified in the *.par* file are ignored.

It is also possible to fix the recombination fraction between adjacent loci for the options **fixed** and **chrompic**, for sex-average or sex-specific analyses, by using the parameter *fixed_dist*.

.dat file and *.ord* file

These files are also constructed from the *.gen* file when the option **prepare** is run. It is not necessary to understand the file structures when performing the analysis with CRI-MAP; the structures of these files are outlined in Green *et al.*, 1990.

The *.ord* file is initialized by **prepare**, but required only for the map building options (**build**, **instant**, and **quick**).

2.2.4.3. Program Options: CEPH pedigrees

There are 10 program options for CRI-MAP. These are **prepare**, **build**, **instant**, **quick**, **all**, **fixed**, **flipsn**, **chrompic**, **twopoint** and **merge**, and are described below.

prepare must be run before any of the analysis options, to create the *.par*, *.dat*, and the *.ord* files. This option enables the user to specify various parameter values that are detailed in the *.par* file above, and select the program option to be used for the analysis.

The option **build** constructs a map by the sequential incorporation of loci, and is the main option used for map construction. The output from **build** consists of a uniquely ordered set of loci, together with a list of possible placements and associated likelihoods for each remaining locus with respect to the uniquely ordered loci. At the beginning of the **build** analysis, the map consists of the "ordered loci" specified by the user in the *.par* file. In the absence of any prior information concerning locus order, such as physical localization data, one usually chooses the "ordered loci" to be a pair of linked and highly informative loci. When **prepare** is used to set up the *.par* file, the two most informative loci are automatically selected. **build** first analyses the "phase known" data, for which maximum likelihood computation is much more rapid than for the full data set, and proceeds to find a set of uniquely ordered loci, starting with the original ordered loci and sequentially adding any additional loci, in order of their informativeness, which have a unique placement. Using this uniquely ordered set of loci as the new "map", the program then proceeds to analyse the full data set, adding loci sequentially as before. Once a set of uniquely ordered loci has been determined, **build** prints out sex-specific and sex-averaged recombination fractions

and the corresponding Kosambi cM for these loci. For each remaining locus not uniquely ordered, it prints out the possible placements with respect to the uniquely ordered loci, along with the $\log_{10}$likelihoods for each placement.

The options **instant** and **quick** are similar to **build**, and each find a uniquely ordered set of loci, and possible placements of remaining loci with respect to them, using only the information in the *.ord* file created by a previous **build** run. The $\log_{10}$likelihood's of the possible placement of loci with respect to the uniquely ordered loci are not computed by **quick**.

The option **all** determines the most likely position of a locus on a predetermined map of loci. The program determines the order with the highest $\log_{10}$likelihood, **L**, and prints out, in decreasing order of $\log_{10}$likelihood, orders whose associated $\log_{10}$likelihoods exceed the amount {**L** - *PUK_LIKE_TOL*}.

The option **fixed** finds the associated maximum likelihood recombination fractions and map distances and the corresponding $\log_{10}$likelihood for a set of loci in a specified order, for either sex-averaged analyses of sex-specific analyses.

The **flips*n*** option of CRI-MAP permutes sets of adjacent loci (the number to be permuted is defined by *n*) within the map, calculates the corresponding likelihoods for these orders, and displays the relative $\log_{10}$likelihood (i.e. the $\log_{10}$likelihood of the reference order, minus that of the permuted order). **flips*n*** can be used to identify for map orders with a likelihood value higher than that of the tentative constructed map. If the *use_ord* file is set to 1 in the *.par* file, permutations incompatible with the orders database are eliminated before calculating likelihoods.

The **chrompic** option displays the grandparental origins of each allele for each child's chromosome of a given family for the phase choice having the highest likelihood; maternal first and then paternal. It determines the number and location of recombinations on each chromosome, and indicates the loci responsible for the recombinations, enabling loci involved in multiple crossover events, such as

isolated double crossovers, to be readily detected, thus flagging candidate data errors.

The option **twopoint** performs twopoint linkage analysis for each pair of loci, providing the female and male recombination fractions and the sex-specific peak LOD (= $\log_{10}$like(f,m)-$\log_{10}$like(.5,.5) at that recombination value. Also included in the output are the values of $\log_{10}$like(f,m)-$\log_{10}$like(0.5,0.5) for r = 0.001. 0.01, 0.05 and all multiples of 0.05 up to 0.5; the analysis can also be performed for the sex-averaged case. Only pairs for which the LODs exceed *PUK_LIKE_TOL* are printed out.

The option **merge** amalgamates two *.gen* files, allowing different sets of loci to be analysed as one data set. **prepare** does not need to precede the **merge** option.

2.2.4.4. Map Construction using CRI-MAP

The first step in constructing a map is the determination of a set of markers that form a linkage group. In the case of chromosome 16, all the loci have been previously assigned to the chromosome, and are at least syntenic. The next step is to determine the probable order of the group of loci. Ideally all *n!/2* orders should be compared for the construction of the map; this is computationally impractical, and all orders cannot be examined. It is reasonable to assume that interference and recombination rates imply that double recombinants occur less frequently than single recombinants, hence the maximum likelihood estimates under each gene order should be constrained accordingly, enabling some of the possible orders to be eliminated. Reliable heuristic algorithms have been designed that search for a subset of orders to find the best supported map, and only a few likely alternatives (Lathrop, 1994). A suitable map function needs to be proposed for the linkage analysis, to consider the occurrence of interference between chiasmata, for the efficient use of multipoint data in determining marker order or map position. For linkage analysis using CRI-MAP, Kosambi's mapping function is used, which is

based on marginal assumptions of interference (See Section 1.6.4). The determination of the most likely order is achieved by comparison of the maximum likelihood values of the possible orders as likelihood ratio statistics, or odds ratios. The best supported order - that with the highest likelihood - is chosen as the probable order. As the process of map construction involves methods of statistical inference, it is impossible to conclude that the best-supported order is in fact the true order, but by comparing the likelihoods, alternative orders with low likelihood values can be rejected. The usual criterion for the rejection of alternative orders, odds of 1000:1 or greater, is somewhat an arbitrary choice, and was initially developed by analogy with a lod score of 3 (odds ratio of 1000:1) used as the critical value to detect linkage between two loci (Morton, 1955; Lathrop, 1994).

The process of map construction using CRI-MAP began with the option **build**. Based on phase-known data, the two most informative loci selected by the **prepare** option were used for the map nucleus and a framework map of loci with odds of 10^4:1 or greater was constructed. Additional loci were then inserted into the map with descending order of informativeness, and the maximum likelihood estimate of the recombination fractions were calculated. The loci were added to the map only if the new order of loci met with the specified tolerance of 10^4:1 or greater. That is, a locus was added to the map if the maximum likelihood estimate was at least 10^4 times more likely than the next most likely order. Any tolerance could be specified, below which loci were rejected for subsequent addition. The orders that met the tolerance criteria were not rejected and were retained as possible orders. This process continued until all loci were placed upon the map and a list of all orders that could not be differentiated with the specified tolerance was produced along with the log likelihood value and genetic map for each order. Local support for order of loci, using the **flips** option of CRI-MAP, was determined by calculating the relative likelihood against the inversion of every adjacent pair of

loci while holding the order of the remaining loci on each side constant. This procedure provides strong support against local inversions. A **flips3** option was performed on this map, and any locus order with odds of less than 10^4:1 was eliminated by removing one marker at a time. After a marker was removed, **flips3** was performed again and the procedure was repeated until odds for all orders were greater than 10^4:1 times as likely as any other order.

A second map was constructed with odds of 10^3:1 using the loci that had resulted from the **flips3** test above, for the map nucleus. The resultant map order was subjected to the same **flips3** testing as the previous map.

As a final check on the validity of the map construction process, all sequences of four consecutive loci were permuted within the structure of the complete map, and all permutations that resulted in $\log_{10}$likelihood differences of less than 3 with the most likely order were reassessed. For the resulting best order, both sex-averaged and sex-specific maps were estimated, with the recombination frequency being converted to centimorgans using the Kosambi Map Function.

After a framework map was constructed in the above manner, the remaining loci were sequentially added to the map by placing each in its favoured position (using the **all** option), regardless of the level of support, until all loci were placed to construct the comprehensive map. Once the order of all the loci was been determined, the recombination fractions that maximise the likelihood function were estimated.

To ensure that the final order determined using the above method is in fact the correct order, the **build** analysis described above was performed several times, using different starting loci. During the map construction process, many factors may influence the outcome from the build analysis such as the presence of data errors. The choice of starting loci for the foundation of the framework map may preclude the addition of other loci at later steps, possibly leading to an erroneous order.

2.2.5. The detection of data errors

The **chrompic** option displays the grand parental origin of each locus along the chromosome. This option of CRI-MAP was used to identify the locations of all cross-overs for each of the maps constructed. Unlikely double recombinants within small intervals were flagged by **chrompic** and were considered as potential data errors if they occurred within a 15 cM interval, and where possible autoradiographs were checked and samples regenotyped if necessary. Double recombinants in small intervals in RFLP loci genotyped by the candidate were checked by regenotyping for the maps constructed in Chapters 3 and 4. In subsequent chapters, double recombinants for RFLP loci within a 15 cM interval that could not be resolved by checking the autoradiograph, were coded as unknown in the CEPH database (O'Connell *et al.*, 1987; 1989; White *et al.*, 1990).

Requests were made to collaborators who submitted RFLP data to the CEPH database to check double recombinants detected within a 15 cM interval. Subsequent corrections were made to the database where possible, or the genotypes were coded as unknown if the submitting collaborator did not respond (O'Connell *et al.*, 1987; 1989; White *et al.*, 1990). The reasoning behind this was that the possible removal of a small amount of correct data was considered inconsequential compared with the alternative failure to remove the majority of errors, since the effect that genotype errors have on map inflation and the selection of the correct locus order can be quite substantial (Buetow, 1991).

Double recombinants detected at PCR formatted loci were checked and regenotyped if necessary. In some instances the double recombinants could not be resolved (even by regenotyping several times, using fresh DNA samples) and no changes were made to the database. Loci for which this was true are detailed in chapter 7. If errors were detected in the data set, the map was reconstructed once the errors were corrected.

2.2.6. Investigation of sex-specific differences in the recombination fraction

Sex specific differences in the recombination rates were investigated for each of the maps constructed in this thesis, at varying levels. Three models were considered:

(i) equal recombination rates in males and females across the entire map,

(ii) sex-average recombination rate between males and females across the entire map,

(iii) sex-specific recombination rates between males and females in each interval on the map under consideration.

The significance of the observed heterogeneity between male and female recombination frequency estimates was determined by calculating the maximum likelihood (-2ln(likelihood)) achieved for each of the models described above and comparing these values. The difference between the likelihood estimates is assymptotically distributed with a chi-square distribution with *n-1* degrees of freedom, where *n* is the difference in the number of parameters estimated for each model (Ott, 1991). Multiple testing was corrected for by choosing the conservative test of Fisher-Bonferroni (ie, divide *P* by *n*, where *n* is the number of intervals) (Hochberg, 1988). The sex difference with respect to recombination was also evaluated separately in each interval on several of the framework maps. The chi-square test statistic was derived by analysing each interval under the assumption of variable or equal recombination in males and females and including three adjacent intervals on each side, in which the recombination frequencies were held equal for males and females in those adjacent intervals (O'Connell *et al.*, 1993), and determining the likelihood in each case. The chi-square test has one degree of freedom.

2.2.7. Utility Programs

Weaver *et al.*, (1992) have made available to the scientific community eight interactive software packages to aid the construction of multipoint linkage maps. These eight programs, PIC/HET, FAMINFO, PREINPUT, INHERIT, TWOTABLE, UNMERGE, GENLINK, and LINKGEN, are written in the programming language C (Kernighan and Ritchie, 1978), to support and complement published linkage mapping software (mainly CRI-MAP). The programs most useful during this project were PIC/HET, TWOTABLE, and UNMERGE; only these programs will be described below.

PIC/HET: This program automatically calculates PIC values, heterozygosity, the maximum number of informative meioses, and allele frequencies for each genetic marker in the *.gen* file (from above), and outputs the information in table format.

TWOTABLE: This program takes the output from the **twopoint** option of CRI-MAP and reports it into an easily read table format.

UNMERGE: This program enables the user to select a subset of the probe/enzyme systems in a *.gen* file, extract them and create a new *.gen* file.

In addition to these programs, one other utility program was used routinely during this project. **mkcrigen**, written by Santosh Mishra (personal communication), was developed to convert the genotype information in the ped.out file from the CEPH database, into the *.gen* file format suitable for use with CRI-MAP.

Two other programs were also used: **mapfun** and **chiprob** (Ott, 1991). The **mapfun** program is designed to convert recombination fractions into map distances and vice versa using seven different map functions. The map functions it uses are those proposed by Haldane (1919), Kosambi (1944), Carter and Falconer (1951), Rao (1973), Felsenstein (1979), Sturt (1976), and the binomial mapping function (Karlin, 1984; Ott, 1991). The **chiprob** program computes the *p*-values for given values of chi-square and number of degrees of freedom.

CHAPTER 3

THE GENETIC MAP ENCOMPASSING *FRA16B*

Table of Contents

Associated Paper (See Appendix A):

Addition of *MT*, *D16S10*, *D16S4*, and *D16S91* to the linkage map within 16q12.1-q22.1. H.M. Kozman, A.K. Gedeon, S. Whitmore, G.K. Suthers, D.F. Callen, G.R. Sutherland, and J.C. Mulley. (1991). *Genomics* **11**:756-759.

3.1. Summary

A 10 point genetic linkage map of the region 16q12.1 to 16q22.1 was constructed using the 40 CEPH reference families. This region encompasses the fragile site *FRA16B*. Four loci, not previously localised on the genetic map, *MT, D16S10, D16S91*, and *D16S4*, were localised on the multipoint linkage map of this region. The order of loci was:

cen - *D16S39* - *MT* - *D16S65* - *D16S10* - *(FRA16B)* - *D16S38* - *D16S4* - *D16S91* - *D16S46* - *D16S47* - *HP* - qter.

By *in situ* hybridisation and southern blot analysis of human/rodent cell lines, and now by linkage analysis, the two loci *D16S10* and *D16S38* were shown to flank *FRA16B*, providing a physical reference point for this multipoint linkage map on the long arm of chromosome 16. The length of the interval between these closest flanking markers (*D16S10* and *D16S38*) is 3.1 cM in males and 2.3 cM in females. These markers demonstrate that the corresponding clones are in close proximity to *FRA16B* and useful as the basis for development of a high resolution physical map of the region, to enable the isolation of *FRA16B* sequences.

3.2. Introduction

The fragile site located at 16q22.1, *FRA16B*, is a distamycin A-inducible fragile site (Sutherland and Hecht, 1985). It is the most common of all rare fragile sites, occurring under certain culture conditions in 5% of German individuals (Schmid *et al.*, 1986). The precise molecular nature of *FRA16B* is not known. The isolation of this fragile site will enable determination of the nature and chemistry of cytogenetic expression of *FRA16B*, and of the group of distamycin A fragile sites. The first step toward the elucidation of the molecular basis of this fragile site is the identification of closely linked markers by the development of a multipoint linkage map encompassing it. This is followed by the construction of a high resolution physical map involving YAC contigs between the closest flanking markers. The fragile site can be sequenced when a clone is identified which contains it. This will ultimately lead to a greater understanding of its nature function and purpose.

In 1986 (Fratini *et al.*, 1986), the known gene order on the 16q was:

cen - *FRA16B* - *HP* - *FRA16D* - *APRT* - qter.

These genes and fragile sites had been regionally assigned by physical mapping methods. The haptoglobin gene (*HP*) was genetically mapped near *FRA16B*, with a genetic distance of 9 cM (Magenis *et al.*, 1970), and was later physically mapped distal to the fragile site (Simmers *et al.*, 1986). This gene has now been further localised to 16q22.1-q22.3 (Natt *et al.*, 1986).

The Metallothionein (*MT*) gene cluster was mapped to chromosome 16 by Karin *et al.* (1984). This cluster contains at least 14 genes and pseudogenes with common sequences separated into two groups, *MT1* and *MT2*. *In situ* hybridisation studies further localised this gene cluster to the distal end of band 16q21 (Sutherland *et al.*, 1987). The 5' flanking region of the *MT2A* gene can be detected by two *Taq*I RFLPs (Hyland *et al.*, 1988); these were used in the development of the linkage map in the current study.

The anonymous DNA fragments, *D16S4* and *D16S10*, were regionally mapped between the breakpoints of CY7 and CY6, in the regions between 16q12.1 and 16q21, by probing southern blots of human/rodent hybrids (Sutherland *et al.*, 1987; Hyland *et al.*, 1988). They were oriented relative to *FRA16B* by *in situ* hybridisation to metaphase chromosomes expressing the fragile site (Callen *et al.*, 1988; Hyland *et al.*, 1988), to further refine the localisations of these two markers.

Thus, from these reports the established physical order of markers was:

cen - *MT,D16S10* - *FRA16B* - *D16S4* - *HP* - qter,

where the precise order of *MT* and *D16S10* was not known.

To determine the genetic distance between these markers, and their placement with respect to the fragile site, *FRA16B*, Mulley *et al.* (1989) performed two point linkage analyses using nine kindreds segregating for *FRA16B*, and the nine Utah pedigrees. Using the nine Utah pedigrees (see Section 2.1.2), they determined a lod score of 8.3 at a recombination frequency of 0.0. between the two markers, *D16S4* and *D16S10*, which flank *FRA16B*. In fact, they observed no recombinants within the *MT, D16S10* - *D16S4* - *HP* cluster.

The first multipoint linkage map for the entire chromosome 16 (including the region under discussion) was reported in 1987 as part of a genetic map of the entire human genome using the CEPH pedigrees (Donis-Keller *et al.*, 1987). In the region encompassing *FRA16B*, the order of markers was:

cen - ***D16S39*** - *(D16S52)* - *(D16S38)* - ***D16S69*** - ***HP*** - qter

The brackets enclose loci whose mutual location cannot be resolved greater than odds of 100:1, and are placed in their most likely positions, given the data; loci in bold face were placed with odds greater than 100:1.

The chromosome 16 genetic map was further developed in 1990, by Julier *et al.* and Keith *et al.* The report by Julier *et al.* included no markers relevant to the present study, and will be discussed in Chapter 6. The Keith *et al.* report is essentially an extension of the Donis-Keller *et al.* report. The order of markers from Keith *et al.* are:

cen - ***D16S39*** - *(D16S59)* - ***D16S65*** -
D16S69, D16S46 - *(D16S38,D16S47)* - ***HP*** - pter.

The order and marker content between these two maps is slightly different. The loci in brackets cannot be placed confidently with odds greater than 100:1 on both maps, and so more information (including physical mapping data) was required to resolve the locus order.

Markers, *MT, D16S4, D16S10* and *D16S91*, were genotyped on the CEPH families and analysed for linkage. They were selected on the basis of previous physical and genetic mapping placing them near *FRA16B* (Callen *et al.*, 1988; Mulley *et al.*, 1989; Chen *et al.*, 1991). The aim was to incorporate these markers into a comprehensive multipoint linkage map surrounding *FRA16B*.

3.3. Materials and Methods

Genotype data from the CEPH pedigrees for markers *D16S39*, *D16S65*, *D16S38*, *D16S46*, and *D16S47* (Database Version 4.0), were used for the linkage analysis. All laboratory methods for DNA analysis were described in Chapter 2. The probe/enzyme systems genotyped by the candidate (*MT*, *D16S4*, *D16S10*, and *D16S91*) and A.K. Gedeon (*MT*, *D16S4*, and *D16S91*) are described in Table 3.1. *D16S39*, *D16S65*, *D16S38*, *D16S46*, and *D16S47*, were extracted from the CEPH database and are described in Table 2.1.

The DNA from the nine UTAH pedigrees was extracted using either of two methods, as described in Section 2.1.4. UTAH DNA was used in preference to

Table 3.1: *Description of probe/enzyme combinations genotyped on the CEPH pedigrees by the candidate.*

Locus	Probe	Number of RFLPs	Enzyme	Alleles Kb	Pic	Physical Location
MT	MT2A	2	TaqI#	7.8/5.3	0.22	16q13-q22.1
			TaqI#	1.7/1.6	0.18	
D16S10	pACHF3.5.1	2	RsaI*	2.6/1.8,0.76	0.33	16q13-q22.1
			MspI*	12.8/10.6	0.16	
D16S4	ACH207	3	TaqI#	5.0,0.55/5.55	0.28	16q22.1
			MspI#	1.7/1.0,0.7	0.32	
			MspI#	2.1/2.9	0.35	
D16S91	LE12	1	RsaI*	2.8/2.6	0.11	16q22.1

* - Genotyped by the candidate.
# - Genotyped by the candidate and A.K. Gedeon

the CEPH DNA where possible to perform the genotyping of our markers, as it was more easily replaced, being cell lines. The method of high salt extraction of cell line DNA generated DNA of higher integrity, and was used in preference to the phenol-chlorophorm method of extraction, which when extracting cell line DNA appeared to produce more sheared DNA.

A 1.8 Kb single copy fragment, provided by J.C. Mulley, was used as the probe for the marker *D16S10*. This fragment was released from the 3.5 kb insert cloned into pSP64 by a *Bgl*II-*Hind*III digest. *D16S91* (*LE12*) is also known as *APOEL1* (Davidson *et al.*, 1987), a related sequence to apolipoprotien located on chromosome 19. The heterozygosity of the *Rsa*I polymorphism low (8%) and an attempt was made (by J.C. Mulley) to search for additional polymorphisms that would render this marker more informative: no polymorphism was detected for *Ava*I, *Ava*II, *Bam*HI, *Ban*I, *Bcl*I, *Bgl*II, *Bst*EII, *Bst*NI, *Bst*XI, *Eco*RI, *Eco*RIV, *Hinc*II, *Hind*III, *Msp*I, *Pst*I, *Pvu*II, *Sac*I, *Stu*I, *Taq*I, *Xba*I, and *Xmn*I using six unrelated individuals. Probes were made from glycerol stocks using the methods described in Section 2.1.5, and inserts generated using the methods described in Section 2.1.6.

Genotyping of the loci describe in Table 3.1 was performed by Southern analysis (Section 2.1.8), radiolabelling of the probe (Section 2.1.9), hybridisation (Section 2.1 10), and subsequent visualisation by autoradiography (Section 2.1.11).

Incorporation of radiolabelled nucleotides were checked using the method described in Section 2.1.9, using TLC cellulose-F precoated paper (Merck) (M. Little, personal communication). The TLC paper was cut to an appropriate size, and the wet tip of a Gilson pipette dotted on to the paper, at the start of the radiolabelling reaction, and again at the end of the radiolabelling reaction. The paper was allowed to dry, and then placed upright in a solution of KH_2PO_4 (in a saturated atmosphere), for 5 minutes. The paper was dried at 37°C for five minutes and exposed to fast film for five minutes at room temperature, and then

developed (Fig. 3.1). Unincorporated nucleotides migrate to the end of the paper, while incorporated nucleotides remain in the same place. If good incorporation occurred, a large dot was seen at the bottom of the TLC paper, and a faint dot at the top. If poor incorporation occurred, then there will be either a large dot at the top, or a smear across the lane. The incorporation of label all four probes shown in Fig. 3.1 (two for *D16S10*, one each for *D16S4* and *D16S91*), was satisfactory, and hybridisation was carried out. After autoradiography, genotypes were scored (Fig. 3.2) and recorded in the CEPH database. Fig. 3.2 demonstrates the inheritance of *D16S4* (*ACH207/TaqI*) in the CEPH pedigree 1416.

Two point and multipoint linkage analysis was performed on an APCIV NEC personal computer, using the computer program package LINKAGE (Version 5.1) (Lathrop *et al.*, 1985) as described in Chapter 2. This package uses Haldane's Mapping Function (1919) and so the possibility of interference was ignored; LINKAGE does not consider interference when the analysis is extended beyond three loci. Lathrop *et al.* (1985) showed that current methodology for linkage analysis in human pedigrees was not sufficiently sensitive to detect interference in analyses involving more than three loci.

Genotypes for each marker were rigorously scanned and checked for double recombinants by haplotype analysis, by hand. Individuals with double recombinants found between markers for which the recombination fraction was less than 15% were regenotyped if they had been genotyped by the candidate, or coded as unknown (White *et al.*, 1990) if they were extracted from the CEPH database (See Chapter 2.2.5).

Sex differences in the recombination rates were assessed by considering three models as described in Chapter 2.2.6. The three models were:

(i) equal recombination rates in males and females,

(ii) sex-average in the recombination rate between males and females,

(iii) sex-specific recombination rates between males and females.

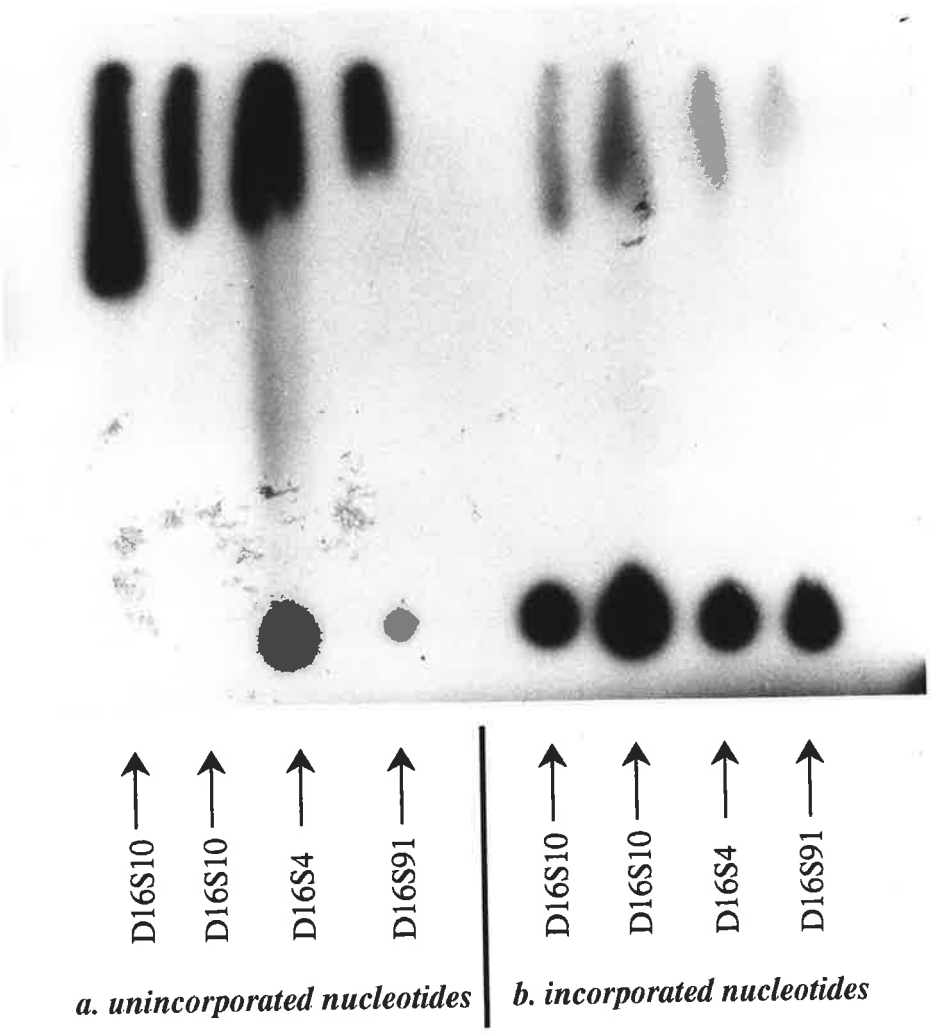

Figure 3.1: *The results testing the incorporation of nucleotides in the radioactive labelling reactions for the markers D16S4 , D16S10 and D16S91, using the TLC paper method (Section 2.1.9).* ***a.*** *unincorporated nucleotides from the start of the reaction migrate to the end of the TLC paper.* ***b.*** *incorporated nucleotides remain as a large dot at the bottom of the paper, if the labelling reaction proceeded normally and good incorporation occurred .*

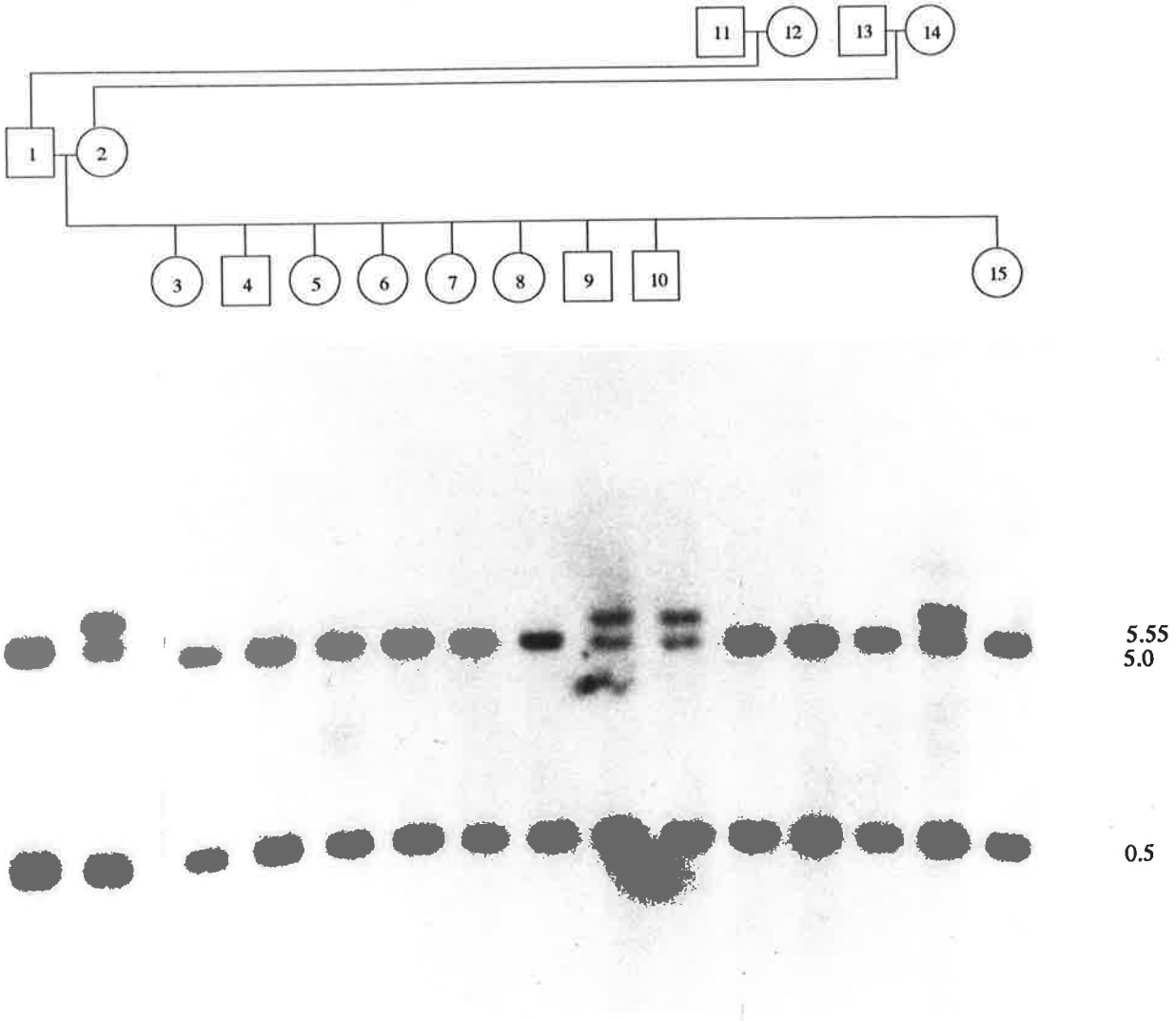

Figure 3.2: *Autoradiograph of the marker ACH207/TaqI, hybridised to DNA from CEPH family 1416, bound to a nylon membrane. The size of the fragments (Kb) are shown, where the 0.5 Kb fragment is a constant band. The inclusion of the pedigree for this family demonstrates the inheritance of this polymorphism.*

3.4. Results

The results of the two point linkage analysis for the 10 marker loci investigated are summarised in Table 3.2. They demonstrate that these 10 loci belong to the same linkage group on chromosome 16. From two point analysis, there was no intragenic recombination within the loci for *D16S4*, *D16S10*, *D16S46*, *D16S65*, and *MT*; data from separate probe/enzyme combinations within each locus were haplotyped by hand by the candidate to reduce the analysis from an 18-point analysis to a 10-point analysis.

An initial analysis was carried out with six of the 10 loci, to determine a foundation map on which to build the final comprehensive map (Keats *et al.*, 1991). *D16S10* and *D16S4* were selected, being reasonably polymorphic and physically mapped. Although *MT* and *D16S91* have been physically mapped, they are not very polymorphic and so were left out of the initial analysis. Markers from the CEPH database, *D16S39*, *D16S65*, *D16S47*, and *HP*, were chosen for the initial analysis as they were physically mapped, polymorphic loci that flanked the new markers to be placed on the multipoint map. The order of these six marker loci is

cen - *D16S39* - *D16S65* - *D16S10* - (*FRA16B*) - *D16S4* - *D16S47* - *HP* - qter,

derived from physical mapping data (Callen *et al.*, 1988; Chen *et al.*, 1991). With marker order established, the analysis was simplified to the estimation of recombination frequencies between adjacent markers. The program CILINK from the LINKAGE package was used for this.

Table 3.3a gives the recombination frequencies, under all three recombination models, between adjacent pairs of loci for the background linkage map of six loci, determined using CILINK. The odds against inverting adjacent loci, under a model of variable recombination, are also given as a measure of support for the most likely order. The odds against inverting the adjacent loci were of the order

Table 3.2: *Summary of Pairwise Recombination Fractions (above the diagonal) and corresponding peak LOD scores (below the diagonal).*

	D16S39	*MT*	*D16S65*	*D16S10*	*D16S38*	*D16S91*	*D16S4*	*D16S46*	*D16S47*	*HP*
D16S39		0.142	0.044	0.176	0.445	ND	0.161	0.116	0.138	0.196
MT	1.71		0.000	0.084	0.077	0.111	0.107	0.000	0.001	0.063
D16S65	8.76	2.41		0.000	0.000	0.000	0.017	0.000	0.000	0.000
D16S10	2.77	9.69	8.43		0.000	0.000	0.020	0.000	0.045	0.061
D16S38	0.02	4.76	3.66	4.21		0.000	0.010	0.000	0.000	0.047
D16S91	2.71	2.69	1.80	7.80	5.41		0.000	0.100	0.000	0.335
D16S4	7.91	4.69	14.12	21.00	9.32	7.22		0.000	0.013	0.025
D16S46	4.80	4.51	7.57	13.53	4.81	4.21	38.53		0.000	0.000
D16S47	3.29	5.11	3.06	13.21	10.22	1.20	19.69	18.7		0.025
HP	4.02	9.40	3.66	15.42	7.33	0.07	23.65	11.12	14.55	

ND - no data

of 5.7 x 10^2 to 1 or greater. This map was consistent with the physical order of marker loci.

The three models for the difference in the recombination between the sexes were tested on this initial map (Table 3.3b). The difference between the maximum likelihood values for the models *(i)* and *(ii)* was significant ($X^2_{(1)} = 10.28$, $P < 0.05$), the difference between models *(i)* and *(iii)* was significant ($X^2_{(5)} = 12.73$, $P < 0.05$), and the difference between models *(ii)* and *(iii)* was not significant ($X^2_{(4)} = 2.45$, $P > 0.05$). Model *(iii)*, a variable difference in the recombination rates between the sexes was chosen as the model on which to base the rest of the analysis.

Once this foundation map had been established, the additional four markers were added to the map; *MT* and *D16S91*, and *D16S38* and *D16S46*. The multipoint map for all ten loci was determined by evaluating the most likely order of all possible orders of loci in the same physical interval, within the background map. The order with the highest likelihood value was selected based on variable differences with respect to recombination between males and females. Once all ten loci had been placed on the map, the relative odds against alternative orders were again determined by the inversion of adjacent loci. The results of the comprehensive map are summarised in Table 3.4.

An ideogram of human chromosome 16q is presented in Fig. 3.3 comparing the data obtained from physical mapping with the genetic map constructed here. The order of loci on both the physical map and the genetic map are in agreement.

The entire male map length between *D16S39* and *HP* on the final map is 11 cM and the corresponding female map length is 46 cM, indicating considerably more recombination in female meioses compared with male meioses for this region of chromosome 16. The distance between the closest markers flanking the fragile site *FRA16B*, *D16S10* and *D16S38*, is 3.1 cM and 2.3 cM in males and females

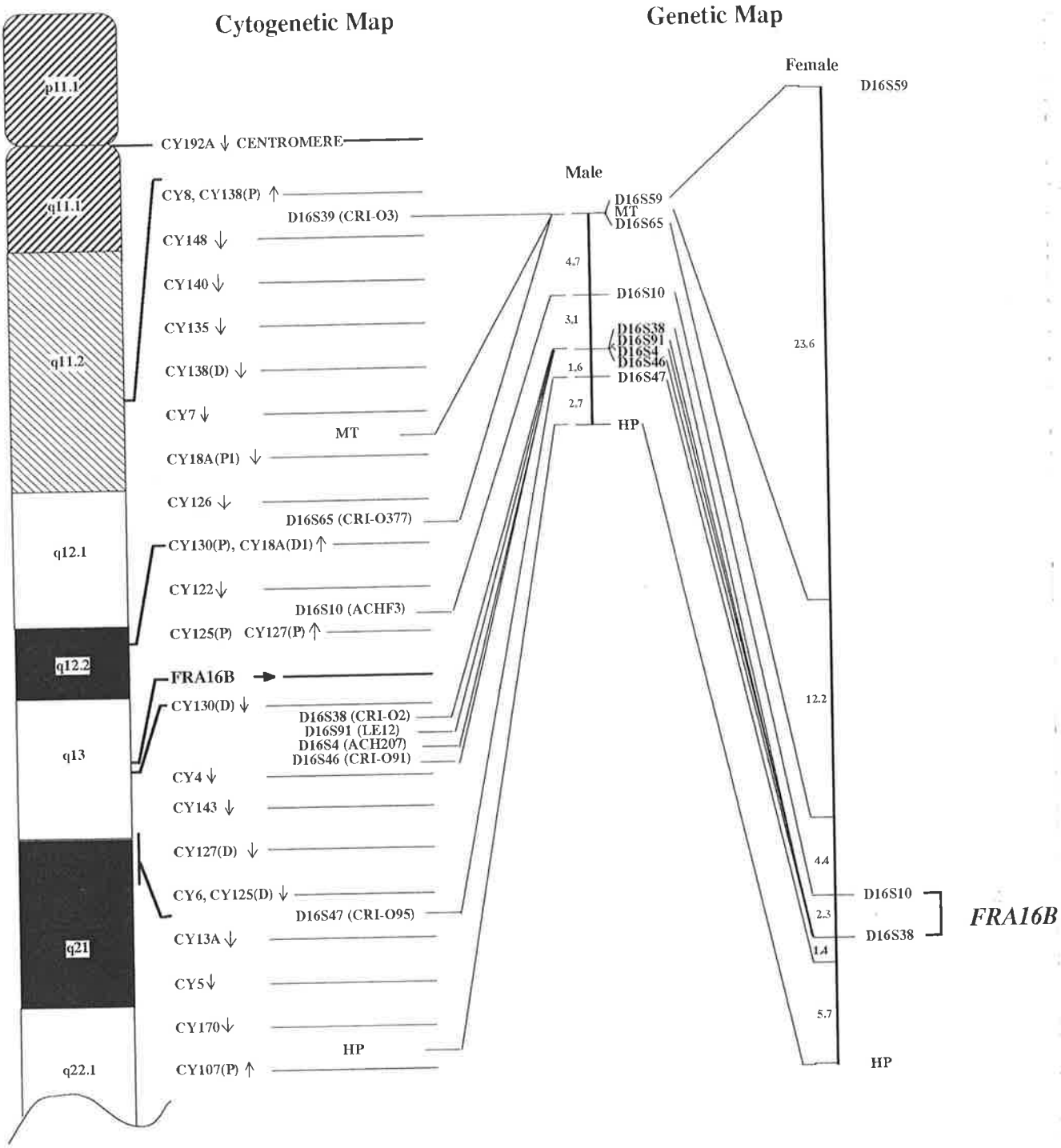

Figure 3.3: *Ideogram of human chromosome 16 showing the cytogenetic locations of he loci used to construct the genetic map (Callen et al., 1988; Chen et al., 1991). The approximate cell line breakpoints are shown. The male and female genetic maps are given with map distances in cM. The position of FRA16B is shown on the map. D, distal; P, proximal.*

Table 3.3

a: *Map Distances in cM (recombination fractions in brackets) between adjacent loci, on the initial map, under different models of sex specific recombination.*

Locus	No Sex Difference	Constant Sex Ratio		Variable Sex Ratio		Odds Ratio*
		Male	Female	Male	Female	
D16S39	17.4 (0.147)	7.0 (0.070)	31.9 (0.236)	0.0 (0.000)	31.6 (0.316)	3.3×10^8
D16S65	0.0 (0.000)	0.0 (0.000)	0.3 (0.003)	6.8 (0.068)	4.6 (0.046)	1.4×10^6
D16S10	1.9 (0.019)	0.7 (0.007)	3.1 (0.031)	4.8 (0.048)	4.0 (0.040)	1.3×10^6
D16S4	0.9 (0.009)	0.4 (0.004)	1.8 (0.018)	1.8 (0.018)	3.9 (0.039)	5.7×10^2
D16S47	1.9 (0.019)	0.7 (0.007)	3.0 (0.030)	1.1 (0.011)	6.5 (0.065)	1.7×10^3
HP						
Total map length (cM)	22.1	8.8	40.0	14.5	50.6	
-2ln(likelihood)	-678.02	-688.30		-690.75		

* - ratios given for the model of variable recombination between the sexes.

Table 3.3 (cont.)

b: *Chi-square values and probabilities under the three different models of sex differences in the recombination for the initial map: (a) there is an equal recombination rate in males and females, (b) there is a constant difference in the recombination rate in males and females, and (c) there is a variable ratio in the recombination rates in each interval between the sexes (O'Connell et al., 1987).*

Models compared	Chi-square Value	Degrees of Freedom	Probability Value
a) and b)	10.28	1	$P < 0.05$
a) and c)	12.73	5	$P < 0.05$
b) and c)	2.45	4	$P > 0.05$

Table 3.4: *Map Distances in cM (recombination frequencies in brackets) between adjacent loci, on the final comprehensive map, under the model of sex-specific recombination.*

Locus	Variable Sex Ratio		Odds Ratio
	Male	Female	
D16S39	0.0 (0.000)	23.6 (0.220)	1.1×10^3
MT	0.0 (0.000)	12.2 (0.120)	7.2
D16S65	4.7 (0.047)	4.4 (0.044)	4.8×10^{10}
D16S10	3.1 (0.031)	2.3 (0.023)	9.9
D16S38	0.0 (0.000)	0.0 (0.000)	1.0
D16S91	0.0 (0.000)	0.0 (0.000)	1.2
D16S4	0.0 (0.000)	0.0 (0.000)	3.4
D16S46	1.6 (0.016)	1.8 (0.018)	7.9×10^2
D16S47	2.7 (0.027)	5.7 (0.057)	1.3×10^4
HP			
Total map length	12.1	50	

respectively. The two point LOD score and sex pooled recombination fraction between these markers are $Z = 4.2$ and $\Theta = 0.0$ respectively.

3.5. Discussion

The map presented defines a linkage group around the fragile site, *FRA16B* (Table 3.4 and Fig. 3.3). The final order of loci along this region is:

cen - *D16S39* - *MT,D16S65* - *D16S10* - *(FRA16B)* -
D16S38,D16S91,D16S4,D16S46 - *D16S47* - *HP* - qter

This map represents 10.5% of the total genetic map of Chromosome 16 in males and 24.4% in females, assuming the total map length of 115 cM in males and 193 cM in females from Keith *et al.* (1990).

In some intervals, the odds against alternative orders for this map, given in Table 3.4, are very low, being of the order of 1:1 between *D16S38* and *D16S91*, for example. The pairs of loci that display odds of less than 100:1 are closely linked, with little or no recombination on a two point analysis evident, and thus can not be ordered with respect to each other by linkage analysis alone. For example, between the pairs of loci *D16S38*, *D16S4*, *D16S91*, and *D16S46*; and between *MT* and *D16S65* (Table 3.2) there is no recombination. Markers with low levels of heterozygosity, such as *MT* and *D16S91*, are also difficult to order because of the limited number of potentially informative meioses for detection of recombinations with closely linked loci. From physical mapping investigations, performed by PFGE, the loci *D16S91* and *D16S4* are physically linked on the same 450 kb *Mlu*I fragment (Lapsys, 1993); thus it would be difficult to determine the exact order of these closely placed markers by genetic analysis alone.

The linear order of the map is compatible with the maps reported by Donis-Keller *et al.* (1987) and Keith *et al.* (1990), with the exception of the marker, *D16S38*. As stated previously, this marker has not been placed with odds greater than

100:1, and has been only placed in its most likely position, which will be affected by the markers incorporated in the specific analysis. The distance between markers in each of the maps are well correlated. For example, Keith *et al.* (1990) reported no recombination between *D16S39* and *D16S65* in males, as has been found on the map generated here. Between *HP* and *D16S47* the recombination frequencies are 2% and 5% in males and females respectively on the map generated by Keith *et al.* (1990), and are 2.7% and 5.7% on the present map, in males and females respectively.

Three models of the variation with respect to recombination between males and females were tested using the initial 6-point map (Table 3.3). Model c), that there exists a variable rate of recombination between males and females for this region, was chosen on which to base the entire analysis. There were several reasons for this. Model c) was significantly different to model a) but not to model b). That is, a variable ratio of female to male map distance over the region studied was actually significantly different from a ratio of 1 to 1, but not significantly different to a constant ratio. Thus, because model b) and model c) are not significantly different, it would have been acceptable to compute the map using model b). However, the analysis was based on model c), as a variable ratio of recombination between the sexes has been previously observed on all human chromosomes (Donis-Keller et al., 1987; Keith *et al.*, 1990; White *et al.*, 1990). Keith *et al.* (1990) demonstrate that the female map is 40% longer than the male map. Donis-Keller *et al.* (1987) observed that throughout the human genome a higher rate of recombination exists in female meioses, with the genetic maps of female autosomes being approximately 90% longer than those of males. More recombination in females has been previously reported in the *HP* region (Reeders and Hildebrand, 1989). Inspection of the map constructed here revealed that there is evidence for a difference in the recombination rate between males and females, the map being 25% longer in females than in males. These observations are in

agreement with Haldane's Rule (Haldane, 1922), that the heterogametic sex has a lower rate of recombination. The phenomenon of sex differences in the frequency of recombination has been documented in other mammals also: in the horse (Anderson and Sandberg, 1984) and in *Mus musculus* (Dunn and Bennet, 1967), where a significant excess of female recombination has been observed in many chromosomal regions.

The latest genetic map for chromosome 16, the CEPH consortium map, has been constructed, and is presented in chapter 8 (Fig. 8.1 and Table 8.5). Since the construction of the map presented in this chapter, many new markers have been isolated from the region near *FRA16B*, including PCR formatted loci. There are now 33 polymorphic loci, of which 16 are PCR formatted, within the interval between *D16S39* and *HP*. The *Msp*I polymorphism for Haptoglobin was used in the construction of the consortium map, whereas the protein polymorphism was used to construct the map presented in this chapter. The distances between the markers in this present study and the CEPH consortium map differ in many regions, due to the addition of the many new markers. The order of these markers is in agreement with the exception of the four loci *D16S38*, *D16S91*, *D16S4*, and *D16S46*. These loci show no recombination on the present map or on the consortium map, and so their order cannot be resolved. The odds for the support for placement of these four loci is only 1:1 and the position of these markers with respect to each other will need to be determined by high resolution physical mapping. On the comprehensive consortium map, the distance across *FRA16B* between the closest flanking markers *D16S10* and *D16S186*, is 1.6 cM in females, and 1.9 cM in males, and the distance between D16S10 and D16S38 is 2.9 cM in females and 2.5 cM in males. The length of the consortium map in this region, between *D16S39* and *HP* is 24.5 cM on the sex average map, and 13.2 cM in males and 37.3 cM in females. These lengths are greater than the lengths presented in this chapter, which are 11 cM in males and 46 cM in females. This

can be explained by the fact that more loci are incorporated on this map, which may introduce additional undetected data errors, which will have the effect of increasing the map length.

From the linkage map constructed above, physical mapping has been initiated, involving YAC and cosmid contig mapping. Fig. 3.4 represents the most current status of the physical map of this region, constructed by Yu Sui, at the WCH. From the current study, the two flanking markers for *FRA16B*, *D16S10* and *D16S38*, are located at either end of the contig map, outside the hybrid cell line break points CY125/CY127 and CY130 (CY125 and CY127 define the same break point). The fragile site lies between the breakpoints of these hybrids. From the contig mapping, *FRA16B* is located within the small region between the sequence tagged sites (STSs), *60H5* and *D16S174*, within 100 kb of *D16S174*.

The known cytogenetic location of fragile sites and translocation breakpoints for many human/mouse somatic cell hybrids allow this genetic map to be anchored to the physical map. The importance of combining information obtained from the physical map with the results of the genetic map to develop a complete portrait of a particular region, especially when small genetic distances are being considered and when markers of low informativeness are being examined, becomes obvious from the results reported here. Firstly, over small genetic distances it is very difficult to order markers by linkage studies with significant confidence. Utilising the order of loci obtained from the high resolution physical map preferably determined independently by PFGE and contig assembly, leaves only the distance between markers to be determined by linkage analysis. Secondly, the genetic map has established close flanking markers for *FRA16B* revealing distances small enough to initiate YAC walking to clone this fragile site; hence, the hunt for the fragile site is approaching its finale.

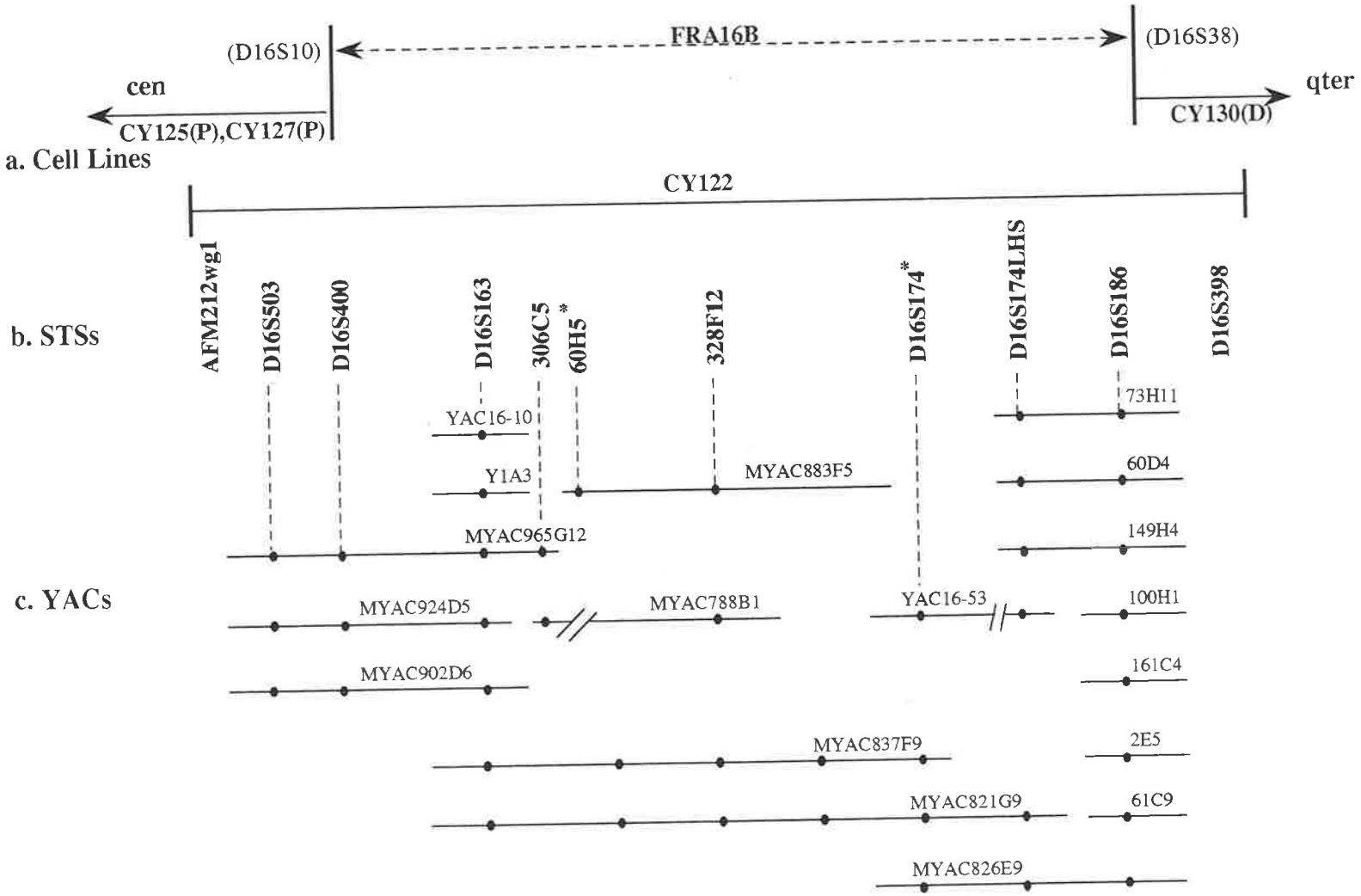

Figure 3.4: *YAC contig map of the FRA16B region.* ***a.*** *The hybrid cell lines flanking the region, with the two RFLPs flanking the fragile site indicated, as determined by the linkage analysis.* ***b.*** *STSs used to isolate YACs to form the contig.* ***c.*** *The YAC contig. The dots indicate overlaps, and locations of the STSs.* * *- STSs that flank the fragile site. Information kindly provided by Yu Sui.*

CHAPTER 4

THE GENETIC MAP ENCOMPASSING *FRA16A*

Table of Contents

Associated Papers (See Appendix A):

A multipoint genetic linkage map around the fragile site *FRA16A* on human chromosome 16. Kozman, H.M., Phillips, H.A., Sutherland, G.R., and Mulley, J.C. 1992. *Genet. (Life Sci. Adv.)* **11**:229-233.

Addition of *MT*, *D16S10*, *D16S4*, and *D16S91* to the linkage map within 16q12.1-q22.1. Kozman, H.M., Gedeon, A.K., Whitmore, S., Suthers, G.K., Callen, D.F., Sutherland, G.R., and Mulley, J.C. 1991. *Genomics* **11**:756-759.

Fragile X syndrome: Genetic localisation by linkage mapping of two microsaatellite repeats FRAXAC1 and FRAXAC2 which immediately flank the fragile site. Richards, R.I., Holman, K., Kozman, H.M., Kremer, E., Lynch, M., Pritchard, M., Yu, S., Mulley, J., and Sutherland, G.R. 1991. *J. Med. Genet.* **28**:818-823.

Fragile X syndrome: diagnosis using highly polymorphic microsatellite markers. Richards, R.I., Shen, Y., Holman, K., Hyland, V.J., Kozman, H.M., Mulley, J., and Sutherland, G.R. 1991. *Am. J. Hum. Genet.* **48**:1051-1057.

4.1. Summary

A detailed genetic linkage map of twelve loci was constructed for the region 16p13.2-16p12.3, using linkage analysis of genotypes from the CEPH reference families. The genetic map was anchored to the folate sensitive fragile site, *FRA16A*. The markers incorporated into the analysis were selected on the basis of their localisation on the cytogenetic-based physical map. Seven of the twelve loci, *D16S292, D16S8, D16S79A, D16S287, D16S96, D16S79B,* and *D16S131*, had not been previously localised on a multipoint linkage map. The remaining five loci, *D16S60, D16S51, D16S64, D16S75*, and *D16S67*, were used as anchor loci for the analysis as they had been previously incorporated into linkage maps. The order of the twelve loci was:

pter - D16S60 - D16S51 - D16S8,D16S292 - D16S79A -
(FRA16A) - D16S287,D16S96,D16S79B - D16S131 -
D16S64 - D16S75 - D16S67 - cen.

The interval across the fragile site *FRA16A*, between *D16S79A* and *D16S287* is 4.0 cM in males and 0.0 cM in females. These closest flanking markers to *FRA16A* are therefore sufficiently close to the fragile site on the genetic map to initiate development of a contig of cloned DNA between these two loci. The order of markers on the genetic map was consistent with that on the cytogenetic map. This provided a foundation for the task of positional cloning of this folate sensitive fragile site in order to further explore the molecular basis of folate sensitive fragile sites, the role of trinucleotide repeats in the genome and their conversion to heritable unstable elements.

4.2. Introduction

The purpose of the study was to ascertain the genetic distances between markers close to *FRA16A*, and identify a marker close enough (approximately 1 Mb away) to *FRA16A* that could form the basis for the development of a YAC contig across the fragile site. The relationship between genetic distance in centimorgans (cM) and physical distance in Mb is not uniform throughout the genome and is sex-specific, although 1 cM is on average approximately equivalent to a physical distance of 1 Mb (Morton, 1991b). A detailed genetic linkage map, spanning the region 16p13.2-p12.3 containing the rare folate sensitive fragile site, *FRA16A* (Sutherland and Hecht, 1985), is presented.

A genetic linkage map encompassing the rare folate sensitive fragile site, *FRAXA*, was constructed (Richards *et al.*, 1991a; 1991b), which identified two microsatellite repeats which immediately flank the fragile site. The identification of these flanking markers ultimately lead to the isolation and cloning of *FRAXA* (Kremer *et al.*, 1991), by YAC contig construction. Three other rare folate sensitive fragile sites *FRAXE* (Knight *et al.*, 1993), *FRAXF* (David Nelson, personal communication), and *FRA11B* (Chris Jones, Alan Tunnacliffe, and Rob Richards, personal communication), have been isolated and cloned. The molecular basis of these four fragile sites is the expansion of an unstable p(CCG)n trinucleotide repeat adjacent to CpG islands which become hypermethylated when the number of copies of the repeat exceed a certain threshold. The cloning of *FRAXA* (Kremer *et al.*, 1991; Verkerk *et al.*, 1991) led to the discovery of a new genetic mechanism for genetic disease based on mutations of heritable unstable DNA sequences (Richards and Sutherland, 1992a; Caskey *et al.*, 1992). Expansion associated with other unstable trinucleotide repeats have also been identified in three other human genetic diseases; spinal and bulbar muscular atrophy (*SBMA*), myotonic dystrophy (*DM*), and Huntington's disease (*HD*). It

now appears likely that unstable trinucleotide repeat sequence mutations are a common cause of human disease (Richards and Sutherland, 1994).

It has been hypothesized that the observed methylation associated with fragile sites is a consequence of the fragile site mutation not a cause of such a mutation, since the *FRA16A* region of chromosome 16 in normal individuals does not exhibit methylation at either the $p(CCG)_n$ repeat or the CpG island adjacent to it, as do the X chromosome fragile sites (Nancarrow *et al.*, 1994).

Unlike the fragile X (*FRAXA*), *FRA16A*, also a heritable folate sensitive fragile site, is not associated with a genetic disease. However, the molecular characterisation of this fragile site will have important implications regarding the molecular basis for folate sensitive fragile sites, and the position of this fragile site in relation to nearby sequences might provide insight into the reason why certain trinucleotide repeats become unstable and why they are not all associated with genetic diseases.

4.3. Methods

Genetic analysis was based on 40 pedigrees in the CEPH reference panel. Seven of the 12 DNA polymorphisms used in this study (Table 4.1) were localised, mapped and genotyped in our laboratory as described previously (Hyland *et al.*, 1988, Phillips *et al.*, 1991a, Callen *et al.*, 1992b, Thompson *et al.*, 1992). Two of these, *D16S292* and *D16S287*, were AC repeat markers. Whole plasmids of *pACHF1.1A6*, *p36-1*, *pVK20A* and *pVK45C6* were isolated using the methods described in Section 2.1.5, and inserts recovered by the methods described in Section 2.1.6. The inserts from these plasmids were labelled and typed by standard Southern analysis methods (Sections 2.1.8 through 2.1.10), by the candidate, HA Phillips, and KA Friend. *pACHF1.1A6* is a subclone of *pACHF1* originally reported by Fratini *et al.*, (1988). For *pACHF1.1A6*, a 600 bp *Bst*NI single copy fragment isolated from within a 1.2 kb *Eco*RI-*Hinc*II fragment cloned

Table 4.1. *Polymorphic loci on chromosome 16p that have been characterised and genotyped at the WCH, and used in the linkage analysis.*

Locus	Probe	Enzyme	Number of Alleles	Het.	PIC
D16S292	*16AC2.3*	AC repeat#	10	0.74	0.71
D16S8	*pACHF1.1A6*	*Pvu*II#	2	0.50	0.38
D16S79A	*p36-1*	*Taq*I*	2	0.34	0.28
D16S287	*16XE81*	AC repeat#	10	0.78	0.76
D16S96	*pVK20A*	*Taq*I,*Msp*I**	4	0.50	0.37
D16S79B	*p36-1*	*Taq*I*	3	0.09	0.09
D16S131	*pVK45C6*	*Taq*I#	2	0.46	0.35

* - genotyped by the candidate
** - genotyped by the candidate and K. Friend
\# - genotyped by HA Phillips

into *pUC18* was used by the candidate as the probe. The marker *D16S79* was found to be duplicated (Callen *et al.*, 1992a) and for the linkage analysis has been examined as two separate loci, *D16S79A* and *D16S79B*. Southern analysis and autoradiography were performed as described in chapter 2. Fig. 4.1 shows the results of autoradiography using the marker *VK20A/TaqI* on the CEPH pedigree 1331.

The five markers extracted from the CEPH database (Version 5.0) were incorporated into the map to provide flanking markers as reference points for the map. These markers were *D16S60*, *D16S51*, *D16S64*, *D16S75*, and *D16S67*, and are described in Table 2.1. They have been previously reported in multipoint genetic maps (Donis-Keller *et al.*, 1987; Keith *et al.*, 1990).

The initial order of the loci in this region was based on that previously determined by physical mapping (Callen *et al.*, 1992a). Cytogenetic mapping was performed using a hybrid cell line panel carrying varying portions of the human chromosome 16. The physical localisations for all the DNA polymorphisms can be visualised on the ideogram of chromosome 16 (Fig. 4.3).

Two point and multipoint linkage analysis was performed using the computer program package CRI-MAP (V2.4) (Lander and Green, 1987; Donis-Keller *et al.*, 1987), on a Sun SPARC station IPC, as detailed in Chapter 2.2.4. The genotypes of all individuals were checked for data errors by identification of apparent double recombinants assuming likely order of loci established by cytogenetic mapping and multipoint linkage analysis, using the **chrompic** option of CRI-MAP, as described in Chapter 2.2.5. Identified recombinants were recoded if found to be incorrect. For the marker *D16S287*, the genotypes for CEPH individuals 1375-11 and 1375-12 were changed from 23 and 14 to 14 and 23 respectively. The genotype for individual 1413-01 for the marker *D16S79A* was corrected to *12* and those for *D16S79B* of individuals 1423-04 and 1423-07 were corrected to *12* and *22* respectively. The genotype for CEPH panel individual 1346-07 for the marker

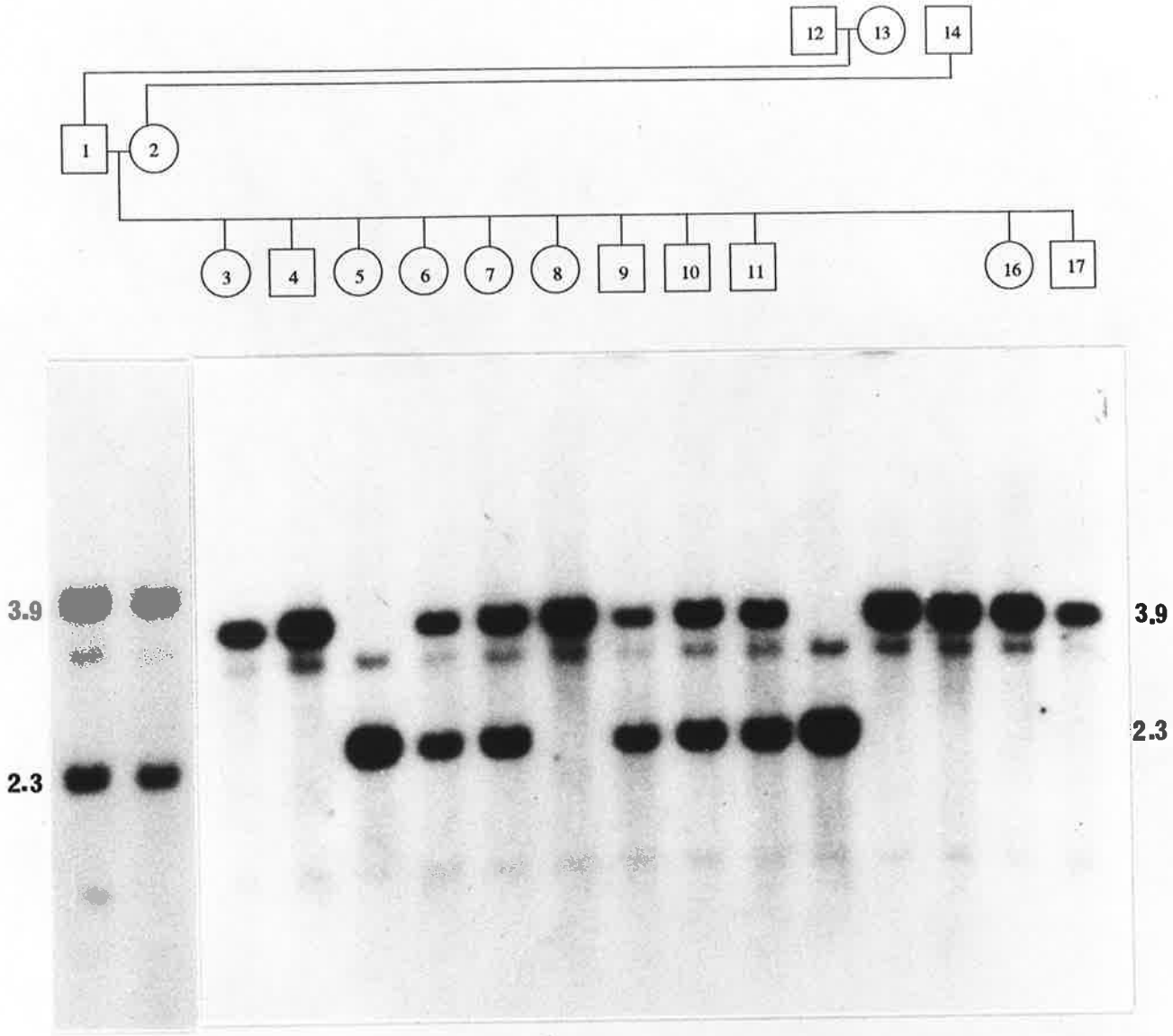

Figure 4.1: *Autoradiograph of the marker VK20A/TaqI, hybridised to DNA from CEPH family 1331, bound to a nylon membrane. The CEPH family DNA was digested with TaqI. The size of the fragments (Kb) are shown. The inclusion of the pedigree for this family demonstrates the inheritance of this polymorphism.*

D16S64 was found to be a double recombinant within 15 cM, and was thus coded as unknown (See Section 2.2.5).

For the marker *D16S287*, many of the genotypes were found to be inconsistent in CEPH family 1333, due to a null allele (Phillips *et al.*, 1992). Thirty nine of the 40 CEPH families showed normal co-dominant segregation, but the marker was shown to have a null allele segregating in the CEPH pedigree 1333. In this family, 4 out of the 8 offspring appeared not to inherit an apparent homozygous allele from their father, who in turn appeared not to inherit an allele from his mother. Mutations within the DNA sequence complementary to the oligoprimers may inhibit or completely prevent their binding, resulting in either reduced or complete loss of the PCR product. Families such as this have no effect on the linkage analysis since the parents appear uninformative.

The map was constructed as described in Chapter 2.2.4.4. Although the order of loci was initially known from physical mapping, the order was validated using the **build** option of CRI-MAP. Firstly, starting with *the two most informative loci, D16S287* and *D16S292*, loci that could be assigned unambiguously were added sequentially to the map. This framework map was validated by comparison with the physical order, and by the inversion of adjacent loci. Loci that could not be placed into a framework map with interval support of greater than 3 were assigned to their most likely position on a comprehensive map using the **all** option of CRI-MAP, regardless of the support for that position. Sex-specific recombination fractions were determined for each interval, and converted to centimorgans (cM) using the Kosambi mapping function.

Tests were performed to determine if there existed a significant difference between males and females with respect to the recombination frequency, using the methods described in Chapter 2.2.6. The two models tested were:

(i) sex average recombination rates along the map,

(ii) sex-specific recombination rates along the map.

A family segregating for *FRA16A* (Fig. 4.2; Simmers *et al.*, 1987) was used to estimate the distance between each of the flanking markers and the fragile site. The LODSCORE option of the LINKAGE program (Chapter 2.2.3.3) was used to determine peak LOD scores and recombination fractions between these markers and the fragile site. The penetrance assumed for this autosomal folate sensitive fragile site was 50% in males and 100% in females (Sherman and Sutherland, 1986). Using LINKMAP (Chapter 2.2.3.3), the position of the fragile site was determined with respect to the background map generated above.

4.4. Results

Two point lod score analysis of these twelve loci revealed that they form a linkage group along a region of the chromosome 16 (Table 4.2).

The multipoint linkage map, with an ideogram of chromosome 16 in the region of *FRA16A*, is detailed in Fig. 4.3 and Table 4.3. The linear order of loci determined by genetic mapping was found to be consistent with the known physical localisations on the cytogenetic-based physical map. Table 4.3 shows the order of loci, the sex-average and sex-specific map distances between loci, and the odds against the inversion of adjacent loci (for the sex-specific model), given as a measure of local support for order.

The entire map length between *D16S60* and *D16S67* is 34.5 cM in males and 48.6 cM in females. The distance between the closest flanking markers (*D16S79A* and *D16S287*) to *FRA16A* is 4.0 cM in males and 0.0 cM in females. No recombination was detected between *D16S287*, *D16S96* and *D16S79B*, and so the order of these loci could not be distinguished by genetic mapping. However, *D16S287* is the closest marker to the fragile site, on the proximal side, by physical mapping results.

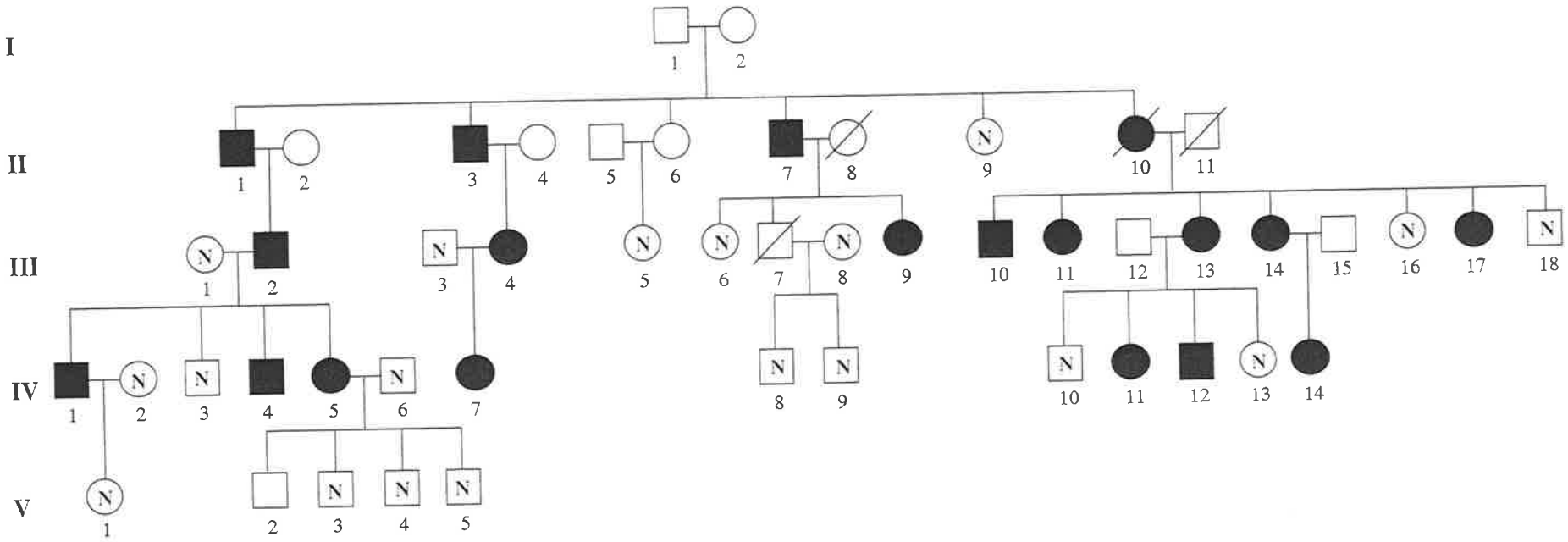

Figure 4.2: *Family segregating for FRA16A. This pedigree has been described previously by Simmers et al., 1987.*

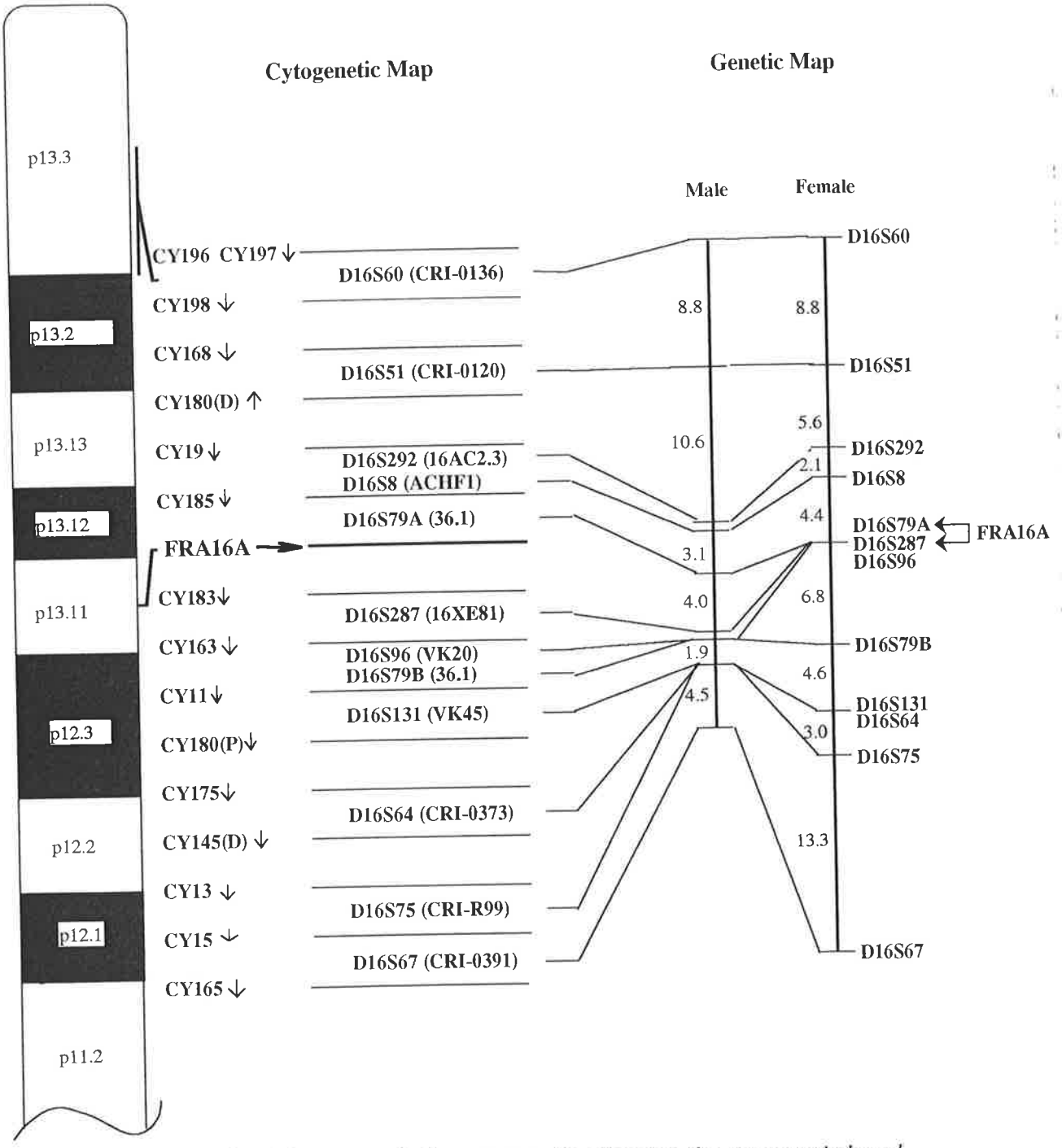

Figure 4.3: *Ideogram of chromosome 16, showing the cytogenetic-based physical map locations and the male and female genetic map of the region encompassing FRA16A. The distances are in cM. D, distal; P, proximal. The genetic map location of FRA16A is indicated.*

Table 4.2: *Summary of pairwise recombination fractions (above the diagonal) and corresponding peak LOD scores (below the diagonal).*

	D16S60	*D16S51*	*D16S292*	*D16S8*	*D16S79A*	*D16S287*	*D16S96*	*D16S79B*	*D16S131*	*D16S64*	*D16S75*	*D16S67*
D16S60	-	0.11	0.15	0.14	0.21	0.18	0.17	0.15	0.20	0.19	0.22	0.34
D16S51	17.20	-	0.10	0.08	0.13	0.13	0.15	0.19	0.18	0.22	0.20	0.33
D16S292	27.88	20.03	-	0.00	0.03	0.06	0.05	0.07	0.09	0.14	0.12	0.21
D16S8	18.41	12.26	64.60	-	0.01	0.05	0.02	0.03	0.08	0.11	0.10	0.18
D16S79A	8.11	6.17	32.33	18.99	-	0.01	0.02	0.03	0.06	0.12	0.16	0.19
D16S287	25.90	17.29	67.31	48.29	35.98	-	0.00	0.03	0.05	0.08	0.11	0.17
D16S96	17.68	9.58	36.96	26.10	20.20	66.09	-	0.00	0.02	0.09	0.11	0.18
D16S79B	5.47	2.93	13.40	5.94	9.71	9.71	15.05	-	0.00	0.04	0.04	0.14
D16S131	9.49	5.32	17.94	12.94	8.07	31.80	16.68	4.82	-	0.00	0.02	0.15
D16S64	11.11	5.01	9.26	5.52	5.95	20.96	13.87	4.82	23.78	-	0.02	0.12
D16S75	9.69	6.48	15.66	10.22	5.30	21.24	16.07	5.37	20.72	31.34	-	0.09
D16S67	3.89	3.47	19.01	12.80	9.67	26.55	15.46	5.89	11.95	19.87	26.09	-

Table 4.3: *A genetic map derived by multipoint tests using CRI-MAP. The distances between loci (in cM) are given in males and females respectively, and then as a measure of support, the odds against the inversion of adjacent loci were calculated.*

Locus	Male Map Distance(cM)	Female Map Distance(cM)	Odds Ratio
D16S60			
	8.8	8.8	2.2×10^2
D16S51			
	10.6	5.6	1.0×10^4
D16S8			
	0.9	2.1	3.9
D16S292			
	3.1	4.4	8.2×10^2
D16S79A			
	4.0	0.0	3.4
D16S287			
	0.7	0.0	4.3
D16S96			
	0.0	6.8	6.5
D16S79B			
	1.9	4.6	1.7
D16S131			
	0.0	0.0	1.0
D16S64			
	0.0	3.0	2.9×10^1
D16S75			
	4.5	13.3	5.6×10^3
D16S67			
Total map length	34.5	48.6	-

Sex-specific differences were assessed, and the male map was determined to be significantly shorter than the female map (chi-square = 32.9, $P < 0.05$, degrees of freedom = 11). Overall there was a 1.4 fold excess of female recombination. Confirmation of tight linkage of this map to *FRA16A* was provided from a kindred segregating with *FRA16A* by two point LOD scores between *FRA16A* and markers from the multipoint map. These LOD scores are given in Table 4.4. The peak LOD score with the closest marker distal to *FRA16A*, *D16S79A*, is $Z = 2.75$, at $\Theta = 0.0$. The genetically unordered cluster of markers *D16S287*, *D16S96* and *D16S79B* are the closest markers proximal to *FRA16A*. The peak LOD score between *FRA16A* and *D16S287* is 4.92 at $\Theta=0.0$, between *FRA16A* and *D16S96* is 0.90 at $\Theta=0.0$, and between *FRA16A* and *D16S79B* is 4.67 at $\Theta=0.08$. These data confirm the proximity of the markers examined to the fragile site by analysis in a family segregating with the fragile site. There was insufficient recombination in the region of the fragile site in this one large *FRA16A* family to enable the closely linked markers to be ordered uniquely or confidently provide estimates for the distances between loci; however, their order and map distances were firmly established from the complementary multipoint genetic and cytogenetic maps.

The LINKMAP analysis of the *FRA16A* family provided a peak location score of 40.9, at $\Theta = 0.0$, for the marker *D16S79A* (Fig. 4.4). This implies that the fragile site shows no recombination with the locus *D16S79A*, and by linkage data, is indeed located on the map generated here. Two point analysis gave a maximum lod score with *D16S287*, since this markers was more informative that *D16S79A*. However, the set of markers closely linked to *D16S79A* were more informative than the set closely linked to *D16S287*, accounting for the peak shifting to *D16S79A*. The data summarised in Fig. 4.4 demonstrates that the *FRA16A* fragile site could be located any where between *D16S292* and *D16S79B*.

The locus *D16S79* has been shown to map to both sides of *FRA16A* by in situ hybridisation (Callen *et al.*, 1992). The 1.0 kb probe *p36-1* detects two *Taq*I

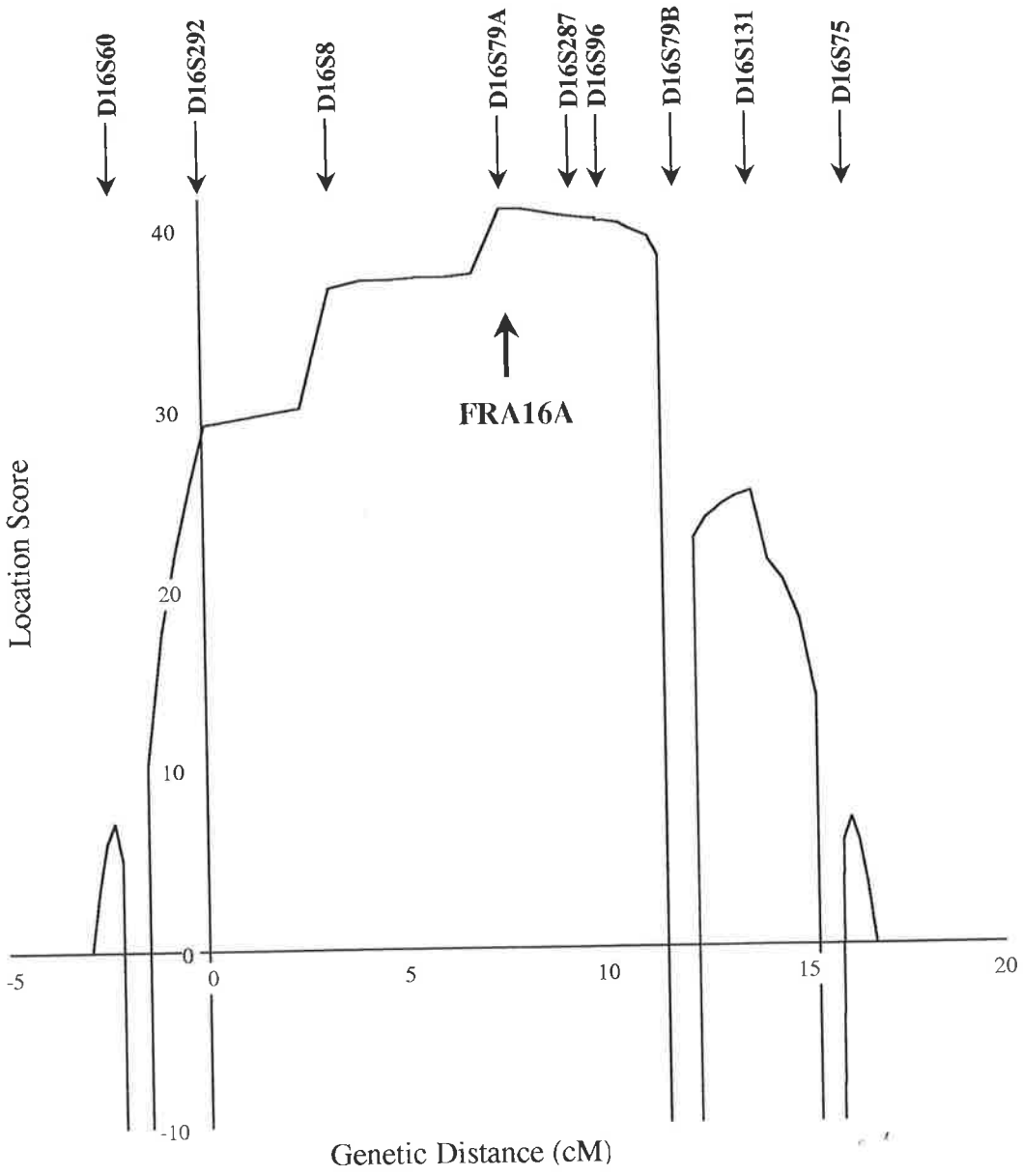

Figure 4.4: *Location scores for the position of FRA16A along the specific region of chromosome 16, determined via LINKMAP analysis of a pedigree segregating for this fragile site (Fig. 4.1). The origin of the map was arbitrarily placed at D16S292. The peak multipoint location score for FRA16A was 40.9, and occurred at 0.0 cM to D16S79A.*

Table 4.4: *Two-point LOD scores for the FRA16A kindred (Previously described by Simmers et al., (1987)).*

Linkage Comparison with *Fra16a*	Lod Score at Recombination Fraction θ									
	0.0	0.01	0.05	0.1	0.2	0.3	0.4	0.5	θ	Z
D16S8	2.74	2.68	2.46	2.18	1.61	1.03	0.46	0.0	0.0	2.74
D16S292	2.44	2.40	2.23	2.02	1.57	1.09	0.56	0.0	0.0	2.44
D16S79A	2.75	2.70	2.50	2.25	1.71	1.08	0.42	0.0	0.0	2.75
D16S287	4.92	4.83	4.46	3.98	2.97	1.88	0.78	0.0	0.0	4.92
D16S96	0.90	0.89	0.84	0.75	0.53	0.30	0.11	0.0	0.0	0.90
D16S79B	-∞	3.95	4.57	4.65	4.03	2.89	1.42	0.0	0.08	4.67
D16S131	0.30	0.31	0.34	0.33	0.26	0.15	0.06	0.0	0.06	0.35
D16S75	-∞	-3.72	-1.64	-0.79	-0.12	0.09	0.10	0.0	0.35	0.11

polymorphisms at this locus in genomic DNA, which appear to be on either side of the fragile site; *D16S79A* detecting one polymorphism on the distal side of *FRA16A*, and *D16S79B* detecting the other polymorphism on the proximal side of *FRA16A*. The genetic linkage map generated here confirms these results, as the two polymorphisms do indeed lie on either side of the fragile site.

The correlation of the genetic and physical map enabled some of the ambiguities of the genetic map to be resolved. The order of loci that were not placed with odds of 1000:1 or greater could, in some instances be determined by comparison with the cytogenetic map. The order of the two loci *D16S287* and *D16S96*, could not be resolved on the genetic map (odds of the inversion of adjacent loci was 10^2:1) (the peak LOD score between these two loci is Z = 66.09 at a recombination fraction of Θ = 0.0), but are ordered on the cytogenetic map, separated by the breakpoint in the hybrid CY163. Pairs of loci *D16S79B* and *D16S131*, and *D16S131* and *D16S64* could not be placed uniquely on the genetic map, but are ordered on the cytogenetic map as they are separated by the hybrids CY11 and CY175, respectively. *D16S79B* is a relatively uninformative marker, and thus difficult to place uniquely by linkage analysis. *D16S8* and *D16S292* could not be uniquely placed on the genetic map, with odds of inversion of adjacent loci of only 10^2:1. On a two point analysis, the peak recombination frequency between *D16S8* and *D16S292* is Θ = 0.02, where there is a maximum LOD score of Z = 58.58, indicating that the loci are tightly linked, with little recombination between them. On the cytogenetic map these loci are not separated by any hybrids, and so are unable to be ordered distinctively by either method.

4.5. Discussion

The map presented in the present study defines a linkage group around the fragile site, *FRA16A*. The final order of loci in the comprehensive map (See Fig. 4.3), along this region is:

pter - D16S60 - D16S51 - D16S8,D16S292 - D16S79A - (FRA16A) - D16S287,D16S96,D16S79B - D16S131 - D16S64 - D16S75 - D16S67 - cen

The largest gap on the map is between *D16S75* and *D16S67*, of 13.3 cM, on the female map. The distance between markers flanking the fragile site is 4.0 cM in males and 0.0 cM in females, indicating that *FRA16A* is likely to be within 1 Mb of at least one of these markers.

This represents a refinement of a preliminary linkage map based on a subset of families and fewer markers (Callen *et al.*, 1989) and is consistent with physical mapping data (Callen *et al.*, 1992). This map represents 30% of the total genetic map of chromosome 16 in males and 25% in females, assuming the total map length of 115 cM in males and 193 cM in females from Keith *et al.*, (1990).

Several alternative orders could not be rejected on the basis of linkage analysis, due to very low odds against the alternative order. Pairs of loci between which there is very little or no recombination cannot be ordered confidently with respect to each other using the likelihood ratio method. Somatic cell hybrids have the potential to provide order independently for those loci separated by chromosomal breakpoints. This approach independently provided the physical order for markers separated by chromosomal breakpoints, enabling the order of loci to be established initially and then validated by genetic mapping. In this study, the pairs of loci: *D16S8* and *D16S292*; *D16S287* and *D16S96*; and *D16S79B* and *D16S131*; and D16S131 and D16S64 could all be ordered definitively on the basis of physical mapping, but were not able to be ordered on the basis of genetic mapping (Fig. 4.3). This demonstrates the value of pursuing more than one form of mapping.

This map is in good agreement with those published elsewhere (Donis-Keller *et al.*, 1987; Keith *et al.*, 1990). The only markers that have been published in other multipoint linkage maps are the five flanking loci extracted from the CEPH

database. The largest discrepancy, being 5.5 cM, is between the loci *D16S51* and *D16S64* in females. The distance between these markers in males and females on the map published by Keith *et al* (1990) is 20 cM and 29 cM respectively. On the map presented here, the distances are 21.2 cM and 23.5 cM respectively. Between these loci on the map presented in this chapter are new different intervening markers. This new information has an effect on genetic map distances. All other discrepancies between this and earlier maps are less than 2 cM. The estimates of genetic distances on the map published here would be expected to be more accurate than previously published maps of this region of chromosome 16 because 8 additional markers, some of which are highly informative, have been characterised and incorporated within the map.

Inspection of the data revealed a tight linkage group around the fragile site. There is an overall 1.4 fold excess in recombination in females which is significant. This has been observed elsewhere (Donis-Keller *et al.*, 1987; Keith *et al.*, 1990; Chapter 3) and in the published CEPH consortium maps (White *et al.*, 1990; Dracopoli *et al.*, 1991; Spurr *et al.*, 1992; Bowcock *et al.*, 1993a; 1993b; Attwood *et al.*, 1994).

An interesting feature of this map is the fact that the *D16S79* locus is duplicated on either side of *FRA16A*. The distance between these loci is approximately 1.8 Mb (Julie Nancarrow, personal comminucation). Other regions of chromosome 16 have been shown to be duplicated also (Germino *et al.*, 1992; European Chromosome 16 Tuberous Sclerosis Consortium, 1993; The European Polycystic Kidney Disease Consortium, 1994). The gene responsible for PKD1 has been confirmed to be *PBP*, and is located adjacent to the *TSC2* locus in a region that is duplicated more proximally on 16p, making isolation of the gene responsible for the disease quite difficult.

The very latest genetic map for this region determined after the isolation of *FRA16A* sequences, can be found in Chapter 8, Fig. 8.1; this map is the CEPH

Consortium Map (Kozman *et al.*, 1994). The distances on the comprehensive map from the consortium map are in good agreement with the map reported here, the largest discrepancy being 3.3 cM between *D16S75* and *D16S67*, on the female map. The sex-average map length on the comprehensive consortium map of this region is 44.1 cM, in comparison with 43.2 cM on the map generated in the current study. On the Consortium Map, there are 41 loci along this region, of which 21 are PCR formatted. The distance across the fragile site is 0.6 cM in females and 2.9 cM in males. A new $(AC)_n$ polymorphism associated with the locus *D16S79A* has been isolated which is more proximal to *FRA16A* than the *Taq*I RFLP of *D16S79A*. Thus on the consortium map, the flanking markers are *D16S79A* defined by the AC repeat polymorphism and *D16S287*.

The known cytogenetic location of fragile sites and translocation breakpoints for many human/mouse somatic cell hybrids allows this genetic map to be anchored to the physical map. For the region near *FRA16A*, this multipoint linkage map has provided the basis for characterisation of the critical region responsible for *FRA16A*, by using closely linked flanking markers to identify and construct YAC and cosmid contigs spanning this fragile site. YACs containing DNA from this region have been isolated, and a YAC contig of the region has been constructed (Nancarrow *et al.*, 1994), and sequences found to be methylation sensitive were located adjacent to a p(CCG)n sequence. Hence *FRA16A* has now been cloned and sequenced (Nancarrow *et al.*, 1994). The structure of FRA16A appears to be remarkably similar to its X-linked counterparts - a normally polymorphic p(CCG)n repeat adjacent to a CpG island with individuals expressing the fragile site having expansion of the repeat sequence, and methylation of the CpG island. It is also apparent that fragile sites can originate from regions not normally associated with methylation, suggesting that the observed methylation associated with fragile site expression, is a consequence of the fragile site mutation, and not a cause of such a mutation (Nancarrow *et al.*, 1994). As other fragile sites are isolated and

characterised, more information can be obtained to determine the relationship between the sequence composition of the fragile site and the chemistry of their induction.

CHAPTER 5

THE DEVELOPMENT OF THE BACKGROUND GENETIC MAP FOR THE LOCALISATION OF BATTEN DISEASE

Table of Contents

Associated Papers (See Appendix A):

Fine genetic mapping of the Batten Disease locus (CLN3) by haplotype analysis and demonstration of allelic association with chromosome 16p microsatellite loci. Mitchison, H.M., Thompson, A.D., Mulley, J.C., Kozman, H.M., Richards, R.I., Callen, D.F., Stallings, R.L., Doggett, N.A., Attwood, J., McKay, T.R., Sutherland, G.R., and Gardiner, R.M. 1993. *Genomics* **16**:455-460

Dinucleotide repeat polymorphism at the D16S288 locus. Shen, Y., Holman, K., Thompson, A., Kozman, H.M., Callen, D.F., Sutherland, G.R., and Richards, R.I.. 1993. *Nucl. Acids Res.* **19**:5445.

Refined genetic localisation for Central Core Disease. Mulley, J.C., Kozman, H.M., Phillips, H.A., Gedeon, A.K., McCure, J.A., Iles, D.E., Gregg, R.G., Hogan, K., Couch, F.J., MacLennan, D.H., and Hann, E.A. 1993. *Am. J. Hum. Genet.* **52**:398-405.

Genetic Mapping of the Batten disease locus (CLN3) to the interval D16S288-D16S383 by analysis of haplotypes and allelic association. Mitchison, H.M., Taschner, P.E.M., O'Rawe, A.M., de Vos, N., Phillips, H.A., Thompson, A.D., Kozman, H.M., Haines, J.L., Schlumpf, K., D'Arigo, K., Boustany, R.-M.N., Callen, D.F., Breuning, M.H., Gardiner, R.M., Mole, S.E., and Lerner, T.J. 1994. *Genomics* **22**:465-468.

Regional localisation of a second non-specific X-linked mental retardation gene (MRX19) to Xp22. Donnelly, A.J., Choo, A., Kozman, H.M., Gedeon, A.K., Danks, D.M., and Mulley, J.C. 1994. *Am. J. Med. Genet.* **15**:581-585.

Clinical and Linkage study of a large family with simple Ectopia Lentis linked to FBN1. Edwards, M.J., Challinor, C.J., Colley, P.J., Roberts, J., Partington, M.W., Holloway, G.E., Kozman, H.M., and Mulley, J.C. 1994. *Am. J. Med. Genet.* In Press.

Autosomal dominant distal myopathy: linkage to chromosome 14. Laing, N.G., Laing, B.A., Meredith, C., Wilton, S.D., Robbins, P.,

Honeyman, K., Dorosz, S., Kozman, H., Mastaglia, F.L., and Kakulas, B.A. 1994. *(Submitted)*.

5.1 Summary

The order of loci on chromosome 16 in the region of the Batten disease, defined on the basis of joint physical and genetic mapping, is:

D16S159-(D16S412-D16S417-D16S294)-D16S403-D16S420-(D16S401-D16S319)-D16S67-D16S295-D16S296-D16S297-D16S148-D16S313-D16S288-(D16S298-D16S299)-(D16S48-SPN-D16S383)-D16S57-D16S300-(D16S285-D16S261)-D16S308

The linkage map of this area includes 27 markers, defining 24 loci, 21 of which are microsatellites, with an average heterozygosity of 73%. The order of the loci in brackets could not be distinguished with odds of greater than 1000:1. The *CLN3* locus, responsible for Batten disease, is localised to the region between flanking markers *D16S288* and *D16S383*, with a sex average distance of 1.6 cM across this interval.

Linkage analysis followed by disequilibrium mapping has been exploited in refining the localisation of Batten disease, to identify close flanking markers, and thus enabling the development of a high resolution cosmid/YAC map to proceed in the selected region containing the *CLN3* disease gene. This will ultimately lead to the cloning and characterisation of this disease locus.

5.2. Introduction

Strategies now exist for the molecular cloning of inherited disease genes known only by their phenotype. The first step is the localisation of the disease mutation to a specific region of the human genome, using genetic linkage analysis. For an autosomal recessive condition this necessitates pooling of information from a number of families on the assumption that all arise from mutations at the same locus. Many disease genes have now been localised to specific regions of the human genome (McKusick & Amberger, 1994). For example, a genetic linkage map of chromosome 14 has enabled the localisation of a gene (*MPD1*) for autosomal dominant distal myopathy to a specific region of this chromosome (Laing *et al.*, 1993). Similarly, background genetic maps have enabled the regional localisation of Ectopia Lentis (*EL*) to chromosome 15 (Edwards *et al.*, 1994), Central Core disease (*CCO*) to chromosome 19 at 19q13.1 (Mulley *et al.*, 1993), and the localisation of the non-specific X-linked mental retardation gene, *MRX19*, to a specific region of the X chromosome, Xp22 (Donnelly *et al.*, 1994). Deletion mapping, homozygosity mapping or disequilibrium mapping (Lander & Botstein, 1987) may be used to further reduce the localisation from that originally defined by the linkage map. Resolution of less than 1 cM for the location of a disease gene is rarely provided by linkage analysis; the fine mapping provided by linkage disequilibrium for example, has the potential to reduce the localisation to a few hundred kb, thus simplifying the search for a candidate gene (Mulley *et al.*, 1993). High resolution physical mapping can then be undertaken by the development of YAC and cosmid contigs in the area. Once the element of the contig containing the disease gene has been identified, the eventual isolation of the gene responsible for the disease is feasible using recently developed methods (Monaco, 1994). The techniques of linkage analysis followed by disequilibrium mapping have been applied to the localisation of the *CLN3* gene for Batten disease.

Batten disease (Juvenile neuronal ceroid lipofuscinosis; *JNCL*) is one of four sub-types of the progressive encephalopathies, neuronal ceroid lipofuscinosis (*NCL*); the juvenile onset sub-type. It is an autosomal recessive neurodegenerative disorder characterised by the accumulation of autofluorescent lipopigments in the neurons and other tissues (Dyken, 1988). Clinically, this disorder is characterised by progressive encephalopathy, loss of vision, seizures, and death. *CLN3* is the gene responsible for *JNCL*.

Genetic linkage studies determined the chromosomal location of *CLN3* to chromosome 16. Initial studies demonstrated linkage between the *CLN3* gene and *HP* (Eiberg *et al.*, 1989). Following this assignment to chromosome 16, multipoint linkage analysis of genetically linked markers was carried out on 42 families segregating for *JNCL*. From these studies the regional localisation of the *CLN3* gene was further refined to an interval between *D16S150* and *D16S148* on the short arm of chromosome 16 (Gardiner *et al.*, 1990), a considerable distance from the *HP* locus. Subsequent multipoint analysis of genetically linked markers enabled a refined localisation to an interval between *D16S67* and *D16S148* (Callen *et al.*, 1991).

Physical mapping of the genetic markers in this interval, using the mouse/human somatic cell hybrid panel of chromosome 16, positioned them between the bands 16p12.1 and 16p12.3; between the breakpoints of CY15 and CY160(D). Additional physical intervals have been defined in this region (Fig. 5.1) (Callen *et al.*, 1992b), providing the opportunity for further refinement of the localisation of *CLN3*.

Cloning and sequencing of cosmid contigs (Stallings *et al.*, 1990; 1992) mapped to 16p12 led to the identification of 10 multiallelic microsatellite marker loci (highly informative polymorphic (AC-TG)n repeats) located in this region. A high resolution genetic map was then constructed using these loci and five traditional

RFLPs, spanning a distance of 44 cM on the sex-average genetic map[1]. The order of the loci incorporated on this map was:

D16S159-D16S294-D16S319-D16S67-D16S295-D16S296-D16S297-D16S148-D16S288-(D16S298-D16S299)-D16S57-D16S285-D16S150-D16S151

D16S159 and *D16S57* were not included in the genetic linkage analysis, but were positioned according to their physical location. The order of loci in brackets could not be distinguished with odds of greater than 1000:1.

Using this background genetic map, 70 JCLN pedigrees, and construction of haplotypes for the analysis of recombinant events, the *CLN3* locus was localised to between *D16S297* and *D16S57*, with a sex-average distance across this interval of 8 cM (Mitchison *et al.*, 1993). Within this interval, strong linkage disequilibrium was detected between the *CLN3* gene and three multiallelic loci, *D16S288*, *D16S298*, and *D16S299*, thus further refining the likely localisation of the disease gene. This disequilibrium was lost after *D16S148* and *D16S57*, and no association was detected between *CLN3* and alleles at any other loci tested outside this interval. The strong allelic association (linkage disequilibrium) identified between *CLN3* and alleles at these three loci, (*D16S288*, *D16S298*, and *D16S299*), predicts that the disease has arisen via a predominant ancestral mutational event (Mitchison *et al.*, 1993), implying a founder effect in juvenile onset neuronal ceroid lipofuscinosis.

The initial finding that the *CLN3* gene was linked to *HP* (situated on the q arm of chromosome 16) was surprising considering the findings of subsequent studies, that the disease gene was actually a considerable distance away on 16p (Callen *et al.*, 1991). This initial linkage was likely to have arisen from the low recombination across the centromere in males (Gardiner *et al.*, 1990) and the inclusion of two Batten families unlinked to this region of 16p, which have now

---

1. This genetic map was constructed by the candidate, and the manuscript (Mitchison *et al.*, 1993) is included in Appendix A. This was an earlier map, to which loci have been added to construct a the map described in this chapter.

been included in a new clinical entity, late juvenile onset neuronal ceroid lipofuscinosis.

Subsequent cosmid contig mapping and isolation of microsatellite markers enabled nine additional microsatellite loci to be physically assigned to this region (Weissenbach *et al.*, 1992; Shen *et al.*, 1994). A total of 27 marker loci (including 21 multiallelic microsatellite markers, with an average heterozygosity of 73%) have now been localised to the *CLN3* region at 16p12 by physical mapping.

Described here is the construction of a genetic linkage map using genotype data generated from the CEPH families, for the entire 27 loci isolated to the *CLN3* region. Using this background map, *CLN3* was subsequently localised to a smaller region on chromosome 16, between two flanking markers, *D16S288* and *D16S383*. This refined localisation is providing the basis for the cloning of the *CLN3* gene.

5.3. Materials and Methods

Table 5.1 describes the loci used in the construction of the background genetic map in the vicinity of the Batten disease. Of the 27 polymorphisms, describing 24 loci, 6 are traditional RFLPs and 21 are microsatellite markers, one of which (*SPN*) represents the gene locus, Sialophorin. All of the loci used in the genetic map were isolated, characterised, physically mapped and genotyped by colleagues, as described in Table 2.1.

Five of the RFLP markers (*D16S159*, *D16S67* (two polymorphisms), *D16S148*, and *D16S48*) were typed using standard Southern hybridisation analysis methods as outlined in Chapter 2 and by Callen *et al.*, (1991). Locus *D16S57* was detected also using Southern analysis, as previously described (Keith *et al.*, 1990). All the RFLP loci were extracted from the CEPH database Version 5. Four dinucleotide repeat loci, *D16S401*, *D16S412*, *D16S417*, and *D16S420*, were described by

Table 5.1: *Loci used in the search for the CLN3 disease gene.*

Location	Locus	Probe/Enzyme	Het.	Type	Inf. Meioses
16p13.11-p12.1	*D16S159*	pCJ52.94T1/TaqI	0.50	RFLP	455
16p13.11-p12.1	*D16S412*	AFM191wb10/PCR	0.76	DINU	101
16p13.11-p12.1	*D16S417*	AFM220xb10/PCR	0.73	DINU	155
16p13.11-p12.1	*D16S294*	16AC1/PCR	0.49	DINU	235
16p13.11-p12.1	*D16S420*	AFM238xb2/PCR	0.82	DINU	149
16p13.11-p12.1	*D16S401*	AFM025tg9/PCR	0.74	DINU	80
16p13.11-p12.1	*D16S319*	16AC7.14/PCR	0.69	DINU	315
16p13.11-p12.1	*D16S67*	16ACCRI-O391/PCR	0.83	DINU	393
	D16S67	CRI-O391/HincII	0.46	RFLP	128
	D16S67	CRI-O391/TaqI	0.50	RFLP	198
16p13.11-p12.1	*D16S295*	16AC62F3/PCR	0.66	DINU	430
16p13.11-p12.1	*D16S296*	16AC62B4/PCR	0.75	DINU	502
16p13.11-p12.1	*D16S297*	16AC15H1H/PCR	0.87	DINU	412
	D16S297	16AC15H1S/PCR	0.69	DINU	142
16p13.11-p12.1	*D16S148*	pCJ52.95M1/MspI	0.39	RFLP	326
16p13.11-p12.1	*D16S313*	MS79/PCR	0.57	DINU	498
16p12.1-p11.2	*D16S288*	16AC7.1/PCR	0.73	DINU	449
16p12.1-p11.2	*D16S298*	16AC3.12/PCR	0.80	DINU	455
16p12.1-p11.2	*D16S299*	16AC6.17/PCR	0.72	DINU	425
16p12.1-p11.2	*D16S48*	CRI-O101/HindIII	0.47	RFLP	144
16p12.1-p11.2	*sialophorin gene*	SPN/PCR	0.96	DINU	463
16p12.1-p11.2	*D16S383*	16AC80B3/PCR	0.45	DINU	235
16p11.2-cen	*D16S57*	CRI-O131/BamHI	0.40	RFLP	79
16p11.2-cen	*D16S300*	16AC1.1/PCR	0.60	DINU	436
16q12.1-q13	*D16S285*	AluGT16/PCR	0.82	DINU	494
16q12.1-q13	*D16S261*	Mfd24/PCR	0.71	DINU	449
16q12.1-q13	*D16S308*	16AC1.18/PCR	0.74	DINU	113

Weissenbach *et al.* (1992), and were extracted from the CEPH database, Version 6. The remaining multiallelic dinucleotide repeat polymorphisms were genotyped by PCR amplification and analysis of denaturing polyacrylamide gels (Thompson *et al.*, 1992; Mitchison *et al.*, 1993; Shen *et al.*, 1991), at the WCH.

Construction and characterisation of the high resolution cytogenetic-based physical map and localisation of all the loci incorporated into this map is described by (Callen *et al.*, 1992), with the exception of the markers SPN and *D16S285, D16S383*, *D16S401*, *D16S412*, *D16S417*, and *D16S420*, for which the physical mapping is described by Shen *et al.* (1994). These loci have been integrated into the physical map by PCR analysis of the chromosome 16 mouse/human hybrid DNA panel (Richards *et al.*, 1991c).

The candidate performed the multipoint linkage analysis using genotypes of the CEPH pedigrees and the linkage program CRI-MAP, as outlined in Chapter 2.2. In brief, the **build** option of CRI-MAP was used to generate a framework genetic map of markers which are placed with odds of 1000:1 or greater. The two most informative loci, *D16S297* and *SPN*, were selected as a pair of tightly linked, highly informative loci to be the foundation on which to build the map. These two loci are separated by a distance of 5 cM (as determined by two-point analysis), with a LOD score of 81. From the construction of this framework map, loci for which there exists more than one marker with no observable intragenic recombinants on this initial map, were considered as a haplotype in the analysis. Loci for which this was true were *D16S67* and *D16S297*.

Once the framework map had been constructed, the option **chrompic** was executed to check for obvious data errors, evident as close double recombinants (See Table 5.2), as described in Chapter 2.2.5.

Using the **all** option of CRI-MAP, the remaining loci not placed on the framework map were incorporated on the map to generate the comprehensive map. The data were checked again for errors using the **chrompic** option (Table 5.2).

Table 5.2: *Individuals with identified genotype errors after the* ***chrompic*** *analysis.*

Locus	Individual	Locus	Individual	Locus	Individual
D16S319	1344-06	*D16S296*	884-07	*D16S300*	1333-04
	1345-03	*D16S288*	1347-07		-09
	-07		1331-05		28-09**
	-09		1362-08	*D16S285*	12-01**
D16S297	1332-06	*D16S299*	35-08		21-03
	-07*		1331-04		21-04
	1334-04*		1345-09#		1345-05**
	-09*	*D16S298*	1362-08		1375-06**
	1362-07		1331-04		1423-01**
	-10	*D16S48*	1413-09«		13291-01**
	-17	*D16S300*	1333-03		

* *D16S297* had a segregating null allele in family ; these individuals could not be conclusively genotyped, and so were coded as unknown.

\# new mutation detected as an allele not segregating in the family.

« RFLPs were not regenotyped and were coded as unknown.

** these individuals were involved in double recombinants, causing the wrong order between *D16S261* and this marker. The new genotypes were difficult to interpret and so these individuals have been coded unknown for this analysis.

After corrections to the database were made, the map was reconstructed using the same pair of starting loci, and also several different pairs; the resultant order of markers was tested against order derived from the cytogenetic-based physical map. Differences in the recombination fractions between males and females were investigated by comparing the likelihood values of the comprehensive map for two models (as described in Chapter 2.2.6):

(i) Sex-averaged recombination values over the entire region and

(ii) Sex-specific values over the entire region.

For further refinement of the *CLN3* gene localisation in this area on chromosome 16, 70 JCLN pedigrees were analysed, by Dr HM Mitchison, under the supervision of Prof RM Gardiner at the Department of Paediatrics, Rayne Institute, London, using the background map determined by the analysis of the CEPH pedigrees.

5.4. Results

Linkage analysis using CRI-MAP, was employed to construct both framework and comprehensive maps for this region (Fig. 5.1, Table 5.3) using genotype data from the CEPH reference families. The framework loci are indicated on Table 5.3. The length of the sex-average comprehensive map is 25.8 cM; the length of the male and female maps are 12.0 cM and 39.1 cM respectively. The order of loci on the genetic map is in complete agreement with the physical order as determined by the cytogenetic-based physical map, and is:

D16S159-*(D16S412-D16S417-D16S294)*-*D16S403-D16S420-(D16S401-D16S319)*-***D16S67-D16S295-D16S296-D16S297-D16S148-D16S313-D16S288***-*(D16S299-***D16S298***)-(D16S48-**SPN**-D16S383)-D16S57-D16S300-(**D16S285**-D16S261)-***D16S308***

The order of loci in brackets could not be distinguished with odds of greater than 1000:1, or by physical mapping. Framework loci are indicated in bold type.

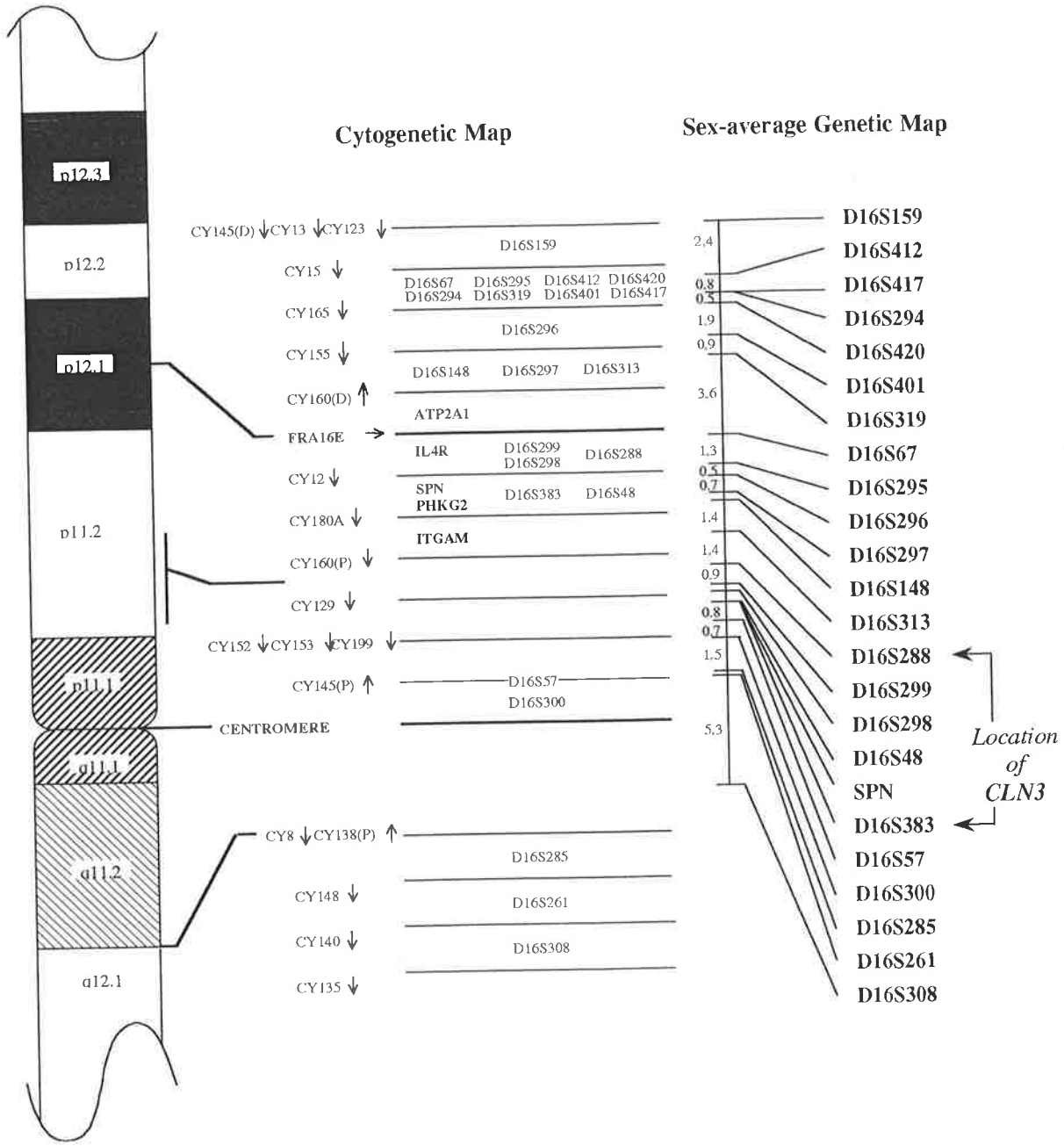

Figure 5.1: *Ideogram of Chromosome 16 showing the cytogenetic localisations of the genetic markers used in the linkage analysis. The comprehensive map is shown, with the distances between loci given in cM. The position of CLN3, the gene responsible for Batten disease is indicated. The genes in the region are also included*

Table 5.3: *Comprehensive map of the Batten disease region. Inversion of adjacent loci has been calculated based on the sex-averaged map. Framework loci are indicated in bold type.*

Locus	Male Map	Female Map	Sex-Average Map	Odds (x:1)
D16S159				
	2.0	3.1	2.4	10^3
D16S412				
	0.0	1.2	0.8	10^2
D16S417				
	0.0	0.0	0.0	1.0
D16S294				
	0.0	0.8	0.5	10^3
D16S420				
	0.0	3.9	1.9	10^{10}
D16S401				
	1.3	0.0	0.9	10^2
D16S319				
	2.9	4.8	3.6	10^{29}
D16S67				
	0.0	2.3	1.3	10^{12}
D16S295				
	1.1	0.0	0.5	10^6
D16S296				
	0.0	1.3	0.7	10^{15}
D16S297				
	0.0	0.9	0.3	10^3
D16S148				
	0.6	2.0	1.4	10^{14}
D16S313				
	0.3	2.5	1.4	10^{22}
D16S288				
	0.3	1.8	0.9	10^{11}
D16S299				
	0.3	0.0	0.3	3.0
D16S298				
	0.0	0.3	0.4	1.0
D16S48				
	0.0	0.8	0.0	1.0
SPN				
	0.0	0.0	0.0	1.0
D16S383				
	0.8	0.7	0.8	1.0
D16S57				
	0.8	0.6	0.7	1.0
D16S300				
	1.6	1.3	1.5	10^7
D16S285				
	0.0	0.5	0.2	10^1
D16S261				
	0.0	10.3	5.3	10^{13}
D16S308				
Totals	12.0	39.1	25.8	

The results from the inversion of adjacent loci, performed to provide support for the placement of loci on the comprehensive sex-average map, are given in Table 5.3. The order of some of the loci not distinguished by odds of 1000:1 or greater on the genetic map, can be resolved by physical mapping information (see fig. 5.1). For example, the loci *D16S298* and *D16S48*, can be ordered on the basis of being separated by the breakpoint of the somatic cell hybrid CY12; *D16S383* and *D16S57* are separated by breakpoint of four hybrid cell lines; and *D16S57* and *D16S300* are separated by the centromere (Callen *et al.*, 1994).

From the analysis using **chrompic**, 28 double recombinant events were detected and are described in Table 5.2. One new mutation was detected in the individual 1345-09 for the locus *D16S299*, and was coded as unknown for the analysis, as the new allele did not segregate in the family. A null allele was detected in the locus *D16S297*, and three individuals were coded as unknown, as their genotypes could not be verified. These individuals were 1332-07, 1334-04 and 1334-09 (Table 5.3). The genetic order of the two loci *D16S285* and *D16S261* were incongruous with the order determined by physical mapping. The physical location of these loci was rechecked and the localisations were found to be correct. Analysis of the genotype data revealed several double recombinants for *D16S285*. Regenotyping for this marker, several times, for the individuals 12-01, 1345-05, 1375-06, 1423-01, and 13291-01 could not be clearly determined on the autoradiographs, and so the genotypes were coded as unknown. This resulted in the loci being placed in the order as determined by physical mapping.

Sex differences in the recombination values were investigated. The length of the male map is 12.0 cM, and the length of the female map is 39.1 cM, representing a 3 fold excess of female recombination to male recombination, which is significant (chi-squared = 58.3, $P < 0.05$, df = 23). Over the entire region there is no interval where male recombination is greater, and in fact there appears to be an overall suppression of male recombination in this region.

5.5. Discussion

Extensive physical and genetic mapping methods have been employed in the construction of the comprehensive map of this region. Nine new microsatellite loci have been added to the previously published background map (Mitchison *et al.*, 1993), which was used to localise the *CLN3* locus to the region between *D16S297* and *D16S57*. Framework and comprehensive maps were constructed for the *CLN3* disease gene region on chromosome 16 (Fig. 5.1, Table 5.2). The length of the sex-average comprehensive map is 25.8 cM; the length of the male and female maps is 12.0 cM and 39.1 cM respectively. The order of the genetic map is in complete agreement with the cytogenetic-based physical map.

Sex differences in the recombination values were investigated, revealing a significant 3 fold increase in recombination in females over males, as has been observed previously for this region of the chromosome (Gardiner *et al.*, 1990; Julier *et al.*, 1990). Over the entire region there is in fact no interval where male recombination is greater, and there appears to be a suppression of male recombination in this region. This suppression of male recombination in the centromeric region has been observed elsewhere (Gardiner *et al.*, 1990; Skolnick, 1991).

Following the linkage analysis, homozygosity mapping and disequilibrium mapping (Lander and Botstein, 1987) were utilised to further refine the localisation of Batten disease (Mitchison *et al.*, 1993; Mitchison *et al.*, 1994). Using this new background comprehensive map, the analysis of disease haplotypes for four microsatellite markers in this interval, *D16S288*, *D16S299*, *D16S298*, and *SPN*, revealed significant allelic association between one allele at each of these loci and *CLN3* (Mitchison *et al.*, 1994). This implies that the *CLN3* gene lies between *D16S288* and *D16S383* over a distance of 1.6 cM on the present map (Mitchison *et al.*, 1994). As the distances between the loci on the genetic map are small, genetic mapping can no longer provide significant additional information

for the further refinement of the disease gene, and so physical mapping techniques must be employed to assemble the contig containing the *CLN3* gene. Genes need to be isolated from this contig (Monaco, 1994) and tested for mutations such as deletions or rare rearrangements in order to identify the gene responsible.

Another report has located the *CLN3* gene to the 16p12.1 region. Lerner *et al.* (1994) detected strong linkage disequilibrium with one *D16S289/D16S299* haplotype, and suggested that the *CLN3* gene lies closer to these two loci than to *D16S288*. *D16S298* and *D16S299* lie approximately 0.3 cM (sex-averaged) apart on the genetic linkage map and about 1 cM from *D16S288* (from results determined by the candidate), although the relative order of these loci has not been determined by Lerner *et al.*

Physical mapping has been performed on this region, and YAC and cosmid contigs are being constructed. A YAC contig of approximately 860 kb has been assembled which contains three of the four loci in linkage disequilibrium (*D16S288*, *D16S298*, and *D16S299*) (Järvelä *et al.*, 1994), delineating the order of these loci as *D16S288*, *D16S298*, and *D16S299* from 16pter to the centromere. Detailed restriction mapping of the region has revealed several CpG islands, which are under investigation to find candidate genes. Three genes known to map to the region have been excluded as candidate genes for the Batten disease (*IL4R*, *UQCRC2* and *SGLT2*); *SPN* is not contained on the 860 kb contig, and is thus not a candidate gene. The gene responsible for Batten disease remains elusive.

The genetic map with 9 additional microsatellite markers (Fig. 5.1) has helped to refine the localisation of *CLN3*. This background map has provided close flanking markers suitable for initiating YAC and cosmid contig construction, demonstrating the power provided by a high resolution genetic map of microsatellite marker loci for the localisation of a rare autosomal recessive disease. This linkage map will improve the reliability of prenatal diagnosis of Battens disease, pending the

identification of the disease gene, as well as providing the starting point for genome walking to isolate and ultimately clone the disease gene.

CHAPTER 6

THE INTEGRATION OF THE GENETIC AND CYTOGENETIC MAPS FOR HUMAN CHROMOSOME 16

Table of Contents

Associated Papers (See Appendix A):

Chromosome 16. In: A comprehensive genetic linkage map of the human genome. (NIH/CEPH Collaborative Mapping Group). Kozman, H.M., Sutherland, G.R., and Mulley, J.C.. (1992). *Science* **258**: 67-162.

Integration of the cytogenentic and genetic linkage maps of human chromosome 16 using 50 physical intervals and 50 polymorphic loci. Kozman, H.M., Phillips, H.A., Callen, D.F., Sutherland, G.R., and Mulley, J.C. (1993). *Cytogenet and Cell Genet* **62**:194-198.

6.1. Summary

A comprehensive genetic linkage map of 50 loci has been anchored to 50 cytogenetically defined intervals on human chromosome 16. The map was constructed by multipoint linkage analysis of genotypes from the CEPH reference pedigrees from 50 loci represented by 68 markers, using the linkage analysis program, CRI-MAP. The genetic loci have all been physically mapped against the human/rodent hybrid cell panel, and the independently derived orders of both the genetic map and the cytogenetic-based physical map are completely compatible. The sex average length is 164.5 cM with an average resolution of 3.3 cM. The length of the male and female maps are 132.8 cM and 201.3 cM respectively. This is the first detailed synthesis of genetic and cytogenetic maps for any human chromosome, and is the first step toward correlating the genetic and physical maps for chromosome 16.

chromosome have already been derived (Kozman *et al.*, 1991; 1993 (Chapters 3 and 4); Mitchison *et al.*, 1993; 1994 (Chapter 5)), as well as independently constructed chromosome maps (Donis-Keller *et al.*, 1987; Julier *et al.*, 1990; Keith *et al.*, 1990).

Similarly, comprehensive cytogenetic-based physical maps of this chromosome have been published (Chen *et al.*, 1991; Callen *et al.*, 1992), based on a panel of human/rodent hybrid cell lines, subdividing the chromosome into 50 intervals averaging 2 Mb in length. Physical mapping is rapid and does not rely on the identification of informative polymorphisms as with genetic methods. The human/rodent hybrid cell line panel provides the simplest approach to the determination of the physical order and localisation of anonymous DNA probes and STS's, and of cDNAs and EST's.

This chapter describes the construction of a genetic linkage map of human chromosome 16, covering 99% of the chromosome, which is closely integrated with the cytogenetic-based physical map, representing the first step toward the development of a complete map of human chromosome 16.

6.3. Materials and Methods

The linkage analysis was based on the 68 polymorphic DNA probes, defining 50 loci, described in Table 6.1. Included in the analysis are 5 marker loci for 3 genes, 6 microsatellite markers of the type $(AC)_n$, and 57 cloned DNA sequences detected by traditional RFLPs. The majority of the loci included in this map have been incorporated in previously published genetic maps for this chromosome (Donis-Keller *et al.*, 1987; Julier *et al.*, 1990; Keith *et al.*, 1990; Kozman *et al.*, 1991; 1993), with the exception of the six microsatellite loci. Nine of the loci incorporated into this genetic map are classified as index markers, with heterozygosities greater than 70% (See Table 6.1; Roberts, 1990). Thirty seven loci previously incorporated in published genetic maps of this chromosome were

Table 6.1: *Chromosome 16 markers used in the analysis.*

Locus	Probe/Enzyme	No. of Alleles	Het. (Obs.)	Chromosome Location
D16S85	3'HVR,5'HVR/RsaI,MspI	9,6	0.97,0.79	16p13.3
D16S83	pEKMDA2I/HinfI	8	0.87	16p13.3
D16S84	pCMM65/EcoRI	2	0.39	16p13.3
D16S94	pVK5B/MspI	2	0.49	16p13.3
D16S63	CRI-O327/HindIII	2	0.38	16p13.3
D16S45	CRI-O90/EcoRI,DraI	2,2	0.62,0.14	16p13.3
D16S56	CRI-O129/BglII,EcoRI	2,2	0.12,0.33	16p13.3
D16S55	CRI-O128/HincII	2	0.24	16p13.3
D16S58	CRI-O133/HindIII	2	0.43	16p13.3
D16S60	CRI-O136/HincII	6	0.79	16p13.2-16p13.3
D16S51	CRI-O120/BamHI	3	0.64	16p13.2
D16S49	CRI-O114/EcoRI	3	0.43	16p13.2
D16S292	16AC2.3/PCR	10	0.78	16p13.12-16p13.13
D16S83	pACHF1.1A6/PvuII	2	0.59	16p13.12-16p13.13
D16S79A	p36-1/TaqI	2	0.34	16p13.11-16p13.12
D16S287	16XE81/PCR	10	0.78	16p13.11
D16S96	pVK20A/TaqI,MspI	4	0.59	16p13.11
D16S131	pVK45C6/TaqI	2	0.39	16p12.3
D16S42	CRI-O66/PstI	2	0.31	16p12.3
D16S64	CRI-O373/HindIII	4	0.64	16p12.3
D16S75	CRI-R99/HindIII,MspI	3,2	0.52,0.19	16p12.3
D16S67	CRI-O391/HincII,TaqI	2,3	0.38,0.60	16p12.2-16p12.3
D16S148	CJ52.95M1/MspI	4	0.35	16p12.1-16p12.2
D16S48	CRI-O101/HindIII	2	0.45	16p11.2-16p12.1

Table 6.1 *(cont.)*: *Chromosome 16 markers used in the analysis.*

Locus	Probe/Enzyme	No. of Alleles	Het. (Obs.)	Chromosome Location
D16S261	MFD24/PCR	6	0.71	16q12.1-16q13
D16S52	CRI-O123/HindIII	2	0.33	16q12.1-16q13
D16S150	CJ52.161/TaqI	2	0.41	16q12.1-16q13
D16S39	CRI-O3/PstI	2	0.52	16q12.1-16q13
D16S65	CRI-O377/HincII,TaqI,BglII	2,3,2	0.31,0.19,0.50	16q13
D16S164	16PHAC-15/PCR	4	0.38	16q21
D16S151	CJ52.209M1/MspI	4,2	0.33,0.28	16q21
D16S265	MFD23/PCR	12	0.78	16q21
D16S10	ACHF3.5.1/RsaI,MspI,TaqI	3	0.59	16q21
D16S186	16PHAC-101/PCR	10	0.57	16q22.1
D16S38	CRI-O2/BamHI	2	0.21	16q22.1
D16S46	CRI-O91/MspI,TaqI	3,4	0.38,0.36	16q22.1
D16S4	ACH207/TaqI,MspI	3	0.63	16q22.1
D16S47	CRI-O95/EcoRI	2	0.50	16q22.1
HP	hp2alpha/MspI	2	0.57	16q22.3-16q23.1
D16S153	CJ52.10T2/TaqI	2	0.50	16q22.3-16q23.1
CTRB	2pEKXp3B/PvuII	3	0.38	16q23.1-16q23.2
D16S50	CRI-O119/TaqI	2	0.55	16q23.3-16q24.1
D16S20	phi 8-9/BglII	3	0.47	16q24.1
D16S157	CJ52.96/TaqI	2	0.22	16q24.1-16q24.2
D16S43	CRI-O84/PstI,HindIII	5,2	0.79,0.33	16q24.1-16q24.2
D16S62	CRI-O149/TaqI	3	0.31	16q24.2-16q24.3
D16S41	CRI-O43/TaqI	4	0.29	16q24.2-16q24.3
D16S7	p79-2-23/TaqI,RsaI	6,7	0.64,0.82	16q24.3
APRT	Aprt/BglII,TaqI	2,2,2	0.50,0.27,0.28	16q24.3
D16S44	CRI-O89/BglII	6	0.62	16q24.3

extracted from the CEPH database (Version 5) (Donis-Keller *et al.*, 1987; Julier *et al.*, 1990; Keith *et al.*, 1990). Six microsatellite markers of the type $(AC)_n$, *D16S164*, *D16S186*, *D16S261*, *D16S265*, *D16S287*, and *D16S292*, not previously incorporated in a genetic map, were genotyped by colleagues at the WCH, SA (see Table 2.1). The determination of the genotypes for these microsatellite loci has been described previously (Weber *et al.*, 1990; Phillips *et al.*, 1991a; 1991b; Thompson *et al.*, 1992). The remaining 13 RFLP markers, from 7 loci, have been described in Table 3.1 and Table 4.1. Methods for determining the genotypes for each member of the CEPH reference families, for 13 RFLP markers genotyped at the WCH, are described in Chapter 2.

The loci incorporated from the report by Julier *et al.* (1990) have been genotyped on the extended CEPH panel of 60 families, as indicated on Table 2.1. The residual loci have been genotyped on the initial 40 CEPH pedigrees only.

All markers incorporated in the linkage analysis have been mapped to the extensive cytogenetic-based physical map of chromosome 16 (Callen *et al.*, 1992). The panel of somatic cell hybrids was constructed by fusing human cell lines containing chromosome 16 translocations and deletions with the mouse hybrid line and selecting for the human APRT gene located at 16q24 (Callen, 1986; Callen *et al.*, 1990). The human cell lines were derived from patients identified in routine clinical cytogenetic investigations (Callen, 1986; Callen *et al.*, 1990; Callen *et al.*, 1992). This panel comprises 60 hybrids, including the breaks defined by the four fragile sites and the centromere, and divides the chromosome into 50 intervals, with an average resolution of 2 Mb.

The microsatellite locus *D16S265* has been previously described and localised to chromosome 16 (Weber *et al.*, 1991), but this localisation needed to be refined. The original primers produced multiple products in the PCR reaction, and attempts to solve the problem by varying the PCR conditions failed. Thus, the candidate redesigned the oligodeoxyribonucleotide primers to give improved specificity for

this locus, and then determined its physical location to chromosome 16. The redesigned primers were synthesized automatically on an Applied Biosystems PCR-mate Synthesiser, purified and quantitated by K Holman, under the supervision of Dr R Richards, at the WCH.

Once appropriate primers had been manufactured, physical localisation was performed by PCR analysis using selected DNAs from the human chromosome 16 somatic cell hybrid panel as templates (Richards *et al.*, 1991c). The primers were challenged against a variety of the human/rodent cell lines, spanning the length of the chromosome 16, to determine the regional localisation of *D16S265*. These cell lines were CY18 (which contains the entire length of chromosome 16 as its only human component), CY185, CY8, CY125, and CY3. Having determined the region, the cell lines CY8, CY132, CY140, CY135, CY7, CY130, CY125, CY4, CY6, CY13A, and CY5 were selected to refine the localisation (Fig 6.1).

The components of the PCR reaction were as follows:

10 μl	2 x PCR reaction mix (consisting of 33 mM $(NH_4)_2SO_4$, 133 mM Tris-HCL (pH 8.8), 13 μM EDTA and freshly added 20 mM ß - mercaptoethanol).
6 μl	20 mM $MgCl_2$[1].
1 μl	Oligo primer pair (150 ng/μl for each oligo).
2 μl	template DNA (200 ng).
1 μl	*Taq* polymerase (0.5 units).

The PCR reaction was performed in a Perkin-Elmer Cetus DNA Thermal Cycler. Denaturation, annealing and extension was performed for each cycle in the PCR reaction as follows:

Cycle 1, repeated 10 times:

---

1. The $MgCl_2$ concentration was optimised at 6.0 mM for this marker

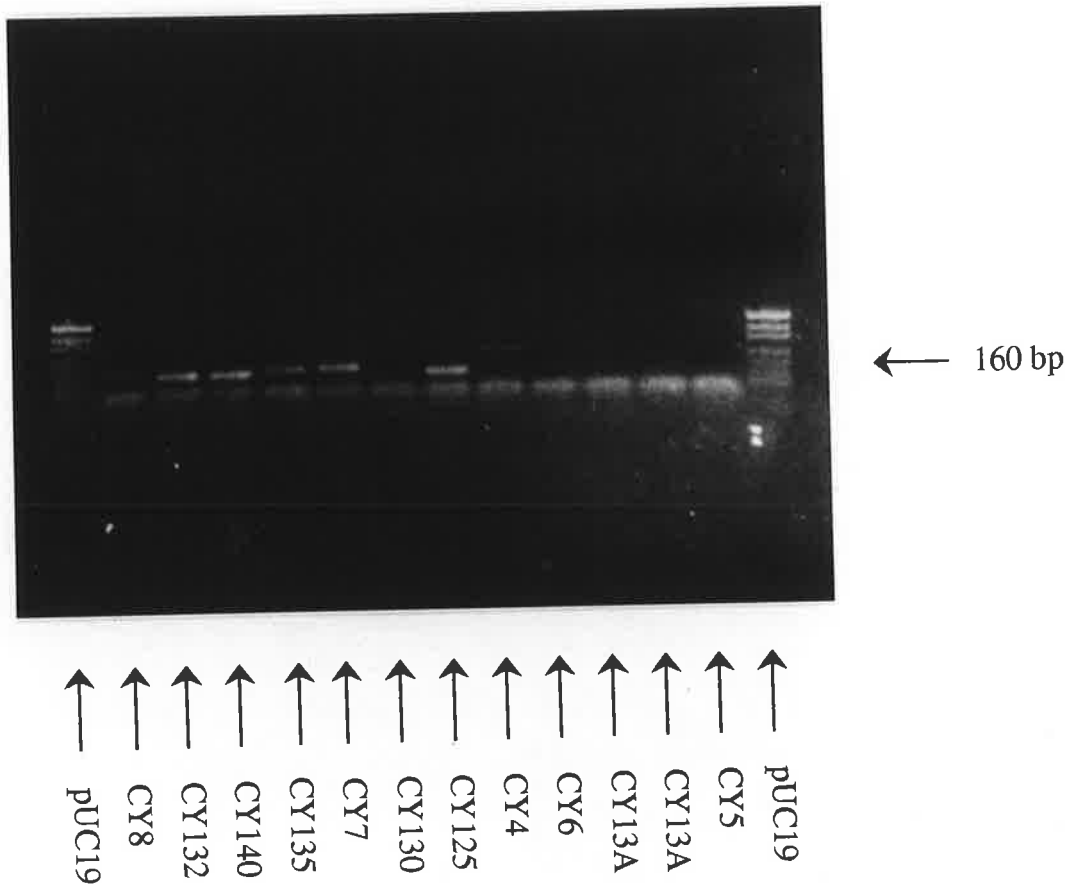

Figure 6.1: *Localisation of D16S265. The hybrid cell lines used for the localisation are shown. pUC19, digested with HpaII was used as the DNA size marker. The positions of the hybrid breakpoints on chromosome 16 are given on Fig. 6.2. The PCR product size for D16S265 is 160 base pairs.*

	94° C	1 minute
	64° C	1½ minutes
	72° C	1½ minutes
Cycle 2, repeated 25 times:		
	94° C	1 minute
	55° C	1½ minutes
	72° C	1½ minutes
Elongation step:		
	72° C	10 minutes

Gel electrophoresis on a 1.4 % agarose gel was used to resolve the PCR products which were then visualised by ethidium bromide staining under ultraviolet light. The marker was mapped to a specific breakpoint interval by virtue of the presence or absence of a PCR product (Fig. 6.1).

Recombination estimates and gene order were determined by likelihood analysis using the linkage package CRI-MAP (V2.4, Lander and Green, 1987), which calculates the maximum likelihood estimate for many loci simultaneously as described in Chapter 2.2.4. Using the **build** option, the two most polymorphic loci, *D16S85* and *D16S83*, were selected as the foundation on which the map was based, and a framework map was constructed. The resultant order of loci on the framework was verified against the order determined by the physical map.

After the framework map was generated, additions of subsequent loci to the map were accepted in the position with the highest likelihood, without regard to the odds of placement, providing their placement agreed with the physical map. The **all** option of the CRI-MAP package was used for this portion of the analysis, thus constructing a comprehensive map. Adjacent loci were inverted to determine the local support for the location of markers, using the **flips** option of CRI-MAP. The placement of loci not ordered on the basis of physical mapping or by odds of at least 1000:1 is considered to be tentative.

The elimination of data errors was attempted by checking for double recombinants within small distances using the **chrompic** option of CRI-MAP as described in

chapter 2.2.5. Markers and individuals responsible for double recombinant events are listed in Table 6.2.

Once all data had been verified, and corrected as necessary, the map was redetermined, several times using different pairs of starting loci to verify that the order of loci was consistent (See Section 2.2.4).

The differences between recombination in males and females, over the length of the map, was assessed under two models; a model of sex-average recombination, and a model of sex-specific recombination as described in Chapter 2.2.6. Each specific interval along the entire length of the map was also assessed for significant differences in the recombination fractions. To determine whether the sex-specific differences in recombination was significant in any interval on the map, two point linkage analyses were performed (sex-specific and sex-average), and tested for significance using the chi-squared test.

6.4. Results

A multipoint linkage map of human chromosome 16 has been constructed using the computer program CRI-MAP. Thirty one loci were incorporated in the framework map (with order support of at least 1000:1), and are indicated in bold type on Fig 6.2 and in Table 6.3. The length of the sex-averaged framework map was 162.7 cM, with an average resolution of 5 cM. There are three gaps larger than 10 cM on this map: between *D16S150* and *D16S65* (13.8 cM), *CTRB* and *D16S20* (15.6 cM), and *D16S157* and *D16S62* (17.4 cM). The sex-average comprehensive map, containing 62 RFLP loci and 6 microsatellite loci, is 164.5 cM, with an average resolution between these loci of 3.3 cM. The length of the male and female maps are 132.8 cM and 201.8 cM respectively (Fig. 6.2, Table 6.4). The largest gap of 13.4 cM is between *D16S157* and *D16S43*, and is the only gap over 10 cM on this map.

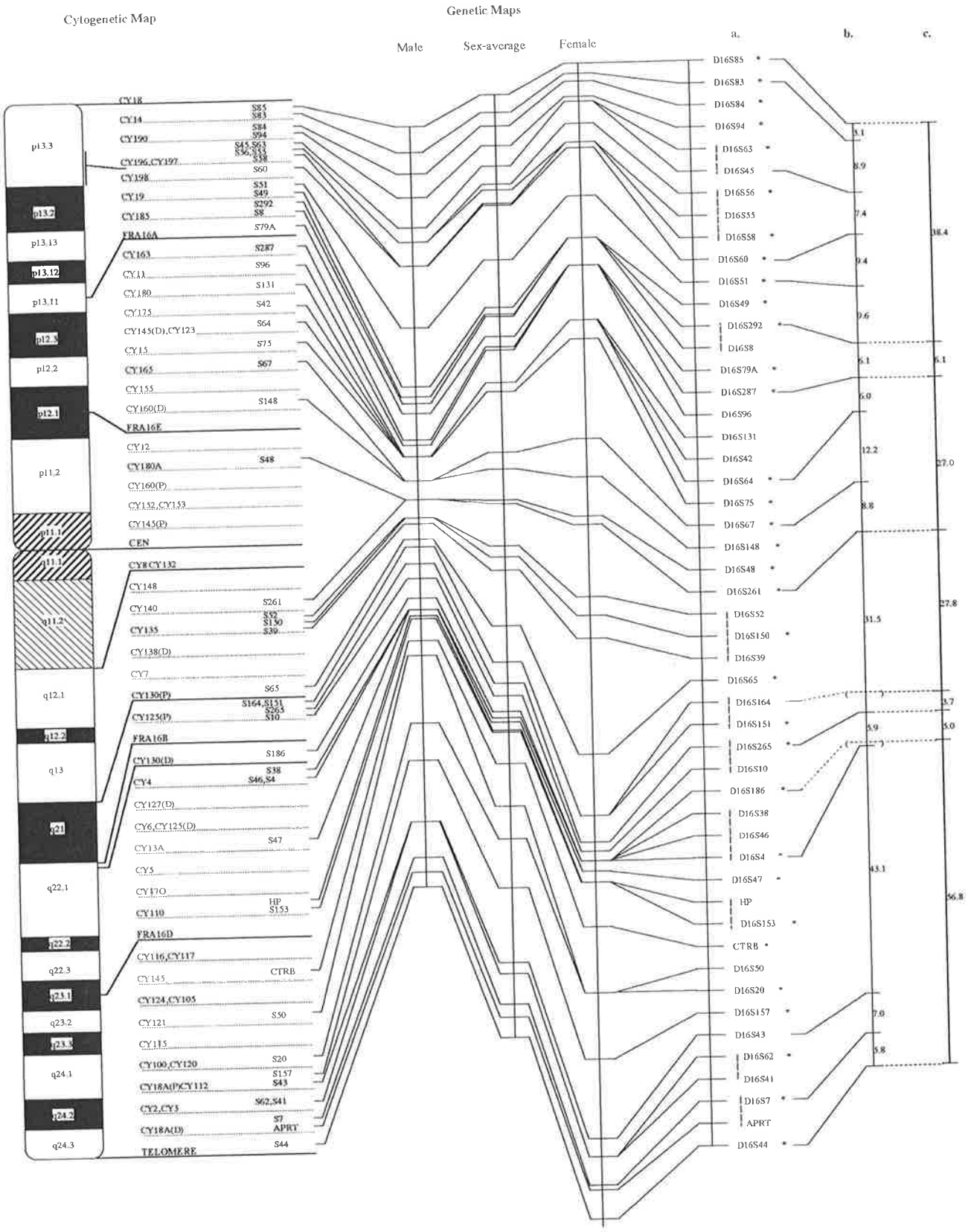

Figure 6.2: *Comprehensive Linkage Map for chromosome 16, is correlated with the cytogenetic-based physical map. a. Order of loci with framework loci indicated in bold type (*). Adjacent loci not uniquely ordered by chromosomal breakpoints and/or with likelihood support of less than 1000:1 are indented and indicated with a dotted line. b. Order and distance between those loci with minimum heterozygosity of 60%. c. Position of the 6 PCR based microsatellite markers.*

Table 6.2: *Genotypes recoded as unknown for this analysis.*

Locus	Individual
D16S83	102-16
	1333-03
D16S84	102-08
	102-11
	102-12
	884-13
D16S94	23-07
	1349-04
	1349-05
	1349-06
	13293-04
D16S8	1347-03
D16S79A	66-05
	66-07
D16S131	66-06
	66-09
	1423-07
D16S150	1345-09
D16S151	1350-07
HP	2-09

Table 6.3: *Sex-average distances (in cM) between the loci in the framework map and likelihood support for order.*

Locus Order	Distances	Odds
D16S85[1]		
	3.0	10^8
D16S83[1]		
	2.6	10^{10}
D16S84[1]		
	2.9	10^8
D16S94[1]		
	3.0	10^8
D16S63[1]		
	3.4	10^5
D16S56[1]		
	3.7	10^{11}
D16S58[1]		
	0.5	10^3
D16S60		
	9.3	10^{37}
D16S51[1]		
	8.5	10^8
D16S49[1]		
	0.9	10^3
D16S292		
	4.0	10^{12}
D16S79A		
	2.3	10^4
D16S287		
	7.2	10^{32}
D16S64		
	1.8	10^6
D16S75		
	8.7	10^{31}
D16S67		
	2.8	10^6
D16S148		
	4.9	10^5
D16S48		
	6.3	10^4
D16S150		
	13.8	10^{13}
D16S65		
	6.3	10^9
D16S151		
	3.7	10^{15}
D16S186		
	0.8	10^3
D16S4		
	0.9	10^4
D16S47		
	1.4	10^3
D16S153		
	7.1	10^{13}
CTRB		
	15.6	10^{22}
D16S20		
	9.1	10^6
D16S157		
	17.4	10^4
D16S62		
	5.4	10^3
D16S7		
	5.3	10^3
D16S44		
Total Length	162.7	

1 - Locus pairs ordered by linkage (with likelihood support of at least 1000:1) but not by chromosomal breakpoints. The remaining locus pairs are ordered by both linkage and by chromosomal breakpoints.

Table 6.4: *Distances (in cM) between the loci in the comprehensive map, including sex-specific and sex-average distances and likelihood support for order, based on the sex-average map.*

Locus Order	Male	Female	Sex Average	Odds
D16S85	4.6	1.7	3.1	10^9
D16S83	3.7	1.4	2.5	10^{11}
D16S84	4.1	2.7	3.3	10^{12}
D16S94	5.3	0.8	3.1	10^{11}
D16S63	0.0	0.0	0.0	1.0
D16S45	2.6	3.8	3.4	10^{15}
D16S56	0.0	3.3	1.4	1.0
D16S55	4.2	0.0	2.2	1.1
D16S58	0.0	0.9	0.5	10^2
D16S60	10.9	8.4	9.4	10^{39}
D16S51	10.5	7.3	8.7	10^8
D16S49	1.8	0.0	0.9	10^2
D16S292	1.0	0.0	0.4	10^2
D16S8	1.7	4.6	3.6	10^9
D16S79A	4.7	0.0	2.2	10^{10}
D16S278	1.0	0.0	0.3	10^2
D16S96	2.1	9.8	5.7	10^{15}
D16S131	0.0	0.0	0.0	1.0
D16S42	0.0	0.0	0.0	1.0
D16S64	0.0	3.4	1.8	10^8
D16S75	3.2	16.3	9.3	10^{34}
D16S67	0.0	6.7	2.8	10^6
D16S148	0.6	7.2	4.9	10^6
D16S48	0.0	9.8	5.2	10^6
D16S52	1.7	2.9	1.8	10^2

Table 6.4 (cont'd): *Comprehensive Map distances (cM)*

Locus Order	Male	Female	Sex Average	Odds
D16S150	3.2	4.4	2.8	10^1
D16S39	0.0	21.2	10.4	10^{13}
D16S65	4.1	11.1	6.5	10^4
D16S164	0.4	0.0	1.6	1.1
D16S151	2.7	1.7	1.6	10^2
D16S10	2.0	3.4	2.8	10^{16}
D16S186	0.6	0.0	0.3	1.0
D16S38	0.7	0.0	0.3	1.0
D16S46	0.0	0.0	0.0	1.0
D16S4	0.0	1.7	0.7	10^4
D16S47	1.3	1.9	1.7	10^4
HP	0.5	0.0	0.5	10^1
D16S153	4.1	8.2	5.8	10^{16}
CTRB	2.4	11.4	7.3	10^2
D16S50	12.0	0.0	5.7	10^3
D16S20	6.3	11.5	8.4	10^6
D16S157	11.0	13.9	13.4	10^9
D16S43	0.0	3.1	1.7	10^3
D16S62	0.0	0.0	0.0	1.0
D16S41	6.2	5.0	5.3	10^6
D16S7	2.8	1.1	2.2	10^2
APRT	2.4	4.8	3.6	10^5
D16S44				
Total Length	132.8	201.8	164.5	

The differences with respect to the recombination fractions between the sexes were investigated. The length of the male comprehensive map was 132.8 cM and the length of the female comprehensive map was 201.8 cM, representing a 1.5 fold excess of female recombination; thus a model of sex-specific differences in the recombination fraction was significantly different from the model of sex-average recombination (Table 6.5a). Each of the 50 intervals in the map were tested for significant differences in the frequency of recombination between the sexes, by comparing the pairwise lod scores of the two models assessed. The difference in the -2ln(likelihood) of the lod scores between these two models for each interval follows a chi-squared distribution, with 1 degree of freedom (Weiffenbach *et al.*, 1991). Using the conservative test of Fisher-Bonferroni (i.e. divide α by 50), the critical value for this chi-squared test is 10.8 (Hochberg, 1988). Using this conservative test, no significant sex differences in the recombination fraction were detected for any interval (Table 6.5b).

The physical localization of *D16S265* was determined by the candidate, using PCR analysis with DNAs from the somatic cell panel used as templates. PCR products were resolved on 1.4% sequence gels and visualised by EtBr staining under ultraviolet lighting. This allowed mapping of this STR marker to a specific breakpoint interval by virtue of the presence or absence of a PCR product. The locus was mapped to 16q21, between the cell lines CY122 and CY125, which is the proximal end of a deletion (Fig. 6.1 and Fig. 6.2).

In several intervals, the sex average distance is either greater than or less than both of the sex-specific estimates. The reason for this anomaly is that CRI-MAP selects the most likely phase for each marker, giving instances where the phase chosen for the sex-averaged map is different to the phase chosen for the sex-specific calculations. If a different phase is chosen, it is possible that the distance between the two markers on the sex-average map is less than the distance on both the male and female maps, or greater than the distances on both the maps. This

Table 6.5a: *Assessment of sex-specific differences in recombination over the entire map.*

Multilocus analysis. **All Intervals**	**Sex-Specific** *-2ln(Likelihood)*	**Sex-Equal** *-2ln(Likelihood)*	X^{2*}	**P**
D16S85 to *D16S44*	-1355.638	-1401.265	209.0	$P < 0.05$

*-Significant when $X^2 > 66.33$ ($P = 0.05$, 48 df).

Table 6.5b: *Assessment of sex-specific differences in recombination for specific intervals. Z is the lod score between the indicated loci. Z(θm,θf) is the lod score for the sex-specific case.*

Locus 1	Locus 2	θm=f	Z	θm	θf	Z(θm,θf)	θm/θf	X2#
D16S85	D16S83	0.03	44.18	0.06	0.00	46.09	-	8.8
D16S83	D16S84	0.02	57.57	0.03	0.01	57.99	3.00	1.9
D16S84	D16S94	0.03	17.18	0.06	0.00	17.60	-	1.9
D16S94	D16S45	0.03	16.96	0.05	0.00	17.80	-	3.9
D16S45	D16S63	0.00	7.22	0.00	0.00	7.22	-	0.0
D16S63	D16S56	0.00	6.92	0.00	0.00	6.92	-	0.0
D16S56	D16S55	0.00	6.02	0.00	0.00	6.02	-	0.0
D16S55	D16S58	0.00	3.01	0.00	0.00	3.01	-	0.0
D16S58	D16S60	0.01	39.80	0.00	0.02	40.23	-	2.0
D16S60	D16S51	0.11	17.20	0.12	0.10	17.24	1.20	0.2
D16S51	D16S49	0.03	11.81	0.05	0.00	11.90	-	0.4
D16S49	D16S292	0.01	30.88	0.01	0.00	31.11	-	1.1
D16S292	D16S8	0.00	64.60	0.01	0.00	64.95	-	1.6
D16S8	D16S79A	0.01	19.28	0.00	0.02	19.43	-	0.7
D16S79A	D16S287	0.02	37.17	0.05	0.00	37.82	-	3.0
D16S287	D16S96	0.00	66.09	0.01	0.00	66.50	-	1.9
D16S96	D16S42	0.05	8.83	0.00	0.08	9.28	-	2.1
D16S42	D16S131	0.00	10.84	0.00	0.00	10.84	-	0.0
D16S131	D16S64	0.00	23.78	0.00	0.00	23.78	-	0.0
D16S64	D16S75	0.03	22.65	0.00	0.04	23.28	-	2.9
D16S75	D16S67	0.12	3.24	0.08	0.17	3.35	0.47	0.5
D16S67	D16S148	0.03	10.33	0.00	0.06	10.79	-	2.1
D16S148	D16S48	0.06	5.15	0.00	0.10	5.56	-	1.9

#- Using the conservative test of Fisher-Bonferoni, the test is significant when $X^2 > 10.8$ ($P < 0.001$, 1 df).

Table 6.5b *(cont.)*: *Assessment of of Sex-Specific Differences in Recombination.*

Locus 1	Locus 2	θm=f	Z	θm	θf	Z(θm,θf)	θm/θf	$X^{2}_{\#}$
D16S48	D16S52	0.03	5.87	0.00	0.05	6.10	-	1.1
D16S52	D16S150	0.04	3.62	0.00	0.05	3.72	-	0.5
D16S150	D16S39	0.03	7.92	0.06	0.00	8.01	-	0.4
D16S39	D16S65	0.04	8.79	0.00	0.11	8.01	-	3.8
D16S65	D16S164	0.00	12.04	0.00	0.00	12.04	-	0.0
D16S164	D16S151	0.00	13.25	0.00	0.00	13.25	-	0.0
D16S151	D16S10	0.02	9.55	0.00	0.04	9.63	-	0.4
D16S10	D16S186	0.03	39.72	0.03	0.03	39.73	1.00	0.0
D16S186	D16S38	0.00	11.74	0.00	0.00	11.74	-	0.0
D16S38	D16S46	0.00	4.82	0.00	0.00	4.82	-	0.0
D16S46	D16S4	0.00	22.28	0.00	0.00	22.28	-	0.0
D16S4	D16S47	0.01	19.34	0.00	0.03	19.72	-	1.7
D16S47	D16S153	0.02	15.81	0.02	0.00	15.93	-	0.6
D16S153	HP	0.02	24.16	0.01	0.03	24.24	0.33	0.4
HP	CTRB	0.07	9.89	0.03	0.13	10.34	0.23	2.1
CTRB	D16S50	0.03	5.88	0.00	0.07	5.96	-	0.4
D16S50	D16S20	0.06	6.28	0.06	0.00	6.38	-	0.5
D16S20	D16S40	0.00	8.13	0.00	0.00	8.13	-	0.0
D16S40	D16S157	0.00	6.62	0.00	0.00	6.62	-	0.0
D16S157	D16S43	0.12	3.19	0.10	0.13	3.20	0.77	0.0
D16S43	D16S62	0.02	13.17	0.00	0.03	13.45	-	1.3
D16S62	D16S41	0.00	21.67	0.00	0.00	21.67	-	0.0
D16S41	D16S7	0.05	7.74	0.00	0.06	7.86	-	0.6
D16S7	APRT	0.01	40.76	0.00	0.00	41.05	-	1.3
APRT	D16S44	0.07	4.43	0.00	0.26	5.58	-	5.3

#- Using the conservative test of Fisher-Bonferoni, the test is significant when $X^2 > 10.8$ (P = 0.001, 1 df).

has been previously documented (Bowcock *et al.*, 1992; 1993b), and is evident in three intervals on the comprehensive map generated here (between *D16S150* and *D16S39*, between *D16S164* and *D16S151*, and between *D16S151* and *D16S10*) (Table 6.4).

At the level of the framework map, 22 loci were independently placed in the same order by linkage analysis and by the analysis of chromosomal breakpoints. The remaining 9 loci in the framework map were ordered by linkage alone, being unresolved on the cytogenetic map where more than one locus mapped to the same physical interval (Fig. 6.2). The placement of loci on the comprehensive map may be locally unreliable with adjacent loci where odds against their inversion are less than 1000:1. If these markers are separated by hybrid breakpoints or fragile sites, then the integration of the comprehensive genetic map and physical maps for this chromosome enables the order of some of these loci to be resolved at the comprehensive map level. The order of 5 of the 19 loci not included on the framework map, but incorporated on the comprehensive could be resolved in this manner (*D16S8*, *D16S96*, *D16S131*, *D16S42*, *D16S50*, and *D16S43*). This comprehensive map will thus provide a selection of markers for refining localisation of disease genes in affected families assigned to this chromosome.

Twenty genotype errors were detected at nine loci, detected as apparent double recombinants using the **chrompic** option of CRI-MAP. Buetow (1991) estimated that an error rate of 0.5% will increase the map length of 1 cM per interval, and an error rate of 1% will increase the map length by 2 cM per interval. The sex average length of the comprehensive map of 49 intervals was 164.5 cM. This implies an average increase of 0.398 cM per interval, assuming the length predicted by chiasma studies of 145 cM for this chromosome. Thus, the estimated error rate of the data included in this map is 0.2%.

6.5. Discussion

Fifty loci on chromosome 16 have now been mapped into a single genetic map and anchored to the cytogenetic map, as all of the loci have been incorporated on the cytogenetic-based physical map (Callen *et al.*, 1992). The framework genetic map incorporates 31 loci and has a sex-average length of 162.7 cM, and an average resolution of 5 cM (Fig. 6.2, Table 6.3). These loci, by definition, have been placed on the map unambiguously with odds of 1000:1 or greater. The remaining loci, being less informative and/or genetically indistinguishable (zero or few recombinants detected between them), were incorporated into the framework map to assemble the comprehensive map. The length of the sex-averaged comprehensive map was 164.5 cM, with an average resolution of 3.3 cM, and lengths of the male and female maps were 132.8 cM and 201.8 cM respectively.

As the genetic map was evolving, so too was the cytogenetic-based physical map (Chen *et al.*, 1991; Callen *et al.*, 1992), subdividing the chromosome into 50 intervals averaging 2 Mb in length. The fifty loci incorporated within the genetic map were also incorporated within the cytogenetic-based physical map, enabling the genetic map to be anchored to the physical map. This cytogenetic-based physical map is the most detailed map available for any of the autosomes (Callen *et al.*, 1992). The genetic map extends from *D16S85*, within 170-430 kb of the telomere on the short arm (Wilkie *et al.*, 1991) to D16S44, within 230 kb of the telomere on the long arm (P. Harris, personal communication).

The order of both the framework loci and the comprehensive loci, as established by linkage analysis, is compatible with the order derived by cytogenetic mapping. Thirty one loci on the framework map were uniquely ordered by linkage, and of these, the order of 22 loci was confirmed independently by physical mapping. Similarly, the order of 9 of the framework loci, not physically mapped to distinctive hybrid intervals was resolved by the genetic map (Table 6.3). The value of this high resolution cytogenetic map became evident during the

construction of the comprehensive map. The comprehensive loci were genetically unresolvable from those loci already placed on the framework map and could not be unambiguously placed by linkage analysis alone. By mapping markers against the panel of human/rodent somatic cell hybrids, 5 of the 19 comprehensive loci could be assigned to distinct physical intervals by the presence of chromosome breakpoints between loci (Fig. 6.2). This demonstrates that the combination of genetic linkage analysis with physical mapping can be extremely helpful in resolving locus order at the level of resolution of the comprehensive map.

The overall female map is 1.5 times longer than the male map, deviating slightly from the genome average observed by Morton (1991b) of 1.7 and the observed genome sex-difference of 1.6 determined by the NIH/CEPH collaborative mapping group (1992). This difference was found to be significant (Table 6.5a). In general, greater recombination in females was observed for most intervals, although no interval was statistically significantly different (Table 6.5b). Sex-specific differences in recombinants have been previously reported for this chromosome (Julier *et al.*, 1990; Keith *et al.*, 1990). In a report of the genetic linkage map of chromosome 21, Tanzi *et al.* (1988) provide evidence where the recombination in females is more than twice as much as that in males. Haldane's rule (1922) is that the heterogametic sex shows less recombination than the homogametic sex, as can be seen in many organisms, for example, in Drosophila (Morgan, 1914), in house mice (Dunn and Bennett, 1967) and in humans (Renwick and Schulze, 1965, Donis-Keller *et al.*, 1987, Warren *et al.*, 1989). Other published data supports the phenomenon of greater recombination in females than males (Stephens *et al.*, 1989), indicating that sex-based differences in recombination rates must be incorporated into an analysis associated with risk assessment for prenatal and presymptomatic diagnostic tests based on linkage.

Detailed linkage maps have previously been published for this chromosome (Donis-Keller *et al.*, 1987; Julier *et al.*, 1990; Keith *et al.*, 1990; Kozman *et al.*,

1991, 1993). For chromosome 16 this is the first reported genetic map accompanied with a detailed physical map, and incorporating $(AC)_n$ markers that are highly informative (*D16S164*, *D16S186*, *D16S261*, *D16S265*, *D16S287*, and *D16S292*). In comparison with the other published maps of chromosome 16, the order of markers is in complete agreement. The genetic distances on the published maps of chromosome 16 and the present study are well correlated, including those determined by Morton (1991b), based on the study of chiasma and interference (Table 6.6).

Table 6.6: *Comparison of map lengths (cM) with other published maps for the entire length of human chromosome 16.*

Group	Male	Female	Sex-Average	No. of loci
Current map	133	202	165	68
Julier *et al.*, 1990	187	226	-	24
Keith *et al.*, 1990	115	193	149	46
Morton, 1991b	120	193	157	-

As can be seen in Table 6.6, the lengths of the comprehensive map established in this study are in good agreement with those previously reported.

The distances on the comprehensive map are in quite good agreement with both the genetic estimates by multiple pairwise analysis (values given above) and by the chiasma counts (providing a sex-averaged distance of 145 cM) (Morton, 1991b). An approximate error rate of 0.5% has been estimated for the CEPH database (White *et al.*, 1990; Dracopoli *et al.*, 1991), which leads to a substantial inflation of genetic map length (Morton and Collins, 1990). The data set used in this analysis would appear to contain few errors, following the correction of errors using **chrompic**, since it agrees closely with the estimates of Morton (1991b). An

error rate of 0.2% has been estimated for these data, based on theoretical considerations proposed by Buetow (1991).

Interest in a correlated genetic and physical map of chromosome 16 has been stimulated as this map becomes saturated with markers. The information obtained from physical mapping has an important role in genetic map validation, as is demonstrated in this study. This is particularly evident at the level of the comprehensive map whose loci cannot be ordered uniquely by linkage analysis alone. Loci placed on the framework map, and additional loci within the comprehensive map ordered on the integrated map by physical criteria alone, are useful for localisation or exclusion of disease genes to human chromosome 16 by linkage. From this analysis it is apparent that genetic linkage analysis is not always the most efficient method of resolving locus order as resolution approaches 1 cM. Therefore a strategy combining classical linkage analysis with physical mapping methods such as PFGE, radiation hybrid mapping, or the use of a hybrid cell panel, is essential for the determination of the order of loci in such situations.

A continuous linkage map, well supported by a cytogenetic map, of 68 loci extending almost the full length of chromosome 16 has been constructed in this chapter, providing a foundation for the generation of a fine structure map for this autosome.

This is the first step in anchoring and confirming the genetic and physical maps together; the ultimate goal being to express the sex-specific recombination fractions (in cM), in terms of the physical distances (in Mb) (Chakravarti, 1991). The present map will facilitate the initial detection of linkage to disease genes and subsequent refinements of their regional localization to enable the eventual characterization of DNA defects by positional cloning (Wicking and Williamson, 1991).

CHAPTER 7

THE CONSTRUCTION OF A HIGH RESOLUTION PCR BASED GENETIC MAP OF HUMAN CHROMOSOME 16

Table of Contents

Associated Papers (See Appendix A):

A PCR based genetic linkage map of human chromosome 16. Shen Y., Kozman, H.M., Thompson, A., Phillips, H.A., Holman, K., Nancarrow, J., Lane, S., Chen, L.-Z., Apostolou, S., Doggett, N.A., Callen, D.F., Mulley, J.C., Sutherland, G.R., and Richards, R.I. (1994). *Genomics* **22**:68-76.

Human chromosome 16, genetic and physical mapping. In: Encyclopaedia of Molecular Biology. Kozman, H.M. and Mulley, J.C. (In Press).

A microsatellite marker within the duplicated *D16S79* locus has a null allele: Significance of linkage mapping. Phillips, H.A., Thompson, A.D., Kozman, H.M., Sutherland, G.R., and Mulley, J.C. (1993). *Cytogenet. Cell Genet.* **64**:131-132.

7.1. Summary

A total of 82 loci, represented by 89 polymorphisms with an average heterozygosity of 70%, have been incorporated into a high resolution genetic linkage map of chromosome 16. The framework map, constructed by multipoint linkage analysis of these markers, consists of 47 PCR formatted loci (average heterozygosity 73%) placed with odds of 1000:1 or greater. The map also included the hypervariable telomeric marker *D16S85*, and two other RFLPs, intralocus markers with the PCR formatted locus, *D16S67*. The length of the sex-average framework map was 152 cM, with a mean distance between markers of 3.2 cM, and a median distance between markers of 2.6 cM. The largest gap between markers is 15.9 cM. The male and female map lengths are 131 cM and 181 cM respectively. All the loci have been placed on the high resolution cytogenetic-based physical map, anchoring the genetic map to the physical map. High resolution genetic maps such as the one constructed here, integrated with the high resolution physical map, are effective tools for precise localisation of disease genes which map to this chromosome.

7.2. Introduction

Linkage analysis is a powerful strategy responsible for the genetic localisation of a large number of human Mendelian disorders. Until now, linkage maps constructed for chromosome 16 have been based mainly on RFLP markers. With the development of the PCR technique and recent availability of highly informative microsatellite markers distributed throughout the human genome, high resolution genetic reference maps have been constructed from the CEPH database. These are the basis for linkage studies attempting to localise genes responsible for human inherited disorders. Whilst the maps based on RFLPs have made an invaluable contribution to the advancement of human genetic research on chromosome 16, RFLPs are less efficacious for the localisation of disease genes and the refinement of map distances for positional cloning, than the highly polymorphic, PCR based, microsatellite markers. Thus, in tandem with the development of a high resolution physical map of cloned contigs, development of a high resolution microsatellite map for human chromosome 16 will be invaluable for the localisation and isolation of defective genes and other interesting anomalies on this chromosome, such as cancer breakpoints, translocation breakpoints, and fragile sites (see Chapters 3 and 4). Accurate high resolution maps are essential for detection of abnormalities related to genomic biological phenomena such as gene conversion or microinversion.

A total of 80 simple tandem repeat (STR) loci and 2 RFLP loci were genotyped on the panel of reference families maintained by CEPH, in Paris. Described here is the construction of a 3.2 cM resolution framework genetic linkage map comprising 47 of the PCR formatted markers (defined by 49 markers) and 1 VNTR marker, *D16S85*. A comprehensive map incorporating all the 82 polymorphic loci (defined by 89 markers) was also constructed. This is the first high resolution map, composed chiefly of highly polymorphic microsatellite markers, to be constructed

Table 7.1: *Loci included on the high resolution STR linkage map.*

Location	Gene	Probe/Enzyme	Het.	No Inf Meioses
16p12.1-p11.2	*SPN*	SPN/PCR	0.96	463
16q24.3	*D16S44*	CRI-O89/BglII	0.50	206
16p13.3-p13.11	*D16S49*	CRI-O114AC/PCR	0.66	113
		CRI-O114/EcoRI	0.39	169
16p13.11-p12.1	*D16S67*	CRI-O391AC/PCR	0.77	395
		CRI-O391/HincII	0.46	128
		CRI-O391/TaqI	0.50	199
16p13.3-p13.11	*D16S79A*	16AC66F3/PCR	0.81	398
16p13.3	*D16S85*	3'HVR/MspI	0.93	573
		3'HVR/RsaI/MspI	0.86	425
		5'HVR/RsaI	0.79	257
16p13.3	*D16S94*	16ACVK5/PCR	0.51	383
		VK5B/MspI	0.49	324
16q21	*D16S164*	16PHAC-15/PCR	0.38	226
16q21	*D16S186*	16PHAC-101/PCR	0.57	413
16q12.1-q13	*D16S261*	MFD24/PCR	0.71	449
16q21	*D16S265*	MFD23/PCR	0.77	505
16q23.2-q24.2	*D16S266*	MFD62/PCR	0.59	114
16p13.3	*D16S283*	SM7/PCR	0.64	407
16q12.1-q13	*D16S285*	AluGT16/PCR	0.83	496
16p13.11-p12.1	*D16S287*	16XE81/PCR	0.78	502
16p12.1-p11.2	*D16S288*	16AC7.1/PCR	0.73	449
16q23.1-q24.2	*D16S289*	16AC7.46/PCR	0.77	396
16p13.3	*D16S291*	16AC2.5/PCR	0.79	519
16p13.2-p13.11	*D16S292*	16AC2.3/PCR	0.74	450
16p13.11-p12.1	*D16S294*	16AC1/PCR	0.49	235
16p13.11-p12.1	*D16S295*	16AC62F3/PCR	0.66	430
16p13.11-p12.1	*D16S296*	16AC62B4/PCR	0.75	502
16p13.11-p12.1	*D16S297*	16AC15H1H/PCR	0.73	412
		16AC15H1S/PCR	0.69	124
16p12.1-p11.2	*D16S298*	16AC3.12/PCR	0.80	455
16p12.1-p11.2	*D16S299*	16AC6.17/PCR	0.72	425
16p11.2-p11.1	*D16S300*	16AC1.1/PCR	0.61	436
16q22.1	*D16S301*	16AC6.21/PCR	0.64	409
16q24.3	*D16S303*	16AC6.26/PCR	0.43	230
16q12.1-q13	*D16S304*	16AC1.14/PCR	0.60	410
16q24.3	*D16S305*	16AC1.15/PCR	0.82	494
16q12.1-q13	*D16S308#*	16AC1.18/PCR	0.77	113
16q21	*D16S310*	MIT-MH20/PCR	0.67	472
16p13.11-p12.1	*D16S313*	MIT-MS79/PCR	0.57	498
16q22.1	*D16S318#*	16AC8.20/PCR	0.54	143
16p13.11-p12.1	*D16S319*	16AC7.14/PCR	0.52	315
16q12.1-q13	*D16S320*	16AC8.52/PCR	0.86	559
16q23.2-q24.3	*D16S332*	16AC305D6/PCR	0.53	324
16q22.1	*D16S347#*	16AC12F8/PCR	0.76	147

- Markers typed on only 8 of the CEPH families.

Table 7.1*(cont'd)*: *Loci included on the high resloution STR linkage map.*

Location	Gene	Probe/Enzyme	Het.	No Inf Meioses
16q12.1-q13	*D16S359*	16AC26E3B/PCR	0.42	321
16q23.2-q24.3	*D16S363#*	16AC51G1/PCR	0.78	164
16p12.1-p11.2	*D16S383*	16AC80B3/PCR	0.45	235
16q21	*D16S389#*	16AC10B3/PCR	0.77	170
16q12.1-q13	*D16S390*	16AC10F5/PCR	0.80	511
16q23.2-q24.3	*D16S392#*	16AC305E9/PCR	0.78	161
16q23.2-q24.3	*D16S393#*	16AC323H4/PCR	0.87	159
16q23.2-q24.2	*D16S395*	16AC33G11/PCR	0.69	467
16q22.1	*D16S397#*	MFD98/PCR	0.70	168
16q22.1	*D16S398*	MFD168/PCR	0.90	559
16q*	*D16S400*	AFM024xg1/PCR	0.60	100
16p13.11-p12.1	*D16S401#*	AFM025tg9/PCR	0.74	80
16q23.1-q24.3	*D16S402#*	AFM031xa5/PCR	0.87	140
16p13.11-p12.1	*D16S403#*	AFM049xd2/PCR	0.86	171
16p13.2-13.11	*D16S404#*	AFM056yf6/PCR	0.82	158
16p13.2-13.11	*D16S405#*	AFM070ya1/PCR	0.78	143
16p13.3-13.11	*D16S406#*	AFM079yh3/PCR	0.82	152
16p13.2-13.11	*D16S407#*	AFM113xe3/PCR	0.86	160
16q12.1-q13	*D16S408#*	AFM137xf8/PCR	0.69	120
16p13.11-p12.1	*D16S410#*	AFM165yb6/PCR	0.57	125
16q12.1-q13	*D16S411#*	AFM186xa3/PCR	0.79	130
16p13.11-p12.1	*D16S412#*	AFM191wb10/PCR	0.76	101
16q24.3	*D16S413#*	AFM196xg1/PCR	0.85	154
16p13.2-13.11	*D16S414#*	AFM205za11/PCR	0.61	137
16q12.1-q13	*D16S415#*	AFM205ze5/PCR	0.74	121
16q12.1-q13	*D16S416#*	AFM210yg3/PCR	0.43	53
16p13.11-p12.1	*D16S417#*	AFM220xb10/PCR	0.73	155
16p13.3-13.11	*D16S418#*	AFM225xd2/PCR	0.83	157
16q12.1-q13	*D16S419#*	AFM225zf2/PCR	0.77	149
16p13.11-p12.1	*D16S420#*	AFM238xb2/PCR	0.82	149
16q22.1	*D16S421#*	AFM240yh6/PCR	0.57	140
16q23.1-q24.2	*D16S422#*	AFM249xc5/PCR	0.80	166
16p13.3	*D16S423#*	AFM249yc5/PCR	0.75	136
16q23.2-q24.3	*D16S449#*	16AC51A4/PCR	0.85	134
16q22.2-q23.1	*D16S450*	16AC80H3/PCR	0.52	232
16q21	*D16S451#*	16AC69F12/PCR	0.84	145
16p13.11-p12.1	*D16S452#*	16AC33A4/PCR	0.68	143
16p13.3	*D16S453*	16AC301G12/PCR	0.45	335
16p13.2-p13.11	*D16S454#*	16AC45G5/PCR	0.75	134
16q22.2-q23.1	*D16S522*	16AC8.21/PCR	0.69	449
16p13.3	*D16S523#*	16AC13H1/PCR	0.68	149
16p13.11-p12.1	*D16S524#*	16AC40A7/PCR	0.76	157
16p13.3	*D16S525#*	16AC308G7/PCR	0.91	180
16q12.1-q13	*D16S531#*	16AC8.15/PCR	0.86	147

* - Physical location not refined.
# - Markers typed on only 8 of the CEPH families.

for chromosome 16. It represents one further step toward the eventual development of the CEPH Consortium map for this chromosome.

7.3. Materials and Methods

The polymorphic loci (82 STRs, 5 RFLPs, and 3 VNTRs) used to construct this PCR based linkage map are described in Table 7.1. These polymorphic loci have been isolated and characterised, physically mapped and genotyped by colleagues from the WCH, and from other laboratories as specified in Table 2.1.

Two approaches were used to isolate the 48 simple tandem repeats from our laboratory. 17 microsatellite markers were isolated using the random isolation approach employed by Thompson *et al.* (1992), from the chromosome 16 cosmid library (Stallings *et al.*, 1990), constructed from the human mouse somatic cell hybrid (CY18) containing chromosome 16 as the only human chromosome (Callen, 1986). Once these 17 randomly isolated markers had been localised to specific regions of the chromosome by physical mapping, an additional 31 microsatellite loci were isolated from specific regions of the chromosome recognised as deficient in STR markers, or as regions of interest, such as those carrying a disease gene or fragile site (Shen *et al.*, 1994). The remaining loci used in the analysis were isolated as referenced in Table 2.1.

All the PCR markers incorporated into the linkage analysis were localised to the cytogenetic-based physical map using the hybrid panel for chromosome 16 (see Fig. 7.1). The markers were mapped to a specific breakpoint interval by virtue of the presence or absence of the PCR product using DNAs from human chromosome 16 hybrid panel as templates in the PCR as described previously (Richards *et al.*, 1991c; Callen *et al.*, 1992; Thompson *et al.*, 1992; Shen *et al.*, 1994).

Fifty five of the STR loci incorporated into the linkage map were genotyped at the WCH, with the genotyping performed by colleagues as specified in Table 2.1. Of these, 39 markers were genotyped on the primary set of 40 CEPH families and 16

markers were genotyped on a subset of 8 of the largest of the CEPH pedigrees (102, 884, 1331, 1332, 1347, 1362, 1413, 1416), as used by Weissenbach *et al.* (1992) (Table 7.1). The remaining PCR based loci were extracted from the CEPH database, version 6. The RFLPs and VNTR loci have been previously described (Chapter 6), and were incorporated in this section as either telomeric markers (*D16S85*, *D16S44*), or as intralocus markers with corresponding STR polymorphisms (*D16S49*, *D16S67*, and *D16S94*). All genotype inconsistencies within a family were checked against the original autoradiograph where possible, and retyped as necessary, prior to the analysis.

Where there exists more than one marker for a locus, and no intralocus recombinants were observed, the distance between them was forced to zero in CRI-MAP. The markers that were considered as a haplotype in this analysis were *D16S49*, *D16S67*, *D16S85*, *D16S94*, and *D16S297*.

The linkage analysis for this chapter was performed entirely by the candidate, as described in chapter 2.2. Locus order and map distances were determined by likelihood analysis using CRI-MAP (Version 2-4) (Lander and Green, 1987; Green *et al.*, 1990) and the Kosambi mapping function.

The framework genetic map was constructed using the **build** option of CRI-MAP. Two loci with the greatest number of informative meioses, separated by a recombination fraction greater than 20% (Buetow *et al.*, 1993) were selected as the foundation for map construction; the two loci selected were *D16S85* (with average heterozygosity 93%) and *D16S320* (with average heterozygosity 86%). Following the determination of a framework map of loci with LOD 3 placement, the local support for the order was determined using **flips** option. Major errors in locus order on the framework map were also tested by assessing the global support for the placement of each locus by removing each locus in turn and testing it against all the remaining intervals (using the **all** option of CRI-MAP). Maximum likelihood support was obtained for each interval, and the odds against placement

in all other intervals on the map compared to the most likely interval were calculated to evaluate if its placement still met the LOD 3 criteria. Loci which did not meet the criteria were removed from the map. This analysis is a powerful means of identifying loci assigned to the wrong region, but tends to give lower likelihood support for local order as it utilizes all loci in the map. The framework map was also rebuilt several times to verify the order, using different pairs of starting loci. The final map was compared to cytogenetic-based physical map to verify the order of loci.

The 33 AC repeat loci and one RFLP marker (*D16S44*) which were not placed on the framework genetic map were regionally localised to intervals between framework markers. The most likely 1000:1 interval for each locus was determined, using the **all** option of CRI-MAP. These loci were also incorporated into their most likely positions on the framework map, thus constructing a comprehensive genetic map. Support for order of loci in the comprehensive map was determined using the **flips** option of CRI-MAP.

After the construction of the framework and comprehensive maps, a series of tests (described below) were performed to identify and correct any erroneous genotypes, and to confirm the preliminary map order. Once these tests had been performed, and all genotypes were satisfactorily certified, the map was reconstructed as described above.

The **chrompic** option of CRI-MAP was used to detect obvious genotype errors, apparent as an excess of double recombinants. When double recombinants were detected, the original data were reexamined and then these DNA samples were regenotyped if necessary. After regenotyping the original DNA samples, which in some cases were typed twice or more, 14 double recombination events within a 15 cM region still remained (Table 7.2). These double recombinants were included in the analysis, as they could not be eliminated on the basis that they were erroneous genotypes.

Table 7.2: *Double recombinant events (occuring within 15 cM) in the genotype data of the STR loci not excluded from the linkage analysis.*

Locus	Probe	Individual	Genetic Distance
D16S261	MFD24	37-04	2.5
D16S261	MFD24	37-04	2.5
D16S265	MFD23	1345-05	4.8
D16S283	SM7	17-09	7.0
		45-09	9.5
D16S285	AluGT16	21-04	2.7
D16S291	16AC2.4	1375-08	5.4
D16S294	16AC1	1344-05,07	3.5
D16S299	16AC6.17	1345-07	0.6
D16S300	16AC1.1	1350-07	3.6
D16S310	MIT-MH20	104-08	11.1
D16S359	16AC26E3B	1408-04	2.3
D16S390	16AC10F5	1344-07	15.0
D16S450	16AC80H3	1334-09	9.5
D16S522	16AC8.21	102-12	6.1
		1349-03	10.8

Map inflation associated with the presence of a locus has been shown to be related to the typing error frequency at that locus, and is estimated at half the average reduction in map length after the removal of each locus in turn from the map (Buetow, 1991). Thus as a further test to assess locus specific error for each of the loci used in this analysis, each locus was removed in turn from the comprehensive map and the total map length recalculated after each removal. This identifies those loci responsible for inflation of the map length (Engelstein *et al.*, 1993).

Sex-specific differences between males and females with respect to the recombination frequency were investigated with respect to the entire framework map as described in Chapter 2.2.6. Sex differences with respect to recombination in each of the 48 intervals on the framework map were evaluated separately, deriving the test statistic by analysing each interval under the assumption of equal or variable recombination in males and females and including three adjacent intervals on each side, in which the recombination values were held fixed (O'Connell *et al.*, 1993). This statistic also follows a chi-square distribution, with 1 degree of freedom. Multiple testing was corrected for by choosing the conservative test of Fisher-Bonferroni (ie, divide P by n, where n is the number of intervals) (Hochberg, 1988).

7.4. Results

All loci incorporated in the genetic map were mapped to the panel of somatic cell hybrids to determine their physical localisations (Table 7.1 and Table 2.1). In total, 80 STR markers mapped to 35 physical intervals (Fig. 7.1). *D16S400* was not physically mapped to the hybrid panel but was included in the genetic map.

Using the raw, uncorrected genotype data from the 82 polymorphic markers, a preliminary sex-average comprehensive map was constructed, and was used to detect potential genotype errors. The length of this preliminary comprehensive

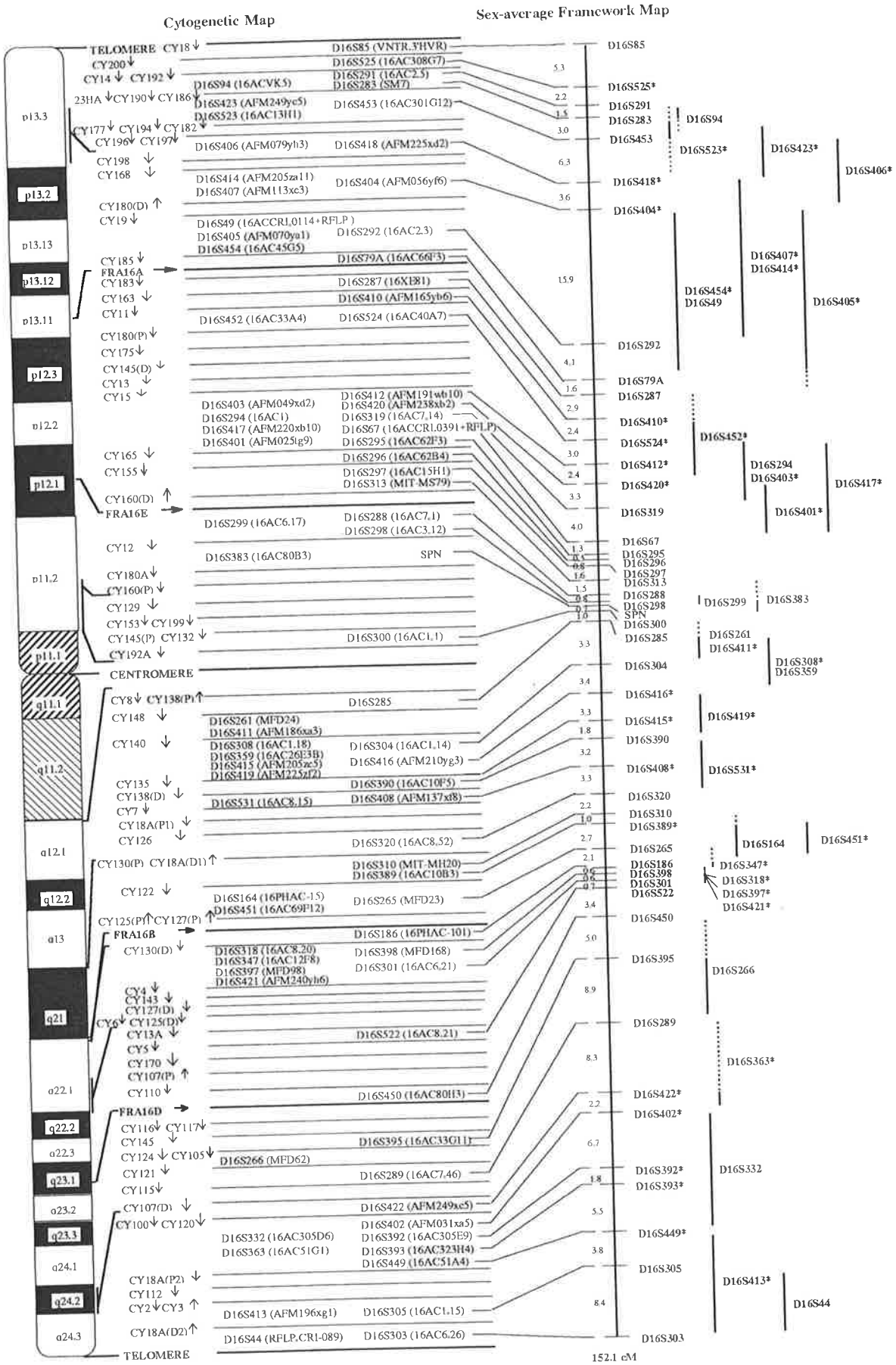

Figure 7.1: Cytogenetic -based physical map and Genetic linkage map based on STR polymorphisms. Comprehensive loci are shown in their most likely locations. The dotted lines indicate reduced localisations for the comprehensive loci, on the basis of physical mapping.

map was 172.3 cM, with an average distance between markers of 2.1 cM (Kozman and Mulley, 1994). Using the **chrompic** option of CRI-MAP, the evidence of existing double or triple recombination events was investigated. Potential data errors were checked and regenotyped as necessary. Individuals whose genotypes could not be altered by regenotyping were left in the analysis, as they could not be eliminated as incorrect genotypes. These are listed in Table 7.2. To supplement the **chrompic** analysis for the detection of potential typing errors, and to determine the effect these errors have on the map length and the order of markers, each non-terminal locus was removed from its position in turn, and the likelihood and map length determined. The mean map length after the removal of each locus in turn was 171.9 cM. The largest reduction in length, of 3.8%, occurred with the removal of *D16S283*, resulting in a length of 165.7 cM (Fig. 7.2 a). Investigation of the output from **chrompic** revealed that there were 5 flanking marker double recombinants due to the marker *D16S283*, 3 of which were corrected after reanalysis of the genotypes. These individuals were 1346-09, 1346-10, and 1349-06. The remaining two double recombinants (individuals 17-04 and 45-09) were left in the analysis (Table 7.2). After the removal of each non-terminal STR locus, the average map length was reduced by 0.23%. If the overall error rate is estimated at one half the average reduction in map length after the removal of each locus in turn, then the overall error rate in the PCR marker typing is estimated at 0.12%, for this preliminary PCR map. If the overall *inflation of map length* due to errors is given by:

$$x_{inf} = 2(n-1)\epsilon,$$

then the total inflation of the preliminary PCR based map would be 19.2 cM. Hence, the length of the error compensated PCR map would be 153.1 cM, only 5.2% longer than that predicted by chiasma studies (145 cM; Morton, 1991b).

From the construction of this preliminary map, the location of three genetic markers was found to be in disagreement with the physical map. These three

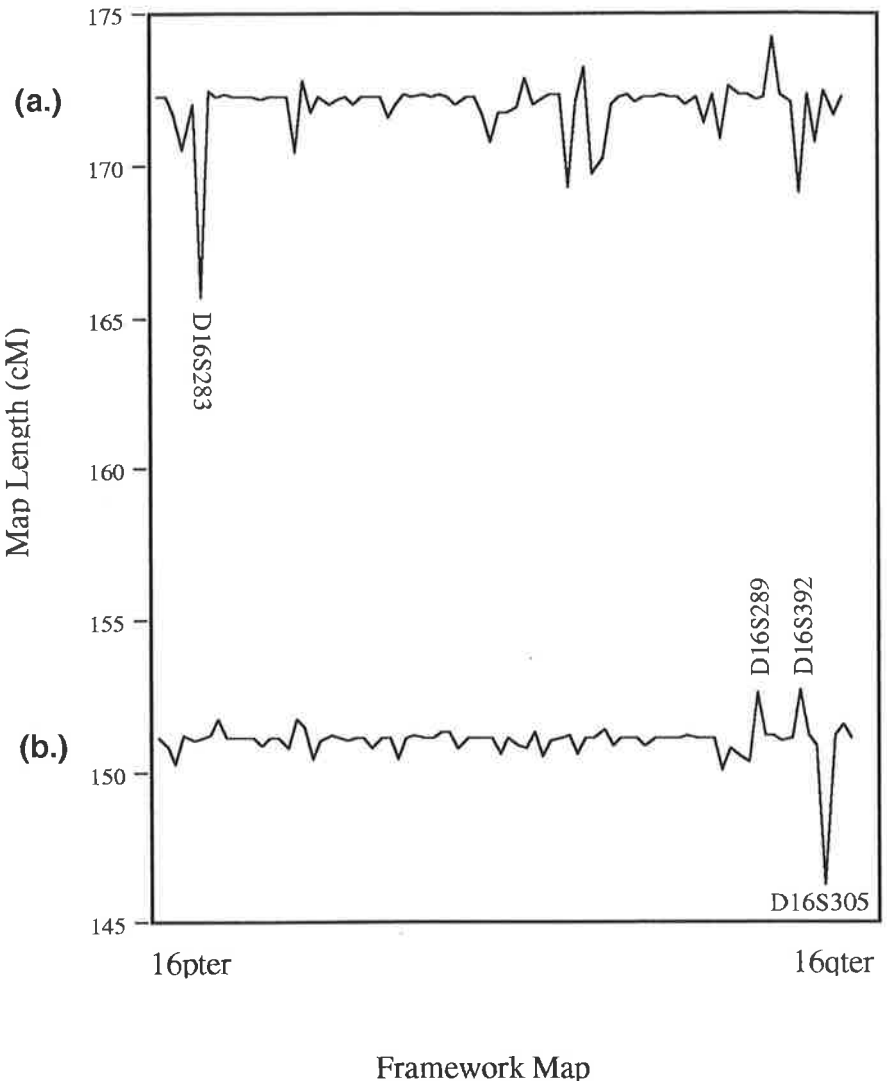

Figure 7.2: *Comparrison of the removal of each non terminal locus between the preliminary framework map (a) and the final error compensated map (b).*

markers were *D16S347,* which physically mapped between CY126 and CY130(P), but genetically mapped between *D16S164* and *D16S168.* The locus *D16S318* mapped between CY4 and CY143 on the physical map, and between *D16S168* and *D16S398* on the genetic map. *D16S363* mapped between *D16S289* and *D16S422* on the genetic map and between CY18A and CY112 on the physical map. To rectify these discrepancies, the location of the three markers was redetermined on both the physical and genetic maps. To test whether the genetic map location was erroneous, the marker was forced into the position on the genetic map as suggested by its physical location to identify the apparent double recombinants. If such observations were due to genotyping errors then this would identify the genotyping errors corrupting the linkage analysis. The physical location was checked against the hybrid panel. The result of these checks determined that the physical map location was incorrect for the marker *D16S347* (Fig. 7.1). Two double recombinants were also identified for this marker, which were regenotyped and corrected, locating this marker between CY130(D) and CY4 on the physical map, and between *D16S186* and *D16S421* on the genetic map, which is in agreement. The individuals found to be incorrect were 102-12, and the paternal grandparents in the family 1416, causing incorrect phase for all the sibs in this family. The physical location of both the STR loci *D16S318* and *D16S363* were changed to agree with the correct genetic localisation (Fig. 7.1).

Once all detected genotype errors had been eliminated, a framework genetic map containing 48 markers was constructed (Fig. 7.1, Table 7.3a). One VNTR marker (*D16S85*) was also incorporated into the framework map, used to anchor the map to the telomeric end of the short arm of chromosome 16. The mean genetic distance between markers on this map was 3.2 cM and the median distance was 2.6 cM, with the largest gap on the chromosome being 15.9 cM. The length of the sex-averaged map was 152 cM and the lengths of the male and female maps were 131 cM and 181 cM respectively. The order of the framework loci as

Table 7.3a: *Map Distances (in Kosambi cM) for the 48 loci in the High Resolution STR Framework Map. Included is the local support for the order of loci.*

Locus	Sex-ave	Female	Male	Odds[a]	Chi-square[b]
*D16S85**	5.3	2.7	8.5	10^9	2.8
D16S525	2.2	0.0	4.7	10^5	8.3
D16S291	1.5	0.8	1.9	10^3	8.3
D16S283	3.0	1.8	4.4	10^{13}	4.6
D16S453	6.3	2.8	9.4	10^{12}	3.7
D16S418	3.6	6.6	1.6	10^6	0.5
D16S404	15.9	12.5	18.2	10^{31}	1.4
D16S292	4.1	4.2	3.8	10^{17}	1.4
D16S79(A)	1.6	0.7	2.2	10^{11}	1.8
D16S287	2.9	5.4	1.2	10^{10}	6.0
D16S410	2.4	4.0	1.2	10^4	7.4
D16S524	3.0	4.5	1.2	10^5	6.9
D16S412	2.4	5.8	0.0	10^4	6.9
D16S420	3.3	6.2	1.2	10^6	0.0
D16S319	4.0	5.8	2.0	10^{25}	5.1
D16S67	1.3	2.1	0.0	10^7	4.1
D16S295	0.5	0.0	1.8	10^6	2.3
D16S296	0.8	1.4	0.0	10^{14}	6.0
D16S297	1.6	3.0	0.0	10^{24}	6.9
D16S313	1.5	2.8	0.3	10^{23}	7.4
D16S288	0.8	1.6	0.0	10^{14}	24.8
D16S298	0.2	0.5	0.0	10^3	2.3
SPN	0.7	0.9	0.3	10^7	1.4
D16S300	1.0	2.2	0.0	10^8	8.7
D16S285	3.3	6.7	0.3	10^{26}	4.1
D16S304	3.4	7.1	0.0	10^4	26.7
D16S416	3.3	4.1	1.5	10^5	27.6
D16S415	1.8	5.4	0.0	10^4	24.4
D16S390	3.2	6.1	0.6	10^8	24.4

* - foundation loci for the **build** analysis.

Table 7.3a *(cont'd)*: *Map Distances (in Kosambi cM) for the 48 loci in the High Resolution STR Framework Map. Included is the local support for the order of loci.*

Locus	Sex-ave	Female	Male	Odds[a]	Chi-square[b]
D16S408	3.3	6.4	0.6	10^{11}	23.9
*D16S320**	2.2	4.9	0.0	10^{17}	17.5
D16S310	1.0	1.9	0.0	10^{3}	3.7
D16S389	2.7	2.8	2.6	10^{6}	2.8
D16S265	2.1	2.1	2.3	10^{17}	0.0
D16S186	0.6	1.0	0.0	10^{6}	2.3
D16S398	0.6	1.0	0.0	10^{8}	0.0
D16S301	0.7	0.4	1.1	10^{5}	0.5
D16S522	3.4	1.8	4.5	10^{16}	1.4
D16S450	5.0	6.7	4.4	10^{13}	1.4
D16S395	8.9	13.1	3.4	1033	7.4
D16S289	8.3	6.5	8.8	1016	0.9
D16S422	2.2	2.2	2.0	103	0.0
D16S402	6.7	6.9	5.9	1016	0.5
D16S392	1.8	3.4	0.0	104	4.6
D16S393	5.5	6.1	4.0	1014	3.2
D16S449	3.8	1.2	11.2	106	0.0
D16S305	8.4	4.6	13.5	106	0.0
D16S303					
Total	152.1	180.7	130.6		

* - foundation loci for the **build** analysis.

a - Odds are calculated from the sex-averaged map.

b - chi-square values calculated for each interval by including three adjacent intervals on each side, keeping Θ constant and equal for males and females in those intervals. With one degree of freedom, chi-square = 3.84 (α = 0.05). If corrected for multiple testing using the conservative test of Fisher-Bonferroni (i.e. divide α by 47), the critical value for chi-square is 10.73.

Table 7.3b: *Differences between males and females with respect to map distances (cM) for specific regions of chromosome 16.*

Region			female	male	Chi-Square	P value
16p	-	**16q**	180.7	130.6	290.3 47 df	$P < 0.05^*$
16pter	-	**S292**	27.2	48.7	31.7 7 df	$P < 0.05^*$
S287	**cen**	**S310**	88.8	12.2	196.0 22 df	$P < 0.05^*$
S422	-	**16qter**	22.2	34.6	11.5 6 df	$P > 0.05$

* Significant.

established by linkage was compatible with the physical order, independently derived using the somatic cell hybrid panel (Fig. 7.1). Major errors in locus order on the map were tested for by removing each locus in turn and placing it in all the remaining intervals and comparing the maximum likelihood, thus testing the global support for order. Support for local order was investigated by calculating the likelihood against the inversion of each pair of adjacent loci. These analyses showed that all loci were assigned to a unique interval, with odds of 1000:1 or greater.

The 34 loci not placed on the framework map are indicated on Fig. 7.1 in their most likely regional intervals. Eight of the comprehensive loci were localised to one specific interval by correlation of the cytogenetic-based physical map and the framework genetic map, and so the local order of some markers could be resolved. The comprehensive loci were also placed in their most likely intervals on the framework map to construct an 82 locus comprehensive map (Table 7.4a); loci were placed in their most likely positions regardless of the level of support. The sex-average length of the comprehensive map was 151.1 cM, with an average interlocus distance of 1.9 cM and a median distance of 1.2 cM. The largest gap, between *D16S414* and *D16S405* is 13.3 cM. The male and female map lengths of the comprehensive map are 127 cM and 179 cM respectively. The map extends from 170 - 230 Kb at pter (*D16S85*) (Wilkie *et al.*, 1991) to *D16S44* which is no further than 230 Kb from 16qter (Harris and Thomas, 1992), although *D16S303* maps distal (Peter Harris, personal communication) to *D16S44*.

The average map length after each locus was removed in turn on the final comprehensive map described above, was 151.01 cM, giving an estimated error rate of 0.03% for the 82 polymorphic loci (Fig. 7.2b). The total inflation of length of the PCR map, due to errors was thus calculated to be 4.4 cM (from the equation above). Thus the final error compensated PCR map length would be 146.7 cM, only 1% longer than the map predicted by chiasma studies. The locus

Table 7.4a: *Distances (in cM) between the loci on the comprehensive map, including sex-specific and sex-average distances and local support for order.*

Locus	Sex-ave	Female	Male	Odds[a]
D16S85	5.2	2.6	8.3	10^{11}
D16S525	2.0	0.0	4.0	10^{5}
D16S291	1.1	0.8	1.2	10^{6}
D16S283	0.4	1.0	0.0	10^{2}
D16S94	1.6	0.6	2.3	10^{3}
D16S523	1.8	1.3	2.6	10^{2}
D16S423	0.0	0.0	0.0	1.0
D16S453	5.8	3.0	8.4	10^{15}
D16S406	0.0	0.0	0.0	1.0
D16S418	4.1	5.6	2.7	10^{13}
D16S404	0.0	0.0	0.0	1.0
D16S407	1.5	2.6	0.0	10^{3}
D16S414	13.6	10.3	17.4	10^{21}
D16S405	0.0	0.0	0.0	1.0
D16S454	0.0	0.0	0.0	1.0
D16S292	0.4	0.0	0.8	1.0
D16S49	4.0	4.1	4.0	10^{15}
D16S79A	1.6	0.8	2.1	10^{11}
D16S287	3.0	5.5	1.2	10^{10}
D16S410	2.0	2.8	0.5	10^{2}
D16S452	0.0	0.0	0.8	1.0
D16S524	3.9	6.5	1.2	10^{7}
D16S412	1.3	3.2	0.0	2.0
D16S403	0.0	0.0	0.0	1.0
D16S417	0.0	0.0	0.0	1.0
D16S294	0.6	1.1	0.0	10^{3}
D16S420	2.4	5.3	0.0	10^{9}
D16S401	1.2	0.0	1.5	10^{2}
D16S319	3.7	5.3	2.5	10^{28}
D16S67	1.3	2.1	0.0	10^{8}

Table 7.4a *(cont'd): Distances (in cM) between the loci in the comprehensive map, including sex-specific and sex-average distances and local support for order.*

Locus	Sex-ave	Female	Male	Odds[a]
D16S295	0.5	0.0	1.9	10^6
D16S296	0.7	1.4	0.0	10^{14}
D17S297	1.6	3.0	0.0	10^{24}
D16S313	1.4	2.6	0.3	10^{23}
D16S288	0.9	1.5	0.3	10^{11}
D16S299	0.3	0.0	0.3	1.6
D16S298	0.3	0.9	0.0	10^4
SPN	*0.0*	0.0	0.1	1.0
D16S383	0.7	0.8	0.2	10^5
D16S300	1.0	1.9	0.0	10^8
D16S285	0.0	0.0	0.2	1.0
D16S411	0.5	0.8	0.3	10^3
D16S261	3.9	7.3	0.3	10^{17}
D16S308	0.0	0.0	0.2	1.0
D16S304	0.4	0.0	0.7	1.9
D16S359	2.7	5.5	0.0	10^8
D16S416	2.2	2.8	1.5	10^7
D16S415	0.0	0.0	0.0	1.0
D16S419	3.5	7.9	3.5	10^{18}
D16S390	2.7	4.9	0.5	10^{10}
D16S408	0.0	0.0	0.3	1.0
D16S531	3.7	7.2	0.4	10^{18}
D16S320	2.2	4.8	0.0	10^{17}
D16S310	1.0	1.9	0.0	10^3
D16S389	1.5	2.4	0.0	10^4
D16S400	0.0	0.0	0.0	1.0
D16S451	1.3	0.3	2.4	10^3
D16S265	0.0	0.0	0.0	1.0
D16S164	2.1	2.1	2.4	10^{14}
D16S186	0.0	0.0	0.0	1.0

Table 7.4a *(cont'd): Distances (in cM) between the loci in the comprehensive map, including sex-specific and sex-average distances and local support for order.*

Locus	Sex-ave	Female	Male	Odds[a]
D16S347	0.6	1.0	0.0	10^5
D16S421	0.0	0.0	0.0	1.0
D16S398	0.0	0.3	0.0	1.0
D16S318	0.0	0.3	0.0	1.0
D16S397	0.6	0.3	0.0	1.9
D16S301	0.7	0.4	1.1	10^5
D16S522	3.2	1.3	4.3	10^{15}
D16S450	5.3	6.9	5.1	10^{15}
D16S395	2.1	3.0	0.0	10^2
D16S266	7.1	10.6	3.5	10^{11}
D16S289	7.1	6.6	7.6	10^{19}
D16S422	0.0	0.0	0.0	1.0
D16S363	1.9	2.2	1.7	10^7
D16S402	5.9	6.8	5.1	10^{18}
D16S392	2.0	4.2	0.0	10^2
D16S332	0.0	0.0	0.0	1.0
D16S393	5.3	5.2	3.8	10^{16}
D16S449	4.5	1.1	9.8	10^{11}
D16S305	1.1	3.7	0.0	10^2
D16S413	6.3	0.4	10.8	10^{11}
D16S44	0.0	0.0	0.0	1.0
D16S303				
Total	151.1	178.9	126.8	

a - Odds are calculated from the sex-averaged map.

Table 7.4b: *Differences between males and females with respect to map distances (cM) for specific regions of chromosome 16 on the comprehensive map.*

Region			female	male	Chi-Square	P value
16p	-	**16q**	178.9	126.8	293.0 81 df	$P < 0.05^*$
16pter	-	**S292**	27.8	46.9	34.7 16 df	$P < 0.05^*$
S287	**cen**	**S310**	85.18	18.7	187.2 35 df	$P < 0.05^*$
S422	-	**16qter**	23.6	31.2	23.8 10 df	$P < 0.05^*$

* Significant.

contributing to the greatest reduction in map length was *D16S305*, with a reduction to 146.3 cM or of 3.2%. This marker is often the most telomeric marker on the map, and thus any recombinants due to this marker are not flagged as data errors by **chrompic**. Recombinants due to this markers were checked if they occurred in a short interval; those that remain in the map are in regions greater than 20 cM. A graphical representation comparing the results of removing each locus in turn for both the preliminary comprehensive map and also the final comprehensive map, are given in Fig. 7.2.

The order of loci on these two maps was not quite identical. Firstly, the two loci *D16S523* and *D16S524*, were not included on the preliminary map, and two RFLP loci *D16S7* and *APRT* were not incorporated onto the final map. The preliminary map order of loci in the region of *FRA16A* was *D16S49*, *D16S405*, *D16S454*, and *D16S292*. The order of these loci on the final map, after the correction of erroneous genotypes in *D16S49* and *D16S454*, is *D16S405*, *D16S454*, *D16S292*, and *D16S49*. On Fig. 7.2 (a and b), a sharp decrease in map length is observed at the 16qter end of the map, which is greater on the final comprehensive map. The locus *D16S305* is responsible for this decrease, and as stated before this locus is often the most telomeric marker and thus double recombinants due to this locus are not flagged as data errors. The two RFLPs, *D16S7* and *APRT*, were included on the initial map, which has obviously had an effect on the recombinants detected for *D16S305*, observed on the error compensated map, .

Differences with respect to the recombination between males and females were investigated on the final framework and comprehensive maps. The $X^2_{(47)}$ test of recombination differences in males and females, for the total length of the framework map, was observed to be significant (significant if $X^2_{(47)} > 64.0$, $P = 0.05$). For this test, $X^2_{(47)} = 290.3$, $P < 0.05$, the hypothesis of a significant difference in the recombination between males and females was not rejected, implying that the overall ratio of 1.4:1 female to male recombination

rates is significant at the 5% level. The next step in quantifying the recombination difference with respect to males and females, was to determine if any of the primary intervals between the end point loci, *D16S85* and *D16S303*, were significantly different. Given 47 comparisons, the stringent multiple comparison P value of 0.00106 (overall P = 0.05) was used, using the conservative test of Fisher-Bonferroni; the test statistic was significant if $X^2_{(1)} > 10.73$; seven of the 47 intervals revealed a significant difference between interval specific male and female recombination. The results of these tests are indicated in Table 7.3a. Comparisons of regions on the framework map was performed to determine if specific regions of the map were responsible for the significant difference with respect to the recombination frequency between males and females (See Table 7.3b). The first region comparing the male and female recombination fraction differences was from 16pter to *D16S292*. The chi-squared value at the 5% level was critical if $X^2_{(7)} > 14.05$; this region showed a significant difference with respect to male and female recombination ($X^2_{(7)} = 31.7$). The next region on the framework map to be investigated was across the centromere, where there is observed to be a 7 fold excess in recombination in females over males, between the markers *D16S287* and *D16S310*. The chi-squared value on 22 degrees of freedom was $X^2_{(22)} = 196.0$. There thus appears to be suppression in the recombination rate in males compared with females across this region which was confirmed statistically. This phenomenon has been reported elsewhere, in relation to Batten disease (Chapter 5; Gardiner *et al.*, 1990), and in general (Skolnick, 1991). The last region to be investigated was at 16qter. The chi-squared value at the 5% level was critical if $X^2_{(6)} > 12.59$, thus this region showed no significant difference with respect to male and female recombination ($X^2_{(6)} = 11.5$).

The overall female to male ratio of recombination fractions in the sex-specific comprehensive map is 1.4:1 (179 cM / 127 cM), which is significant ($X^2_{(81)} =$

293.0, $P < 0.05$). For the specific intervals investigated, as for the framework map above, from pter to *D16S292*, the male map length is 1.7 fold longer than the female map, and from *D16S422* to 16qter, the male map is 1.3 fold longer than the female map. These differences are significant (Table 7.4b). Across the centromeric region of the comprehensive map there is a 4.6 fold excess in female recombination; this excess is significant (Table 7.4b).

There appears to be a number of clusters and gaps of the simple tandem repeats on the framework genetic linkage map (Fig. 7.1). The largest gap occurs on the male genetic map, and is 18.2 cM between *D16S404* and *D16S292*. This is the only gap on the genetic map in excess of 10 cM. On the comprehensive map, the largest gap, occurring in the same region on the male genetic map, is 17.4 cM. At each of the telomeres, on the sex-average framework map, the mean distance between loci is 4.7 cM and 4.1 cM for 16pter and 16qter respectively, as compared to the 3.2 cM average for the entire length of the map. Across the centromere there appears to be a clustering of markers, where the average interlocus distance, between *D16S67* and *D16S416* (a region containing 12 of the 82 STR markers), is only 1.2 cM, as opposed to the 3.2 cM mean distance for the entire map. Thus, it appears that there is a clustering of loci around the centromeric region, and a lack of loci at each of the telomeres. It would be difficult to determine if this were really the case, unless physical distances could be compared. As there appears to be definite suppression of recombination in the centromeric region (Skolnick, 1991), and increased recombination at the telomeres (NIH/CEPH Collaborative Mapping Group, 1992), this would account for the apparent clustering of loci in the centromeric region, and sparsity of loci at the telomeres.

All loci used in the construction of this map have been accurately localised on the cytogenetic map and the high resolution framework and comprehensive genetic maps were correlated with the high resolution cytogenetic-based physical map,

demonstrating excellent accordance between them. It is apparent that the combination of genetic linkage analysis and physical mapping can be advantageous in resolving locus order at the resolution of the comprehensive map. For example, in Fig. 7.1, it can be seen that the comprehensive loci have been placed on the map in regions where they lie with odds of 1000:1 over any other interval or region in the map. Combining this information with their physical location enables the refinement of location of ten of the comprehensive loci, and the placement of eight comprehensive loci (*D16S94*, *D16S523*, *D16S383*, *D16S261*, *D16S411*, *D16S347*, *D16S266*, and *D16S363*) to a single interval. A good example of this is the STR marker, *D16S363*. Its genetic location is between *D16S402* and *D16S289* on the framework map. Its physical location, between the hybrids CY100, CY120 and CY18A(P), implies that the *D16S363* must be located between the framework markers *D16S422* and *D16S402*. Additional loci to which this rationale can be applied are indicated on the figure.

It is also possible to define the order of ambiguous loci on the physical map using the genetic map. Some markers are not separated by hybrids on the physical map, but are uniquely ordered on the framework genetic map with odds of 1000:1 or greater. As an example, *D16S291* and *D16S283* are not ordered with respect to the physical map. However, they are ordered on the genetic map with odds of 9 x 10^3. This is also true of the 5 loci (*D16S412, D16S420, D16S319, D16S67, D16S295*) located between the hybrids CY15 and CY165. On the genetic map these loci are uniquely ordered with odds of nearly 10^5 or greater:

	2.4		3.3		4.0		1.3	
D16S412		*D16S420*		*D16S319*		*D16S67*		*D16S295*
	$9x10^4$		$3x10^6$		$3x10^{25}$		$1x10^7$	

This demonstrates the usefulness of developing the physical and genetic map concurrently, and correlating the two as an on going progression of map validation. Other loci whose order can be determined uniquely on the genetic map

but not on the physical map are *D16S291* and *D16S283*, *D16S297* and *D16S313*, *D16S288* and *D16S298*, *D16S304* and *D16S416*, *D16S310* and *D16S389*, *D16S398* and *D16S301*, and the four loci between the hybrids CY100 and CY18A, *D16S402*, *D16S392*, *D16S393*, and *D16S449* (see Fig 7.1).

7.5. Discussion

The PCR based map of human chromosome 16 described provides a rapid means of screening this chromosome for the presence of disease genes. The framework map contains a total of 48 STR polymorphisms, and one VNTR polymorphism, *D16S85*, anchoring the map at 16pter. The polymorphisms are separated by an average distance of 3.2 cM, and extends 131 cM in males, 181 cM in females; the sex-average length of the map is 152 cM. With the average heterozygosity of these markers being 73%, an index map can potentially be delineated. The loci qualifying as index markers with heterozygosity greater than 70%, PCR formatted, and separated by approximately 10 cM, are indicated on Fig. 7.1. Approximately 99% of the chromosome is covered as *D16S85* maps 170 kb to 230 Kb from 16pter (Wilkie *et al.*, 1991), and *D16S303* maps less than 230 Kb from 16qter (Peter Harris, personal communication). Two non-STR markers, *D16S85* and *D16S44* were included in the map construction as anchors to the telomeric ends of the chromosome.

The length of the sex-average, male and female comprehensive maps were 151 cM, 127 cM and 179 cM respectively. Each marker has been localised on the high resolution cytogenetic-based physical map, which was divided into 66 breakpoint intervals (on average 1.5 Mb per interval) by a panel of 67 hybrids (Callen *et al.*, 1992a; Callen *et al.*, 1994). This enabled the framework and comprehensive maps to be anchored to the cytogenetic map and represents the integration of the physical and genetic maps, as a form of map verification; the two maps are in complete congruence.

The female framework map is 1.4 times longer than the male framework map, which is significant. This observed excess of female recombination is similar to the genome average, of 1.6, as estimated from the NIH/CEPH collaborative mapping group (1992). The distal regions, at 16pter and 16qter, show the previously reported trend of having a excess of male recombination over female recombination, where this excess is significant at 16pter. Published linkage maps on chromosome 16 (Donis-Keller *et al.*, 1987; Julier *et al.*, 1990; Keith *et al.*, 1990; Kozman *et al.*, 1993; Shen *et al.*, 1994) also observed the general trend of an excess of female recombination across the entire chromosome, and the same excess of male recombination at both telomeric ends of the chromosome.

The framework map, generated using the **build** option of CRI-MAP, contains loci with order support of 1000:1 or greater. Local and global support for the placement of each of the loci has been investigated, ensuring that the order presented here is indeed the correct order of loci, also supported completely by the order determined on the cytogenetic-based physical map. During the construction of the map, many factors may influence the outcome from the build analysis. The choice of starting loci for the foundation of the framework map may preclude the addition of other loci at later steps, leading to an erroneous order. Therefore the map was recomputed several times using different pairs of loci as the basis on which to build the map. Each analysis produced the same order of loci. It is also possible that the later addition of one locus may reduce the support for other loci previously placed on the map. This reduction is often dramatic enough to result in insufficient support for unique localisation (Buetow *et al.*, 1993). Hence, there are some instances where after a build run the resulting support of $<$ 1000:1 is obtained. In the map present here, there were some instances where markers were incorporated into the framework map with odds of less than 1000:1, but after determining the support for local and global order, these loci were excluded from

the framework map, resulting in the generation of a map of loci all with unique placement.

A locus on the map with a large portion of genotyping errors may preclude the placement of additional loci adjacent to it due to these data errors (Beutow *et al.*, 1993). Furthermore, it is possible that the effects of genotype errors are exacerbated by the use of a reduced CEPH panel, as the effect of one recombinant will be greater in the fewer families. Increasingly higher-resolution maps require even larger data sets to guarantee that a sufficient number of recombination events will be observed to order the loci. Buetow (1991) performed simulation studies that imply that the extended CEPH panel should have sufficient power to derive order for a small high resolution map composed of highly informative loci. This indicates that not only are the subset of eight families insufficient to interpret the order of loci, but it is possible that the genotypes of the 40 families will also be insufficient. On the comprehensive map (Table. 7.4), it is already evident that at this resolution (1.9 cM), the order of loci is difficult to determine, possibly exacerbated by having a reduced data set. It may not be necessary to genotype more than the families typed for this study, if maps are constructed within the bounds set by the order on the physical map, as is the case here.

Buetow (1991) and Lasher *et al.* (1991) examined in detail the effects of undetected genotyping errors on marker order and inflation of map distance. Lasher *et al.* (1991) provided a theoretical model to demonstrate the magnitude of the effect that genotyping errors have on the length of a single chromosome genetic map. For even a small error rate of 1%, with a map density of 2 cM, the relative error can be as great as 100%, leading to a doubling of the chromosome length.

The majority of errors in high-resolution genetic maps will appear as apparent recombinants with flanking markers; the observation of 2 or more recombinants within a short distance in a given meiosis (Lasher *et al.*, 1991). The vast

proportion of true gametes are nonrecombinant, and even a low frequency of genotyping errors (1% or less) proportionally represents a large fraction of true recombinant events (Buetow, 1991). After aggressive error checking of the data used to construct this high resolution STR based genetic map, many such recombination event were indeed detected during the analysis and the large majority of these were corrected following regenotyping; in 14 instances (listed in Table 7.2) double recombinant events could not be excluded as genotype errors after regenotyping, and the possibility remains that these are in fact either new mutations, true double or triple recombinants, or even gene conversion events.

A comparison of the length of the genetic map reported here with those previously published and with the chiasma map implies that although there may be some double or triple recombination events present in the data, with some affecting the total map length, they may be real events as opposed to errors in the data. The length of the sex-average PCR comprehensive map is 151 cM, and the length of map based on cytogenetic counts of chiasma (Morton, 1991b) is 145 cM. The estimated rate of error in genotypes is 0.03%, resulting in a corresponding error corrected map length of 146.7 cM, which is 1% longer than the length estimated by chiasma counts for this chromosome, indicating that there may be few errors remaining in the data.

Typing a given marker on only a subset of the 40 CEPH families appears, as expected, to reduce the chance of their incorporation into the framework map, compared with less informative markers typed on the full set of 40 families. The reason for this is a significant reduction in the number of potentially informative meioses for detection of the recombinants needed for ordering loci. Of the 39 markers genotyped only on the 8 largest CEPH families, 33% (16/48) were incorporated in the framework map and 68% (23/34) were incorporated at the comprehensive map level. Of the STR loci typed on the 40 CEPH families, 67% (32/48) were incorporated into the framework map, and 29% (10/34) were

incorporated at the comprehensive map level. The average heterozygosity of both groups of loci, those typed on eight families and those typed on 40 families are 76% and 65% respectively. Thus, genotyping markers on only eight families reduces the chance of being incorporated on the framework map, in spite of their level of heterozygosity.

Microsatellite markers have relatively high mutation rates, as high as 10^{-3} per generation (Weissenbach *et al.*, 1992), which has important implications for genetic mapping. When the mutation results in conversion to an allele already present in the family, it may be detected as an apparent double recombinant, and thus corrupt the analysis since regenotyping merely confirms the original result. Fourteen mutation events were detected as inconsistent genotypes within families, as a result of genotyping the 55 STR markers, leading to an estimated mutation rate of 4.0×10^{-4} per locus per gamete per generation (Shen *et al.*, 1994). This mutation rate is 2.5 - 3 times lower than those estimated previously by Weber and May (1989) with an estimate of 1.2×10^{-3}, and Weissenbach *et al.* (1992) with an estimate 1.0×10^{-3}. However, if the 14 double recombinants, detected above in the **chrompic** analysis, were in fact the result of a new mutation, that is a mutation resulting in a genotype compatible with the parental genotypes, but not fitting in with the phase for the family, then the estimate of the mutation rate would be 1.0×10^{-3} per locus per gamete per generation, a very similar result to those previously obtained (Weissenbach *et al.*, 1992). The possibility exists that a proportion of these double recombinants are in fact real events, and some may be due to a mutation.

An intriguing aspect to this resource of PCR based markers, is the number of microsatellite markers segregating with null alleles. PCR depends upon the binding of specific primers to either side of the polymorphic repeat sequence to generate a PCR product. If one of the primers fails to bind, no PCR product will be formed - and this is referred to a null allele. Null alleles were found in seven

STR markers (Callen *et al.*, 1993). Four of the loci (*D16S287, D16S294, D16S304,* and *D16S305)* showed a null allele in a single sibship, and three of the loci (*D16S67, D16S79A,* and *D16S297)* showed a null allele in many of the CEPH families (Callen *et al.*, 1993; Phillips *et al.*, 1992). Callen *et al.* (1993) commented that null alleles can occur at appreciable frequency at a locus but their presence will not corrupt a linkage analysis, either in the construction of maps, or in application to diagnosis. This is because the parent will appear uninformative, with only one detectable allele suggesting homozygosity. The presence of a segregating null allele results in loss of information due to the reduction in informative meioses.

Development of this 3 cM resolution map of highly informative STR polymorphisms has occurred within the time frame specified by the first five year goal of the human genome project. The goal was the attainment of a resolution of between 2 to 5 cM for genetic maps (Collins and Galas, 1993). The physical map of this chromosome provides information enabling the resolution of the order of loci at the comprehensive map level to be refined. The map constructed in this chapter has been aggressively evaluated for genotype errors in the loci, and provides support for the order of loci, and the distances between them, and is closely integrated with the physical map. The cytogenetic-based physical map is continuing to develop (Callen *et al.*, 1994), and now divides the chromosome into 67 unique physical intervals, with an average resolution of 1.3 Mb.

The availability of an accurate linkage map of chromosome 16 with many polymorphic markers is important for genetic studies involving this chromosome. It will provide an efficient means for regional localisation of genetic disorders located on chromosome 16, for the detection of loss of heterozygosity in cancers, for the evaluation of linkage disequilibrium and disease causing mutations, and for the analysis of multifactorial diseases. It will enable chromosome 16 to be screened for disease genes entirely with PCR typeable polymorphisms. Integrated

fine structure maps loci will eventually be applied to the nucleotide sequencing of this chromosome, but for the present they greatly simplify the isolation and characterisation of genes and disorders isolated on chromosome 16.

CHAPTER 8

THE CEPH CONSORTIUM LINKAGE MAP OF HUMAN CHROMOSOME 16

Table of Contents

Associated Paper (See Appendix A):

The CEPH consortium map of human chromosome 16 Kozman, H.M., Keith, T.P., Donis-Keller, H., White, R.L., Weissenbach, J., Dean, M., Vergnaud, G., Kidd, K., Gusella, J., Royle, N.J., Sutherland, G.R., and Mulley, J.C. (1994). Genomics, In Press.

8.1. Summary

A Centre d'Etude du Polymorphisme Humain (CEPH) consortium map of human chromosome 16 has been constructed. The map contains 158 loci defined by 191 different primer pairs or probe/restriction enzyme combinations, of which 7 represent gene loci. The marker genotypes, contributed by 9 collaborating laboratories, were obtained from the CEPH families DNA. A total of 60 loci, with an average heterozygosity of 68%, have been uniquely placed on the framework genetic map. The length of the sex average map is 165 cM, with a mean genetic distance between loci of 2.8 cM; the median distance between markers is 2.0 cM. The male map length is 136 cM and the female map length is 197 cM. The map virtually covers the entire chromosome, from *D16S85*, within 170 to 430 Kb of the 16p telomere, to *D16S303* at 16qter. All loci incorporated in the linkage map have been extensively physically mapped on a partial human chromosome 16 somatic cell hybrid panel, thus anchoring the genetic map to the cytogenetic-based physical map, and enabling confirmation of the map order. This is by far the most detailed genetic map ever provided for chromosome 16, collating all data gathered prior to July 31, 1994.

8.2. Introduction

CEPH consortium maps have been published for six human chromosomes: 1, 2, 9, 10, 13, and 15q (Dracopoli *et al.*, 1991; Spurr *et al.*, 1992; Attwood *et al.*, 1994; White *et al.*, 1990; Bowcock *et al.*, 1993a; 1993b). The features of each of these consortium maps are summarised in Table 8.1.

This chapter describes the construction of the CEPH consortium linkage map of human chromosome 16, built from 191 polymorphic markers defining 158 loci, of which 79 are PCR formatted markers, 102 are classical RFLP markers, and 10 are VNTR markers. The genotype data were contributed to the consortium database by nine collaborating laboratories (Table 8.2), and represent genotypic data obtained from the families in the CEPH reference panel (Dausset *et al.*, 1990).

Previous multipoint linkage maps of chromosome 16 based on the CEPH family genotypes have incorporated many of the markers in this study (Donis-Keller *et al.*, 1987; Julier *et al.*, 1990; Keith *et al.*, 1990; NIH/CEPH collaborative mapping group, 1992; Weissenbach *et al.*, 1992; Kozman *et al.*, 1993; Shen *et al.*, 1994).

Localised genetic maps, surrounding three of the 16 disease genes known to be located on chromosome 16 (See Fig. 1.1), for *CLN3* (Mitchison *et al.*, 1993; 1994), *PKD1* (Hyland *et al.*, 1990; Keith *et al.*, 1990), and *MEF* (Aksentijevich *et al.*, 1993) have been previously published, as have genetic maps around two of the rare fragile sites on this chromosome, FRA16A and FRA16B (Kozman *et al.*, 1991; 1992). An extensive background genetic map, such as the CEPH consortium linkage map, will provide opportunities for the regional localisation of disease genes not already finely mapped on chromosome 16, and will be the foundation for the generation of maps in regions of other disease genes, not previously mapped. These studies are essential for the initiation of high resolution physical mapping, involving YAC or cosmid isolation and characterisation which

Table 8.1: *Summary of CEPH Consortium Maps*

Chromosome	No.Loci/ No.Markers	No. Framework Markers	Mean Distance	No. PCR Markers	No. Labs	Genetic Length(cM)a female	male	ave	Physical Length(Mb)	Loci/Mb	Reference
1	101/ 146	58	6.7	4	11	477.7 (362	308 201	389.7 282)	263	0.4	Dracopoli *et al*, 1991
2	60/ 73	36	9.1	1	14	430 (337	260.6 185	327.8 261)	255	0.2	Spurr *et al*, 1992
9	97/ 124	42	4.3	59	14	237 (164	176 129	209 146)	145	0.7	Attwood *et al*, 1994
10	39/ 52	28	11	-	5	309 (205	214 129	- 167)	144	0.3	White *et al*, 1990
13	59/ 94	25	7	-	9	203 (160	158 100	178 130)	114	0.5	Bowcock *et al*, 1993
15q	41/ 45	21	8	3	10	190 (195	127 100	158 148)	106	0.4	Bowcock *et al*, 1993
16	158/ 191	60	2.3	79	9	196 (179	136 111	165 145)	98	1.6	This study

a - The values in brackets are those predicted by Morton (1991b).

Table 8.2: *Contributors to the CEPH Chromosome 16 Database*

Laboratory Number	Collaborator and Institution	Markers Contributed
1	Dr H. Donis-Keller Department of Surgery Washington University, School of Medicine St. Louis, Missouri, U.S.A.	58
8	Dr K.K. Kidd Department of Human Genetics Yale University School of Medicine New Haven, CT, U.S.A.	2
10	Dr J. Gusella Neurogenetics, Research 3 Massachusetts General Hospital, Boston, MA, U.S.A.	1
20	Dr R. White Howard Hughes Medical Institute University of Utah, Salt Lake City, Utah, U.S.A.	29
31	Dr A.J. Jeffreys Department of Genetics University of Leicester Leicester, U.K.	1
42	Dr J. Weissenbach Génétique Moléculaire Humaine Institut Pasteur Paris, FRANCE	22
43	**Dr G.R. Sutherland** **Department of Cytogenetics** **and Molecular Genetics,** **Women's and Children's Hospital** **North Adelaide, AUSTRALIA**	**73**
46	Dr G.V. Vergnaud Centre d'Etudes du Bouchet, Vert le Petit, and Institut de Biologie, Nantes, FRANCE	2
55	Dr M. Dean National Cancer Institute Frederick Cancer Research Facility Frederick, MD, U.S.A.	3
	Total	191

will lead to the eventual isolation of the genes responsible for diseases on chromosome 16.

Chromosome 16 has been extensively mapped by somatic cell hybrid analysis using Southern or PCR analysis, or *in situ* hybridisation. This cytogenetic-based physical map is based on human/mouse somatic cell hybrids, utilizing the *APRT* gene (at 16q24) as a selectable marker, which have been constructed from rearrangements of chromosome 16. The chromosome is divided by these hybrids, the breaks defined by the fragile sites and the centromere, to represent over 66 distinct physical intervals (Callen *et al.*, 1992). For chromosome 16, the genetic and cytogenetic maps have been refined simultaneously, and 158 of the 191 polymorphic markers in the chromosome 16 consortium database have been localised to specific physical intervals on the cytogenetic-based physical map. This enables a detailed correlation between the physical and genetic maps, as a natural progression of map validation. The marker order on the framework genetic map shows no discrepancies with the order of these markers on the cytogenetic based physical map. This is the first consortium map to be tightly anchored to the physical map.

8.3. Materials and Methods

The 191 polymorphic systems (defining 158 loci) in the chromosome 16 CEPH consortium database were submitted by 9 collaborating laboratories (Tables 8.2 and 8.3). All genotypes were obtained from the database, except where indicated on Table 8.3, where the genotyping was performed by the candidate (see also Chapters 3 and 4).

The database includes all data contained within Version 5 of the CEPH database, with the exception of HP (the protein polymorphism), and *D16S149* (CJ59.27), and with the addition of 53 newly genotyped PCR markers from collaborators 42 and 43 submitted for the construction of the consortium map.

Table 8.3: *Markers included in the Chromosome 16 Consortium Map.*

Location	Gene/Locus	Probe/Enzyme	Het.	No inf meioses	CEPH Lab.
16q24.3	APRT	M13-APRT/TaqI	0.28	192	8
		M13-APRT/TaqI	0.27	150	8, 20
		Aprt/BglII	-	27	20
16q12.1-q13	CETP	pCETP.11A/TaqI	0.47	98	20
16q23.1-q24.2	CTRB1	2pEKXp3B/PvuII	0.38	311	20
16p13.3	HBZP1	HBZP1/BglII	0.47	130	01
		HBZP1/HincII	0.80	246	01
16q22.1-q23.1	HP	hp2-alpha/MspI	0.50	336	01
16q12.1-q13	MT2A	MT2AA/TaqI##	0.25	118	43
		MT2AB/TaqI##	0.20	-	43
16p12.1-p11.2	SPN	SPN/PCR	0.96	463	43
16q22.1	D16S4	ACH207/TaqI##	0.36	361	43
		ACH207/MspI##	0.44	-	43
		ACH207/MspI##	0.47	-	43
16q23.1-q24.2	D16S5	ACH224/RsaI**	0.44	61	43
16q24.3	D16S7	p79-2-23/TaqI	0.64	201	01
		p79-2-23/RsaI	0.82	677	20
16p13.2-p13.11	D16S8	ACHF1/PvuII	0.50	354	43
16q21	D16S10	ACHF3/TaqI##	0.47	348	43
		ACHF3/RsaI##	0.41	-	43
		ACHF3/MspI##	0.18	-	43
16q23.1-q24.2	D16S20	phi8-9/BglII	0.47	406	20
16q22.1	D16S38	CRI-O2/BamHI	0.35	73	01
16q12.1-q13	D16S39	CRI-O3/PstI	0.49	173	01
16q23.1-q24.2	D16S40	CRI-O15/TaqI	0.50	172	01
16q24.2-q24.3	D16S41	CRI-O43/TaqI	0.36	85	01
16p13.11-p12.1	D16S42	CRI-O66/PstI	0.36	103	01
16q23.1-q24.2	D16S43	CRI-O84/HindIII	0.33	111	01
		CRI-O84/PstI	0.69	237	01
16q24.3	D16S44	CRI-O89/BglII	0.50	206	01
16p13.3	D16S45	CRI-O90/EcoRI	0.49	304	01
		CRI-O90/DraI	0.14	52	01
16q22.1	D16S46	CRI-O91/MspI	0.36	137	01
		CRI-O91/TaqI	0.38	104	01
16q21-q23.1	D16S47	CRI-O95/EcoRI	0.49	158	01
16p12.1-p11.2	D16S48	CRI-O101/HindIII	0.47	144	01
16p13.3-p13.11	D16S49	CRI-O114AC/PCR	0.66	113	43
		CRI-O114/EcoRI	0.39	169	01
16q23.1-q24.2	D16S50	CRI-O119/TaqI	0.46	162	01
16p13.3-p13.11	D16S51	CRI-O120/BamHI	0.60	220	01

- Markers typed on only 8 of the CEPH families.
- Markers typed by the candidate, as outlined in Chapters 3 and 4.
* - Physical location not refined.
** - Markers not previously incorporated into a published genetic linkage map of chromosome 16.

Table 8.3 (cont.): *Markers included in the Chromosome 16 Consortium Map.*

Location	Gene/Locus	Probe/Enzyme	Het.	No inf meioses	CEPH Lab.
16q12.1-q13	D16S52	CRI-O123/HindIII	0.41	92	01
16p*	D16S53	CRI-O125/PstI	0.49	175	01
16p*	D16S54	CRI-O126/HincII	0.20	53	01
		CRI-O126/EcoRI	0.42	161	01
16p13.3	D16S55	CRI-O128/HincII	0.30	72	01
16p13.3	D16S56	CRI-O129/EcoRI	0.38	235	01
		CRI-O129/BglII	0.20	75	01
16p11.2-p11.1	D16S57	CRI-O131/BamHI	0.40	79	01
16p13.3	D16S58	CRI-O133/HindIII	0.32	248	01
16q*	D16S59	CRI-O134/MspI	0.33	122	01
16p13.3-p13.11	D16S60	CRI-O136/HincII	0.83	444	01
16p*	D16S61	CRI-O144/EcoRI	0.46	198	01
16q24.2-q24.2	D16S62	CRI-O149/TaqI	0.45	106	01
16p13.3	D16S63	CRI-O327/HindIII	0.38	276	01
16p13.11-p12.1	D16S64	CRI-O373/HindIII	0.46	195	01
16q12.1-q13	D16S65	CRI-O377/BglII	0.41	148	01
		CRI-O377/HincII	0.50	84	01
		CRI-O377/TaqI	0.24	56	01
16p*	D16S66	CRI-O383/PstI	0.47	189	01
16p13.11-p12.1	D16S67	CRI-O391AC/PCR	0.77	395	43
		CRI-O391/HincII	0.46	128	01
		CRI-O391/TaqI	0.50	199	01
16p12-qter*	D16S68	CRI-O393/BglII	0.47	165	01
		CRI-O393/PstI	0.28	100	01
16q22.1-q24.3	D16S69	CRI-P84/BamHI	0.38	132	01
16p*	D16S70	CRI-P130/MspI	0.48	195	01
16q*	D16S71	CRI-P400/BamHI	0.48	139	01
16p*	D16S72	CRI-P403/MspI	0.46	178	01
16p*	D16S73	CRI-P477/TaqI	0.38	128	01
16p*	D16S74	CRI-P85/BglII	0.49	196	01
16p13.11-p12.1	D16S75	CRI-R99/HindIII	0.51	162	01
		CRI-R99/MspI	0.31	88	01
16p*	D16S76	CRI-L223/TaqI	0.46	160	01
16q*	D16S77	CRI-L922/HincII	0.36	112	01
		CRI-L922/HindIII	0.21	72	01
16p13.3-p13.11	D16S79A	16AC66F3/PCR	0.81	398	43
		36-1A/TaqI##	0.34	245	43
16p13.11-p12.1	D16S79B	36-1B/TaqI##	0.15	83	43
16p13.3	D16S80	24-1/TaqI**	0.34	256	55
16p13.3*	D16S83	pEKMDA2.1/HinfI	0.87	752	20

\# - Markers typed on only 8 of the CEPH families.
\## - Markers typed by the candidate.
* - Physical location not refined.
** - Markers not previously incorporated into a published genetic linkage map of chromosome 16.

Table 8.3 (cont.): *Markers included in the Chromosome 16 Consortium Map.*

Location	Gene/Locus	Probe/Enzyme	Het.	No inf meioses	CEPH Lab.
16p13.3*	D16S84	pCMM65/EcoRI	0.39	357	20
16p13.3	D16S85	3'HVR/MspI	0.93	573	1
		3'HVR/RsaI/MspI	-	425	1
		5'HVR/RsaI	0.79	257	1
16q22.1	D16S91	LE12/RsaI##	0.11	46	43
16p13.3	D16S94	16ACVK5/PCR	0.51	383	43
		VK5B/MspI	0.49	324	43
16p13.11-p12.1	D16S96	VK20A/TaqI##	0.49	347	43
		VK20A/MspI##	0.36	-	43
		VK20B/MspI##	0.59	-	43
16p13.11-p12.1	D16S131	VK45C6/TaqI	0.46	240	43
16p*	D16S137	pKKA22/PstI	0.60	163	20
16p13.11-p12.1	D16S148	CJ52.95M1/MspI	0.39	326	20
16q12.1-q13	D16S150	1CJ52.161/TaqI	0.38	281	20
		2CJ52.161/TaqI	0.04	31	20
16q21	D16S151	1CJ52.209M/MspI	0.32	266	20
		2CJ52.209M/MspI	0.25	178	20
16q22.1	D16S152	1CJ52.1/TaqI	0.31	158	20
		2CJ52.1/TaqI	0.50	236	20
16q22.2-q23.1	D16S153	CJ52.10T2/TaqI	0.46	466	20
16q24.2-q24.3	D16S154	CJ52.105/TaqI	0.69	466	20
16q22-qter*	D16S155	CJ52.199/MspI	0.43	344	20
16q22-qter*	D16S156	1CJ52.197/TaqI	0.27	244	20
		2CJ52.197/TaqI	0.23	141	20
16q23.1-24.3	D16S157	CJ52.96/TaqI	0.25	148	20
16p13.2-q21	D16S158	CJ52.112/TaqI	0.15	113	20
16p13.11-p12.1	D16S159	CJ52.94T1/TaqI	0.50	481	20
16q*	D16S160	CJ52.196M1/MspI	0.50	278	20
16q21	D16S164	16PHAC-15/PCR	0.38	226	43
16q21	D16S186	16PHAC-101/PCR	0.57	413	43
16p13.3*	D16S246	218EP/PvuII**	0.46	265	55
16p13.3*	D16S252	pCMM103/PvuII	0.51	218	20
16q12.1-q13	D16S261	MFD24/PCR	0.71	449	43
16q21	D16S265	MFD23/PCR	0.77	505	43
16p13.3*	D16S282	14C19/HaeIII**	0.73	421	46
16p13.3	D16S283	SM7/PCR	0.64	407	55
16q12.1-q13	D16S285	AluGT16/PCR	0.83	496	10
16p13.11-p12.1	D16S287	16XE81/PCR	0.78	502	43
16p12.1-p11.2	D16S288	16AC7.1/PCR	0.73	449	43
16q23.1-q24.2	D16S289	16AC7.46/PCR	0.77	396	43

- Markers typed on only 8 of the CEPH families.
- Markers typed by the candidate.
* - Physical location not refined.
** - Markers not previously incorporated into a published genetic linkage map of chromosome 16.

Table 8.3 (cont.): *Markers included in the Chromosome 16 Consortium Map.*

Location	Gene/Locus	Probe/Enzyme	Het.	No inf meioses	CEPH Lab.
16p13.3	D16S291	16AC2.5/PCR	0.79	519	43
16p13.2-p13.11	D16S292	16AC2.3/PCR	0.74	450	43
16p13.11-p12.1	D16S294	16AC1/PCR	0.49	235	43
16p13.11-p12.1	D16S295	16AC62F3/PCR	0.66	430	43
16p13.11-p12.1	D16S296	16AC62B4/PCR	0.75	502	43
16p13.11-p12.1	D16S297	16AC15H1H/PCR	0.73	412	43
		16AC15H1S/PCR	0.69	124	43
16p12.1-p11.2	D16S298	16AC3.12/PCR	0.80	455	43
16p12.1-p11.2	D16S299	16AC6.17/PCR	0.72	425	43
16p11.2-p11.1	D16S300	16AC1.1/PCR	0.61	436	43
16q22.1	D16S301	16AC6.21/PCR	0.64	409	43
16q24.3	D16S303	16AC6.26/PCR	0.43	230	43
16q12.1-q13	D16S304	16AC1.14/PCR	0.60	410	43
16q24.3	D16S305	16AC1.15/PCR	0.82	494	43
16p*	D16S307	pMS637/HaeIII**	0.56	401	46
16q12.1-q13	D16S308#	16AC1.18/PCR	0.77	113	43
16p13.3*	D16S309	pMS205/MboI**	0.97	516	31
16q21	D16S310	MIT-MH20/PCR	0.67	472	43
16p13.11-p12.1	D16S313	MIT-MS79/PCR	0.57	498	43
16q22.1	D16S318#	16AC8.20/PCR	0.54	143	43
16p13.11-p12.1	D16S319	16AC7.14/PCR	0.52	315	43
16q12.1-q13	D16S320	16AC8.52/PCR	0.86	559	43
16q23.2-q24.3	D16S332	16AC305D6/PCR	0.53	324	43
16q22.1	D16S347#	16AC12F8/PCR	0.76	147	43
16q12.1-q13	D16S359	16AC26E3B/PCR	0.42	321	43
16q23.2-q24.3	D16S363#	16AC51G1/PCR	0.78	164	43
16p12.1-p11.2	D16S383	16AC80B3/PCR	0.45	235	43
16q21	D16S389#	16AC10B3/PCR	0.77	170	43
16q12.1-q13	D16S390	16AC10F5/PCR	0.80	511	43
16q23.2-q24.3	D16S392#	16AC305E9/PCR	0.78	161	43
16q23.2-q24.3	D16S393#	16AC323H4/PCR	0.87	159	43
16q23.2-q24.2	D16S395	16AC33G11/PCR	0.69	467	43
16q22.1	D16S397#	MFD98/PCR	0.70	168	43
16q22.1	D16S398	MFD168/PCR	0.90	559	43
16p13.11-p12.1	D16S401#	AFM025tg9/PCR	0.74	80	42
16q23.1-q24.3	D16S402#	AFM031xa5/PCR	0.87	140	42
16p13.11-p12.1	D16S403#	AFM049xd2/PCR	0.86	171	42
16p13.2-13.11	D16S404#	AFM056yf6/PCR	0.82	158	42
16p13.2-13.11	D16S405#	AFM070ya1/PCR	0.78	143	42
16p13.3-13.11	D16S406#	AFM079yh3/PCR	0.82	152	42

- Markers typed on only 8 of the CEPH families.
- Markers typed by the candidate.
* - Physical location not refined.
** - Markers not previously incorporated into a published genetic linkage map of chromosome 16.

Table 8.3 (cont.): *Markers included in the Chromosome 16 Consortium Map.*

Location	Gene/Locus	Probe/Enzyme	Het.	No inf meioses	CEPH Lab.
16p13.2-13.11	D16S407#	AFM113xe3/PCR	0.86	160	42
16q12.1-q13	D16S408#	AFM137xf8/PCR	0.69	120	42
16p13.11-p12.1	D16S410#	AFM165yb6/PCR	0.57	125	42
16q12.1-q13	D16S411#	AFM186xa3/PCR	0.79	130	42
16p13.11-p12.1	D16S412#	AFM191wb10/PCR	0.76	101	42
16q24.3	D16S413#	AFM196xg1/PCR	0.85	154	42
16p13.2-13.11	D16S414#	AFM205za11/PCR	0.61	137	42
16q12.1-q13	D16S415#	AFM205ze5/PCR	0.74	121	42
16q12.1-q13	D16S416#	AFM210yg3/PCR	0.43	53	42
16p13.11-p12.1	D16S417#	AFM220xb10/PCR	0.73	155	42
16p13.3-13.11	D16S418#	AFM225xd2/PCR	0.83	157	42
16q12.1-q13	D16S419#	AFM225zf2/PCR	0.77	149	42
16p13.11-p12.1	D16S420#	AFM238xb2/PCR	0.82	149	42
16q22.1	D16S421#	AFM240yh6/PCR	0.57	140	42
16q23.1-q24.2	D16S422#	AFM249xc5/PCR	0.80	166	42
16p13.3	D16S423#	AFM249yc5/PCR	0.75	136	42
16q23.2-q24.3	D16S449#	16AC51A4/PCR	0.85	134	43
16q22.2-q23.1	D16S450	16AC80H3/PCR	0.52	232	43
16q21	D16S451#	16AC69F12/PCR	0.84	145	43
16p13.11-p12.1	D16S452#	16AC33A4/PCR	0.68	143	43
16p13.3	D16S453	16AC301G12/PCR	0.45	335	43
16p13.2-p13.11	D16S454#	16AC45G5/PCR	0.75	134	43
16q22.2-q23.1	D16S522	16AC8.21/PCR	0.69	449	43
16p13.3	D16S523#	16AC13H1/PCR	0.68	149	43
16p13.11-p12.1	D16S524#	16AC40A7/PCR	0.76	157	43
16p13.3	D16S525#	16AC308G7/PCR	0.91	180	43
16q12.1-q13	D16S531#	16AC8.15/PCR	0.86	147	43
16p*	**	Hag1/HinfI	-	241	20
16p*	**	69A4/PstI	-	313	20
16q*	**	76BC6/MspI	-	490	20
16q*	**	LC9/MspI	-	243	20

- Markers typed on only 8 of the CEPH families.
- Markers typed by the candidate.
* - Physical location not refined.
** - Markers not previously incorporated into a published genetic linkage map of chromosome 16.

Genotypic data were obtained primarily from the 40 families in the CEPH reference panel. In addition, markers from contributing laboratory 20 were genotyped on 18 of the 25 families from the extended CEPH panel. All markers contributed by laboratory 42, and 16 markers from laboratory 43, were genotyped on 8 of the CEPH families only. The 8 families are: 102, 884, 1331, 1332, 1347, 1362, 1413, 1416.

Most of the loci (129 loci, 158 markers) included in the database have been physically mapped (Callen *et al.*, 1992; 1994; Shen *et al.*, 1994). The 29 loci not physically mapped are: *D16S53*, *D16S54*, *D16S59*, *D16S61*, *D16S66*, *D16S68*, *D16S70*, *D16S71*, *D16S72*, *D16S73*, *D16S74*, *D16S76*, *D16S77*, *D16S83*, *D16S84*, *D16S137*, *D16S155*, *D16S156*, *D16S158*, *D16S160*, *D16S246*, *D16S252*, *D16S282*, *D16S307*, *D16S309*, *69A4*, *76BC6*, *Hag1*, and *LC9*.

Most of the data included in this consortium map have been previously published in linkage maps of chromosome 16, and thus checked for genotype errors or inconsistencies. There were 10 new markers not previously incorporated in a genetic map: *D16S5*, *D16S80*, *D16S246*, *D16S282*, *D16S307*, *D16S309*, *Hag1*, *69A4*, *76BC6*, and *LC9*. Error checking was performed on all of the data as described bellow, and as far as possible, all potential errors in the database were identified.

The candidate performed the entire linkage analysis for the chromosome 16 consortium database, utilising the computer program CRI-MAP (which uses the Kosambi mapping function) on a Sun SPARC station IPC. To perform the preliminary error check, a comprehensive map was constructed, as described in Chapter 2.2. Closely spaced double recombinants were identified and flagged using the **chrompic** option of CRI-MAP. Changes made to the database are indicated in Table 8.4. The PCR markers used in the analysis have been rigorously checked, as described in Chapter 7; some apparent double recombinants were not altered as a result of regenotyping and were included in the analysis

Table 8.4: *Corrections made to the database*

Gene/Locus	CEPH Id	cM (Sex average)[a]	Coded[b]
CTRB1	1346-09,10	7.3	00
	28-07	9.9	00
D16S4	1347-06	4.7	00*
D16S7	12-03	5.5	00
	104-07	8.8	00
D16S44	1332-12	3.1	In PCR
	1345-04	8.6	In PCR
D16S48	1413-09	3.0	00
D16S60	1375-05,06	7.0	00
D16S64	1346-06	10.0	00
D16S67	1346-09	5.9	00
	1346-10	4.4	00
D16S79A	1346-10	4.4	00*
	102-14	4.4	00*
D16S83	21-03	6.7	00
	102-12,16	8.8	00
	1333-03	6.4	00
D16S84	884-13	4.2	00
	102-12	6.5	00
D16S137	1349-03	2.3	00
D16S152	13293-09	3.3	00
	1345-07	3.7	00
	1345-09	3.7	00
	1377-09	3.7	00
	28-07	6.0	00
D16S153	102-08	5.5	00
D16S150	1345-09	7.3	00
D16S159	102-08,10	4.3	00
D16S261	37-04	2.5	In PCR
D16S265	1345-05	4.8	In PCR

Table 8.4 (cont.): *Corrections made to the database*

Gene/Locus	CEPH Id	cM (Sex average)[a]	Coded[b]
D16S283	17-09	2.5	00**
	45-09	2.5	00**
D16S285	21-04	2.6	In PCR
D16S294	1344-05,07	3.5	In PCR
D16S299	1345-07	0.6	In PCR
D16S300	1333-all	--	00**
	28-09	8.7	In PCR
	1350-07	3.6	In PCR
D16S303		In PCR	
D16S308	884-09	10.0	In PCR
D16S309	21-08	7.9	00
	102-12	6.5	00
D16S310	104-08	6.4	In PCR
D16S359	1408-04	4.4	In PCR
D16S390	1344-07	15	In PCR
D16S450	1334-09	9.5	In PCR
D16S522	102-12	6.1	In PCR
	1349-03	3.5	In PCR
69A4	1340-07	5.7	00
	1346-10	5.7	00
	1349-05	5.7	00
	1375-13	5.7	00
	1377-07	8.0	00
	1377-09	7.4	00
	1416-09	5.2	00
	1420-04,05	6.3	00
	21-04,05	8.0	00
	2-08	8.7	00
76BC6	1418-03	6.5	00

Notes:

a. These represent the distance across the interval where the double recombination event occured; only those distances > 10 cM were considered.

b. Genotypes of PCR markers involved in a double recombination event in > 10 cM have been checked; if the code was unchanged, it was left in the analysis (Shen *et al*, 1994). All RFLP loci involved in a double recombination event in > 10 cM have been coded as unknown. In addition, all genotypes for 1416-10 are coded as unknown; suspect individual for laboratory 43.

* Autoradiograph checked, but not clear; genotype coded as unknown

** re-genotyped three times, genotypes still not clear; genotypes coded as unknown.

(Table 8.4). It is possible they are either real recombinant events, or have occurred as a result of an undetected new mutation to an allele already present in the family.

From this preliminary analysis, intralocus systems with no recombinations observed between them were considered as a haplotype, and the distance between them forced to zero in CRI-MAP. The systems were: *APRT*, *HBZ1*, *D16S7*, *D16S43*, *D16S45*, *D16S46*, *D16S49*, *D16S54*, *D16S56*, *D16S65*, *D16S67*, *D16S68*, *D16S75*, *D16S77*, *D16S85*, *D16S94*, *D16S150*, *D16S151*, *D16S152*, *D16S159*, and *D16S297*. *D16S79A* has two intralocus markers, *16AC66F3* and *36-1A*, but these were not treated as a haplotype as *D16S79A* has been found to be involved in a duplication with *D16S79B* (Phillips *et al.*, 1992; Callen *et al.*, 1992). These two RFLP loci hve been estimated to be approximately 1.8 Mb apart (Julie Nancarrow, personal communication). The four systems *MT2A*, *D16S4*, *D16S10*, and *D16S96* were reduced from eleven probe/enzyme systems by haplotyping by hand from Chapter 3, and 4.

Once all the data had been checked, a framework map was constructed, using the **build** option of CRI-MAP. For the purpose of this study, the analysis was divided into two parts - the short arm, 16p, and the long arm, 16q, with overlap either side of the centromere. The reason for this was that the 191 markers included in the analysis made it prohibitive to construct the map in one analysis. Any markers not previously localised on chromosome 16 were provisionally mapped to one arm using a combined framework map from Shen *et al* (1994) and a preliminary framework map constructed for the consortium, in conjunction with the **all** option of CRI-MAP. Once all markers had been localised to a specific arm of the chromosome, the building of the consortium 16p and 16q genetic maps were performed as two separate analyses. Six PCR formatted markers were used as overlapping markers; these were *D16S285*, *D16S304*, *D16S416*, *D16S415*, *D16S390*, and *D16S408* for the 16p analysis, and *D16S288*, *D16S298*, *D16S299*,

SPN, D16S383, and *D16S300* for the 16q analysis. *D16S300* and *D16S285* flank the centromere, and as no centromeric marker has been isolated for chromosome 16, the distance between these flanking markers (both included on the framework map) was included in the genetic length of 16p. The methods for generating a map for each arm were the same, and are described below.

For each arm, a pair of highly informative linked loci, separated by a distance of approximately 10 cM, were selected to be the foundation upon which to build a framework map. The loci on 16p were *D16S287* (with heterozygosity 78%) and *D16S296* (with heterozygosity 75%), with a maximum likelihood distance between them on a two point analysis of $\Theta = 0.18$ ($Z = 29.67$). On 16q, the two loci used as the foundation for the map were *D16S320* (with heterozygosity 86%) and *D16S398* (with heterozygosity 90%). These two loci were separated by a recombination fraction of $\Theta = 0.07$ cM with a maximum LOD score of $Z = 83.89$, as determined by a two point analysis.

These loci were used as the starting point for the **build** analysis of loci on 16p and on 16q. Subsequently, ranked markers were added to these two loci only if the odds of their placement exceeded 1000:1 in one position compared with the next most likely position, thus generating an initial "framework" map. Map validation by performing all adjacent two- and three-locus permutations to assess the support for local orders was carried out. Only orders that had likelihood support of 1000:1 or greater were retained. 82% of the markers included in the consortium map are physically mapped; the validity of this framework genetic map was tested by close examination of marker order on the cytogenetic-based physical map.

Using the **all** option of CRI-MAP, the additional 98 loci not placed on the framework map, were assigned to intervals between framework markers within which they are likely to lie with 1000:1 odds or greater. Comprehensive maps for each of the chromosome arms were also obtained by manually adding to each map

loci that could not be placed on the framework map with odds greater than 1000:1, and thus into the best position on the framework map.

A final error check to examine recombination events was performed using **chrompic**. Changes were made to the database as necessary, and the map was reconstructed as above. Table 8.4 summarises all changes made to the database, and also indicates double recombinants in microsatellite markers whose genotypes remained in the analysis after confirmation (Chapter 7; Shen *et al.*, 1994).

Once the framework maps were constructed and the remaining markers were located in their most likely interval, the placement of all loci was reexamined against the physical map. Discordant localisation was observed for the locus *D16S152*. This marker was initially placed on the genetic map below *D16S522* (see Fig. 8.1). The genetic map position (by way of forcing the physical map order and detecting apparent genotypes as potential double crossovers or errors) and the physical map position (by repeating the mapping to the somatic cell hybrid panel) were redetermined. The position on the physical map was checked and confirmed. It was found that genotypes for four individuals caused apparent double recombinants within a 10 cM interval (Table 8.4), and were the cause of misplacement on the genetic map. These were coded as unknown, using the criteria described above, and the map position redetermined. On this subsequent analysis, the location of *D16S152* shifted to the correct, predetermined physical interval.

Final maps were validated by running the **flips3** option and also by rerunning **build**, using different pairs of starting loci to initiate the map construction each time. The distances across the centromere, and surrounding intervals, were verified by performing an analysis involving the framework markers between *D16S287* to *D16S398*.

Evidence for sex-specific differences in recombination rates was investigated under two models: no sex difference (combined male and female recombination rates)

and variable sex difference, for the framework map, as described in Chapter 2.2.6. The sex difference with respect to recombination was also evaluated in each of the 59 intervals on the framework map separately, as described in Chapter 2.2.6.

8.4. Results

The final consortium framework map is shown in Fig. 8.1. Comprehensive loci that could not be placed on this framework map with odds of 1000:1 or greater have been assigned to intervals within which they are likely to lie with greater than 1000:1 odds, as indicated on Fig. 8.1. The physical localisation of each of the 158 markers is also shown on this figure. Also indicated on Fig. 8.1 are loci suitable to be classified as index markers, with heterozygosity > 70% and at intervals of approximately 10 cM on this map. There are some gaps greater than 10 cM and these will need to be targeted in the future.

Table 8.5a gives the distances between the framework markers on the sex-average, female and male maps, along with the local support for order, and with the chi-squared values for the investigation of significant differences in recombination between the sexes for each interval. The average heterozygosity of the 60 loci on the framework map was 68%, and 46 of these loci were PCR formatted.

The sex-average genetic length of the framework map is 165 cM, with an average distance between markers of 2.8 cM and a median distance between markers of 2 cM. The sex-specific length in males is 136 cM and in females is 197 cM. The maximum distance between any pair of loci on the sex-averaged framework map is 11.2 cM, between the loci *D16S404* and *D16S292*, and is the only gap larger than 10 cM.

Seventy eight PCR formatted markers were included in the analysis, with 43 of these being incorporated within the framework map. Of the 43 PCR formatted markers included on the framework map, 11 (with an average heterozygosity value

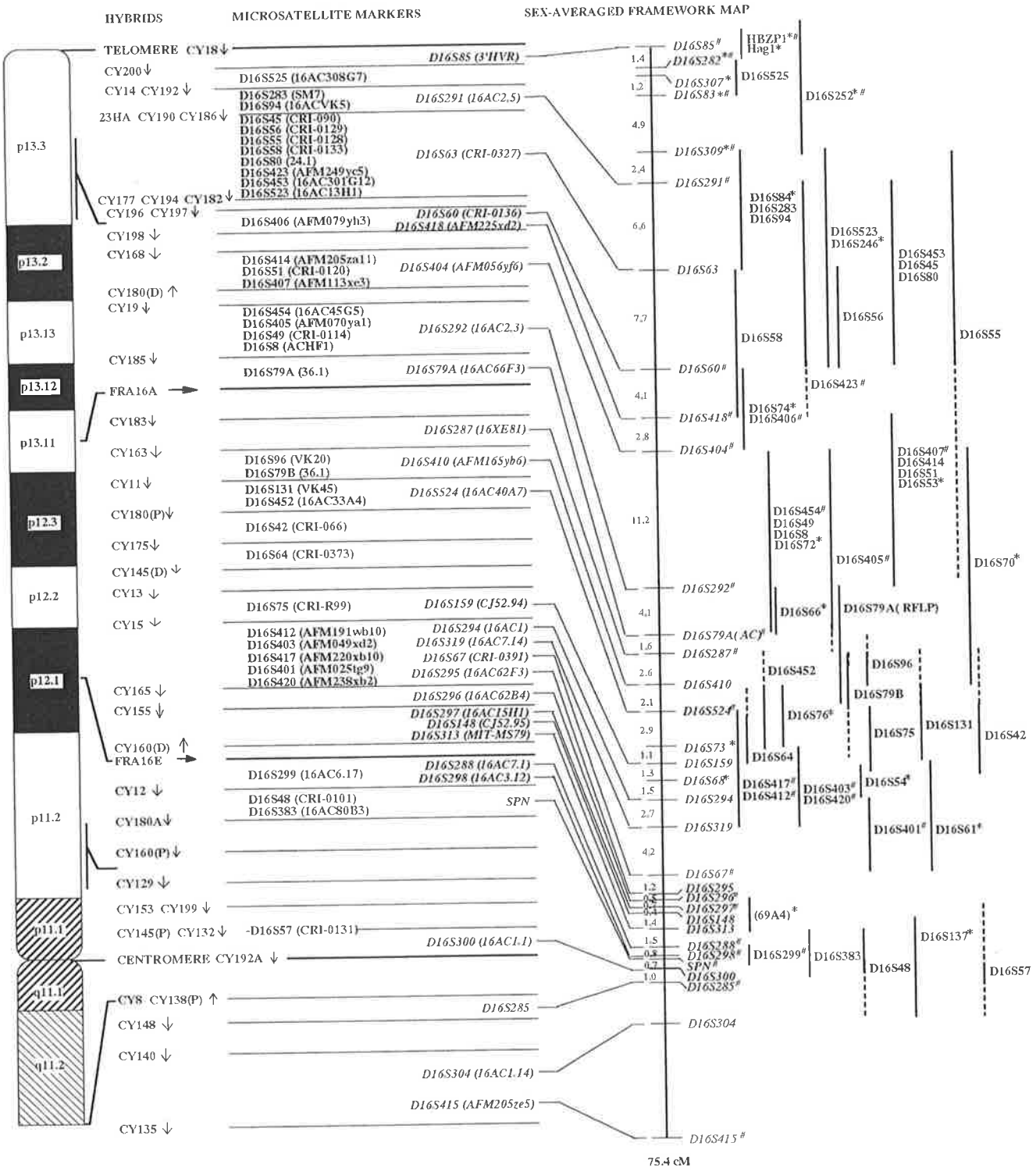

Figure 8.1: *Physical Map and Consortium Genetic Map of Human Chromosome 16p. The likely locations of the Comprehensive markers are indicated to the right of the Framework map.* * - *Physical localisation not refined.* # - *Markers with heterozygosity > 70%. Dashed lines indicate the reduced likely locations as determined by physical mapping using the hybrid panel*

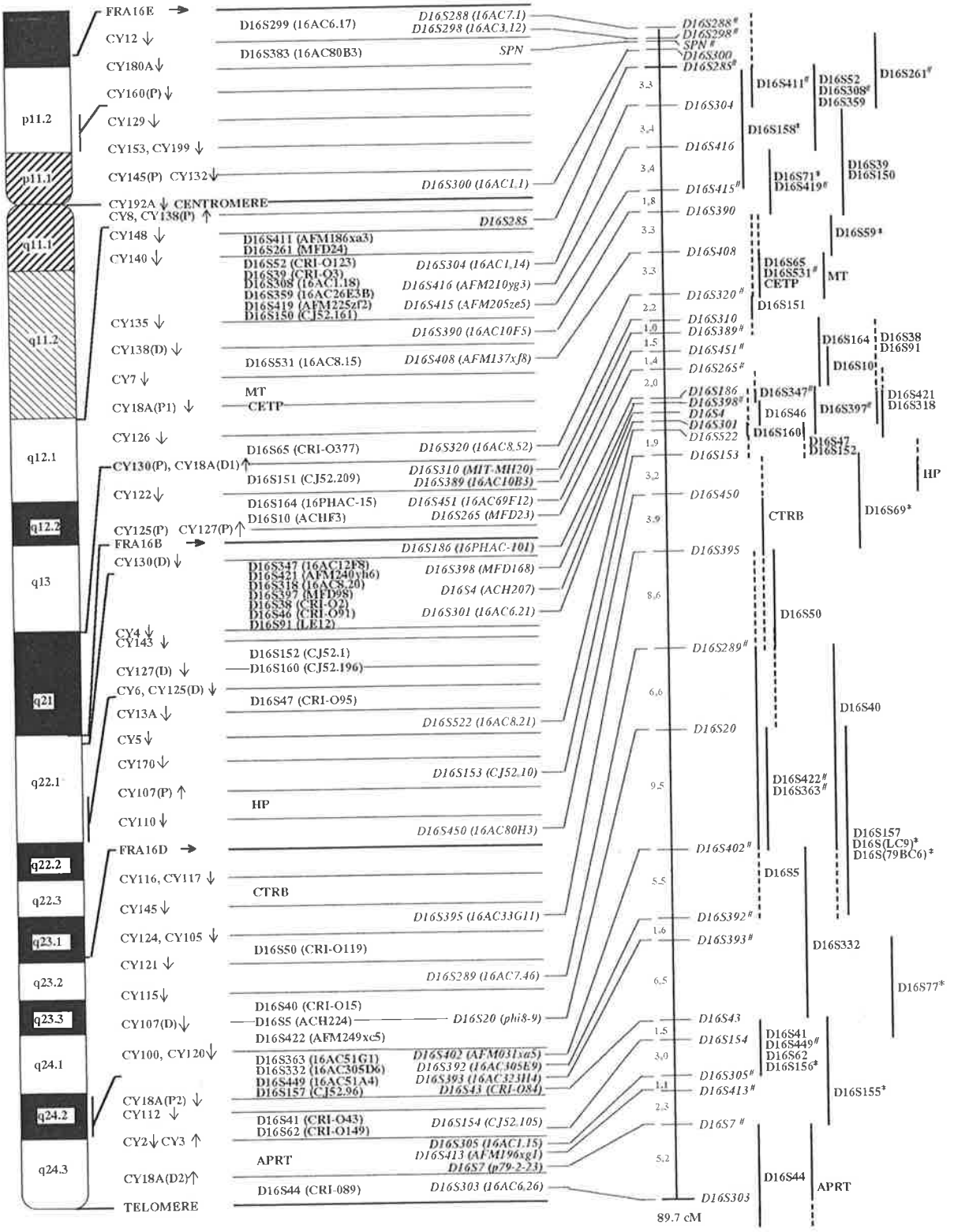

Figure 8.1: *Physical Map and Consortium Genetic Map of Human Chromosome 16q. The likely locations of the Comprehensive markers are indicated to the right of the Framework map. * - Physical localisation not refined. # - Markers with heterozygosity > 70%. Dashed lines indicate the reduced likely locations as determined by physical mapping using the hybrid panel*

Table 8.5a: *Map Distances (in Kosambi cM) for the 30 loci in the CEPH Consortium Framework Map. Included are the odds against inverting adjacent loci, and values for tests of significance in recombination differences between males and females.*

Locus	Sex-Ave	Female	Male	Odds[a]	Chi-Squared[b]
D16S85	1.4	0.7	2.3	10^{11}	2.8
D16S282	0.6	0.0	1.3	10^{4}	4.6
D16S307	1.2	0.1	2.1	10^{13}	8.3
D16S83	4.9	5.1	4.0	10^{17}	6.9
D16S309	2.4	2.2	2.6	10^{15}	0.0
D16S291#	6.6	9.1	3.2	10^{9}	8.3
D16S63	7.7	6.2	9.7	10^{22}	4.6
D16S60	4.1	1.7	6.5	10^{10}	3.7
D16S418#	2.8	4.9	1.3	10^{7}	0.5
D16S404#	11.2	8.1	13.9	10^{30}	1.4
D16S292#	4.1	4.2	4.0	10^{17}	1.4
D16S79(A)	1.6	0.7	2.2	10^{12}	1.8
*D16S287**	2.6	5.6	0.7	10^{10}	6.0
D16S410	2.1	3.9	0.7	10^{5}	7.4
D16S524#	2.9	4.7	0.9	10^{4}	6.9
D16S73	1.1	2.5	0.0	10^{3}	6.9
D16S159	1.3	1.5	1.0	10^{5}	0.0
D16S54	1.5	1.0	2.1	10^{3}	0.0
D16S294	2.7	4.1	0.8	10^{6}	2.3
D16S319	4.2	6.0	2.6	10^{30}	5.1
D16S67#	1.2	2.1	0.0	10^{9}	4.1
D16S295	0.5	0.0	1.1	10^{6}	2.3
*D16S296**	0.7	1.4	0.0	10^{15}	6.0
D16S297	0.4	0.9	0.0	10^{3}	4.1
D16S148	1.4	2.0	0.6	10^{14}	3.7
D16S313	1.5	2.8	0.3	10^{23}	7.4
D16S288	0.8	1.6	0.0	10^{14}	6.9
D16S298	0.2	0.5	0.0	10^{3}	2.3
SPN#	0.7	0.9	0.3	10^{7}	1.4
D16S300	1.0	2.2	0.0	10^{3}	8.7
D16S285#	3.3	6.7	0.3	10^{26}	24.8
D16S304	3.4	7.1	0.0	10^{4}	26.7

Table 8.5a (cont.):

Locus	Sex-Ave	Female	Male	Odds[a]	Chi-Squared[b]
D16S416	3.4	4.2	1.5	10^5	27.6
D16S415#	1.8	5.4	0.0	10^4	24.4
D16S390	3.3	6.2	0.5	10^8	24.4
D16S408	3.3	6.4	0.7	10^{11}	23.9
*D16S320#**	2.2	4.9	0.0	10^{17}	17.5
D16S310	1.0	1.9	0.0	10^3	3.7
D16S389	1.5	2.4	0.0	10^6	2.8
D16S451	1.4	0.4	2.6	10^3	0.5
D16S265	2.0	2.1	2.2	10^{17}	0.5
D16S186	0.5	0.9	0.0	10^5	2.3
*D16S398#**	0.7	0.6	0.6	10^4	0.0
D16S4	*0.6*	1.2	0.0	10^4	3.2
D16S301	0.7	0.5	1.1	104	0.5
D16S522	1.9	1.8	2.4	10^6	1.4
D16S153	3.2	0.4	4.2	10^{12}	0.9
D16S450	3.9	6.5	2.9	10^{12}	1.4
D16S395	8.6	12.0	4.0	10^{37}	7.4
D16S289#	6.6	3.2	10.5	10^{10}	0.9
D16S20	9.5	11.2	7.8	10^8	0.0
D16S402#	5.5	6.6	4.5	10^{13}	0.5
D16S392	1.6	3.5	0.0	10^4	4.6
D16S393#	6.5	4.2	8.9	10^{12}	3.2
D16S43	1.5	2.1	1.2	10^4	0.0
D16S154	3.0	1.8	4.3	10^5	2.3
D16S305#	1.1	2.6	0.0	10^3	0.0
D16S413	2.3	0.0	3.7	10^4	1.4
D16S7	*5.2*	3.4	7.8	10^9	1.8
D16S303#					
Total 16p (cM)	75.4	86.8	64.3		
Total 16q (cM)	89.7	110.1	71.8		
Total (cM)	165.1	196.9	136.1		

a Odds are calculated from the sex-averaged map.

b was calculated for each interval by keeping Θ constant and equal for males and females in three adjacent intervals on each side. Chi-squared = 3.84 (α = 0.05). If corrected for multiple testing by the conservative test of Fisher-Bonferroni (i.e. divide α by 59), the critical value for chi-squared is 11.2.

* indicates the starting loci for the final build run for the 16p and 16q maps.

\# markers incorproated into an "INDEX" map.

Table 8.5b: *Regional analysis of differences in recombination between males and females.*

Region	male	female	Chi-Squared (df)	*P*
16pter - D16S292	46.9 cM	38.1 cM	18.86 (10),	$P < 0.05^*$
D16S287 - cen - D16S451	14.1 cM	88.9 cM	214.8 (27),	$P < 0.05^*$
D16S393 - 16qter	25.9 cM	17.6 cM	7.8 (6),	$P > 0.05$

* - significant

of 76%) of them were typed on only 8 of the CEPH families (29%, 11/38, (38 of the 78 PCR loci typed on 8 families)), where as 80% (32/40, (40 of the 78 PCR loci typed on 40 families)) of the loci typed on the full forty families (with an average heterozygosity of 67%) were incorporated on the framework map. This implies that genotyping only a small subset of the reference families has placed these useful PCR markers at a disadvantage regarding being selected as framework markers, because they have a smaller number of crossover events with closely linked loci, making determination of the correct marker order more difficult.

The length of the sex-averaged comprehensive map was 179 cM, with an average resolution between loci of 1.1 cM. The length of the male and female maps was 147 cM and 212 cM respectively. The distances between loci and the local support for loci, based on the sex-average map, are given in Table 8.6.

The comprehensive map covers more than 99% of the entire length of the chromosome, extending from *D16S85*, which is within 170-430 Kb of pter (Wilkie *et al.*, 1991) to *D16S44*, which is within 230 kb of qter. The most distal 16q marker, genetically and physically is *D16S303*, being closed to 16qter than *D16S44* (P. Harris, personal communication). No centromeric marker has been isolated for chromosome 16, so it is difficult to determine the exact distance of the centromere to each of its flanking markers, *D16S300* and *D16S285*. The distance between these two markers on the sex-average framework map is 1.0 cM.

For the short arm, 16p, there were 97 polymorphisms incorporated in the analysis, which were reduced to 81 loci after haplotyping. The framework map for 16p consists of 30 loci, and extends 75 cM on the sex-average map. The mean distance between loci is 2.5 cM and the median distance between loci is 1.5 cM. The length of the female map is 87 cM and the length of the male map is 64 cM. When comparing the sex-specific results with Morton's predictions (1991b), there is quite good correlation. For 16p, Morton predicted that the male map length was 60 cM and female map length was 96 cM. The female map is approximately

Table 8.6: *Comprehensive consortium map. The distance between loci is given in cM, and the odds for local support for order were determined using the sex-average map.*

Locus	Male	Female	Sex-Average	Odds
HBZ1	0.0	0.0	0.0	1.0
Hag1	0.0	0.0	0.0	1.0
D16S85	2.3	0.7	1.5	10^{20}
D16S282	1.3	0.0	0.6	10^{6}
D16S307	0.9	0.0	0.3	10^{3}
D16S525	1.4	0.7	1.1	10^{14}
D16S83	0.7	4.1	1.3	1.1
D16S252	2.7	0.0	2.7	10^{8}
D16S309	0.8	0.9	0.8	10^{6}
D16S84	2.5	1.4	2.0	10^{4}
D16S291	1.3	1.1	1.3	10^{6}
D16S283	0.0	0.9	0.4	10^{2}
D16S94	0.7	3.7	1.8	10^{5}
D16S246	4.6	4.1	4.7	10^{4}
D16S523	0.0	0.0	0.0	1.0
D16S453	0.0	0.0	0.0	1.3
D16S80	1.3	0.9	1.0	10^{3}
D16S63	0.0	0.0	0.0	1.0
D16S45	2.6	2.9	2.8	10^{17}
D16S56	3.5	2.7	2.9	10^{13}
D16S58	0.0	0.0	0.0	1.0
D16S423	0.0	0.6	0.5	1.0
D16S55	0.0	0.0	0.0	1.0
D16S60	3.5	0.7	1.9	10^{8}
D16S74	3.4	0.0	1.7	1.4
D16S418	0.0	1.7	0.0	1.0
D16S406	1.6	5.8	4.3	10^{8}
D16S404	0.0	0.0	0.0	1.0
D16S407	2.4	0.0	1.6	10^{4}
D16S51	0.9	1.1	1.0	10^{5}
D16S53	0.0	0.0	0.0	1.0
D16S414	0.0	0.0	0.0	1.0

Table 8.6 (cont.): *Comprehensive consortium map.*

Locus	Male	Female	Sex-Average	Odds
D16S66	10.3	6.1	8.2	10^{17}
D16S72	1.6	0.0	0.7	10^{3}
D16S49	0.8	0.0	0.3	1.2
D16S405	0.0	0.0	0.0	1.0
D16S454	0.0	0.0	0.0	1.0
D16S292	0.9	0.0	0.4	10^{1}
D16S8	2.8	3.7	3.5	10^{10}
D16S79A	0.4	0.0	0.0	1.0
D16S79AC	2.5	0.6	1.8	10^{18}
D16S287	0.5	0.0	0.3	1.7
D16S96	0.0	3.4	1.4	10^{2}
D16S79B	0.8	0.9	0.9	1.2
D16S410	0.8	2.2	1.5	10^{5}
D16S76	0.0	1.4	0.6	10^{2}
D16S452	0.0	0.0	0.0	1.0
D16S524	0.7	2.4	1.6	10^{5}
D16S131	0.0	0.0	0.0	2.1
D16S42	0.0	1.3	0.7	10^{2}
D16S73	0.0	0.0	0.0	1.7
D16S64	0.0	2.2	1.2	10^{12}
D16S75	0.0	1.8	1.0	10^{3}
D16S159	1.3	0.0	0.0	1.0
D16S412	0.0	2.3	1.5	6.3
D16S54	0.0	0.0	0.0	1.0
D16S417	0.0	0.0	0.0	1.0
D16S403	0.0	0.0	0.0	1.0
D16S61	0.0	0.0	0.0	1.0
D16S68	1.5	0.5	0.9	10^{6}
D16S294	0.0	0.7	0.5	10^{2}
D16S420	0.0	3.4	2.0	10^{9}
D16S401	2.1	0.0	1.0	10^{2}
D16S319	2.3	5.3	3.8	10^{31}
D16S67	0.0	2.0	1.1	10^{9}

Table 8.6 (cont.): *Comprehensive consortium map.*

Locus	Male	Female	Sex-Average	Odds
D16S295	1.2	0.0	0.5	10^6
D16S296	0.0	1.4	0.7	10^4
69A4	0.0	0.0	0.0	1.0
D16S297	0.3	2.2	1.1	10^4
D16S148	0.3	1.2	0.9	10^4
pKKA22	0.3	1.7	1.0	10^3
D16S313	0.3	2.3	1.3	10^{15}
D16S288	0.3	1.5	0.8	10^{10}
D16S299	0.3	0.0	0.3	1.3
D16S298	0.0	0.9	0.3	10^4
SPN	0.3	0.0	0.0	1.0
D16S383	0.0	0.0	0.0	1.0
D16S48	0.0	0.7	0.4	1.0
D16S57	0.0	0.1	0.3	1.0
D16S300	0.0	2.2	1.1	10^9
D16S285	0.1	0.0	0.0	1.0
D16S411	0.5	0.8	0.5	10^3
D16S261	0.6	7.1	3.8	10^8
D16S158	0.0	0.0	0.0	1.0
D16S52	0.0	0.0	0.0	1.0
D16S304	0.0	0.0	0.0	1.0
D16S308	0.6	0.0	0.4	1.4
D16S359	0.0	3.3	1.9	10^8
D16S416	2.6	2.3	1.8	1.7
D16S150	0.0	3.4	2.0	8.9
D16S415	0.0	0.0	0.0	1.1
D16S419	0.0	1.1	0.5	10^2
D16S39	0.0	3.8	1.0	10^1
D16S71	1.7	3.5	3.4	10^2
D16S390	1.7	1.6	1.2	10^3
D16S59	0.0	1.8	1.0	10^7
D16S408	0.0	0.0	0.0	1.0
D16S531	0.0	9.1	4.8	10^7

Table 8.6 (cont.): *Comprehensive consortium map.*

Locus	Male	Female	Sex-Average	Odds
MT	0.0	0.0	0.0	1.0
D16S65	0.8	0.0	1.2	1.6
CETP	0.0	5.4	2.0	10^6
D16S320	0.4	1.8	1.2	10^3
D16S151	0.3	2.9	1.4	10^3
D16S310	0.0	1.3	0.7	10^3
D16S389	0.0	2.0	1.3	10^7
D16S451	2.5	1.5	1.7	4.2
D16S164	0.0	0.0	0.0	1.0
D16S265	0.0	0.0	0.0	1.0
D16S10	2.0	1.6	1.7	10^{17}
D16S186	0.0	0.0	0.0	1.0
D16S347	0.0	0.8	0.5	10^5
D16S398	0.0	0.0	0.0	1.0
D16S397	0.0	0.0	0.0	1.0
D16S421	0.0	0.0	0.0	1.0
D16S318	0.5	0.5	0.6	10^5
D16S46	0.0	0.0	0.0	1.0
D16S38	0.0	0.0	0.0	1.0
D16S91	0.0	0.0	0.0	1.0
D16S4	0.0	1.2	0.7	10^9
D16S301	0.0	1.6	0.7	1.6
D16S160	0.5	0.0	0.1	1.0
D16S152	0.0	0.6	0.3	10^2
D16S47	0.8	0.0	0.3	2.2
D16S522	2.3	0.0	1.4	10^8
D16S153	0.6	1.1	0.7	10^2
HP	4.1	0.1	2.1	10^{15}
D16S450	0.0	0.0	0.0	1.0
D16S69	3.6	6.8	4.8	10^{10}
CTRB	0.0	2.2	1.6	10^1
D16S395	3.3	10.9	5.3	10^{10}
D16S50	3.3	0.0	3.4	10^{39}

Table 8.6 (cont.): *Comprehensive consortium map.*

Locus	Male	Female	Sex-Average	Odds
D16S289	5.8	4.1	5.7	10^{11}
D16S20	0.0	6.7	1.2	1.1
D16S5	2.4	1.5	3.0	1.8
D16S40	1.3	0.6	1.5	2.4
D16S422	0.0	0.0	0.0	1.0
D16S363	0.0	1.5	0.0	1.0
76BC6	0.0	1.0	0.5	1.1
LC9	2.7	0.5	2.3	3.6
D16S402	0.1	0.7	0.6	1.0
D16S157	6.4	8.8	7.3	10^{6}
D16S392	0.0	4.1	2.1	10^{2}
D16S332	0.0	0.0	0.0	1.0
D16S393	8.4	5.7	7.7	10^{19}
D16S449	2.2	0.0	0.0	1.0
D16S77	0.0	0.0	0.0	1.0
D16S43	2.2	1.4	2.0	10^{4}
D16S155	0.0	0.0	0.0	1.0
D16S41	0.0	0.0	0.0	1.0
D16S156	0.0	0.0	0.0	1.0
D16S154	1.5	0.0	0.0	1.0
D16S62	2.0	0.9	2.0	10^{3}
D16S305	0.0	2.1	1.3	10^{3}
D16S413	3.5	0.0	1.8	10^{6}
D16S7	3.2	0.0	1.8	10^{6}
APRT	2.7	3.1	2.7	10^{5}
D16S44	0.0	0.0	0.0	1.0
D16S303				
Total	147.4 cM	211.8 cM	179.3 cM	

10% shorter than predicted, and the male map is approximately 10% longer (Table 8.7).

For chromosome 16q, 94 polymorphisms reduced to 77 loci after haplotyping, were incorporated into the analysis. The framework map for 16q consists of 30 loci, and the sex-average map extends 90 cM. The average distance between loci is 3 cM and the median distance between loci is 2.3 cM. On the sex-specific map, the length of the female map is 110 cM and on the male map, is 72 cM. Morton predicted, for 16q, that the male map length was 56 cM and the female map length was 90 cM. These distances are actually less than those he predicted for the short arm, hence they are not in such good agreement as for 16p. Both Morton's estimates of the length of 16p and 16q are approximately 20% shorter than obtained here (Table 8.7).

Table 8.7: *Comparison of lengths (in cM) between the chiasma map (Morton, 1991b) and the consortium framework map.*

		Male	Female
Morton (1991b)	*16p*	60	96
	16q	56	90
Current	*16p*	64	87
	16q	72	110

Morton estimated that the short arm was 39 Mb, or 40% of the length of the entire 98 Mb chromosome. The male framework map for 16p is 47% of the total, and the female map for 16p is 44% of the total length. Morton also estimated that the long arm consisted of 59 Mb, or 60% of the entire chromosome. The genetic maps for 16q for males and females represent 53% and 56% respectively. These results appear to correlate with the physical lengths of the chromosome quite well. Scrutiny of the data for both the framework map and the comprehensive map, reveals that in some intervals, the sex-average estimate is either larger than both

the sex-specific estimates, or less than both the sex-specific estimates. CRI-MAP selects the phase with the highest likelihood. In some cases, the phase chosen for the sex-average calculations differs from the phase chosen for the sex-specific calculations. Due to this difference in phase choice, instances where the sex-average map the distance between two loci was less than or greater than the distance on the male and female maps were detected on the frameworkl map between *D16S265* and *D16S186*; and *D16S398* and *D16S4* (for detail see Table 8.5a). For this consortium map we identified one pedigree (family 17) where this phenomenon occurred, with the markers *D16S265*, *D16S186*, *D16S398*, and *D16S4*. There were 10 incidences where this phenomenon occurred on the comprehensive map, which were between: *D16S246* and *D16S523*, *D16S57* and *D16S300*, *D16S416* and *D16S150*, *D16S390* and *D16S59*, *D16S65* and *CETP*, *D16S318* and *D16S46*, *D16S50* and *D16S289*, *D16S289* and *D16S20*, *D16S5* and *D16S40*, and *D16S40* and *D16S422* (Table 8.6). There were six pedigrees where a different phase was selected in the sex-average calculations to that chosen for the sex-specific calculations. The families where this occurred were 1330, 1334, 1344, 1359, 1377, 13291. This phenomenon has been previously noted in two consortium maps 13 and 15q (Bowcock *et al.*, 1992; 1993)

Sex-specific differences in the recombination frequency in males and females were investigated for each of the maps constructed (16p and 16q). As a general trend, greater recombination is observed in female meioses compared with male meioses, which is significant (Table 8.5b). Over all, the female/male ratio of map distances for the consortium map, averaged 1.5 across the entire length of chromosome 16. The sex difference with respect to recombination was evaluated in each of the 59 intervals on the framework map separately. From these tests, seven single intervals were shown to exhibit a significant difference in the recombination rates between males and females (Table 8.5a).

The region encompassing the centromere on chromosome 16 shows a greatly expanded distance on the female map compared with that in males due to excess female recombination. From the region *D16S287* to *D16S451* female map length is 88.9 cM while the male map length is 14.1 cM, representing a 6.3 fold excess of female recombination. For 27 degrees of freedom, the chi-squared test is significant if chi-squared > 40.1. Thus, as chi-squared = 214.8 for this region, this six-fold difference in the recombination frequency is significant (See Table 8.5b). In contrast to this trend, the female/male ratio of recombination frequency is reversed in the two distal regions of the chromosome. Between *D16S85* and *D16S307*, at 16pter, there is a 1.6 fold excess in male recombination, which is significant. In the interval between *D16S393* and *D16S303*, at 16qter, there is a 1.8 fold excess of male recombination compared with female recombination was observed, which is not statistically significant (Table 8.5b). The expansion of the subtelomeric regions in male maps has also been documented for the consortium map of chromosome 13, and for linkage maps of other chromosomes; 9, 10, 11, 12, 15, 17, 19 (See O'Connell *et al.*, 1989 for discussion).

Table 8.4 outlines the double recombinants that were detected during the construction of this map. All of the double recombinant events were coded as unknown with the exception of some of the double recombinants detected at PCR loci, which remain in the analysis (Shen *et al.*, 1994).

Many of the markers within the consortium map have been assigned to a cytogenetic interval, thus enabling the consortium map to be anchored to the high resolution cytogenetic map, which is divided into 66 breakpoint intervals (Fig. 2). Comparison of the genetic and physical maps during their development has been invaluable for validation of marker order. There was one example, the locus *D16S152*, where the importance of the development of the genetic map and the cytogenetic map concurrently became evident. The location of this marker relative to adjacent markers differed on the cytogenetic map compared with the genetic

map. This conflict was resolved by checking both the genetic map location and the physical map location, and it was discovered that four erroneous genotypes were responsible for a genetic mislocalisation. Coding these genotypes as unknown enabled agreement between the physical map and the genetic map (see Fig. 8.1).

The combination of the genetic map and the cytogenetic map can be extremely beneficial in refining locus order at the level of the comprehensive map. For example, the genetic location of the marker CTRB is between *D16S153* and *D16S289* (Fig. 8.1). However, the physical location is between breakpoints CY116/CY115 and CY145, which narrows the genetic localisation to between *D16S450* and *D16S395*. Localisation can be refined in this way for 29 comprehensive markers, or 30% of the markers not placed on the framework map (See Figure 8.1).

8.5. Discussion

The CEPH consortium framework map for chromosome 16 has now been constructed. The sex-average length of the framework map, 165 cM, is about 10% longer than predicted by chiasma studies (145 cM; Morton, 1991b). The sex-specific maps are also approximately 10% longer in this analysis, than predicted independently by chiasma studies. The difference found between the length of the map determined by chiasma counts and the genetic distance determined in the present study could be due to (if the chiasma map is the standard) a small number of undetectable errors. Chiasma studies involve cytogenetic counts of recombination events, and by nature assume interference and also underestimate the true recombination fraction due to conservative scoring (Morton, 1991b). Thus, the predicted length of chromosome 16 by counts of chiasmata, of 145 cM, may indeed be an underestimate of the chromosomes true genetic length. Conversely, genetic map distances could be inflated and the

occurrence of multiple recombination events could be exaggerated by neglecting interference (Morton *et al.*, 1986) during construction of the genetic map. This implies that the apparent 10% discrepancy between these two types of maps may, in reality, be less.

In addition to the framework map, a comprehensive map of 158 loci was also constructed. The length of the sex-averaged comprehensive map was 179 cM, with an average resolution between loci of 1.1 cM. The length of the male and female maps was 147 cM and 212 cM respectively. The distances between loci, and the local support for loci, based on the sex-average map are given in Table 8.6.

As a general trend, greater recombination is observed in females on the consortium map. This observation has been previously reported (Donis-Keller, *et al.*, 1987; Julier *et al.*, 1990; Keith *et al.*, 1990; NIH/CEPH collaborative mapping group, 1992; Kozman *et al.*, 1993). Donis-Keller *et al.* (1987) observed that the genetic length of the autosomes is approximately 90% longer in females than in males. However, more recently, maps have shown that a smaller ratio exist between the sexes (NIH/CEPH collaborative mapping group, 1992), where the estimated female/male ratio of map distances was 1.6; the estimate of 1.5 reported in this chapter does not deviate significantly from this. In contrast, greater recombination frequency in males is observed in the two distal regions of the chromosome; from 16pter, (*D16S85* to *D16S307*) there is a 1.6 fold excess in male recombination which is significant, and in the interval between *D16S393* to *D16S303* at the long arm telomere, and there is a 1.8 fold excess of male recombination, which is not significant (Table 8.5b).

The region encompassing the centromere on chromosome 16 shows a significant increase in female map length over the length of the male map. From the region *D16S287* to *D16S395* female map length is 89 cM while the male map length is 14 cM representing a 6.3 fold excess of female recombination (Table 8.5b). Overall

across this region encompassing the centromere, there appears to be a suppression of recombination, where the average resolution on the sex-average map, between loci in the interval from *D16S67* to *D16S395* is 1.6 cM, as opposed to the resolution of the entire map of 2.8 cM. The reverse is true at the terminal ends of the chromosome, where the average resolution in the intervals investigated above were 3.9 cM at 16pter, and 2.8 cM at 16qter. These observations are supported by those made by Morton *et al.*, 1977, where they based their assumptions on:

"1.) the centromere suppresses crossing-over in its vicinity, and

2.) there is an excess of crossing-over at the termini of arms and a deficiency in the subterminal regions."

Each specific interval on the framework map was investigated for evidence of a significant difference in the rate of recombination between males and females; no interval was shown to exhibit a significant difference (Table 8.5a).

Beutow (1991) estimated that an error rate of 0.5% will increase the map length by 1 cM per interval, and an error rate of 1.0% will increase the map length of 2 cM per interval. From an analysis of chiasma counts, chromosome 16 is expected to span a distance of 145 cM (Morton, 1991b). If there is an error rate in the database of 0.5%, then we expect that the length of the map containing 59 intervals, would be 204 cM. Since the sex-average length of the map is only 165 cM then an approximate error rate of 0.17% is estimated to be present in the consortium database using the chiasma map as the standard. As stated above, the assumption of no interference may also have inflated the map length, and hence, a smaller error rate than 0.17% may indeed exist for this data. Loci that were coded as unknown as a result of error checking are listed in Table 8.4.

The first linkage map spanning the length of the chromosome was described by Donis-Keller *et al.*, in 1987. This map was developed further in 1990 (Julier *et al.*; Keith *et al.*), and then integrated with the cytogenetic based physical map in 1993 (Kozman *et al.*). The male map reported by Donis-Keller *et al.*, (1987) was

164 cM long. The female map length was 237 cM. All the loci incorporated into that map have been used in the consortium map. Julier *et al.*, (1990) published a map of chromosome 16 of 24 polymorphic loci, all of which are incorporated on the consortium map, with the exception of *D16S49* and the protein polymorphism for Haptoglobin, *HP*. The male map length was 187 cM, and the female map length was 226 cM.

A map of 46 DNA polymorphisms was described by Keith *et al* (1990), all of which are included in the consortium map. The distances of the male and female maps were 115 cM, and 193 cM respectively. The map by Kozman *et al.*, (1993), incorporates 67 polymorphic loci defining 50 loci, had been localised on the cytogenetic based physical map. The map extended the same length as the present map (from *D16S85* to *D16S44*). It has since been determined that the marker *D16S303* is closer to the qter than *D16S44* (P Harris, personal communication). On this map the length of the framework sex-average map was 162 cM; while on the female and male maps the length was 199 cM and 133 cM respectively. All of the markers incorporated on this map are incorporated into the consortium map. Comparisons between intervals on each of these maps, where possible, show that the lengths between loci on each of these maps are in remarkable agreement.

The 80 locus map of microsatellite markers for chromosome 16 (Shen *et al.*, 1994), has been integrated into the consortium map. This map of chromosome 16 has a sex-average framework map length of 152 cM; 13 cM shorter than the consortium map. The female map has a length of 181 cM and the male map has a length of 131 cM. The order of the PCR formatted markers common to both these framework maps is identical and in general the distance between the markers is very similar, even when there are intervening RFLP loci. However, there were several intervals where the difference between the map intervals was greater than 5 cM. This occurred in the following intervals:

1.

PCR				10.8 cM				
	D16S291	-	(*D16S63*	-	*D16S60*)	-	*D16S418*	
CONSORTIUM		6.6 cM		7.7 cM		4.1 cM		

2.

PCR		10.5 cM				
	D16S289	-	(*D16S20*)	-	*D16S402*	
CONSORTIUM		6.6 cM		9.5 cM		

3.

PCR				3.8 cM				
	D16S393	-	(*D16S43*	-	*D16S154*)	-	*D16S305*	
CONSORTIUM		6.5 cM		1.5 cM		3.0 cM		

The difference in length is possibly due to undetected errors in the intervening RFLPs.

The order of loci on the framework and comprehensive maps are congruent with the cytogenetic-based physical map. The majority of loci on the framework map were independently ordered on the genetic and physical maps. There were several instances where the order of loci was determined by linkage alone (as the loci were not separated by a hybrid cell line), thus resolving the order of loci not resolved on the physical map. At the level of the comprehensive map, the order of 30 loci not resolved on the genetic map alone were resolved by consideration of physical and genetic mapping data. In some instances, the position of the locus could be refined to a specific interval between adjacent markers. The loci for which this is true are: *APRT*, *CTRB*, *CETP*, *HP*, *D16S47*, *D16S50*, *D16S65*, *D16S96*, *D16S151*, *D16S152*, *D16S160*, *D16S347*, *D16S363*, *D16S411*, *D16S422*, and *D16S531*. This demonstrates the value of combining different mapping techniques to confirm and refine locus order.

The relationship between the physical map and the genetic map becomes clearer as the resolution on the genetic map approaches 1 cM. Although on average one centiMorgan is approximately equivalent to one megabase, this relationship varies

across the genome, within chromosomes, and varies with sex. Recombination is suppressed near the centromere and increased near the telomeres. Recombination in males appears to be decreased near the centromere and is in some cases greater than in the females in the telomeric regions (Skolnick, 1991).

The aim of the Human Genome project has been to develop linkage maps with average resolution of 2-5 cM, using highly informative PCR-formatted loci, with an average heterozygosity greater than 70% (Roberts, 1992). The framework consortium map presented here comprises 49 PCR formatted loci and has an average resolution of 2.8 cM, which is within an objective of the genome project. In addition to this high resolution linkage map, which will be very useful for the localisation of disease genes to this chromosome, a comprehensive map has been constructed with an interlocus resolution of 1.1 cM. The value of this dense background map of genetic markers will be that the number of informative meioses needed to establish linkage will be more easily attained. Many genes responsible for human hereditary diseases can now be localised to small chromosomal regions to facilitate the task of isolating disease genes.

CHAPTER 9

CONCLUDING REMARKS

The aims of this thesis as set out in Chapter 1.8, have been attained. Genetic linkage maps of chromosome 16 with increasing resolution have been constructed and described. The ultimate genetic map is now the CEPH consortium map, covering 99% of the chromosome, with an average interlocus resolution of 2.8 cM on the framework map. A summary of the evolution of the linkage map of chromosome 16, over the past 4 years is described in Table 9.1.

Table 9.1: *The development of the genetic linkage map of regions of human chromosome 16, and of the entire chromosome, by this project.*

Chapter	No. of Markers	No. of STRs	Chromosome Coverage	Sex-Average Length (cM)	% error[a]
3	18	0	16q12.1-16q23.1	22.1#	-
4	16	2	16p13.3-16p12.1	43.2*	-
5	27	21	16p13.11-16q11.2	25.8*	-
6	68	6	16pter-16qter	165#	0.20
7	82	80	16pter-16qter	152#	0.03
8	191	80	16pter-16qter	165#	0.17

framework map
* comprehensive map
a where estimated

The discovery of restriction fragment length polymorphisms (RFLPs) (Botstein *et al.*, 1980) brought with it a proliferation of polymorphic genetic markers suitable for the construction of human genetic maps. Using these markers, multipoint linkage maps of chromosome 16 were reported (Donis-Keller *et al.*, 1987; Julier *et al.*, 1990; Keith *et al.*, 1990; Kozman *et al.*, 1991 (Chapter 3)).

Since that time, a new generation of highly polymorphic genetic markers, based on di-, tri-, and tetra- nucleotide repeat sequences of the type $(AC)_n$ have been isolated and characterised (Weber and May, 1989; Litt and Luty, 1989). These have streamlined the construction of *high resolution* genetic linkage maps. These

markers are PCR formatted, thus more easily typed than RFLPs. They were found to be on average more polymorphic than RFLPs, and so the probability of being informative for linkage studies, including detection of linkage and defining marker order, is increased.

Maps containing PCR markers have been constructed (Kozman *et al.*, 1992 (Chapter 4); Mitchison *et al.*, 1993 (Shutter 5); NIH/CEPH collaborative mapping group, 1992 (Chapter 6), including a high resolution PCR-based genetic map (Chapter 7). The CEPH consortium map of chromosome 16 integrated the RFLP and PCR maps (Chapter 8) described above. Gyapay *et al.* (1994) recently produced an independent map of the human genome incorporating more than 2,000 polymorphic markers, with an average spacing of 2.9 cM. The map of chromosome 16 generated by Gyapay *et al.* (1994), when fully integrated with the PCR map presented in Chapter 7 will represent a powerful tool for genetic mapping for this chromosome. The markers of Gyapay *et al.* (1994) which had been included in the earlier Weissenbach *et al.* (1992) are already integrated and have been included in the maps presented in Chapter 7 and 8. These PCR formatted markers represent STS's, which in most cases have been physically ordered along the chromosome, are useful for developing the physical map, and allow easy integration of the genetic and cytogenetic-based physical map.

The power of a genetic map increases with the density of its markers and the accuracy with which they can be placed. These maps are of sufficient density to identify the map location of most Mendelian phenotypes to within a few centiMorgans (Schmitt and Goodfellow, 1994). The choice of markers for genome screening is greater, and once linkage is detected, numerous additional markers from the region may be useful for precise regional localisation.

The availability of a complete, high-resolution linkage map of human chromosome 16 will provide close flanking markers for use as starting points to initiate the cloning of the disease gene and enable the characterisation of the gene responsible

at the molecular level. This will ultimately allowing afflicted families to benefit from more accurate prenatal diagnosis, and possible gene therapy.

Finally, linkage maps of high density and constructed with highly polymorphic markers will allow more detailed study of phenomena involving recombination, sex-specific recombination differences, chromosomal region-specific differences, chiasma interference, and the non-linear relationship between genetic and physical distances (Chakravarti, 1991).

References

Aksentijevich, I., Pras, E., Gruberg, L., Shen, Y., Holman, K., Helling, S., Prosen, L., Sutherland, G.R., Richards, R.I., Ramsburg, M., Dean, M., Pras, M., Amos, C., and D.L. Kastner. 1993. Refined Mapping of the gene causing Familial Mediterranean Fever by linkage and homozygosity studies. *Am. J. Hum. Genet.* **53**:451-461.

Anderson, L. and Sandberg, K. 1984. Genetic linkage in the horse. II Distribution of male recombination estimates and the influence of age, brees and sex on recombination frequency. *Genetics* **106**:109-122.

Armour, J.A., Wong, Z., Wilson, V., Royle, N.J., and Jeffreys, A.J. 1989. Sequences flanking repeat arrays. *Nucl. Acid. Res.* **17**:4925-4935.

Armour, J.A., Crosier, M., and Jeffreys, A.J. 1992. Human minisatellite alleles detectable only after PCR amplification. *Genomics* **12**:116-124.

Attwood, J., Chiano, M., Collins, A., Donis-Keller, H., Dracopoli, N., Fountain, J., Falk, C., Goudie, D., Gusella, J., Haines, J., Armour, J.A.L., Jefferys, A.J., Kwiatkowski, D., Lathrop, M., Matise, T., Northrup, H., Pericak-Vance, M.A., Phillips, J., Retief, A., Robson, E., Shields, D., Slaugenhaupt, S., Vergnaud, G., Weber, J., Weissenbach, J., White, R., Yates, J., and Povey, S. 1994. CEPH consortium map of chromosome 9. *Genomics* **19**:203-214.

Baker, E., Guo, X.-H., Orsbsorn, A.M., Sutherland, G.R., Callen, D.F., Hopwood, J.L. and Morris, P.C. 1993. The Morquio A syndrome (Mucopolysaccharidosis IVA) gene maps to 16q24.3. *Am J Hum Genet* **52**:96-98.

Banchs, I., Bosch, J., Guimerà, J., Làzaro, C., Puig, A., and Estivill, X. 1994. New alleles at microsatellite loci in CEPH families mainly arise from somatic mutations in the lymphoblastoid cell lines. *Hum. Mut.* **3**:365-372.

Barnard, G.A. 1949. Statistical Inference. *J. R. Statist. Soc.* **11**:115-139.

Beckmann, J.S., Tomfohrde, J., Barnes, R.I., Williams, M., Broux, O., Richard, I., Weissenbach, J., and Bowcock, A.M. 1993. A linkage map of human chromosome 15 with an average resolution of 2 cM and containing 55 polymorphic microsatellites. *Hum Mol Genet* **2**:2019-2030.

Beggs, A.H., Phillips, H.A., Kozman, H.M., Mulley, J.C., Wilton, S.D., Kunkel, L.M., and Laing, N.G. 1992. A (CA)n repeat polymorphism for the human skeletal muscle α-actinin gene *ACTN2* and its localisation on the linkage map of chromosome 1. *Genomics* **13**:1314-1315.

Birnboim, H.C. and Doly, J. 1979. A rapid alkaline extension procedure for screening recombinant pliasmid DNA. *Nucl. Acid. Res.* **7**:1513-1522.

Bishop, D.T., and Thompson, E.A. 1988. Linkage information and bias in the presence of interference. *Genet. Epidemiol.* **5**:107-119.

Blin, N., and Stafford, D.W. 1976. Isolation of high molecular weight DNA. *Nucl. Acid. Res.* **3**:2303.

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**:314-331.

Bowcock, A., and Cavalli-Sforza, L. 1991. The study of variation in the human genome. *Genomics* **11**:491-498.

Bowcock, A.M., Barnes, R.I., White, R.L., Kruse, T.A., Tsipouras, P., Sarfarazi, M., Jenkins, T., Viljoen, C., Litt, M., Kramer, P.L., Murray,

J.C., and G. Vergnaud. 1992. The CEPH consortium linkage map of human chromosome 15q. *Genomics* **14**:833-840.

Bowcock, A., Osborne-Lawrence, S., Barnes, R., Chakravarti, A., Washington, S. and Dunn, C. 1993a. Microsatellite polymorphism linkage map of human chromosome 13q. *Genomics* **15**:376-386.

Bowcock, A.M., Gerken, S.C., Barnes, R.I., Shiang, R., Jabs, E.W., Warren, A.C., Antonarakis, S., Retief, A.E., Vergnaud, G., Leppert, M., Lalouel, J.-M., White, R.L., and L.L. Cavalli-Sforza. 1993b. The CEPH Consortium Linkage Map of Human Chromosome 13. *Genomics* **16**:486-496.

Breuning, M.H., Saris, J.J., Wapenaar, M.C., den Dunnen, J.T., van Ommen, G.J.B., and Pearson, P.L. 1989. The isolation of DNA probes from chromosome 16. for the diagnisis of polycystic kidney disease. In: Approaches to the pathogenisis of polycystic kidney disease, Carone, F.A. (eds).

Breuning, M.H., Snijdwint, F.G.M., Brunner, H., Verwest, A., Ijdo, J.W., Saris, J.J., Dauwerse, J.G., Blonden, L., Keith, T., Callen, D.F., Hyland, V.J., Xaio, G.H., Scherer, G., Higgs, D.R., Harris, P., Bachner, L., Reeders, S.T., Germino, G., Pearson, P.L., and van Ommen, G.J.B. 1990. Map of 16 polymorphic loci on the short arm of chromosome 16 close to the polycystic kidney disease gene (PKD1). *J. Med. Genet.* **27**:603-613.

Breuning, M.H., Dauwerse, H.G., Fugazza, G., Saris, J.J., Spruit, L., Wijnen, H., Tommerup, N., van der Hagen, C.B., Imazumi, K., Kuroki, Y., van der Boogaard, M.-J., de Pater, J.M., Mariman, E.C.M., Hamel, B.C.J., Himmelbauer, H., Frischauf, A.-M., Stallings, R.L., Beverstok, G.C., van Ommen ,G.-J., and Hennekam, R.C.M. 1993. Rubinstein-Taybi syndrome caused by submicroscopic deletions within 16p13.3. *Am J Hum Genet* **52**:249-254.

Bridges, C.B., and Morgan, T.H. 1919. The second chromosome group of mutant characters of Drosophila melanogaster. Carnegie Inst. Washington Publ. **278**:123-304.

Bridges, C.B., and Morgan, T.H. 1923. The third chromosome group of mutant characters of Drosophila melanogaster. Carnegie Inst. Washington Publ. **327**:12-30.

Brook, J.D., McCurrach, M.E., Harley, H.G., Buckler, A.J., Church, D., Aburatani, H., Hunter, K., Stanton, V.P., Thirion, Hudson, T., Sohn, R., Zemelman, B., Snell, R.G., Rundle, S.A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., Juvonen, V., Johnson, K., Harper, P., Shaw, D.J., and Housman, D.E. 1992, Molecular basis of myotonic dystrohpy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. Cell 68:799-808.

Buetow, K.H. 1991. Influence of aberrant observations on high-resolution linkage analysis outcomes. *Am. J. Hum. Genet.* **49**:985-994.

Buetow, K.H., Shiang, R., Yang, P., Nakanura, Y., Lathrop, G.M., White, R., Wasmuth, J.J., Wood, S., Berdahl, L.D., Leysens, N.J., Ritty, T.M., Wise, M.E., and Murray, J.C. 1991. A detailed multipoint map of human chromosome 4 provides evidence for linkage heterogeneity and position specific recombination rates. *Am. J. Hum. Genet.* **48**:911-925.

Buetow, K.H., Duggan, D., Yang, B., Ludwigsen, S., Puck, J., Porter, J., Budarf, M., Spielman, R., and Emanuel, B.S. 1993. A microsatellite-based multipoint index map of human chromosome 22. *Genomics* **18**:329-339.

Buetow, K.H., Weber, J.L., Ludwigsen, S., Scherpbier-Heddema, T., Duyk, G.M., Sheffield, V.C., Wang, Z., and Murray, J.C. 1994. Integrated human genome-wide maps constructed using the CEPH reference panel. *Nature Genet.* **6**:391-393.

Burke, D.T., G.F. Carle, and Olson, M.V. 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial hromosome vectors. *Science* **236**:806-812.

Callen, D.F. 1986. A mouse-human hybrid cell panel for mapping human chromosome 16. *Ann. Genet.* **29**:235-239.

Callen, D.F., Hyland, V.J., Baker, E.G., Fratini, A., Simmers, R.N., Mulley, J.C., and Sutherland, G.R. 1988. Fine mapping of gene probes and anonymous DNA fragments to the long arm of chromosome 16. *Genomics* **2**:144-153.

Callen, D.F., Hyland, V.J., Baker, E.G., Fratini, A., Gedeon, A.K., Mulley, J.C., Fernandez, K.E.W., Breuning, M.H., and Sutherland, G.R. 1989. Mapping the short arm of human chromosome 16. *Genomics* **4**:348-354.

Callen, D.F., Baker, E., Eyre, H.J., and Lane, S.A. 1990. An Expanded mouse-human hybrid cell panel for mapping human chromosome 16. *Ann Genet* **4**:190-195.

Callen, D.F., Baker, E., Lane, S., Nancarrow, J., Thompson, A., Whitmore, S.A., Maclennan, D.H., Berger, R., Cherif, D., Jarvela, I., Peltonen, L., Sutherland, G.R. and Gardiner, R.M. 1991. Regional mapping of the Batten Disease locus (CLN3) to human chromosome 16p12. *Am J Hum Genet* 49:1372-1377.

Callen, D.F., Doggett, N.A., Stallings, R.L., Chen, L.Z., Whitmore, S.A., Lane, S.A., Nancarrow, J.K., Apostolou, S., Thompson, A.D., Lapsys, N.M., Eyre, H.J., Baker, E.G., Shen, Y., Holman, K., Phillips, H., Richards, R.I., and Sutherland, G.R. 1992a. High-resolution cytogenetic-based physical map of human chromosome 16. *Genomics* **13**:1178- 1185.

Callen, D.F., Hildebrand, C.E., Reeders, S. 1992b. Report of the second international workshop on human chromosome 16 mapping. *Cytogenet. Cell Genet.* **60**:157-176.

Callen, D.F., Thompson, A.D., Shen, Y., Phillips, H.A., Richards, R.I., Mulley, J.C., and G.R. Sutherland. 1993a. Incidence and origin of "Null" alleles in (AC)n microsatellite markers from chromosome 16. *Am. J. Hum. Genet.* **52**:922-927.

Callen, D.F., Eyre, H., Lane, S., Shen, Y., Hansmann, I., Spinner, N., Zackai, E., McDonald-McGinn, D., Schuffenhauer, S., Wauters, J., Van Thienen, M-N., Van Roy, B., Sutherland, G.R., and Hann, E.A. 1993b. High resolution mapping of interstitial long arm deletions of chromosome 16: relationship to phenotype. *J. Med. Genet.* **30**:828-832.

Callen, D.F., Doggett, N.A., Apostolou, S., Dolman, G., Lane, S.A., Whitmore, S.A., and Sutherland, G.R. 1994. Use of a high resolution hybrid cell panel for physical mapping of chromosome 16. In: The report of the third international workshop on human chromosome 16. Submitted to *Cytogenet. Cell. Genet.*

Carter, T.C., and Falconer, D.S. 1951. Stocks for detecting linkage in the mouse and the theory of their design. J. Genet. 50:307-323.

Caskey, C.T., Pizzuti, A., Fu, Y.-H., Fenwick, R.G., and Nelson, D.L. 1992. Triplet repeat mutations in human disease. *Science* **256**:784-789.

Ceccherini, I., Romeo, G., Lawrence, S., Breuning, N.H., Harris, P.C., Himmelbauer, H., Frischauf, A.M., Sutherland, G.R., Germino, G.G., Reeders, S.T. and Morton, N.E. 1992. Construction of a map of chromosome 16 by using radiation hybrids. *Proc Natl Acad Sci USA* **89**:104-108.

Chakravarti, A. 1991. Short Communication: A graphical representation of genetic and physical maps: The Marey map. *Genomics* **11**:219-222.

Chen, L.Z., P.C. Harris, S. Apostolou, E. Baker, K. 2Holman, S.A. Lane, J.K. Nancarrow, S.A. Whitmore, R.L. Stallings, C.E. Hildebrand, R.I. Richards, G.R. Sutherland, and D.F. Callen. 1991. A refined physical map of the long arm of human chromosome 16. *Genomics* **10**:308-312.

Chumakov, I., Rigault, P., Guillou, S., Ougen, P., Billaut, A., Guascon, G., Gervy, P., Legall, I., Soularue, P., Grinas, L., Bougueleret, L., Bellane-Chantelot, C., Lacroix, B., Barot, E., Gesnouin, P., Pook, S., Vaysseix, G., Frelat, G., Schmitz, A., Sambucy, J.L., Bosch, A., Estivill, X., Weissenbach, J., Vignal, A., Reithman, H., Cox, D., Patterson, D., Gardiner, K., Hattori, M., Sakaki, Y., Ichikawa, H., Ohki, M., Le Paser, D., Heg, R., Antonarakis, S.E., and Cohen, D. 1992. Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359**:380-386.

Collins, F. and Galas, D. 1993. A new five-year plan for the U.S. human genome project. *Science* **262**:43-46.

Conneally, P.M., and Rivas, M.L. 1980. Linkage analysis in man. In: Harris, Hirschhorn, Eds. Advances in Human Genetics 10. New York: Plenum Press.

Cook, P.J.L., Robson, E.B., Buckton, K.E., Jacobs, P.A., Polani, P.E. 1974. Segregation of genetic makers in families with chromosome polymorphisms and structural rearrangements involving chromosome 1. *Ann Hum Genet* **37**:261-274.

Cottingham, R.W., Idury, R.M., and Schaffer, A.A. 1993. Faster sequential genetic linkage computations. *Am. J. Hum. Genet.* **53**:252-263.

Cox, D.R., Burmeister, M., Price, E.R., Kim, S., and Myers. R.M. 1990. Radiation hybrid mapping: A omatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250**:245-250.

Cox, T.C., Kozman, H.M., Raskind, W.H., May, B.K., and Mulley, J.C. 1992. Identification of a highly polymorphic marker within intron 7 of the ***ALAS2*** gene and suggestion of at least two loci for X-linked sideroblastic anemia. *Hum. Mol. Genet.* **8**:639-641.

Crowe, J.F. 1988. A diamond anniversary: The first chromosome map. *Genetics* **118**:1-3.

Dausset, J., Cann, H., Cohen, D., Lathrop, G.M., and White, R. 1990. Centre d'Etude du Polymorphisme Humain (CEPH): Collaborative genetic mapping of the human genome. *Genomics* **6**:575-577.

Davidson, P.J., Jobsis, J., Davis, M., Wallis, S., and Humphries, S.E. 1987. Identification of a gene on chromosome 16, which is related to that for apolipoprotien E (apoE) on chromosome 19. *Cytogenet. Cell Genet.* **46**:604-605 [A1013].

Dawson, E.S., Shaikh, S., Weber, J.L., Wang, Z., Weissenbach, J., Powell, J.F., and Gill, M. 1993. A continuous linkage map of 22 short tandem repeat polymorphisms on human chromosome 12. *Genomics* **17**: 245-248.

Dean, M., Drumm, M.L., Stewart, C., Gerrard, B., Perry, A., Hidaka, N., Cole, J.L., Collins, F.S., Iannuzzi, M.C. 1989. Approaches to localizing disease genes as applied to cystic fibrosis. *Nucl. Acids Res.* **18**:345-350.

Decker, R.A., Moore, J., Ponder, B., and Weber, J.L. 1992. Linkage mapping of human chromosome 10 microsatellite polymorphisms. *Genomics* **12**:604-606.

Dempster, A.P., Laird, N.M., and Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.* **39**:1-38.

Doggett, N.A., Callen, D.F., Duessing, L.A., Tesmer, J.G., Meincke, L.J., Bruce, D.C., Altherr, M.R., Ford, A.A., Torney, D.C., Sutherland, R.D., Lowenstein, M.G., Mundt, M.O., Bruno, W.J., Knill, E.H., Deaven, L.L., Sutherland, G.R., Richards, R.I., and Moyzis, R.K. 1994. An integrated physical/genetic/cytogenetic map of human chromosome 16. In: The report of the third international workshop on human chromosome 16. Submitted to *Cytogenet. Cell. Genet.*

Donahue, R.P., Bias, W.B., Renwick, J.H., and McKusick, V.A. 1968. Probable assignation of the Duffy blood group locus to chromosome 1 in man. *Proc. Natl. Acad. Sci. USA* **61**:949-955.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R., Lander, E.S., Botstein, D., Akots, G., Rediker, K.S., Gravius, T., Brown, V.A., Rising, M.B., Parker, C., Powers, J.A.,Watt, D.E., Kauffman, E.R., Bricker, A., Phipps, P., Muller-Kahle, H., Fulton, T.R., Ng, S., Schumm, J.W., Braman, J.C., Knowlton, R.G., Barker, D.F., Crooks, S.M., Lincoln, S.E., Daly, M.J., and Abrahamson, J. 1987. A genetic linkage map of the Human Genome. *Cell* **51**:319-337.

Donnelly, A., Kozman, H., Gedeon, A.K., Webb, S., Lynch, M., Sutherland, G.R., Richards, R.I., and Mulley, J.C. 1994. A linkage map of microsatellite markers on the human X chromosome. *Genomics* **20**:363-370.

Donnelly, A.J., Choo, A., Kozman, H.M., Gedeon, A.K., Danks, D.M., and Mulley, J.C. Regional localisation of a second non-specific X-linked mental retardation gene (*MRX19*) to Xp22. 1994. *Am. J. Med. Genet.* **15**:581-585.

Dracopoli, N.C., O'Connell, P., Elsner, T.I., Lalouel, J.-M., White, R.L., Buetow, K.H., Nishimura, D.Y., Murray, J.C., Helms, C., Mishra, S.K., Donis-Keller, H., Hall, J.M., Lee, M.K., King, M.-C., Attwood, J., Morton, N.E., Robson, E.B., Mahtani, M., Willard, H.F., Royce, N.J., Patel, I., Jeffreys, A.J., Verga, V., Jenkins, T., Weber, J.L., Mitchell, A.L., and A.E. Bale. 1991. The CEPH consortium linkage map of human chromosome 1. *Genomics* **9**:686-700.

Dunn, L.C., and Bennet, D. 1967. Sex differences in recombination of linked genes in animals. *Genet. Res.* **9**:211-220.

Dyken, P.R. 1988. Reconsideration of the classification of the neuronal ceroid lipofucinosis. *Am. J. Med. Genet. Suppl.* **5**:69-84.

Economou, E.P., Bergen, A.W., Warren, A.C., and Antonarakis, S.E. 1990. The polydeoxyadenylate tract of Alu repetitive elements is polymorphic in the human genome. *Proc Natl Acad Sci USA* **87**:2951-2954.

Edwards, A.W.F. 1984. *Likelihood* Cambridge University Press.

Edwards, M.J., Challinor, C.J., Colley, P.J., Roberts, J., Partington, M.W., Holloway, G.E., Kozman, H.M., and Mulley, J.C. 1994. Clinical and

Linkage study of a large family with simple Ectopia Lentis linked to *FBN1*. *Am. J. Med. Genet.* In Press.

Eiberg, H., Gardiner, R.M., and Mohr, J. 1989. Batten disease (Spielmeyer-Sjogren disease) and haptoglobin (HP): Indication of linkage and assignment to chromosome 16. *Clin. Genet.* **36**:217-218.

Elston, R.C., and Stewart, J. 1971. A general model for the analysis of pedigree data. *Hum. Hered.* **21**:523-542.

Eneglstein, M., Hudson, T.L., Lane, J.M., Lee, M.K., Leberone, B., Landes, G.M., Peltonen, L., Weber, J.L. and Dracopoli, N.C. 1993. A PCR-based linkage map of human chromosome 1. *Genomics* **15**: 251-258.

European Chromosome 16 Tuberous Sclerosis Consortium, 1993, Identification and characterisation of the tuberous sclerosis gene on chromosome 16. *Cell* **75**:1305-1315.

Feinberg, A.P., and Vogelstein, B. 1983. A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* **132**:6-13.

Felsenstein, J. 1979. A mathematically tractable family of genetic mapping functions with different amounts of interference. Genetics 91:769-775.

Fisher, R.A. 1935. The detection of linkage. *Ann. Eugen.* **6**:187-201.

Fisher, R.A. 1970. Statistical Methods for research workers. 14th ed. New York: Hafner Press.

Fitch, D.H.A., Mainone, C., Goodman, M., and Slightom, J.L. 1990. Molecular history of gene conversions in the primary fetal τ-globin genes. *J. Biol. Chem.* **265**:781-793.

Fratini, A., Simmers, R.N., Callen, D.F., Hyland, V.J., Tischfield, J.A., Stambrook, P.J., and Sutherland, G.R. 1986. A new location for the human adenine phosphoribosyltransferase gene (APRT) distal to the haptoglobin (HP) and fra(16)(q23) (FRA16D) loci. *Cytogenet. Cell Genet.* **43**:10-13.

Fratini, A., Baker, E.G., Callen, D.F., Reeders, S.T., Hyland, V.J., and Sutherland, G.R. 1988. *pACHF1.1 (D16S8)* detects a common *Pvu*II RFLP and maps at 16p13.3-16p13.11. *Nucleic Acids Res.* **16**:9065.

Fujiwara, M., Yoshimoto, T., Morita, Y., and Kamada, M. 1992. Interstitial deletion of chromosome 16q: 16q22 is criticl for 16q- syndrome. *Am. J. Med. Genet.* **43**:561-564.

Gardiner, M., Sandford, A., Deadman, M., Poulton, J., Cookson, W., Reeders, S., Jokiaho, I., Peltonen, L., Eiberg, H., and Julier, C. 1990 Batten disease gene (*CLN3*) maps to human chromosome 16. *Genomics* **8**:387-390.

Germino, G.G., Weinstat-Saslow, D., Himmelbauer, H., Gillespie, G.A.J., Somlo, S., Wirth, B., Barton, N., Harris, K.L., Frischauf, A-M., and Reeders, S.T. 1992. The gene for autosomal dominant polycystic kidney disease lies in a 750-kb CpG-rich region. *Genomics* **13**:144-151.

Green, E.D. and Green, P. 1991. Sequence tagged site (STS) content mapping of human chromosomes: Theoretical considerations and early experiences. *PCR Meth. Appl.* **1**:77-90.

Green, P. 1992. Construction and comparrison of chromosome 21 radiation hybrid and linkage maps using CRI-MAP. *Cytogenet. Cell Genet.* **59**:122-124.

Green, P., C. Helms, B. Weiffenbach, K. Stephens, T. Keith, D. Bowden, D. Smith, and H. Donis-Keller. 1989. Construction of a linkage map of the

human genome, and its application to mapping genetic diseases. ***Clin. Chem. 35/7(B):B33-B37.***

Green, P., Falls, K., and Crooks, S. 1990. Documentation for CRI-MAP, Version 2.4 (3/26/90).

Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., Marc, S., Bernardi, G., Lathrop, M., and Weissenbach, J. 1994. The 1993-94 Genethon human genetic linkage map. *Nature genetics* 7:246-333.

Haines, J.L. 1992. Chromlook: An interactive program for error detection and mapping in referrence linkage data. *Genomics* **14**:517-529.

Haldane, J.B.S. 1919. The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.* **8**:299-309.

Haldane, J.B.S. 1922. Sex ratio and unisexual sterility in hybrid animals. *J. Genet.* **12**:101-109.

Haldane, J.B.S., and Smith, C.A.B. 1947. A new estimate of the linkage between the genes for colour blindness and haemophilia in man. *Ann. Eugen.* **14**:10-31.

Hamada, H., Petrino M.G. and Kakunaga T. 1982. A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. *Proc Natl Acad Sci USA* **79**:6465-6469.

Harris, H., Hopkinson, D.A., and Edwards, Y.H. 1977. Polymorphism and the subunit structure of enzymes: A contribution to the neutralist-selectionist controversy. *Proc. Natl. Acad. Sci. USA.* **74**:698-701.

Harris, P., Thomas, S., Ratcliffe, P.J., Breuning, M.H., Coto, E., and Lopez-Larrea, C. 1991. Rapid genetic analysis of families with polycystic kidney disease 1 by means of a microsatellite marker. *Lancet* **338**:1484-1487.

Harris, P.C. and Thomas, S. Length polymorphism of the subtelomeric regions of both the short (p) and long (q) arms of chromosome 16. 1992. *Cytogenet. Cell Genet.* **60**:171[ABST].

Hazan, J., Dubay C., Pankowiak M.-P., Becuwe N., and Weissenbach J. 1992. A genetic linkage map of human chromosome 20 composed entirely of microsatellite markers. *Genomics* **12**:183-1.

Henske, E.P., Ozelius, L., Gusella, J.F., Haines, J.L., and Kwiatkowski, D.J. 1993. A high-resolution linkage map of human 9q34.1. **Genomics 17**:587-591.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75:800-802.

Hoffman, E. 1994. The evolving genome project: current and future impact. *Am. J. Hum. Genet.* **54**:129-136.

Hudson, T.J., Engelstein M., Lee M.K., Ho E.C., Rubenfield M.J., Adams C.P., Housman D.E., and Dracopoli N.C. 1992. Isolation and chromosomal assignment of 100 highly informative human simple sequence repeat polymorphisms. *Genomics* **13**:622-629.

Hyland, V.J., Grist, S., Callen, D.F. and Sutherland, G.R. 1988. Anonymous DNA probes to human chromosome 16 derived from a flow-purified library. *Am. J. Hum. Genet.* **42**:373-379.

Hyland, V.J., Fernandez, K.E.W., Callen, D.F., MacKinnon, R.N., Baker, E.G., Friend, K., and Sutherland, G.R. 1989a. Assignment of anonymous

DNA probes to specific intervals of human chromosomes 16 and X. *Hum. Genet.* **83**:61-66.

Hyland, V.J., Callen D.F., Fernandez, K.E.W., MacKinnon, R.N., Friend, K., Mulley, J.C., Fratini, A.F., Baker, E.G., Breuning, M.H., Keith, T., and Sutherland, G.R. 1989b. Anonymous DNA probes to specific intervals of human chromosome 16. *Cytogenet. Cell. Genet.* **51**:1017.

Hyland, V.J., Friend, K., and Sutherland, G.R. 1989c. A TaqI RFLP detected by the probe VK45C6 (D16S131) at 16p13.11. *Nucl. Acids Res.* **17**:6430.

Hyland, V.J., Suthers, G.K., Friend, K., MacKinnon, R.N., Callen, D.F., Breuning, M.H., Keith, T., Brown, V.A., Phipps, P., and G.R. Sutherland. 1990. Probe, VK5B, is located in the same interval as the autosomal dominant adult polycystic kidney disease locus, PKD1. *Hum. Genet.* **84**:286-288.

Järvelä, I.E., Mitchison, H.M., Gardiner, R.M., Callen, D.F., Breuning, M.H., Lerner, T.J., Mole, S.E., and the Batten Disease Consortium. 1994. Long range restriction map of a YAC contig in the Batten disease region of chromosome 16 and exclusion of potenetial candidate genes from this region. In: The report of the third international workshop on human chromosome 16. Submitted to *Cytogenet. Cell. Genet.*

Jeffreys, A.J., Wilson, V., and Thein, S.L. 1985. Individual-specific "fingerprints" of human DNA. *Nature* **316**:76-79.

Julier, C., Nakamura, Y., Lathrop, M., O'Connell, P., Leppert, M., Mohandas, T., Lalouel, J.-M., and R. White. 1990. A primary map of 24 loci on human chromosome 16. *Genomics* **6**:419-427.

Kandt, R.S., Haines J.L., Smith M., Northrup H., Gardner R.J.M., Short M.P., Dumars K., Roach E.S., Steingold S., Wall S., Blanton S.H., Flodman P., Kwiatkowski D.J., Jewell A., Weber J.L., Roses A.D., and Periccak-Vance M.A. 1992. Linkage of an important gene locus for tuberous sclerosis to a chromosome 16 marker for polycystic kidney disease. *Nature Genetics* **2**:37-41.

Karin, M., Eddy, R.L., Henry, W.M., Haley, L.L., Byers, M.G., and Shows, T.B. 1984. Human metallothionein genes are clustered on chromosome 16. *Proc. Natl. Acad. Sci. USA.* **81**:5494-5498.

Karlin, S. 1984. Theoretical aspects of genetic map functions in recombination processes. In Human Population Genetics: The Pittsburgh symposium, edited by A. Chakravarti, 209-228. New York: Van Nostrand Reinhold.

Keats, B., Ott, J., and Conneally, M. 1989. Report of the committee on linkage and gene order. Cytogenet. Cell Genet. 51:495-502.

Keats, B.J.B., S.L. Sherman, N.E. Morton, E.B. Robson, K.H. Buetow, P.E. Cartwright, A. Chakravarti, U. Francke, P.P. Green, and J Ott. 1991a. Guidelines for human linkage maps: An international system for human linkage maps (ISLM, 1990). (Meeting Report) ***Genomics 9:557-560.***

Keats, B.J.B., Sherman, S.L., and Ott, J. 1991b. Report to the committee of linkage and gene order. Cyto & cell genet. 58:1097-1102.

Keith, T., Green, P., Reeders, S., Brown, V.A., Phipps, P., Bricker, A., Knowlton, R., Nelson, C., and Donis-Keller, H. 1987. A linkage map of chromosome 16 with 41 RFLP markers. *Cytogenet. Cell Genet.* **46**:638. [A1267]

Keith, T.P., Green, P., Reeders, S.T., Brown, V.A., Phipps, P., Bricker, A., Falls, K., Rediker, K.S., Powers, J.A., Hogan, C., Nelson, C., Knowlton, R., and H. Donis-Keller. 1990. Genetic linkage map of 46 DNA markers on human chromosome 16. *Proc. Natl. Acad. Sci. USA.* **87**:5754-5758.

Kernighan, B., and Ritchie, D. 1978. *The C Programming Language*. Prentice-Hall.

Knight, S.J.L., Flannery, A.N., Hirst, M.C., *et al.* 1993. Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. *Cell* **74**:127-134.

Koenig, M., Hoffman, E.P., Bertelson, C.J., Monaco, A.P., Feener, C., and Kunkel, L.M. 1987. Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organisation of the DMD gene in normal and affected individuals. Cell 50:509-517.

Kohara, Y., Akiyama, K., and Isono, K. 1987. The physical map of the whole *E.coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50**:495-508.

Konradi, C., Ozelius, L., Yan, W., Gusella, J.F., and Breakefield, X.O. 1991. Dinucleotide repeat polymorphism (*D16S285*) on human chromosome 16. *Nucl. Acids Res.* **19**:5449.

Koorey, D.J., Bishop G.A., and McCaughan G.W. 1993. Allele non-amplification: a source of confusion in linkage studies employing microsatellite polymorphisms. *Hum. Mol. Genet.* **2**: 289-291.

Kosambi, D.D. 1944. The estimation of map distances from recombination values. Ann. Eugen. 12:172-175.

Kozman, H.M., Gedeon, A.K., Whitmore, S., Suthers, G.K., Callen, D.F., Sutherland, G.R., and Mulley, J.C. 1991. Addition of *MT*, *D16S10*, *D16S4*, and *D16S91* to the linkage map within 16q12.1-q22.1. *Genomics* **11**:756-759.

Kozman, H.M., Phillips, H.A., Sutherland, G.R., and Mulley, J.C. 1992. A multipoint genetic linkage map around the fragile site *FRA16A* on human chromosome 16. *Genet. (Life Sci. Adv.)* **11**:229-233.

Kozman, H.M., Phillips, H.A., Callen, D.F., Sutherland, G.R., and Mulley, J.C. 1993. Integration of the cytogenentic and genetic linkage maps of human chromosome 16 using 50 physical intervals and 50 polymorphic loci. *Cytogenet. and Cell Genet.* **62**:194-198.

Kozman, H.M., Phillips, H.A., Callen, D.F., Sutherland, G.R., and J.C. Mulley. 1993. Integration of the Cytogenetic and genetic linkage maps of human chromosome 16 using 50 physical intervals and 50 polymorphic loci. *Cytogenet. and Cell Genet.* **62**:*194-198.*

Kozman, H.M., and Mulley, J.C. 1994. Human chromosome 16, genetic and physical mapping. In: Encyclopaedia of Molecular Biology. (In Press).

Kozman, H.M., Keith, T.P., Donis-Keller, H., White, R.L., Weissenbach, J., Dean, M., Vergnaud, G., Kidd, K., Gusella, J., Royle, N.J., Sutherland, G.R., and Mulley, J.C. 1994. The CEPH consortium map of human chromosome 16. *Genomics*, In Press.

Kremer, E.J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S.T., Schlessinger, D., Sutherland, G.R., and Richards, R.I. 1991. Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)*n*. *Science* **252**:1711-1714.

Kwiatkowski, D.J., Hinske E.P., Weimer K., Ozelius L., Gusella J.F. and Haines J. 1992. Construction of a GT polymorphism map of human 9q. *Genomics* **12**:229-240.

Laing, N.G., Laing, B.A., Meredith, C., Wilton, S.D., Robbins, P., Honeyman, K., Dorosz, S., Kozman, H., Mastaglia, F.L., and Kakulas, B.A. 1994. Autosomal dominant distal myopathy: linkage to chromosome 14. *(Submitted)*.

Lalouel, J.-M. 1979. GEMINI: A computer program for optimisation of general non-linear functions. *University of Utah, Department of Medical Biophysics and Computing, Technical Report No. 14 Salt Lake City, Utah.*

Lander, E.S., and Botstein, D. 1986. Mapping complex genetic traits in humans: New methods using a complete RFLP linkage map. *Cold Spring Harbour Symposia on Quantitative Biology*. **51**:49-62.

Lander, E.S. and Botstein, D. 1987. Homozygosity mapping: a way to map human recessives traits with the DNA of inbred children. *Science* **236**:1567-1570.

Lander, E.S., and Green, P. 1987. Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. U.S.A.* **84**:2363-2367.

Lander, E.S., Green, P., Abrahamson, J., Barlow, A., Daly, M.J., Lincoln, S.E., and Newburg, L. 1987. Mapemaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1**:174-181.

Lange, K., Weeks, D., and Boehnke, M. 1988. Programs for pedigree analysis: MENDEL, FISHER, and dGENE. *Genet. Epidemiol.* **5**:471-472.

Lapsys, N. 1993. The FRA16B locus: long range restriction mapping of 16q13-16q22.1. PhD thesis, University of Adelaide.

Lasher, L., Reefer, J., and Chakravarti, A. 1991. Effects of genotyping errors on the estimation of chromosome map length. *Am. J. Hum. Genet.* **49**(supplement):369.

Lathrop, G.M. 1994, Computing the genetic map. *Nature Genetics* **6**:326-328.

Lathrop, G.M., and J.-M. Lalouel. 1984. Easy calculation of lod scores and genetic risks on small computers. *Am. J. Hum. Genet.* **36**:460-465.

Lathrop, G.M., and J.-M. Lalouel. 1988. Efficient computations in multilocus linkage analysis. *Am. J. Hum. Genet.* **42**:498-505.

Lathrop, G.M., Hooper, A.B., Huntsman, J.W., and Ward, R.H. 1983. Evaluating pedigree data. I. The estimation of pedigree error in the presence of marker miss typing. *Am. J. Hum. Genet.* **35**:241-262.

Lathrop, G.M., Lalouel, J.-M., Julier, C., and Ott, J. 1984. Strategies for multilocus linkage analysis in humans. *Proc. Natl. Acad. Sci. U.S.A* **81**:3443-3446.

Lathrop, G.M., Lalouel, J-M., Julier, C., and Ott, J. 1985. Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *Am. J. Hum. Genet.* **37**:482-498.

Lathrop, G.M., Lalouel, J.-M., and White, R.L. 1986. Construction of human linkage maps: Likelihood calculations for multilocus linkage analysis. *Genet. Epidem.* **3**:39-52.

Lathrop, G.M., Leppert, M., Nakamura, Y., O'Connell, P., Litt, M., Maslen, S., Cartwright, P., Lalouel, J-M., and White, R. 1987. A primary genetic linkage map of chromosome 16. *Cytogenet. Cell Genet.* **46**:643. [A1347]

Lauthier, V., Mariat, D., Zoroastro, M., and Vergnaud, G. 1991. A synthetic probe STR14C19, detects a new polymorphic locus at 16pter (D16S282). *Nucleic Acids Research* **19**:4015.

Lawrence, S., Morton, N.E., and Cox, D.R. 1991. Radiation hybrid mapping. *Proc. Natl. Acad. Sci. USA.* **88**:7477-7480.

Lerner, T.J., D'Arigo, K.L., Haines, J.L., Doggett, N.A., Buckler, A., and the Batten Disease Consortium. 1994. In: The report of the third international workshop on human chromosome 16. Submitted to *Cytogenet. Cell. Genet.*

Lincoln, S.E., and Lander, E.S. 1992. Systematic detection of errors in genetic linkage data. *Genomics* **14**:604-610.

Litt, M., and Luty, J.A. 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44**:397-401.

Litt, M., Kramer, P., Hauge, X.Y., Weber, J.L., Wang, Z., Wilkie, P.J., Holt, M.S., Mishra, S., Donis-Keller, H., Warnich, L., Retief, A.E., Jones, C. and Weissenbach, J. 1993. A microsatellite-based index map of human chromosome 11. *Hum. Mol. Genet.* **2**: 909-913.

Magenis, R.E., Hetch, F., and Lovrien, E.W. 1970. Heritable fragile site on chromosome 16: Probable localisation ofhaptoglobin locus in man. *Science* **170**:85-87.

Maniatis, T., Fritsh, E.F., and Sanbrook, J. 1982 Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Habor, New York.

Matise, T.C., Perlin, M., Chakravarti, A. 1994. Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. *Nat. Genet.* **6**:384-390.

McDonnell, M.W., Simon, M.N., and Studier, F.W. 1977. Analysis of restriction fragments of T7 DNA and determination of molecular weights by electrophoresis in neutral and alkaline gels. *J. Mol. Biol.* **110**:119.

McInnis, M.G., Chakravarti, A., Blaschak, J., Petersen, M.B., Sharma, V., Avramopoulos, D., Blouin, J.-L., Konig, U., Brahe, C., Matise, T.C., Warren, A.C., Talbot, C.C., Van Broeckhoven, C., Litt, M., and Antonarakis, S.E. 1993. A linkage map of human chromosome 21: 43 PCR markers at average intervals of 2.5 cM. *Genomics* **16**:562-571.

McKusick, V.A., and Amberger, J.S. 1994. The morbid anatomy of the human genome: chromosomal location of mutations causing disease. J. Med. Genet. 31:265-279.

Meyers, D.A., Conneally, P.M., Lovrien, E.W., Magenis, E., Merritt, A.D., Norton, J.A., Palmer, C.G., Rivas, M.L., Wang, L., and Yu, P.L. 1976. Linkage group I: The simultaneous estimation of recombination and interference. *HGM* **3**:335-339.

Meyers, D.A., Conneally, P.M., Hecht, F., Lovrien, E.W., Magenis, E., Merritt, A.D., Palmer, C.G., Rivas, M.L., and Wang, L. 1975. Linkage group I: Multipoint mapping. *HGM* **2**:381-389.

Miller, S.A., Dykes, D.D., Pulesky, H.F. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucl. Acid. Res.* **16**:1215.

Mills, K.A., Buetow, K.H., Xu, Y., Weber, J.L., Altherr, M.R., Wasmuth, J.J., and Murray, J.C. 1992. Genetic and physical maps of human chromosome 4 based on dinucleotide repeats. *Genomics* **14**:209-219.

Mitchison, H.M., Thompson,A.D., Mulley, J.M., Kozman, H.M., Richards, R.I., Callen, D.F., Stallings, r., Doggett, N., Attwood, J., McKay, T.R., Sutherland, G.R., Gardiner, R.M. 1993. Fine genetic mapping of the Batten disease locus (CLN3) by haplotype analysis and demonstration of allelic association with chromosome 16 microsatellite loci. *Genomics* **16**:455-460.

Mitchison, H.M., Taschner, P.E.M., O'Rawe, A.M., de Vos, N., Phillips, H.A., Thompson, A.D., Kozman, H.M., Haines, J.L., Schlumpf, K., D'Arigo, K., Boustany, R.-M.N., Callen, D.F., Breuning, M.H., Gardiner, R.M., Mole, S.E., and Lerner, T.J. 1994. Genetic Mapping of the Batten disease locus (*CLN3*) to the interval *D16S288-D16S383* by analysis of haplotypes and allelic association. *Genomics* **22**:465-468.

Monaco, A.P. 1994. Isolation of genes from cloned DNA. *Current Opinion in Genet. and Dev.* **4**:360-365.

Morgan, T.H. 1911. Random segregation verses coupling in Mendelian inhetitance. *Science* **34**:384.

Morgan, T.H. 1928. *The Theory of Genes*. New Haven, Conn.: Yale University Press.

Morgan, T.H., Sturtevant, A.H., and Bridges, C.B. 1925. The genetics of Drosophila. *Bibl. Genet.* **2**:1-262.

Morton, N.E. 1955. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7**:277-318.

Morton, N.E. 1988. Multipoint mapping and the emperors clothes. *Ann. Hum. Genet.* **52**:309-318.

Morton, N.E. 1991a. Gene maps and location databases. *Ann. Hum. Genet.* **55**:235-241.

Morton, N.E. 1991b. Parameters of the human genome. *Proc. Natl. Acad. Sci. USA.* **88**:7474-7476.

Morton, N.E. and Andrews, V. 1989. MAP, an expert system for multiple pairwise linkage analysis. *Ann. Hum. Genet.* **53**:263-269.

Morton, N.E. and Collins, A. 1990. Standard maps of chromosome 10. *Ann. Hum. Genet.* **7**:277-318.

Morton, N.E., Rao, D.C., Lindsten, J., Hulten, M., and Yee, S. 1977. A chiasma map of Man. *Hum. Hered.* **27**:38-51.

Morton, N.E., C.J. MacLean, R. Lew, and S. Yee. 1986. Multipoint linkage analysis. *Am. J. Hum. Genet.* **38**:868-883.

Mulley, J.C. and Sutherland, G.R. 1993. Integrating maps of chromosome 16. *Current Opinion in Genetics and Development* **3**: 425-431.

Mulley, J.C., V.J. Hyland, A. Fratini, L.J. Bates, A.K. Gedeon, and G.R. Sutherland. 1989. A linkage group with FRA16B (the fragile site at 16q22.1). *Hum. Genet.* **82**:131-133.

Mulley, J.C., Kozman, H.M., Phillips, H.A., Gedeon, A.K., McCure, J.A., Iles, D.E., Gregg, R.G., Hogan, K., Couch, F.J., MacLennan, D.H., and Hann, E.A. 1993. Refined genetic localisation for Central Core Disease. *Am. J. Hum. Genet.* **52**:398-405.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, T.H., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. and White, R. 1987. Variable number of tandem repaet (VNTR) markers for human mapping. *Science* **235**:1616-1622.

Nakamura, Y., Carlson, M., Krapcho, K., Kanamori, M. and White, R. 1988. New approach for isolation of VNTR markers. *Am. J. Hum. Genet.* **43**:854-859.

Nancarrow, J.K., Kremer, E.J., Holman, K., Eyre, H., Doggett, N.A., Le Paslier, D., Callen, D.F., Sutherland, G.R., and Richards, R.I. 1994. Implications of the structure of autosomal fragile site, *FRA16A*, for the mechanism of fragile site genesis. *Science* 264:1938-1941.

Natt, E., Kao, F.T., Rettenmeier, R., and Scherer G. 1986. Assignment of the human tyrosine aminotransferase gene to chromosome 16. *Hum. Genet.* **72**:225-228.

NIH/CEPH Collaborative Mapping Group. 1992. A Comprehensive Genetic Linkage Map of the Human Genome. *Science 258*:67-86.

O'Connell, P., Lathrop, G.M., Law, M., Leppert, M., Nakamura, Y., Hoff, M., Kumlin, E., Thomas, W., Elsner, T., Ballard, L., Goodman, P., Azen, E., Sadler, J.E., Cai, G.Y., Lalouel, J.-M., and R. White. 1987. A primary genetic linkage map for human chromosome 12. *Genomics* **1**:93-102.

O'Connell, P., Lathrop, G.M., Leppert, M., Makamura, Y., Muller, U., Lalouel, J.-M., and White, R. 1988. Twelve loci form a continuous linkage map for human chromosome 18. *Genomics* **3**:367-372.

O'Connell, P., Lathrop, G.M., Makamura, Y., Leppert, M., Ardinger, R.H., Murray, J.L., Lalouel, J-M., and White, R. 1989. Twenty-eight loci form a continuous linkage map of markers for human chromosome 1. *Genetics* **4**:12-20.

O'Connell, P.,Plaetke, R., Matsunami, N., Odelberg, S., Jorde, L., Chance, P., Leppert, M., Lalouel, J.-M., White, R. 1993. An extended Genetic Linkage Map and an "Index" Map for Human Chromosome 17. *Genomics* **14**:38-47.

Ott, J. 1974. Estimation of the recombination fraction in human pedigrees: Efficient computation of the likelihood for human linkage studies. *Am. J. Hum. Genet.* **26**:588-597.

Ott, J. 1977. Counting methods (EM algorithm) in human pedigree analysis: Linkage and segregation analysis. *Ann. Hum. Genet.* **40**:443-454.

Ott, J. 1991. "Analysis of Genetic Linkage," John's Hopkins Press, Baltimore.

Pascoe, L.L. and Morton, N.E. 1987. The use of map functions in multipoint mapping. *Am. J. Hum. Genet.* **40**:174-183.

Penrose, L.S. 1935. The detection of autosomal linkage in data which consists of pairs of brothers and sisters of unspecified parentage. *Ann. Eugen.* **6**:133-138.

Petrukhin, K.E., Speer, M.C., Cayanis, E., Bonaldo, M.F., Tantravahi, U., Soares, M.B., Fischer, S.G., Warburton, D., Gelliam, T.C., and Ott, J. 1993. A microsatellite genetic linkage map of human chromosome 13. *Genomis* **15**:76-85.

Phillips, H.A., Harris, P., Richards, R.I., Sutherland, G.R., and Mulley, J.C. 1991a. Dinucleotide repeat polymorphisms at the *D16S164*, *D16S168* and *D16S186* loci at 16q21-q22.1. *Nucl Acid Res* **19**: 6964.

Phillips, H.A., Hyland, V.J., Holman, K., Callen, D.F., Richards, R.I., and Mulley, J.C. 1991b. Dinucleotide repeat polymorphism at *D16S287*. *Nucl. Acids Res.* **19**:6664.

Phillips, H.A., Thompsom, A.D., Kozman, H.K., Sutherland, G.R., and Mulley, J.C. 1993. A microsatellite marker within the duplicated D16S79 locus has a null allele: significance for linkage mapping. *Cytogenet. Cell Genet.* **64**:131-134.

Race, R.R., and Sanger, R. 1968. Blood groups in man. Blackwell scientific publications, Oxford.

Rao, C.R. 1973. Linear Statistical Inference and its applications. New York: John Wiley.

Rao, D.C., Morton, N.E., Lindsten, J., Hulten, M., and Yee, S. 1977. A mapping function for man. *Hum. Hered.* **27**:99-104.

Reeders S.T., Keith, T., Green, P., Germino, G.G., Barton, N.J., Lehmann, O.J., Brown, V.A., Phipps, P., Morgan, J., Bear, J.C., and Parfrey, P. 1988. Regional localisation of the autosomal dominant Polycystic Kidney Disease Locus. *Genomics* **3**:150-155.

Reeders, S.T., and Hildebrand, C.E. 1989. Report of the committee on the genetic constitution of chromosome 16. *Cytogenet. Cell Genet.* **51**:299-318.

Reeders, S.T., Hildebrand, C.E., and Sutherland, G.R. 1991. Report of the committee on the genetic constitution of chromosome 16. *Cytogenet. Cell Genet.* **58**:643-685.

Renwick, J.H. and Bolling, D.R. 1971. An analysis procedure illustrated on a triple linkage of use for prenatal diagnosis of Myotonic dystrophy. *J. Med. Genet.* **8**:399-406.

Renwick, J.H. and Schulze, J. 1965. Male and female recombination fraction for the nail-patella: ABO linkage in man. *Ann. Hum. Genet.* **28**:379-392.

Richards, R.I., Holman, K., Kozman, H.M., Kremer, E., Lynch, M., Pritchard, M., Yu, S., Mulley, J., and Sutherland, G.R. 1991a. Fragile X syndrome: Genetic localisation by linkage mapping of two microsaatellite repeats FRAXAC1 and FRAXAC2 which immediately flank the fragile site. *J. Med. Genet.* **28**:818-823.

Richards, R.I., Shen, Y., Holman, K., Kozman, H., Hyland, V.J., Mulley, J.C. and Sutherland, G.R.. 1991b. Fragile X Syndrome: Diagnosis using highly polymorphic microsatellite markers. *Am J Hum Genet* 48:1051-1057.

Richards, R.I., Holman, K., Lane, S., Sutherland, G.R. and Callen, D.F. 1991c. Human chromosome 16 physical map: mapping of somatic cell hybrids using multiplex PCR deletion analysis of sequence tagged sites. *Genomics* **10**:1047-1052.

Richards, R.I., Holman, K., Shen, Y., Kozman, H., Harley, H., Brook, D. and Shaw, D. 1991d. Human glandular Kallikrein genes: genetic and physical mapping of the KLK1 locus using a highlpolymorphicmicrosatellite PCR marker. *Genomics* **11**:77-82.

Richards, R.I., and Sutherland, G.R. 1992. Heritable unstable DNA sequences. *Nature Genet.* **1**:7-9.

Richards, R.I., and Sutherland, G.R. 1994. Simple repeat DNA is not replicated simply. *Nature Genet.* **6**:114-116.

Rigby, P.W.J., Diecjmann, M., Rhodes, C., Berg, P. 1977. Labelling deoxyribonucleic acid to high specific activity in vitro by nick translation with DNA polymerase I. *J. Mol. Biol.* **113**:237-251.

Roberts, L. 1990. The genetic map is back on track after delays. *Science* **248**:805.

Roberts, L. 1992. NIH takes new tack on Gene Mapping. *Science 258*:1573.

Roberts, R.G., Gardner, R.J., and Bobrow, M. 1994. Searching for the 1 in 2,400,000: a review of Dystrophin gene point mutations. Hum. Mut. 4:1-11.

Rogaev, E.I. and Keryanov, S.A. 1992. Unusual variability of the complex dinucleotide repeat block at the SPN locus. *Hum Mol Genet* **1**:657.

Royle, N.J., Clarkson, R.E., Wong, Z., and Jeffreys, A.J. 1988. Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. *Genomics* 3:352-360.

Royle, N.J., Armour, A.J., Webb, M., Thomas, A., and Jeffreys, A.J. 1992. A hypervariable locus *D16S309* located at the distal end of 16p. *Nucl. Acid Res.* **20**:1164.

Schmid, M., Feitchtinger, W., Jebberger, A., Köhler J., Lange, R. 1986. The fragile site (16)(q22). I. Induction by AT-specific DNA-ligands and population frequency. *Hum. Genet.* **74**:67-73.

Schmitt, K. and Goodfellow, P.N. 1994. Predicting the furture. *Nature Genet.* **7**:219.

Sealey, P.G., Whittaker, P.A., and Southern, E.M. 1985. Removal of repeat sequences from hybridisation probes. *Nucl. Acid. Res.* **13**:1905-1922.

Sefton, L. and Goodfellow, P.N. 1992. The human genetic map. *Current Opinion in Genetics and Development* **2**:387-392.

Shen, Y., Holman, K., Thompson, A., Kozman, H., Callen, D.F., Sutherland, G.R., and Richards, R.I. 1991. Dinucleotide repeat polymorphism at the D16S288 locus. *Nucl. Acids Res.* **19**:5445.

Shen, Y., Thompson, A.T., Holman, K., Callen, D.F., Sutherland, G.R., and Richards, R.I. 1992. Four dinucleotide repeat polymorphisms on human chromosome 16 at D16S289, D16S318, D16S319 and D16S320. *Hum. Mol. Genet.* **1**:773.

Shen, Y., Holman, K., Dogget, N.A., Callen, D.F., Sutherland, G.R., and Richards, R.I. 1993a. Five dinucleotide repeat polymorphisms on human chromosome 16q24.2-q24.3. *Hum. Mol. Genet.* **2**: 1504.

Shen, Y., Holman, K., Dogget, N.A., Callen, D.F., Sutherland, G.R., and Richards, R.I. 1993b. Six dinucleotide repeat polymorphisms on human chromosome 16q12.1-q24.1. *Hum. Mol. Genet.* **2**:1505.

Shen, Y., Holman, K., Dogget, N.A., Callen, D.F., Sutherland, G.R., and Richards, R.I. 1993c. Three dinucleotide repeat polymorphisms on human chromosome 16p13.11-p13.3. *Hum. Mol. Genet.* **2**:1506.

Shen, Y., Holman, K., Dogget, N.A., Callen, D.F., Sutherland, G.R., and Richards, R.I. 1993d. Four dinucleotide repeat polymorphisms on human chromosome 16. *Hum. Mol. Genet.* **2**:1745.

Shen, Y., Kozman, H.M., Thompson, A.D., Phillips, H.A., Holman, K., Nancarrow, J.K., Lane, S.A., Chen, L.Z., Apostolou, S., Doggett, N.A., Callen, D.F., Mulley, J.C., Sutherland, G.R., and R.I. Richards 1994. A

PCR-based genetic linkage map of human chromosome 16. *Genomics* **22**:68-76.

Shen, Y., Holman, K., Dogget, N.A., Callen, D.F., Sutherland, G.R. and Richards, R.I. 1994 Dinucleotide repeat polymorphisms at the D16S525, D16S359, D16S531 and D16S522 loci. *Hum. Mol. Genet.* **3**:210.

Sherman, S.L., and Sutherland, G.R. 1986. Segregation analysis of rare autosomal fragile sites. *Hum. Genet.* **72**:123-128.

Shohat, M., Bu, X., Shohat, T., Fischel-Ghodsian, N., Magal, N., Nakamura, Y., Schwabe, A.D., Schlezinger, M., Danon, Y. and Rotter, J.I. 1992. The gene for Familial Mediterranean Fever in both armenians and non-Ashkenazi Jews is linked to the a-globin complex on 16p: evidence for locus homogeneity. *Am J Hum Genet* **51**:1349-1354.

Silver, L.M. 1992. Bouncing off microsatellites. *Nature Genetics* **2**:8-9.

Simmers, R.N., I. Stupans, and G.R. Sutherland. 1986. Localization of the human haptoglobin genes distal to the fragile site at 16q22 using in situ hybridization. *Cytogenet. Cell Genet.* **41**:38-41.

Simmers, R.N, Mulley, J.C., Hyland, V.J., Callen, D.F., and Sutherland, G.R. 1987. Mapping the human α globin gene complex to 16p13.2-pter. *J. Med. Genet.* **24**:761-766.

Skolnick, M. 1991. Workshop report: Attributes of markers on linkage and physical maps. *Cytogenet. Cell Genet.* **58**:1839-1840.

Smith, C.A.B. 1986. The development of human linkage analysis. *Ann. Hum. Genet.* **50**:293-311.

Somlo, S., Wirth, B., Germino, G.G., Weinstat-Saslow, D., Gillespie, G.A.J., Himmelbauer, H., Steevens, L., Coucke, P., Willems, P., Bachner, L., Coto, E., Lopez-Larrea, C., Peral, B., San Millan, J.L., Saris, J.J, Breuning, M.H., Frishauf, A.-M., and Reeders, S.T. 1992. Fine genetic localisation for the gene for autosomal dominant polycystic kidney disease (PKD1) with respect to physically mapped markers. *Genomics* **13**:152-158.

Southern, E.M. 1975. Measurement of DNA length by gel electrophoresis. *Anal. Biochem.* **100**:319-323.

Spurr, N.K., Cox, S., Bryant, S.P., Attwood, J., Robson, E.B., Shields, D.C., Steinbrueck, T., Jenkins, T., Murray, J.C., Kidd, K.K., Summar, M.L., Tsipouras, P., Retief, A.E., Kruse, T.A., Bale, A.E., Vergnaud, G., Weber, J.L., McBride, O.W., Donis-Keller, H., and R.L. White. 1992. The CEPH Consortium linkage map of human chromosome 2. *Genomics* **14**:1055-1063.

Stallings, R.L., Tornsy, D.C.S., Hildebrand, C.E., Longmire, J.L., Deaven, L.L., Jett, J.H., Doggett, N.A. and Moyzis, R.K. 1990. Physical mapping of human chromosomes by repetitive sequence finger printing. *Proc Natl Acad Sci USA* **87**: 6218-6222.

Stallings, R.L., Ford, A.F., Nelson, D., Torney, D.C., Hildebrand, C.E. and Moyzis, R.K. 1991. Evolution and distribution of (GT)n repetitive sequences in mammalian genomes. *Genomics* **10**:807-815.

Stallings, R.L., Doggett, N.A., Callen, D.F., Apostolou, S., Chen, L.Z., Nancarrow, J.K., Whitmore, S.A., Harris, P., Mitchison, H., Breuning, M., Saris, J., Fickett, J., Cinkosky, M., Torney, D., Hildebrand, C.E., and Moyzis, R.K. 1992. Evaluation of a cosmid contig physical map of human chromosome 16. *Genomics* **13**:1031-1039.

Starck,, J., Bouhass, R., and Morle Godet, J. 1990. Extent and high frequency of a short gene conversion between the human Aτ and Gτ fetal globin genes. *Hum. Genet.* **84**:179-184.

Stephens, K., Green, P., Riccardi, V.M., Ng, S., Rising, M., Barker, D., Darby, J.K., Falls, K.M., Collins, F.S., Willard, H.F., and Donis-Keller, H. 1989. Genetic analysis of eight loci tightly linked to neurofibromatosis 1. *Am. J. Hum. Genet.* **44**:13-19.

Straub, R.E., Speer, M.C., Luo, Y., Rojas, K., Overhauser, J., Ott, J., and Gilliam, T.C. 1993. A microsatellite genetic linkage map of human chromosome 18. *Genomics* **15**:48-56.

Sturt, E. 1976. A mapping function for human chromosomes. *Ann. Hum. Genet.* **40**:147-163.

Sturtevant, A.H. 1913. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *J. Exp. Zool.* **14**:43-59.

Sutherland, G.R. 1993. *Human Fragile Sites*. In "Genetic Maps. Locus Maps of Human Maps" (S.J. O'Brien, ED), Sixth edition, Cold Spring Harbour Laboratory Press, New York).

Sutherland, G.R., and Hecht, F. 1985. *"Fragile sites on human chromosomes."* Oxford Univ. Press, New York.

Sutherland, G.R., Reeders, S., Hyland, V.J., Callen, D.F., Fratini, A., and Mulley, J.C. 1987. Molecular genetics of human chromosome 16. *J. Med. Genet.* **24**:451-456.

Suthers, G.K. 1992. The human gene map near the fragile X. PhD thesis, University of Adelaide.

Tanzi, R.E., Haines, J.L., Watkins, P.C., Stewart, G.D., Wallace, M.R., Hallewell, R., Wong, C., Wexler, N.S., Conneally, P.M., and Gusella, J.F. 1988. Genetic linkage map of human chromosome 21. *Genomics* **3**:129-136.

The European Polycyctic Kidney Disease Consortium, 1994. The polycystic kidney disease 1 gene encodes a 14 kb transcript and lies within a duplicated region on chromosome 16. *Cell* **77**:881-894.

The Huntington's Disease Collaborative Research Group, 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72**:971-983.

Thompson, A.D., Shen, Y., Holman, K., Sutherland, G.R., Callen, D.F, Richards, R.I. 1992. Isolation and characterisation of $(AC)_n$ microsatellite genetic markers from human chromosome 16. *Genomics* **13**:402-408.

Thompson, E.A. 1984. Information gain in joint linkage analysis. *IMA J. Math. Appl. Med. Biol.* **1**:31-49.

Tomfohrde, J., Wood, S., Schertzer, M., Wagner, M.J., Wells, D.E., Parrish, J., Sadler, L.A., Blanton, S.H., Daiger, S.P., Wang, Z., Wilkie, P.J., and Weber, J.L.. 1992. Human chromosome 8 linkage map based on short tandem repeat polymorphisms: effect of genotying errors. *Genomics* **14**:114-152.

Urabe, K., Kimura, A., Harada, F., Iwanaga, T., and Sasazuki, T. 1990. Gene conversion in steriod 21-hydroxylase genes. **Am. J. Hum. Genet. 46**:1178-1186.

Vallada, H.P., Collins, J.E., Dunham, I., Dawson, E., Morray, R.M., Gill, M., Collier, D.A. 1994. Genetic mapping of 14 short tandem repeat polymorphisms on human chromoroms 22. *Hum. Genet.* **93**:688-690.

Verkerk, A.J.M.H., Pieretti, M., Sutcliffe, J.S., Fu, Y.-H., Kuhl, D.P.A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F., Zhang, F., Eussen, B.E., van Ommen, G.J.B., Blonden, L.A.J., Riggins, G.J., Chastain, J.L., Kunst, C.B., Galjaard, H., Caskey, C.T., Nelson, D.L., Oostra, B.A., Warren, S.T. 1991. Identification of a gene (FMR-1) containing a CCG repeat coincident with a breakpoint cluster region exhibitig length variation in fragile X syndrome. *Cell* **65**:905-914.

Wang, Z. and Weber, J.L. 1992. Continuous linkage map of human chromosome 14 short tandem repeat polymorphisms. *Genomics* **13**:532-536.

Warren, A.C., Slaugenhaupt, S.A., Lewis, J.G., Chakravarti, A., and Antonarakis, S.E. 1989. A genetic linkage map of 17 markers on human chromosome 21. *Genomics* **4**:579-591.

Weaver, R., Helms, C., Mishra, S.K., and Donis-Keller, H. 1992. Software for analysis and manipulation of genetic linkage data. *Am. J. Hum. Genet.* **50**:1267-1274.

Weber, J.L. 1990. Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. *Genomics* **7**:524-530.

Weber, J.L. and May, P.E. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**:388-396.

Weber, J.L. and Wong, C. 1993. Mutation of human short tandem repeats. *Hum Mol Genet* **2**: 1123-1128.

Weber, J.L., Kwited, A.E., and May, P.E. 1990. Dinucleotide repeat polymorphisms at the *D16S260*, *D16S261*, *D16S265*, *D16S266* and *D16S267* loci. *Nucl Acid Res* **18**:4034.

Weber, J.L., Polumeropoulos, M.H., May, P.E., Kwitek, A.E., Xioa, H., McPherson, J.D., and Wasmuth, J.J. 1991. Mapping of human chromosome 5 microsatellite DNA polymorphisms. *Genomics* **11**:695-700.

Weeks, D.E., Lathrop, G.M., and Ott J. 1993. Multipoint mapping under genetic interference. *Hum. Hered.* **43**:86-97.

Weiffenbach, B., Falls, K., Bricker, A., Hall, L., McMahon, J., Wasmuth, J., Funanage, V., and Donis-Keller, H. 1991. A genetic linkage map of human chromosome 5 with 60 RFLP loci. *Genomics* **10**:173-185.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., Lathrop, M. 1992. A Second Generation linkage map of the human genome. *Nature* **359**:794-801.

White, R.L. 1985. DNA sequence polymorphisms revatilize linkage, approaches in human genetics. *Trends in Genet.* **2**:177-181.

White, R.L. and Lalouel, J.-M. 1986. Investigation of genetic linkage in human families. *Adv. Hum. Genet.* **16**:121-228.

White, R.L. and Lalouel, J.-M. 1988. Sets of linked genetic markers for human chromosome. *Ann. Rev. Genet.* **22**:259-279.

White, R.L., Leppert, M., Bishop, D.T., Barker, D., Berkowitz, J., Brown, C., Callahan, P., Holm, T., and Jerominski, L. 1985. Construction of linkage maps with DNA markers for human chromosomes. *Nature* **313**:101-105.

White, R.L., Lalouel, J.-M., Nakamura, Y., Donis-Keller, H., Green, P., Bowden, D.W., Mathew, C.G.P., Easton, D.F., Robson, E.B., Morton, N.E., Gusella, J.F., Haines, J.L., Retief, A.E., Kidd, K.K., Murray, J.C., Lathrop, G.M., and H.M. Cann. 1990. The CEPH consortium primary linkage map of human chromosome 10. *Genomics* **6**:393-412.

Whitmore, S.A., Apostolou, S., Lane, S., Nancarrow, J.K., Phillips, H.A., Richards, R.I., Sutherland, G.R., and Callen, D.F. 1993. Isolation and charaterisation of transcribed sequences from a chromosome 16 hn-cDNA library and the physical mapping of genes and transcribed sequences using a high resolution somatic cell panel of human chromosome 16. *Genomics* **20**:169-175.

Wicking, C. and Williamson, B. 1991. From linked marker to gene. *Trends Genet.* **7**:288-293.

Wilkie, A.O.M., Higgs, D.R., Rack, K.A., Buckle, V.J., Spurr, N.K., Fischel-Ghodsian, N., Ceccherini, I., Brown, W.R.A., Harris, P.C. 1991. Stable Length Polymorphism of up to 260 Kb at the tip of the short arm of human chromosome 16. *Cell* **64**:595-606.

Wilkie, P.J., Krizman, D.B., and Weber, J.L. 1992. Linkage map of human chromosome 9 microsatellite polymorphisms. *Genomics* **12**:607-609.

Wilkie, P.J., Polymeropoulos, M.H., Trent, J.M., Small, K.W., and Weber, J.L. 1993. Genetic and physical map of 11 short tandem repeat polymorphisms on human chromosome 6. *Genomics* **15**:225-227.

Willard, H.F., Waye, J.S., Skolnick, M.H., Schwartz, C.E., Powers, V.E., and England, S.B. 1986. Detection of restriction fragment length polymorphisms at the centromeres of human chromosomes by using chromosome-specific alpha-satellite DNA probes: implication for development of centromere-based genetic linkage maps. *Proc. Natl. Acad. Sci. USA* **83**:5611-5615.

Williamson, R., Bowcock, A., Kidd, K., Pearson, P., Schmidtke, J., Ceverha, P., Chipperfield, M., Cooper, D.N., Coutelle, C., Hewitt, J., Klinger, K., Langley, K., Beckmap, J., Tolley, M., and Maidak, B. 1991. Report for the DNA committee and catalogues of cloned and mapped genes, markers formatted for PCR and DNA polymorphisms. *Cytogenet. Cell Genet.* **58**:1190-1832.

Wilson, S.R. 1988. A major simplification in the preliminary ordering of linked loci. *Genetic Epidemiology* **5**:75-80.

Wu, C.F.J. 1983. On the convergence properties of the EM algorithm. *Ann. Stat.* **11**:95-103.

Yan, W., Boustany, R.-M.N., Konradi, C., Ozelius, L., Lerner, T., Trofatter, J.A., Julier, C., Breakefield, X.O., Gusella, J.F. and Haines, J.L. 1993. Localization of Juvenile, but not late-infantile, Neuronal Ceroid Lipofuscinosis on chromosome 16. *Am J Hum Genet* **52**: 89-95.

APPENDIX A

PUBLICATIONS AND MANUSCRIPTS

The majority of the material presented in this thesis has been published, is "in press", or has been submitted. These are detailed below where the role of the candidate is specified. Reference to the appropriate chapters in the thesis is indicated, where appropriate. Publications and manuscripts are included in the remainder of this appendix.

Addition of *MT*, *D16S10*, *D16S4*, and *D16S91* to the linkage map within 16q12.1-q22.1. Kozman, H.M., Gedeon, A.K., Whitmore, S., Suthers, G.K., Callen, D.F., Sutherland, G.R., and Mulley, J.C. 1991. *Genomics* **11**:756-759.

The candidate genotyped the 40 CEPH families for the loci *MT*, *D16S4*, *D16S10*, and *D16S91*, performed the multipoint linkage analysis for the map encompassing *FRA16B* and wrote the manuscript (chapter 3).

Fragile X Syndrome: Diagnosis using highly polymorphic microsatellite markers. Richards RI, Shen Y, Holman K, Kozman H, Hyland VJ, Mulley JC and Sutherland GR. 1991b. *Am J Hum Genet* 48:1051-1057.

The candidate constructed the background linkage map in the region of the fragile X syndrome.

Fragile X syndrome: Genetic localisation by linkage mapping of two microsatellite repeats FRAXAC1 and FRAXAC2 which immediately flank the fragile site. Richards, R.I., Holman, K., Kozman, H.M., Kremer, E., Lynch, M., Pritchard, M., Yu, S., Mulley, J., and Sutherland, G.R. 1991a. *J. Med. Genet.* **28**:818-823.

The candidate constructed the background linkage map in the region of the fragile X syndrome.

Dinucleotide repeat polymorphism at the D16S288 locus. Shen, Y., Holman, K., Thompson, A., Kozman, H., Callen, D.F., Sutherland, G.R., and Richards, R.I. 1991. *Nucl. Acids Res.* **19**:5445.

The candidate placed the *D16S288* locus onto the genetic map.

A (CA)n repeat polymorphism for the human skeletal muscle α-actinin gene ACTN2 and its localisation on the linkage map of chromosome 1. Beggs, A.H., Phillips, H.A., Kozman, H.M., Mulley, J.C., Wilton, S.D., Kunkel, L.M., and Laing, N.G. 1992. *Genomics* **13**:1314-1315.

The candidate performed the linkage analysis for the localisation of the $(AC)_n$ repeat polymorphism.

A multipoint genetic linkage map around the fragile site *FRA16A* on human chromosome 16. Kozman, H.M., Phillips, H.A., Sutherland, G.R., and Mulley, J.C. 1992. *Genet. (Life Sci. Adv.)* **11**:229-233.

The candidate genotyped *D16S79* and *D16S96* on the 40 CEPH families, performed the multipoint linkage analysis for the map encompassing *FRA16A* and wrote the manuscript (chapter 4).

Chromosome 16. In: A comprehensive genetic linkage map of the human genome. (NIH/CEPH Collaborative Mapping Group). Kozman, H.M., Sutherland, G.R., and Mulley, J.C.. 1992. *Science* **258**: 67-162.

The candidate constructed the linkage map of chromosome 16 for this collaborative report (chapter 6), and wrote the chromosome 16 section.

Identification of a highly polymorphic marker within intron 7 of the *ALAS2* gene and suggestion of at least two loci for X-linked sideroblastic anemia. Cox, T.C., Kozman, H.M., Raskind, W.H., May, B.K., and Mulley, J.C. 1992. *Hum. Mol. Genet.* **8**:639-641.

The candidate constructed the background linkage map in the region of the *ALAS2* gene.

Integration of the Cytogenetic and genetic linkage maps of human chromosome 16 using 50 physical intervals and 50 polymorphic loci. Kozman, H.M., Phillips, H.A., Callen, D.F., Sutherland, G.R., and J.C. Mulley. 1993. *Cytogenet. and Cell Genet.* **62***:194-198.*

The candidate performed the multipoint linkage analysis for the construction of the genetic linkage map of chromosome 16 (chapter 6) and wrote the manuscript.

Refined genetic localisation for Central Core Disease. Mulley, J.C., Kozman, H.M., Phillips, H.A., Gedeon, A.K., McCure, J.A., Iles, D.E., Gregg, R.G., Hogan, K., Couch, F.J., MacLennan, D.H., and Hann, E.A. 1993. *Am. J. Hum. Genet.* **52**:398-405.

The candidate constructed the background linkage map in the region of the central core disease and performed the localisation studies using the affected family.

A microsatellite marker within the duplicated *D16S79* locus has a null allele: significance for linkage mapping. Phillips, H.A., Thompsom, A.D., Kozman, H.K., Sutherland, G.R., and Mulley, J.C. 1993. *Cytogenet. Cell Genet.* **64**:131-134.

The candidate performed the linkage analysis for this microsatellite marker, to confirm its localisation adjacent to *FRA16A* (chapter 7).

Fine genetic mapping of the Batten disease locus (CLN3) by haplotype analysis and demonstration of allelic association with chromosome 16 microsatellite loci. Mitchison, H.M., Thompson,A.D., Mulley, J.M., Kozman, H.M., Richards, R.I., Callen, D.F., Stallings, r., Doggett, N.,

Attwood, J., McKay, T.R., Sutherland, G.R., Gardiner, R.M. 1993. *Genomics* **16**:455-460.

The candidate constructed the background linkage map in the region of the *CLN3* gene (chapter 5).

Genetic Mapping of the Batten disease locus (CLN3) to the interval D16S288-D16S383 by analysis of haplotypes and allelic association. Mitchison, H.M., Taschner, P.E.M., O'Rawe, A.M., de Vos, N., Phillips, H.A., Thompson, A.D., Kozman, H.M., Haines, J.L., Schlumpf, K., D'Arigo, K., Boustany, R.-M.N., Callen, D.F., Breuning, M.H., Gardiner, R.M., Mole, S.E., and Lerner, T.J. 1994. *Genomics* **22**:465-468.

The candidate constructed the background linkage map in the region of the *CLN3* gene (chapter 5).

Regional localisation of a second non-specific X-linked mental retardation gene (MRX19) to Xp22. Donnelly, A.J., Choo, A., Kozman, H.M., Gedeon, A.K., Danks, D.M., and Mulley, J.C. 1994. *Am. J. Med. Genet.* **15**:581-585.

The candidate constructed the background linkage map in the region of the *MRX19* gene.

A linkage map of microsatellite markers on the human X chromosome. Donnelly, A., Kozman, H.M., Gedeon, A.K., Webb, S., Lynch, M., Sutherland, G.R., Richards, R.I., and Mulley, J.C. 1994. *Genomics* **20**:363-370.

The candidate performed the linkage analysis for this PCR map of the X chromosome.

A PCR-based genetic linkage map of human chromosome 16. Shen Y., **Kozman, H.M.**, Thompson, A.D., Phillips, H.A., Holman, K., Nancarrow, J.K., Lane, S.A., Chen, L.Z., Apostolou, S., Doggett, N.A., Callen, D.F., Mulley, J.C., Sutherland, G.R., and R.I. Richards. 1994. *Genomics* **22**:68-76.

The candidate performed the multipoint linkage analysis for the high resolution PCR-based linkage map of chromosome 16 (chapter 7). This linkage map was incorporated into the thesis submitted by Yang Shen, who characterised and genotyped many of the markers. The thesis was entitled: *A High-Resolution Genetic Map of Human Chromosome 16 and Localisation of the MEF Gene* and was submitted to the University of Adelaide.

The CEPH consortium map of human chromosome 16. Kozman, H.M., Keith, T.P., Donis-Keller, H., White, R.L., Weissenbach, J., Dean, M., Vergnaud, G., Kidd, K., Gusella, J., Royle, N.J., Sutherland, G.R., and Mulley, J.C. 1994. *Genomics*, In Press.

The candidate coordinated the error checking among the chromosome 16 collaborating laboratories, carried out the multipoint linkage analysis for the construction of the CEPH consortium map of chromosome 16 (chapter 8), and wrote the manuscript.

Clinical and Linkage study of a large family with simple Ectopia Lentis linked to FBN1. Edwards, M.J., Challinor, C.J., Colley, P.J., Roberts, J., Partington, M.W., Holloway, G.E., Kozman, H.M., and Mulley, J.C. 1994. *Am. J. Med. Genet.* In Press.

The candidate constructed the linkage map in the region of the FBN1 gene, and assisted in the localisation of Ectopia Lentis to this region.

Autosomal dominant distal myopathy: linkage to chromosome 14. Laing, N.G., Laing, B.A., Meredith, C., Wilton, S.D., Robbins, P., Honeyman, K., Dorosz, S., Kozman, H., Mastaglia, F.L., and Kakulas, B.A. 1994. *(Submitted).*

The candidate performed the multipoint linkage analysis for the localisation of the dominant distal myopathy gene to chromosome 14.

Human chromosome 16, genetic and physical mapping. In: Encyclopaedia of Molecular Biology. Kozman, H.M., and Mulley, J.C. 1994. (Submitted).

The candidate collated the information for this article from current published material, and wrote the article. In addition, the candidate performed the multipoint linkage analysis for the linkage maps included in this article (chapter 7).

Kozman, H.M., Gedeon, A.K., Whitmore, S., Suthers, G.K., Callen, D.F., Sutherland, G.R. and Mulley, J.C. (1991) Addition of *MT*, *D16S10, D16S4*, and *D16S91* to the linkage map within 16q12.1-q22.1
Genomics, v. 11 (3) pp. 756-759

Richards, R.I., Shen, Y., Holman, K., Kozman, H., Hyland, V.J., Mulley, J.C., and Sutherland, G.R. (1991) Fragile X syndrome: diagnosis using highly polymorphic microsatellite markers

American Journal of Human Genetics, v. 48 (6), pp. 1051-1057

Richards, R.I., Holman, K., Kozman, H., Kremer, E., Lynch, M., Pritchard, M., Yu, S., Mulley, J., and Sutherland, G.R. (1991) Fragile X syndrome: genetic localisation by linkage mapping of two microsatellite repeats FRAXAC1 and FRAXAC2 which immediately flank the fragile site.
Journal of Medical Genetics, v. 28 (12), pp. 818-823

Shen, Y., Holman, K., Thompson, A., Kozman, H., Callen, D.F., Sutherland, G.R., and Richards, R.I., (1991) Dinucleotide repeat polymorphism at the D16S288 locus. *Nucleic Acids Research, v. 19 (19), pp. 5445*

Beggs, A.H., Phillips, H.A., Kozman, H., Mulley, J.C., Wilton, S.D., Kunkel, L.M., and Laing, N.G. (1992) A (CA)n repeat polymorphism for the human skeletal muscle alpha-actinin gene ACTN2 and its localization on the linkage map of chromosome 1. *Genomics, v. 13 (4), pp. 1314-1315*.

Kozman, H.M., Phillips, H.A., Sutherland, G.R., and Mulley, J.C., (1993) A multipoint genetic linkage map around the fragile site FRA16A on human chromosome 16.
Genetics Life Science Advances, v. 11, pp. 229-233.

Cox, T.C., Kozman, H.M., Raskind, W.H., May, B.K., and Mulley, J.C. (1992) Identification of a highly polymorphic marker within intron 7 of the *ALAS2* gene and suggestion of at least two loci for X-linked sideroblastic anemia.
Human Molecular Genetics, v. 1 (8), pp. 639-641.

Kozman, H.A., Sutherland, G.R., and Mulley, J.C. (NIH/CEPH Collaborative Mapping Group) (1992) A comprehensive genetic linkage map of the human genome. *Science, v. 258 (5079), pp. 67-86*

Kozman H.M., Phillips H.A., Callen D.F., Sutherland G.R., and Mulley J.C. (1993)
Integration of the cytogenetic and genetic linkage maps of human chromosome 16 using 50 physical intervals and 50 polymorphic loci.
Cytogenetics and Cell Genetics, v. 62 (4), pp. 194-198

Mulley, J.C., Kozman, H.M., Phillips, H.A., Gedeon, A.K., McCure, J.A., Gregg, Hogan, R.G., Couch, F.J.MacLennan, D.H. et al (1993) Refined genetic localization for central core disease.
American Journal of Human Genetics, v. 52 (2), pp. 398-405

Phillips, H.A., Thompson, A.D., Kozman, H.M., Sutherland, G.R., and Mulley, J.C. (1993) A microsatellite marker within the duplicated D16S79 locus has a null allele; significance for linkage mapping.
Cytogenetics and Cell Genetics, v. 64 (2), pp. 131-132.

Thompson, A.D, Mulley, J.M., Kozman H.M., Richards, R.I., Callen D.F., Stallings, R., Doggett, N., Attwood, J., McKay, T.R., Sutherland, G.R., and Gardiner, R.M. (1993) Fine genetic mapping of the Batten diseasae locus (CLN3) by haplotype analysis and demonstration of allelic association with chromosome 16 microstellite loci.
Genetics, v. 16 (2), pp. 455-460

Mitchison, H.M., Taschner, P.E., O'Rawe, A.M., DeVos, N., Phillips, H.A., Thompson, A.D., Kozman, H.M., et al (1994) Genetic mapping of the Batten disease locus (CLN3) to the interval D16S288-D16S383 by analysis of haplotypes and allelic association.
Genomics, v. 22 (2), pp. 465-468.

Donnelly, A.J., Andy Choo, K.H., Kozman, H.M., Gedeon, A. K., Danks, D.M., Mulley, J.C. (1994) Regional localisation of a non-specific x-linked mental retardation gene *(MRXl9)* to xp22.
American Journal of Medical Genetics, v. 51 (4), pp. 581-585.

Donnelly, A., Kozman, H., Gedeon, A.K., Webb, S., Lynch, M., Sutherland, G.R., Richards, R.I., Mulley, J.C. (1994) A linkage map of microsatellite markers on the human X chromosome.
Genomics, v. 20 (3), pp. 363-370.

Shen, Y., Kozman, H.M., Thompson, A., Phillips, H.A., Holman, K., Nancarrow, J., Lane, S., Chen, L.Z., Apostolou, S., Doggett, N.A., et al (1994) A PCR-based genetic linkage map of human chromosome 16.
Genomics, v. 22 (1), pp. 68-76.

Kozman, H.M., Keith, T.P., Donis-Keller, H., White, R.L., Weissenbach, J., Dean, M., Vergnaud, G., Kidd, K., Gusella, J., Royle, N.J., et al (1995) The CEPH consortium linkage map of human chromosome 16.
Genomics, v. 25 (1), pp. 44-58.

Edwards, M.J., Challinor, C.J., Colley, P.W., Roberts, J., Partington, M.W., Hollway, G.E., Kozman, H.M., Mulley, J.C. (1994) Clinical and linkage study of a large family with simple ectopia lentis linked to FBN1.
American Journal of Medical Genetics, v. 53 (1), pp. 65-71..

Autosomal dominant distal myopathy: linkage to chromosome 14.

N.G. Laing, B.A. Laing, C. Meredith*, S.D. Wilton, P. Robbins#, K. Honeyman*, S. Dorosz, H. Kozman^, F.L. Mastaglia, B.A. Kakulas. Australian Neuromuscular Research Institute, QEII Medical Centre, Perth, Australia; *Edith Cowan University, Perth, Australia; #Department of Pathology, QEII Medical Centre, Perth, Australia; ^Department of Cytogenetics and Molecular Genetics, Adelaide Children's Hospital, Adelaide, Australia.

Running head: Autosomal dominant distal myopathy

Corresponding author:

Dr Nigel G Laing, Australian Neuromuscular Research Institute, 4th Floor 'A' Block, QEII Medical Centre, Perth, Western Australia 6009, Australia.

Telephone: 61-9-389-3487; FAX: 61-9-389-3487; Email: nlaing@uniwa.uwa.edu.au

Summary

We have studied a family segregating a form of autosomal dominant distal myopathy (MIM 160500) containing nine living affected individuals. The myopathy in this family is closest in clinical phenotype to that first described by Gowers in 1902. A search for linkage was conducted using microsatellite, VNTR and RFLP markers. In total, 87 markers on all chromosomes except chromosome 12, were run. Positive linkage was obtained with 14 out of 15 markers tested on chromosome 14 with little indication of linkage elsewhere in the genome. Maximum two point LOD scores of 2.60 at recombination fraction Θ = 0.00 were obtained for the markers MYH7 and D14S64 - the family structure precludes a two-point LOD score of 3 or greater. Recombinations with D14S72 and D14S49 indicate that this distal myopathy locus (MPD1) should lie between these markers. A multipoint analysis assuming 100% penetrance using the markers D14S72, D14S50, MYH7, D14S64, D14S54 and D14S49 gave a LOD score of exactly 3 at MYH7. Analysis at a penetrance of 80% gave a LOD score of 2.8 at this marker. This probable localisation of a gene (MPD1) for distal myopathy on chromosome 14 should allow other investigators studying distal myopathy families to test this region for linkage in the various distal myopathies to consolidate the linkage or demonstrate the likely heterogeneity.

Introduction

The distal myopathies are a genetically heterogeneous group of disorders showing both autosomal dominant and autosomal recessive inheritance, with diverse phenotypic features and pathological changes (Mastaglia, 1991). Gowers (1902)