GENOMICS OF CHROMOSOME 11:

EXCLUSION MAPPING OF EPISODIC ATAXIA ON HUMAN CHROMOSOMES 6p AND 17

CHARACTERIZATION OF DNA PROBES WHICH DETECT RESTRICTION FRAGMENT LENGTH POLYMORPHISMS ON HUMAN CHROMOSOME 11q

A DATABASE MANAGEMENT SYSTEM FOR THE ENTRY, ORGANIZATION, AND TRANSFER OF FAMILY GENOTYPES FOR LINKAGE ANALYSIS

by

Christopher J. Dubay

A DISSERTATION

Presented to the Department of Medical Genetics and the Oregon
Health Sciences University School of Medicine
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

May 1990

APPROVED:

[redacted]f Thesis)

[redacted]

(Chairman, Graduate Council)

DEDICATION

This thesis is dedicated to the spirit of research, as exemplified in the Litt Lab: where there is always someone to share an idea or problem with, to support or critique, and to talk and laugh. And especially to Mike Litt who has helped me become a better scientist, and critical thinker.

This work must also be inscribed with the names and deeds of all those who supported its creation.

To Laura, who supported me in every way possible, this thesis could not have been done without you. Your help as a friend, and lab assistant, is truly excellent. You have helped make the going, and not just the getting there, good.

To my family who have been mentors for my lifetime. All of the encouragement and perspective you offer have been a comfort, joy, and inspiration.

To the professors at OHSU who have taken the time to teach courses, and provide education. The didactic process course work you help provide is the basis for any contributions to the scientific community I will make. You have demonstrated to me the importance of teaching, and of being open to learning, in classrooms, offices, and hallways.

ACKNOWLEDGMENTS

Litt Lab:

Mike Litt
Ruth Litt
Marilyn Jones
Thas Phromchotikul
Xue Ya Lowe
Zheng Guo
Musen Al-Dhalimy
Jeff Luty

Cheryl Maslen
Nancy Payne
Jill Kellogg
Sophia Kondoleon

NIH Grant: R01-NS25370

NIH Grant: R01-GM32500

The Tartar Trust

Oregon Medical Research Foundation

Glen Evans

Mark Lathrop, CEPH

Jurg Ott

The Yale Human Gene Mapping Library

CONTENTS

LIST of TABLES

NOTE: All tables appear at the end of the section they are referenced in.

Chapter 1:

Chapter 2:

List of Figures

NOTE: All figures appear at the end of the section they are referenced in.

Introduction:

Chapter 1:

Chapter 2:

Abbreviations

ATCC	American Type Culture Collection
CEPH	Centre d'Etude du Polymorphisme Humain
HGML	Human Gene Mapping Library
PMMS	Pedigree/Marker Management System
PIC	Polymorphic Information Content
RFLP	Restriction Fragment Length Polymorphism
VNTR	Variable Number of Tandem Repeats
t	Theta (Recombination Fraction)
cM	centiMorgans (Unit of map distance)
LOD	Log of Odds
YAC	Yeast Artificial Chromosome
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
mRNA	Messenger RNA
EA	Episodic Ataxia
SCA	Spinocerebellar Ataxia
TSC	Tuberous Sclerosis
AT	Ataxia Telangectasia
MEN	Multiple Endocrine Neoplasia type 1
NOD mouse	Nonobese Diabetic mouse
SCA	Spinocerebellar Ataxia
bp	Base Pairs (of DNA)
kb	Kilo-Base Pairs (1000bp)
Mb	Mega-Base Pairs (1000kb)

Thesis Introduction

ABSTRACT

The scientific community has identified a goal: mapping the human genome. In pursuit of this goal the development and application of appropriate technologies in many fields is being approached, and a new word has been coined for this process: Genomics. Genomics is the search for keys to the structure and organization of a genome. One method of describing the organization of a genome is creating a genetic map. Genetic mapping determines the general location of genes and polymorphic loci on chromosomes and their position relative to each other based on the observed segregation patterns of these markers. Genetic mapping, together with techniques which produce physical maps, forms the basis for mapping genomes. The genomics of human chromosome 11q, approached with the techniques of genetic mapping, is the underlying topic of this thesis. Three chapters form the body of this thesis:

1. Determination of a genetic location for episodic ataxia (EA) in three study families has been approached using the technique of exclusion mapping. A mapped gene locus for spinocerebellar ataxia (SCA), and regions of chromosome 17 for which linkage maps exist, have been excluded as locations for EA genes in three study families. Chromosome 11q in the region of 11q24-qter, has been identified by others in our lab as a putative location for a locus associated with Episodic Ataxia (EA) in a family affected with the disease.

2. Cosmids from a chromosome 11q library have been screened for their ability to detect restriction fragment length polymorphisms (RFLPs). Six cosmids have been found which allow determination of 8 polymorphic genetic loci. These loci have been typed in CEPH reference families, and added to the genetic linkage map of chromosome 11q.

3. A database management system has been developed to facilitate the collection, processing, and analysis of genotype data. The Pedigree/Marker Management System (PMMS) is a menu-driven database management system that allows for entry and reporting of phenotypes and family structures. It also allows for the translation of these data between the many formats it is used in: datasets for use with the LINKAGE suite of genetic analysis programs, datasets for inclusion with the CEPH genetic database format, and datasets entered on laboratory workstations. The PMMS is user friendly, and provides documentation including step-by-step procedures for its use in various applications.

I. Introduction

A. Overview of Thesis

This thesis consists of an introduction, three chapters, and a discussion. The work in this thesis is a result of research carried out in Dr. Mike Litt's laboratory at the Oregon Health Sciences University. It began with a gene mapping study for autosomal dominant episodic ataxia

(EA), and when Marilyn Jones working in our lab found an indication of linkage of EA to the long arm of chromosome 11 (11q), our attention turned to developing polymorphic markers on 11q to better map EA.

I will begin the introduction with a description of genomics, as a conceptual framework for this thesis. I will then describe my focus in this thesis: the genomics of human chromosome 11q, as a foundation for the significance of this thesis. I will then relate the development of methodologies for the study of the specific areas of genomics I have worked on, as background for the approach taken in each of the chapters.

B. Genomics

1. Mapping the Human Genome

The scientific community has identified a goal: mapping the human genome. This goal has been deemed worthy of funding on a large scale by the United States government. In pursuit of this goal, the development and application of appropriate technologies in many fields is being approached. A new word has been coined for this process: Genomics. The word genomics, in it's "ics" suffix, refers to keys to the riddles providing understanding of a branch of study, as the Greeks referred to in "mathematics" and "metaphysics"; a method of attacking problems. Genomics is the search for keys to the structure and organization of a genome. The two main approaches used in this search are genetic mapping and physical mapping.

Genetic mapping determines the general location of genes on chromosomes and their positions relative to each other. One method of genetic mapping called linkage analysis is based on observed segregation patterns of genes and other genetic markers in families. Physical mapping is concerned with ordering of physical markers: creating restriction maps, defining cytogenetic landmarks using chromosomal staining patterns to describe "bands", and relating these to gene locations and observed anomalies (e.g. translocations, breakpoints, etc.). Physical mapping at the most basic level is the determination of the nucleotide sequence for regions of the genome. This sequence can then be scanned in an attempt to identify sequences that describe genes, in terms of their coding and control regions. Having sequences available would greatly assist in the important task of identifying disease genes that have been mapped to a given region.

Although the sequence of the genome is the ultimate goal, the two operations of mapping and sequencing must go hand-in-hand; blind sequencing of the human genome would not be as useful or interesting as sequencing regions of expressed genes. Currently technological issues of how we can efficiently determine, store, and analyze the sequence make genetic mapping the working edge of the genome project. Relating the emerging genetic map to physical maps will hopefully provide insights into how our genome is organized, and allow relation of genetic to physical distances.

T.H. Roderick of the Jackson Laboratory coined the term genomics (59) to describe a discipline that combines cell and molecular biology with classical genetics, and which is fostered by computational science. All of these approaches to understanding development and disease should be viewed as parts of the same analytic process. The interdisciplinary approach of genomics requires molecular biologists and biochemists, human and somatic cell geneticists, cytogeneticists, population and evolutionary biologists, genetic epidemiologists, clinical geneticists, and computer scientists to come together to approach topics including:

- Chromosomal assignments of genes and DNA fragments by genetic and physical mapping.

- Nucleic acid sequences of cloned genes and other interesting portions of the genome.

- Spatial distribution of gene families and homologous gene sequences that define amino acid sequence domains.

- Comparative analysis of genomes to yield structural, functional, and evolutionary insights.

- Patterns of organization within genomes giving insights into gene regulation and development.

- Methods for genomic cloning, restriction mapping, and DNA sequencing.

- Computational methodologies for manipulation and analysis of DNA and protein sequence data.

- Analysis of genetic linkage data in pursuit of information on inherited disorders.

- Development of experimental, computational, and database management techniques of broad applicability for obtaining or using data on genome organization.

In this thesis I will focus on applying the first and last two of these approaches to the genomics of chromosome 11q.

2. Genomics of Human Chromosome 11q

The long arm of human chromosome 11 (11q) represents 2.76% of the human genome in terms of cytogenetically observed length, and 2.3% of the male autosomal genome in terms of genetic length. So of the ~100,000 genes postulated to exist in our genome, roughly 2500 genes could be expected to reside on 11q. Some of these genes have been identified and shown to play important roles in the human phenotype. Variant alleles of these genes are responsible for a number of diseases. I will now describe some of these genes and diseases in an effort to underscore the importance of this region of the genome.

Multiple endocrine neoplasia (MEN) is a predisposition to hyperplasia of the parathyroid glands, and to hyperplasia or tumors of the anterior pituitary and the endocrine pancreas. MEN type 1 is inherited as an autosomal dominant trait, and has been mapped to the long arm of chromosome 11 (11q12.12)(60). This region of 11q has shown to be consistently lost in 7 MEN1-associated parathyroid tumors which were typed for multiple RFLPs from the region, suggesting that an anti-oncogene may be present in this region.

Ataxia Telangiectasia (AT) (97,98) is a disorder of childhood which is characterized by progressive cerebellar ataxia, hypersensitivity of fibroblasts to ionizing radiation (99), a sixty-one fold increase in incidence of cancer in Caucasian patients (184 fold increase in blacks), characteristic chromosomal rearrangements in lymphocytes (usually involving chromosomes 7 and 14), absent or hypoplasic thymus with cellular and humoral deficiencies, elevated serum alpha fetoprotein, premature ageing, and endocrine disorders (e.g. insulin-resistant diabetes mellitus). Individuals heterozygous for the AT gene are generally healthy, but their cells show a sensitivity to ionizing radiation which is intermediate between that of normal individuals and affected homozygotes (100). Heterozygous females are at increased risk for breast cancer (101). The carrier frequency of the allele causing the AT phenotype is estimated to be between 0.5 and 5.0%. Four clinically indistinguishable complementation groups exist among AT affecteds (A,C,D and E) based on inter-group fusion cell lines exhibiting normal radiation sensitivity. Linkage analysis of five group A families produced a maximum LOD score of 3.63 at a zero recombination distance with THY1, a gene localized to chromosome 11q22.3 that reveals a two allele RFLP. When 3 non-group A and 20 unassigned AT families were included, the maximum LOD score with THY1 increased to 4.33 at theta of 0.1 centiMorgans, and linkage to anonymous DNA marker pYNB3.12 (also from 11q22-q23) was also shown (LOD=5.58 at theta 0.08) (62). There has been description of genetic heterogeneity in AT (i.e. complementation groups), but so far only one genetic locus has been characterized.

Tuberous sclerosis (TSC) is an autosomal dominant disorder characterized by hamartomas (i.e. abnormal cell and tissue arrangements) in skin, brain, and kidneys (102). The frequency of TSC is estimated to be 1 in 10,000, with new mutations being observed at the rate of 1 in 60,000 to 1 in 84,000 live births. Linkage of TSC with 11q was demonstrated in 15 affected families. TSC showed a maximum LOD of 3.26, at theta equal to 8cM, with probe MCT128.1 (D11S144) localized to 11q22-11q23 (63).

Asthma and rhinitis are symptoms of responses to commonly inhaled antigens termed atopy. Autosomal dominant inheritance of atopy has been observed in seven study families, and a linkage with probe pMS.51 (103) localized to 11q, has been described (LOD=5.58 at 10.5% recombination fraction) (64).

The nonobese diabetic mouse (NOD mouse) is an animal model for insulin-dependent diabetes. A polygenic basis for susceptibility to this disease has been demonstrated in NOD mice by the determination of three recessive loci required for development of overt diabetes (65). One of the loci is localized proximally to the Thy-1/Alp-1 cluster on mouse chromosome 9, which is homologous to the human THY-1 and ALP-1 region on human chromosome 11q (66,75). This suggests that 11q is a potential site for a diabetogenic locus analogous to the NOD mouse locus. Mapping of genes in this region of human chromosome 11q will thus help in the mapping of the mouse genome (125).

The most common non-Robertsonian constitutional translocation t(11;22)(q23-q11), and other cytogenetically similar but unique rearrangements which are associated with Ewing sarcoma, and peripheral neuroepithelioma, involve breakpoints in the region 11q23-q24 (71). Also t(11;14) and t(9;11) translocations associated with leukemia occur in this region (71). Jacobson's syndrome (70) is a described clinical entity characterized by moderate dysmorphic features and retardation, caused by monosomy at sub-band 11q24.1.

Many members of the immunoglobulin supergene family localize to 11q22-23 (72,73,74). Genes for a neural cell adhesion molecule, and for the variable region related cell surface antigen Thy-1, have demonstrable homology with immunoglobulin T3 chains, and are also located in this same region. This series of genes is maintained in a homologous region of the mouse genome (75). These observations make 11q23-q24 a region of outstanding interest with regard to the evolution of the immunoglobulin superfamily.

Other known diseases which have been shown to be linked to 11q include acute intermittent porphyria (67), and apolipoprotein complex A1-C3-A4 abnormalities (67). Other known genes which have been mapped to chromosome 11q include the dopamine D2 receptor (11q22-q23)(104), which is a candidate gene for many neuropsychiatric disorders including alcoholism and drug addiction, and subunit VIII of cytochrome C oxidase (68).

We can see that chromosome 11q contains some important and interesting regions of the genome. This synopsis represents only a small fraction of the total number of genes, and diseases related to them, expected to be found on chromosome 11q. I submit that pursuing the genomics of chromosome 11q will prove to have significant benefits. I will now describe some the methodologies which may be applied in this pursuit.

C. Methodologies

1. Overview

In this section of the introduction I will describe some of the methodologies which may be applied in genomics, many of which I have used in my work on this thesis. I will begin with an introduction to linkage analysis as a strategy for genetic mapping, and a description of a basic tool of this strategy: genetic markers. I will then discuss how linkage analysis can be applied to mapping genetic diseases. Next, I will describe a process important in facilitating mapping of genes and genetic diseases: the construction of genetic maps. And finally, I will discuss genetic databases, and specifically a database system I have developed for supporting the application of these methodologies: the Pedigree/Marker Management System (PMMS).

2. Linkage Mapping

The task of gene mapping is to determine on which chromosome a given gene resides, and to what region of that chromosome it is localized. To address this task one may use a technique called linkage analysis.

Linkage analysis is concerned with establishment of linkage between two genetic loci. Consider two gene loci, each having two alleles: Aa and Bb. There are 4 possible haplotypes (i.e. sets of genes received from each parent, one from each locus: AB Ab aB ab), and 10 possible genotypes (i.e. sets of two haplotypes). In general for a two locus system, locus 1 having n_1 alleles, and locus 2 having n_2 alleles, there are n_1n_2 possible haplotypes and, $((n_1n_2) \times ((n_1n_2) + 1)) / 2$ possible genotypes (of which n_1n_2 are homozygous) (22).

An individual who is heterozygous at both loci with the haplotype set AB/ab, can produce 4 possible gametes: AB aB Ab and ab. If the two loci are unlinked, the expected ratio of these gametes is 1:1:1:1. If the two loci are linked, gametes Ab and aB are recombinants, that is they represent a recombination of the portions of an individual's genetic composition received from each haplotype.

Genetic linkage is indicated when few recombinants occur. The closeness of the linkage is indicated by the recombination fraction: theta. Theta is the proportion of recombinants among all haplotypes produced by a parent. If two loci are unlinked, theta equals 1/2, indicating that the two loci are just as likely to be inherited separately as together. If two loci are tightly linked, theta equals 0, indicating that they are always inherited together. It is important to note that theta is calculated from observations of products of meioses in families, so linkage studies involve families.

One can not always distinguish between recombinant and non-recombinant haplotypes. The phase of the parent (i.e. which alleles of the two loci were inherited as a haplotype from a grandparent) must be known to determine if recombination has occurred. A phase-known double backcross (e.g. AB/ab X ab/ab) is a fully informative mating, allowing direct counting of recombinants in offspring. In 1955 Morton (13) developed a set of tables to allow estimation of the recombination fraction in all possible matings, and to calculate a likelihood that the estimated theta was correct based on the observations. Various estimations of theta are made for an observed family, and likelihood of each is calculated, to give a most likely theta.

Once theta has been determined at the maximum likelihood, one can translate the value into a map distance. Many algorithms exist for this computation, and some of them are quite complicated. The simplest mapping function is distance equals theta, but this is true only for small values of theta, and is not additive for thetas between two adjoining marker pairs due to the effect of multiple crossing over. To account for multiple crossovers, the formula $t_{12}=t_1+t_2-(2 \times t_1 t_2)$, was used by Haldane (79) to calculate a genetic distance equal to $-1/2 \times \ln(1-2t)$. This mapping function provides a calculation of map distance, expressed in map or crossover units called centiMorgans, corresponding to percentage recombination. Kosambi (28) extended Haldane's assumptions to account for the effects of positive interference (i.e. a crossover inhibiting the formation of subsequent crossovers in its neighborhood). Figure 1 compares these mapping functions.

It is important to note that map distances thus calculated refer to an average number of points for exchange per chromosome, and do not necessarily reflect physical distance. Parameters governing numbers of crossovers per physical distance are not consistent throughout the genome. There are hot spots for recombination (108), and other factors, including sex (109,110), can play a role.

If a large number of equally dispersed gene loci spanning the entire genome were known, the sum of the distances between them would equal the total map length of the human genome. Based on counts of the chiasmata formed at crossover sites during meioses, Renwick (80) has estimated the total map size of the human genome as 27.5 Morgans for females and 38.5 for males, giving the often quoted average of 33 Morgans. Dividing by the number of base pairs of DNA in our genome (3.3×10^9) we can give a rough estimate of a million base pairs per 0.01 Morgans (1 centiMorgan or 1 cM).

In 1947 Haldane and Smith (81) analyzed the likelihood of segregation of two diseases in seventeen pedigrees. They showed how to calculate the correct probability of occurrence of the observed multigenerational segregation taking into account gene frequencies and mutations. They described a probability P as a function of F and theta, where F is the collection of phenotypes in the family, and theta is the recombination frequency. When calculating P one can use various values of theta in an attempt to maximize P. The ratio of P(F,theta)/P(F, 1/2) is called the odds ratio of linkage (probability linkage at theta divided by probability of non-linkage where theta equals 0.5). The logarithm of this ratio was given the term "lods".

In 1955 Morton (13) presented a streamlined approach for calculating "lods", and the maximum likelihood recombination fraction. He used the term LOD score for the likelihood ratio, which he denoted Z. Morton's tables for Z of theta for various types of two-generation families represented a real breakthrough. Instead of going through an explanation of the calculation of these scores, suffice it to say that a computer program can calculate them (25) based on variable assumptions of mode of inheritance, inbreeding, gene frequencies, epistasis (the dependence of one polymorphism upon another), and degree of penetrance for definable liability classes.

Since it is assumed that the phenotypes in one family occur independently of those in other unrelated families, LOD scores for different families can be added up for a total LOD score (this is equivalent to the multiplication of probabilities). In reporting LOD scores one usually presents horizontally a series of recombination fractions (e.g. 0.01, 0.05, 0.1, 0.2, 0.3, and 0.4), and vertically the LOD scores for each family at each fraction, with the sum of the LODs at the bottom. The recombination fraction with the highest LOD is the most likely. A LOD score of 1 to 2 is interesting, from 2 to 3 is suggestive, and greater than 3 is considered proof of linkage (22). Alternatively, if the LOD score is less than negative 2, this proves the exclusion of linkage within an interval of the recombination fraction. A good rule of thumb for an autosomal dominant disease, with a fully informative mating, is that each offspring will add 0.3 to the LOD score if the markers are linked with no observed recombination. Exceptions to the generally accepted significance of a LOD score greater than 3 do exist when many independent markers are being tested against an unknown trait locus (30). LOD scores of 5 or more must be reached to prove linkage when the numbers of markers tested for linkage approaches 100.

In reporting LOD scores it is also important to consider a confidence interval for the estimation of theta. The standard confidence interval corresponds to the region of theta values defined by the intersection of the likelihood curve over all values of theta (0 to 0.5) with a line described by the maximum LOD score minus one. (Figure 2)

3. Genetic Markers

In using linkage mapping to localize a genetic disease we study the cosegregation of the disease phenotype in affected families with genetic marker loci. A genetic marker is a system that allows determination of a phenotype which allows inference of a genotype at a genetic locus, preferably at a locus with a known genetic location (I will discuss how markers are mapped in section 5.b).

The main requirement for a marker to be useful in a linkage study is that it be polymorphic. A marker can be considered polymorphic if it has a frequency of 95% or less for its most common allele (91). Since we want as many people to be informative (i.e. heterozygous) at that locus in our study families as possible, the more polymorphic the better. Two measures of a marker's informativeness are commonly used: heterozygosity, which is the observed frequency of heterozygotes among unrelated individuals, and polymorphic information content (PIC) (31), which is the probability that any given offspring will be informative at the locus. Markers with PICs of 0.5 or greater are considered highly informative.

There are many types of genetic marker loci. The first linkage studies were performed using polymorphic blood cell antigens and blood proteins as markers. Cytogenetically observable chromosome variations can be used as markers, and provided the first autosomal assignment of a marker in 1968 (the Duffy blood group with a cytogenetic chromosome 1 heteromorphism (78)). Currently the most developed source of polymorphisms in the human genome are restriction fragment length polymorphisms (RFLP).

A restriction fragment length polymorphism is an observed difference in sizes of DNA fragments due to variations in the DNA sequence between individuals. To determine a phenotype at a locus using a RFLP, DNA from an individual is incubated with type II restriction endonucleases (reviewed in 82), commonly referred to as restriction enzymes, that cause cleavage of DNA at specific base sequences, resulting in fragments of defined length. The fragments are separated by electrophoresis on agarose gels, and transferred to nitrocellulose or nylon membranes, creating Southern blots (39). These blots are then hybridized with radioactively labeled probes. The patterns of hybridization, visualized with autoradiography, describe alleles for the genetic (or "marker") locus which is detected by the probe. This is the phenotype of the individual at that marker locus (Figure 3).

Differences between individuals in the lengths of restriction fragments detected by a probe for a marker locus could result from many kinds of genotypic differences: one or more individual bases could differ resulting in the loss or formation of a cleavage site, or, insertion or deletion of blocks of DNA within a fragment could alter its size. RFLP alleles are always expressed in a co-dominant manner.

In a RFLP linkage study, phenotypes determined in this manner at various marker loci are combined with pedigree information and affection status to determine the likelihood for co-segregation of a disease phenotype and the marker loci phenotypes. There is a basic assumption inherent in linkage studies that all phenotypes must be known with certainty. To minimize errors in typing individuals which would contradict this assumption, all RFLP results are entered into a laboratory workstation, printed out, and then independently re-checked against the original autoradiograph from which the phenotypes were initially scored.

As stated previously, if co-segregation is observed, and is found to be significant (LOD score >3), we can state that the locus and the putative disease gene are genetically linked. This linkage analysis also works to exclude the gene from regions of the genome surrounding the marker in the cases where co-segregation is not observed (LOD score <-2). Many closely spaced marker loci are typed in families segregating for the disease to extend these areas of detection to regions of, and eventually the entire length of, the genome.

When the evidence for linkage or exclusion is inconclusive (i.e. LOD scores for a given marker are less than 3 and greater than -2), three methods of improving them may be approached. The first is to get more families segregating for the disease, and to type the markers in these new families, since LOD scores are additive between families. For example, if two families each gave a positive LOD score of 2 between a marker and disease locus, one could report a cumulative LOD of 4, which establishes linkage. A second method is to try typing more markers from the region in an attempt to find some which are more informative in the families being studied.

The third method which can increase significance of a LOD score in a linkage study is multipoint linkage analysis (25). In multipoint linkage analysis one uses a set of closely linked markers with known genetic distances between them, and determines likelihoods for locations of a locus with respect to the set (Figure 4). The requirement for being able to perform multipoint analysis is a genetic map, which gives the distances between marker loci. In section 5.b I will examine the process of creating genetic maps.

4. Disease Mapping

Since Botstein et al. first described a strategy for mapping disease genes using RFLPs and linkage analysis (31), the technique has been applied with success in a number of studies: Huntington's disease (41), cystic fibrosis (42), neurofibromatosis type 1 (43), chronic granulomatous disease (44), and others (reviewed in 11) including Ataxia Telangectasia which resides on chromosome 11q (62). The mapping of a disease gene is significant in providing a tool for prenatal or early diagnosis of the disease (32), and for carrier detection. It is also an important first step in obtaining a clone of that gene.

The feasibility of disease gene mapping is increasing as more and more probes are being generated and placed on linkage maps. Three groups have made great strides towards a linkage map of the entire genome. Ray White's laboratory has distributed preliminary linkage maps incorporating many highly informative probes spanning 23 chromosomes, and has published a number of detailed maps for specific chromosomes (17,18,37). Collaborative Research Inc. has published a map consisting of 403 polymorphic loci, including 393 RFLPs, which they estimate to provide detectable linkage with 95% of the human genome (19). The Centre d'Etude du Polymorphisme Humain (CEPH) is currently working to combine linkage map results from many researchers. CEPH members are provided with DNAs from many large three generation families, which they are obliged to type with any markers they have developed. Results of these typings are then returned to CEPH, which compiles them, and makes them available as a computerized dataset to its members for use in linkage studies (23,107). CEPH has also used this data set to compile collaborative maps, the first of which has recently been published for chromosome 10 (53). Two linkage maps of probes from the 11q region are available (54)(55).

One stumbling block which exists in this approach is that there is very little or no coordination of linkage map publication between the various groups which make probes available. This makes for multiple maps, none of which gives relative genetic positions for all the available markers, thus hampering our ability to perform multipoint linkage analyses. As mentioned before, efforts to overcome this limitation are in progress at CEPH and other centers. Advances in collaboration technology, and the development of well-connected genetics workstations to facilitate data sharing will help alleviate this problem.

Factors influencing a disease's amenability to the gene mapping process include: the existence of large families available for DNA sampling, with multiple affected members, a clear inheritance pattern, complete penetrance (optimally with a test for heterozygotes), and an early age of onset. Discovery of linkage indicates the presence of a major gene in the etiology of a disease (20,21).

Moving from a closely linked marker locus to a disease gene (a strategy termed reverse genetics), is not a trivial process. Applications of techniques involving long-range restriction mapping (33), new DNA sequencing technology (120,121), artificial yeast chromosomes (119), somatic cell genetics (12), and improved chromosome walking strategies using linking and jumping libraries (47) have made headway in going from marker to gene, and provide hope that the task will be simplified in the coming years. Success in cloning the genes for Duchenne's muscular dystrophy (48), cystic fibrosis (49,50,51), and retinoblastoma (52) have demonstrated that this process is feasible and useful.

Upon isolation of a cloned gene, one may begin to ask questions regarding the nature of the protein product, its biochemical properties, and its developmental and tissue specific expression. This process will hopefully lead to answers regarding the biochemical lesion and its relation to the primary genetic defect. Many new techniques are becoming available for characterization of single gene defects from a clone of the gene. Expression of cloned genes can be studied using Northern blotting, in-situ hybridization techniques, and antibodies made against synthetic peptides. Additionally, transgenic mice, heterozygous for a dominant disease gene linked to an inducible promoter transfected into their genome, can be used as animal models in the development of treatments for a disease.

The success of reverse genetics illustrates the significance of mapping a gene. Clearly those genes which have been mapped, with many demonstrated closely-linked marker loci, will be among the first cloned. Having a genetic map of available marker loci has been essential in the successful application of this technique in genomics.

5. Making Genetic Maps

a. Marker Development

In this section I will discuss the construction of a genetic map of the human genome in terms of strategies for its development, and the state of the current map today. I will describe my approach to extending this map in the region of chromosome 11q, and demonstrate how collaboration becomes a focal point for development of genetic maps. As I said before, the sequence map is the ultimate map, but without the computing resources to even use such a map if it existed today, we are currently working to develop another form of map, known as a genetic map.

Linkage analysis is a powerful tool for gene mapping. But to use linkage analysis most efficiently, a map of many genetic loci must be developed. Genetic maps can only be generated by the time-consuming process of typing many families with a series of markers. The results of the typing are first used to determine the genetic distance between pairs of markers using two-point linkage analysis. When enough closely-grouped markers are defined, they can be subjected to multipoint analysis to determine the most likely order of markers and relative distances between them, based on observations made of many meioses. This data may then be combined with results from other markers, typed in the same families, and new most likely distances and orders of the markers may be recalculated. Data on genetic maps is dynamic. Often sections of the map will be re-oriented on the chromosome, and distances between markers will change. Additionally, differences in the frequency of recombination in males and females require that each region of the genome have two maps: one for each sex. Figure 5 shows a typical genetic map (from 54).

In 1980 Botstein, White, Skolnick, and Davis (31) published a strategy for construction of a genetic linkage map of the human genome using RFLPs for the purpose of mapping genetic diseases. They proposed development of a "perfect" linkage map of RFLPs evenly spaced throughout the genome at 20cM intervals, allowing theoretical detection of any linkage within 10cM of a mapped marker. Their strategy is outlined below:

- o Develop single copy DNA probes to detect DNA sequence polymorphisms. These define genetic marker loci.

o If necessary, expand loci to include more polymorphism.

o Test linkage to other probes, arrange loci on a map.

o Analyze disease pedigrees, link traits to marker loci.

o Use the markers to predict the occurrence of a trait.

A number of questions arise upon consideration of the feasibility of this task. How many markers are needed? How polymorphic must each marker be? How many families are required to establish linkage? How much polymorphism might one expect in the human genome?

The general formula for the number of markers one would require to span the entire genome is 3300/d, where d is the acceptable distance in centiMorgans between two markers, which sets the number of markers at about 165 for a map with a 20 centiMorgan resolution.

In using genetic markers for linkage studies, the question of how much polymorphism is required is based on how the degree of polymorphism at a marker locus influences the probability of detecting linkage with another locus. The informativeness of a marker in this context can be represented by the probability that a child who has inherited a rare dominant allele from a parent at one locus will be informative for linkage between that locus and marker locus. This probability is the polymorphism information content, or PIC (as described previously), of this marker. PICs greater than 0.5 are considered highly informative. In terms of how much polymorphism is required the rule is the more (i.e. the higher the PIC) the better.

Estimation of the number of families required to establish a linkage are based on the pedigree structure and family size (83). Multigenerational pedigrees (i.e. with available parents, grandparents, and children) and large sibships give the most information in linkage analysis (111). The expected LOD score from a family varies with the recombination fraction, such that to establish linkage with a LOD score of 3 would require 300 individuals at a recombination fraction of 0.3, and as few as 21 individuals (in three five-child families) between two tightly linked markers (e.g. theta equal to 0.001).

The frequency of polymorphisms in DNA has been approached from many directions. 28% of 71 proteins studied in European populations showed polymorphism based on visible electrophoretic variations (84). Assuming that these represent about a third of the nucleic acid sequence changes (i.e. those altering the net charge on the protein), and that many third base changes in codons are not reflected due to the degeneracy of the genetic code, it has been estimated that the frequency of DNA sequence polymorphism is about 0.001 per base pair among protein coding sequences. This can be considered as a lower limit of polymorphism, since sequences coding for mRNAs have been shown to diverge more slowly than the bulk of DNA (85). An upper limit for the extent of DNA polymorphism in mammals due to base pair changes can be derived from observation that no melting temperature depression is observed when DNA samples from wild-type and inbred mice are hetero-annealed, indicating heterozygosity of less than 0.5% (86). Using an approximation assuming no selection of sequences, with a conservative minor allele frequency of 10%, this corresponds to a 1.1% frequency of polymorphism in DNA due to base pair changes (87).

If we define S as the probability that a given restriction site is not polymorphic (such that $S=(1-p)^n$, where p equals the fraction of sites that are polymorphic, and n is the number of bases in the restriction enzyme recognition site), accounting for average restriction fragment length, appearance and disappearance of sites, and other factors, Upholt showed that the fraction of restriction fragments showing no polymorphism is approximately equal to $S^2(2-S)$ (88). For a four-cutter restriction enzyme this formula gives a range of 12.3% to 0.34% chance that a given restriction fragment is polymorphic, based on our upper and lower estimates of base polymorphism (and 17.5% to 0.5% for a six-cutter). Polymorphism in DNA can also be due to insertions and deletions, which are under-represented in the above estimate and serve to increase the expected amount of DNA polymorphism.

There is an increased rate of mutation at 5-methylcytosine (112), which has been theorized to cause the observed increase in frequency of RFLPs detected with restriction enzymes containing the dinucleotide sequence CpG in their recognition sight (most notably TaqI and MspI) (113,114). This makes TaqI and MspI restriction enzymes especially efficient for the detection of RFLPs.

With these questions answered, we can begin to develop strategies for fulfilling the goals of Botstein's proposal. The issues involved with these strategies include: a source of marker DNA, screening techniques for RFLPs, selection for highly informative RFLPs, and efficient map production.

Sources of probes for detecting RFLPs initially were libraries of DNA in plasmid, cosmid, or bacteriophage vectors derived from total human DNA as a source of anonymous DNA segments, or from cDNA reverse transcribed from total human mRNA preparations as a source of translated DNA segments presumably corresponding to functional genes. Plasmids hold inserts of a few base pairs to thousands of bases, cosmids can contain inserts of up to 40 kb, and bacteriophage can hold replacement inserts of up to 20 kb. Vectors carrying large genomic DNA inserts should reveal more RFLPs than vectors carrying cDNAs (usually about 500bp to 5kb) since they contain inserts many times larger, which contain more sequence variation than transcribed regions (85). But large genomic inserts expected to contain a higher frequency of RFLPs due to their size are also more likely to contain interspersed repetitive sequences which may account for up to 10% of the genome (89). These interspersed-repeat DNA segments hybridize to fragments from many regions of the genome, making the patterns of fragments revealed by these probes too complex for easy interpretation. The development of techniques for blocking these repeats by pre-hybridizing the probes and target DNAs on Southern blots with a vast excess of total human DNA, which at the correct stringencies pair with the repetitive sequences and leave the single copy unique sequences free for probe/target hybridization, has greatly enhanced the usability of these large insert vectors (90).

By using human/rodent somatic cell hybrids (115), flow sorted chromosomes, or yeast artificial chromosomes (YACs), where only a single chromosome or fraction of a chromosome is represented, as sources for probes, the efficiency of marker discovery in a given region can be increased by focusing on a single chromosome or chromosome region. This "divide and conquer" strategy helps in the development of these markers.

These libraries are then screened for RFLPs. This process involves typing a series of unrelated individuals, whose DNA has been digested with a series of restriction enzymes, with many probes containing human inserts which have been isolated from the library, and looking for variable bands in the resulting restriction fragment pattern. As mentioned previously, TaqI and MspI are restriction enzymes which contain the mutation hot spot CpG dinucleotide in their recognition site, making them useful enzymes for such a screening. Another enzyme useful for screening because it has been empirically shown to reveal a high degree of polymorphism is RsaI (19). 40% of all probes listed in the Yale human gene mapping library reveal RFLPs with one of these three enzymes.

Once a polymorphism has been found in random individuals, in order for it to be useful in linkage studies, verification is needed that it is inherited in a Mendelian fashion. This is accomplished by typing the marker in families informative for the marker (e.g. at least one parent heterozygous), and observing that alleles are inherited correctly from parents to children.

The polymorphism is then characterized in terms of the sizes of the restriction fragments representing each allele, as well as the fragments that appear as constant bands. The frequency of each allele is calculated in a population of random, unrelated individuals. Care must be taken if these allele frequencies are used to calculate expectations of phenotypes in populations, since observed frequencies of alleles may vary between populations with different ethnic backgrounds.

If two polymorphic loci are revealed by a single probe, or by two probes known to be closely linked, alleles from the two systems should be tested for linkage disequilibrium (95,96). Linkage disequilibrium between alleles of two closely linked marker loci is often observed as the absence or reduced frequency of a haplotype or haplotypes in the population. If the two systems are not in total disequilibrium, they may be combined by using haplotypes of the two systems for typing in linkage studies, creating a compound locus with a greater overall heterozygosity.

The next step in making a new marker really useful is to determine its localization in the genome. If the marker came from a library of known chromosomal origin this task is simplified to verification of its localization, or localizing it to a sub-region of the original library source. Localization of markers to chromosomes can be approached using physical and genetic techniques. Physical techniques include use of somatic-cell hybrid mapping panels for both chromosomal (116) and regional (117) localization, and the use of cytogentic in-situ mapping techniques (118,58). A genetic approach to creating maps of marker loci is the topic of the next section.

b. Map Construction

In this section I will describe genetic maps in terms of how they are developed and used to map genetic markers and diseases.

The genetic map is a linear map connecting DNA markers which reside on the same chromosome, or chromosome region. As stated previously, a marker can be an anonymous segment of DNA or a region of a cloned gene, and it must demonstrate polymorphism which follows Mendelian inheritance patterns. The map describes two properties of each marker: first, the genetic distance (in Morgans) it lies from its two closest flanking markers, and second, the statistical odds that it is in the correct position with regard to flanking markers (Figure 5, Panels A & B). Additionally a map may describe the genetic distances between adjacent probes as observed in males and females, and when the distances for sexes are averaged. In general the amount of observed recombination in females is usually greater, making the female map longer, but there are regions where the opposite is true.

Since the genetic map is developed using linkage analysis, it is called a linkage map; and is subject to the same uncertainties regarding the confidence of a location that linkage analysis is. It can be very misleading to believe that these linkage maps are exact; on the contrary, in Panel C of Figure 5 one can see the uncertainty that is associated with each marker location.

One method of creating genetic maps is to determine phenotypes in a series of families for the set of markers to be mapped, and establish genetic distances between them using linkage analysis. New markers can be added to the map in relation to other previously mapped markers that have been typed in the same families. This work is facilitated by large multigenerational families in which many markers have been typed. The Centre d'Etude du Polymorphisme Humain or CEPH acts as a repository for 40 such families (currently being expanded to 60), and makes DNA samples from them available to researchers constructing genetic maps (69,93,107).

For a family to be useful in mapping two markers it must be informative at both marker loci, with the same parent heterozygous for both markers. As stated before, families are most informative when the phase of the markers, which defines the set of alleles in each parental haplotype allowing determination of recombination events, is known. Knowing the phenotypes of the grandparents makes it possible to infer the phase of the parents.

To determine phenotypes of CEPH families, the marker is first typed in the CEPH parents, to determine which families are informative. The marker is then typed in all members of families which have at least one heterozygous parent.

By typing many informative families with a series of markers we can create a dataset which when entered into a computer can be used to perform a linkage analysis. This analysis will take all available information into account and give a maximum likelihood recombination distance between each set of two markers, along with a likelihood (i.e. LOD score) that the calculated distance is correct given the observations. As stated previously, for the maximum likelihood recombination distance to be accepted as evidence of linkage, it must have a LOD score of greater than three. Since LOD scores are additive between families, by using the panel of CEPH families a significant likelihood can usually be reached for closely linked markers (e.g. theta less than 20 centiMorgans).

As mentioned previously, three groups have produced a number of linkage maps in an effort towards a complete linkage map of the human genome. Ray White's laboratory has published a number of chromosomal maps (17,18,37). Collaborative Research Inc. has published a map which they estimate to provide detectable linkage with 95% of the human genome (19). CEPH is currently working to combine linkage map results from many researchers which have been generated by running markers through the large three generation families CEPH makes available to its members and other investigators (23,107). CEPH has recently published its first collaborative map (53). Two linkage maps of probes from human chromosome 11q are currently available (54)(55).

However there are problems with these maps: they often contain no markers in common, and distances between markers on different maps cannot be combined. Overcoming this lack of coordination in linkage map publication between the various groups which make probes available will be achieved through advances in collaboration technology and genetic informatics. These advances involve the development of standardized datasets for organizing and reporting linkage analyses and related data to support this coordination. Such a dataset is the topic of the next section.

6. Genomic Database

a. PMMS

We have reviewed the fundamentals of two aspects of genomics; now I wish to turn to an aspect that I feel is extremely significant to the success of all genomical approaches: genetic informatics. Genetic informatics is concerned with the development of experimental, computational, and database management techniques of broad applicability for obtaining or using data on genome organization. In this discussion, the data on genome organization will be data associated with a linkage study. I will detail what data is associated with a linkage study, and describe a systems approach to its management. Then I will describe my pedigree/marker management system (PMMS) (27), and show how it is designed to address this approach. I will also describe a vision of what systems might extend PMMS in the near future.

In a linkage study, data on phenotypes is combined with family information to create datasets for linkage analysis by computers. The results of the analysis also form a dataset of distances and likelihoods based on a given model.

Phenotype data, at the most basic level, consists of the alleles observed for an individual at a given genetic locus. There is a large amount of data associated with a RFLP phenotype for an individual: information about the probe, the blot the observation was made on, and the genetic locus. The probe needs to be catalogued in terms of the vector used, the size and site of the DNA insert, where DNA from this vector is stored in the lab, who prepared it and when, who provided the probe, and any special instructions for its use. The blot used in typing needs to have each lane defined as to whose DNA is present, what restriction enzymes were used to digest each DNA sample, size standards used, volt hours the gel was run, and the percentage agarose in the gel, as well as information on how well the blot is working, and how often it has been used. Information on the genetic locus includes the expected sizes of the various alleles if the locus is a RFLP, the frequency of those alleles in the population, and the restriction enzyme and probe combination that reveals those alleles. The final aspect of phenotype information is the phenotype itself: which alleles were observed for which individual, on which blot, and with what probe.

Family data is comprised of information on each individual and their relationships in the family. Individual information includes: lab number, sample number, sample date, sample storage location, sex, date of birth, and if they are from a study family, information on their affection status, and dates of diagnoses. Family information is reduced to who an individual's mother and father are, all other biological relationships can be inferred from this. Other data for study families is disease specific: what disease phenotypes are required to consider an individual affected, age of onset, penetrance of disease, liability class, and mode of inheritance.

When a linkage analysis is run, using as input portions of the datasets previously described, the output is a series of likelihoods (usually LOD scores) for each family, recombination fraction, and model. The model for the analysis describes the mode of inheritance, penetrances (for each liability class), information on the markers for which linkage is being tested, including allele frequencies, what recombination fractions are to be tested, and if sex-specific recombination fractions are to be used. If the analysis is multipoint, information on fixed distances between mapped markers is also included.

Edwards has proposed a systems approach to linkage analysis (29) that provides a useful structure for framing and ordering the datasets I have described. Figure 6 shows the flow of family information and samples from the field to the lab where blots are made and probed, the results of which are translated into phenotypes and analyzed along with family information to produce LOD scores, which are summarized to produce a linkage map. I have developed PMMS as a tool to help with the information flow between each phase of this process.

Family information and samples gathered in the field are entered into a FAMILY database (See Figure 7 for database formats). Blots are designed on the computer, and their production coordinated by a BLOTS database. Appropriate markers to be run are chosen from a database of PROBES. When the blots have been hybridized with the probes and the autoradiographs developed, the genotypes are entered into a RESULTS database. The results file also serves an important function in error checking, since all genotype results are printed out after entry and independently re-checked against the original autoradiographs. All of these data entry steps have been facilitated on the Macintosh because of its visual nature and the resulting ease of learning and use. Detailed descriptions of each of these databases can be found in the PMMS documentation, which is the third section of this thesis.

The final inference step which combines data from each of these data files, and produces the LOD scores from which one may draw conclusions regarding linkage, uses computer programs which run on IBM-PC type microcomputers. The programs which perform them are, like so much IBM computer software, cryptic at best in their operation.

The PMMS serves as a bridge between the Apple Macintosh computers used in our lab for data collection, and the IBM-PC computers used for linkage analysis. In this situation, PMMS can be considered as an interface between these two areas, as well as a database management system in its own right. Its main goal is to organize and combine data gathered in the field and the laboratory into data ready for analysis by the LINKAGE program suite (24). Results from the LINKAGE package are then transferred into spread-sheets for use in tables and graphs. I plan to extend PMMS to automate this process: to track linkage analyses that have been performed, and to summarize their output in a database.

In closing I want to present a vision of what a genetics workstation might do for geneticists in the future. Imagine a workstation connected to a network that can access a central genetics database. To add data to this database you place an autoradiograph on a scanning device, and the workstation displays an electronic image of the result, and enhances it. The workstation uses information on the probe/blot combination you identify the autoradiograph as representing to display a pedigree of the individuals on the blot, with their interpreted phenotypes, allowing you to view its assignments and check them for consistent data. Once validated, the data is uploaded to the central database. You then open a display window on the workstation that has settings for all the parameters associated with a linkage analysis, select a set of markers and families to check linkage in, and receive an almost instantaneous graphic interpretation of the linkage calculations, which can be stored along with the settings for the analysis. This scenario would greatly simplify the process which is currently required to perform a computer linkage analysis, and is really not so far off. New detection systems for polymorphisms which allow phenotypes to be directly read by computers will also help increase the efficiency of this process. One of the main benefits of this scenario is that any researcher on the network could access the latest databased calculation of a genetic distance between two markers for use in their own study. And the increase in the quality and availability of genetic information would be reflected in increased quality and volume of genetic linkage results.

D. Application

1. Episodic Ataxia

In this section I will describe the process of applying the methodologies we have discussed in my research. I will relate the story of my work, and describe the underlying thread that brings my research together.

In summary, my first research objective was to exclude the location of a putative disease gene for Episodic Ataxia (EA) in a family segregating for the disease from two regions of the genome: the short arm of chromosome 6, and chromosome 17. The first because of a reported linkage of a similar neurological disease in the HLA region of this chromosome which could be allelic with EA. And the second because a good linkage map existed for markers on chromosome 17, allowing us to use multipoint linkage analysis to increase the regions of that chromosome we could exclude as compared with two-point analyses. I created the PMMS to help me in this research. As I was nearing the completion of the exclusion mapping, Marilyn Jones working in our lab found a suggestive linkage of EA to a marker on chromosome 11q (marker phi2-25 (D11S38), theta=0.06, LOD=2.68). After testing more than 30 other probes from chromosome 11 in an unsuccessful attempt to establish a more significant linkage, we decided to try and develop more genetic markers in this region of the genome.

2. Chromosome 11q Marker Development

Dr. Cheryl Maslen working in our lab had started this work before the EA linkage was found. She described seven new markers on 11q, and published linkage data, which I helped generate, that was extended in the CEPH families to put her markers on the most recent chromosome 11 map (54). I have worked to extend her results, and have characterized eight new RFLPs on 11q.

I conducted a search for more RFLPs by screening twenty cosmids from a library created by G.A. Evans and K.A. Lewis, derived from cell line TG 5D1-1, which contains 11q13.1-11qter as its only human component (57). My procedure was to make cesium chloride-banded DNA preps from each of the cosmids, and use whole cosmids as probes on southern blots containing DNA from six or more unrelated individuals, digested with at least three restriction enzymes (TaqI, MspI and RsaI) for an initial screening, since 40% of the 1500 probes in the HGML are polymorphic with one of those three enzymes. Three cosmids which were found to map to the region of our EA linked markers were screened for RFLPs with an additional six restriction enzymes.

All polymorphisms found were checked for Mendelian inheritance (Figure 8). The cosmid probes were mapped on a somatic cell hybrid mapping panel (Figure 9). To add these loci to the genetic map, I first ran them on blots containing DNA from parents of CEPH families to determine which families were informative for the marker systems (Figure 10). I ran all 8 markers in parents of 36 of the CEPH families. I have tabulated the frequencies of alleles fro each marker in this set of unrelated individuals, and calculated both percent heterozygosity and polymorphism information content (PIC) for each marker based on these frequencies.

I have calculated the linkage disequilibrium between alleles in marker loci revealed by the same cosmid. The informativeness of these markers not in significant disequilibrium can be increased by haplotyping individuals at these two loci.

Next, I ran the markers through the informative CEPH families to produce a dataset of phenotypes. I have merged this dataset into the CEPH data format, and have sent the resulting dataset to CEPH for inclusion in their Version 4 database. I have done two-point linkage analysis and interval mapping of my markers with markers in the CEPH version 3 database, and markers in a dataset provided by Mark Lathrop (54). All results of this work are presented in chapter 2.

In the course of this work I have applied most of the methodologies discussed previously. Chapters one and two of this thesis will form the basis for manuscripts to be submitted for publication to document this work.

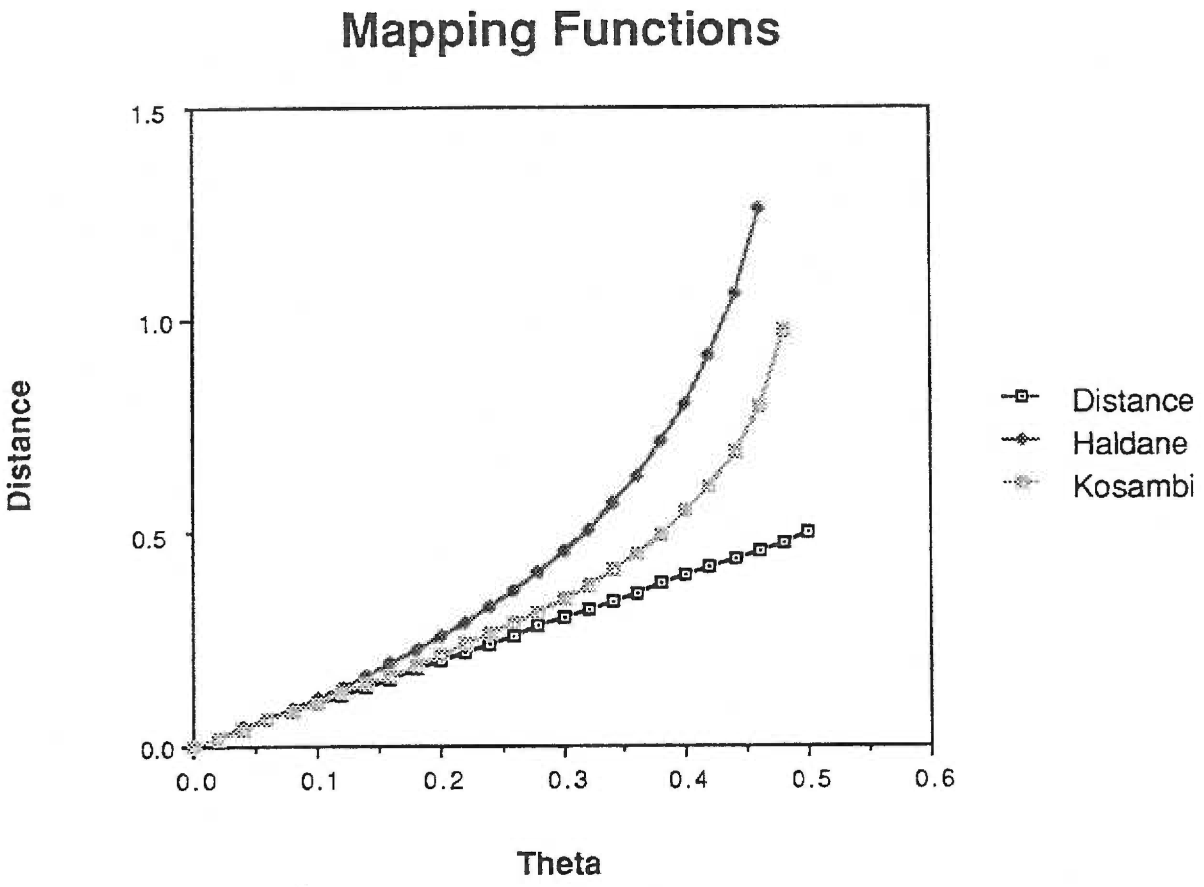


Figure 1. The above graph compares three mapping functions which relate the recombination fraction theta to genetic distance measured in Morgans.

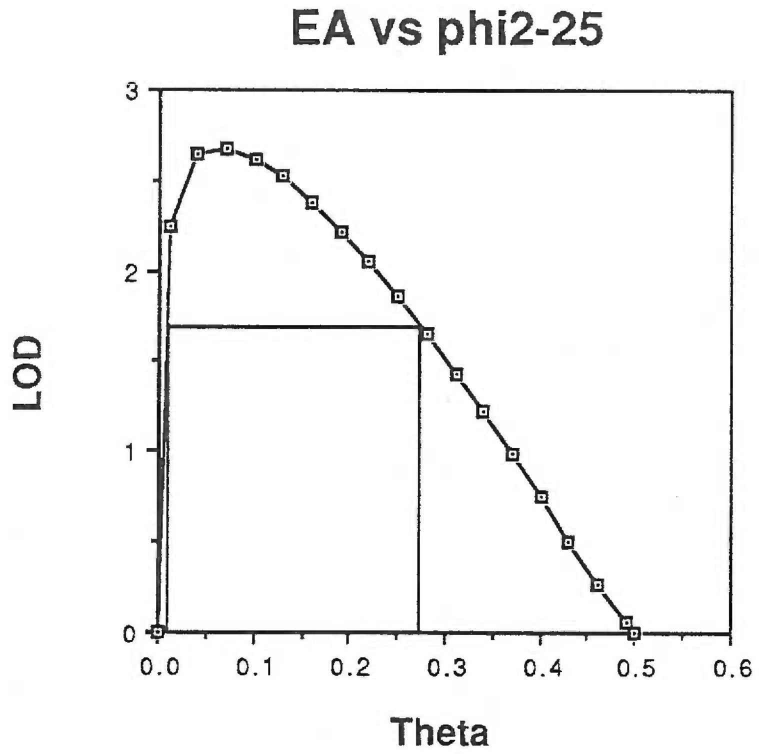


Figure 2. The above graph shows a typical LOD vs Theta graph with a LOD-1 confidence interval shown. The Maximum LOD occurs at theta of 0.63, and the confidence interval for theta is between 0.02 and 0.27.

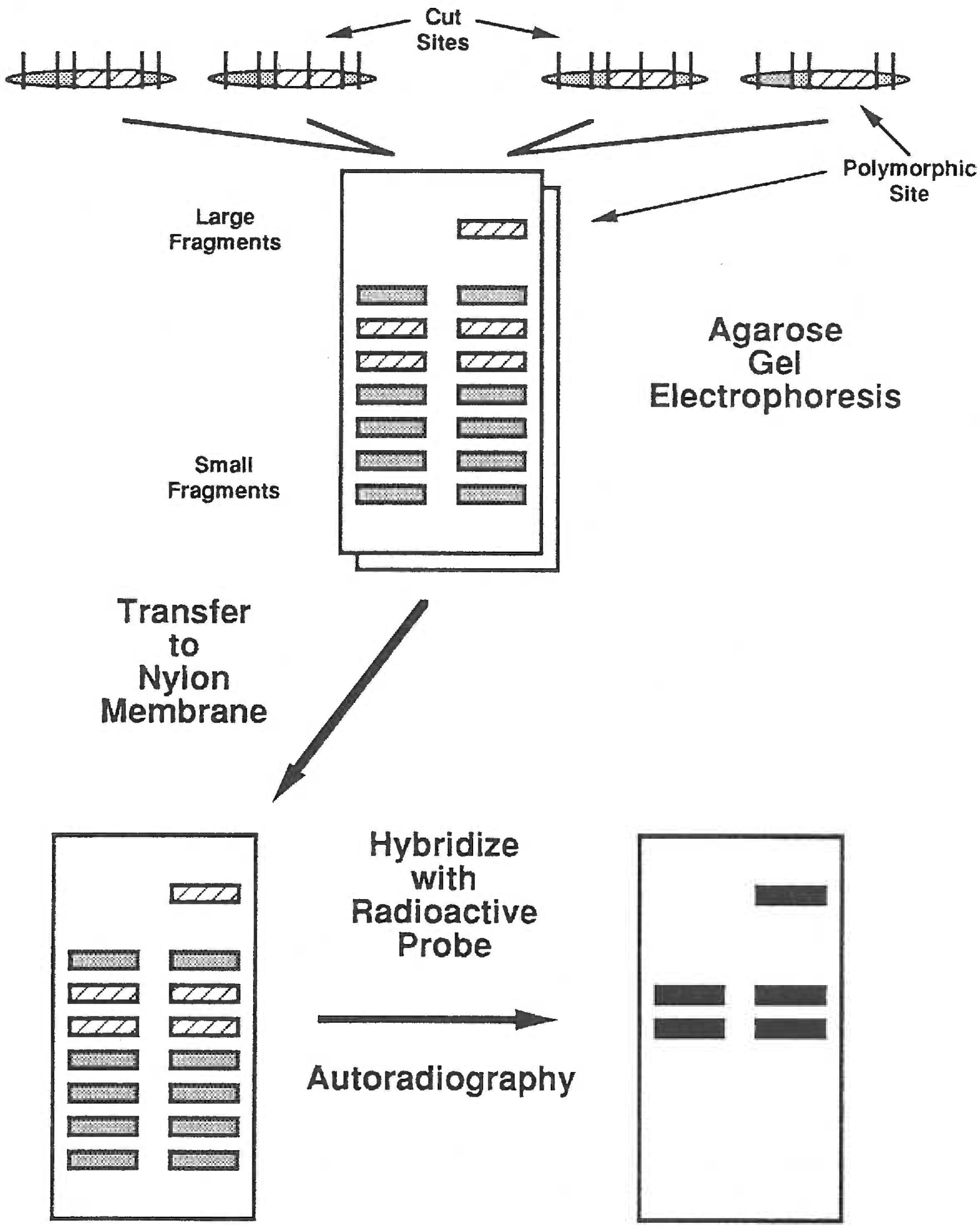


Figure 3. This is a diagramatic representation of how RFLP phenotypes are determined. DNA (shown as a pair of chromosomes each for two individuals) is cut with restriction enzymes which cleave the DNA at the labeled cut sites. The DNA fragments are then size fractionated by electrophoresis, transfered to a nylon membrane, probed with radioactively labeled RFLP probe DNA, corresponding to the light regions of the chromosomes shown at top, and the phenotype is revealed by autoradiography.

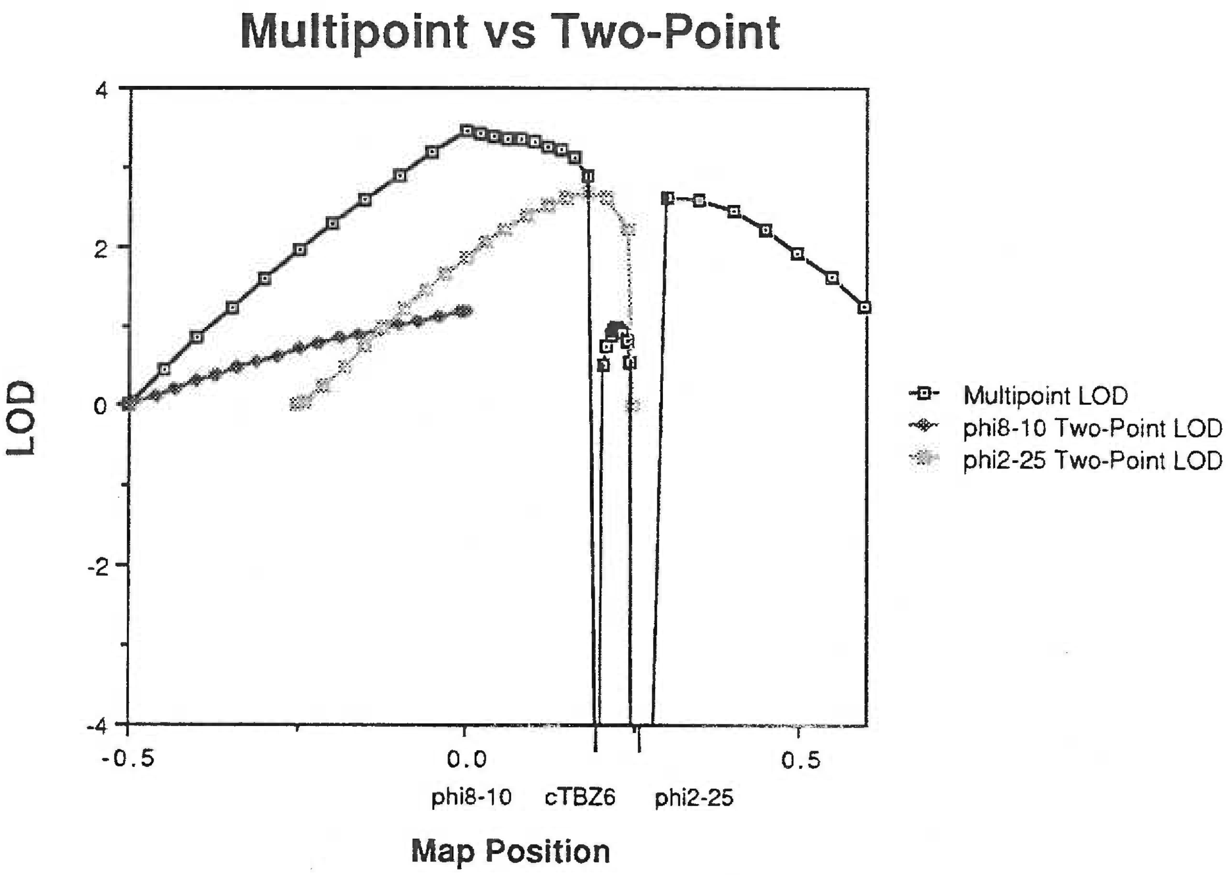


Figure 4. The above graph compairs a multipoint linkage analysis of EA with three markers from chromosome 11q, to two-point linkage analysis for two of the markers. The multipoint graph has been overlayed with the results from two-point analyses. Note the increase in LOD score multipoint analysis gives over the two-point maximum LODs in this example.

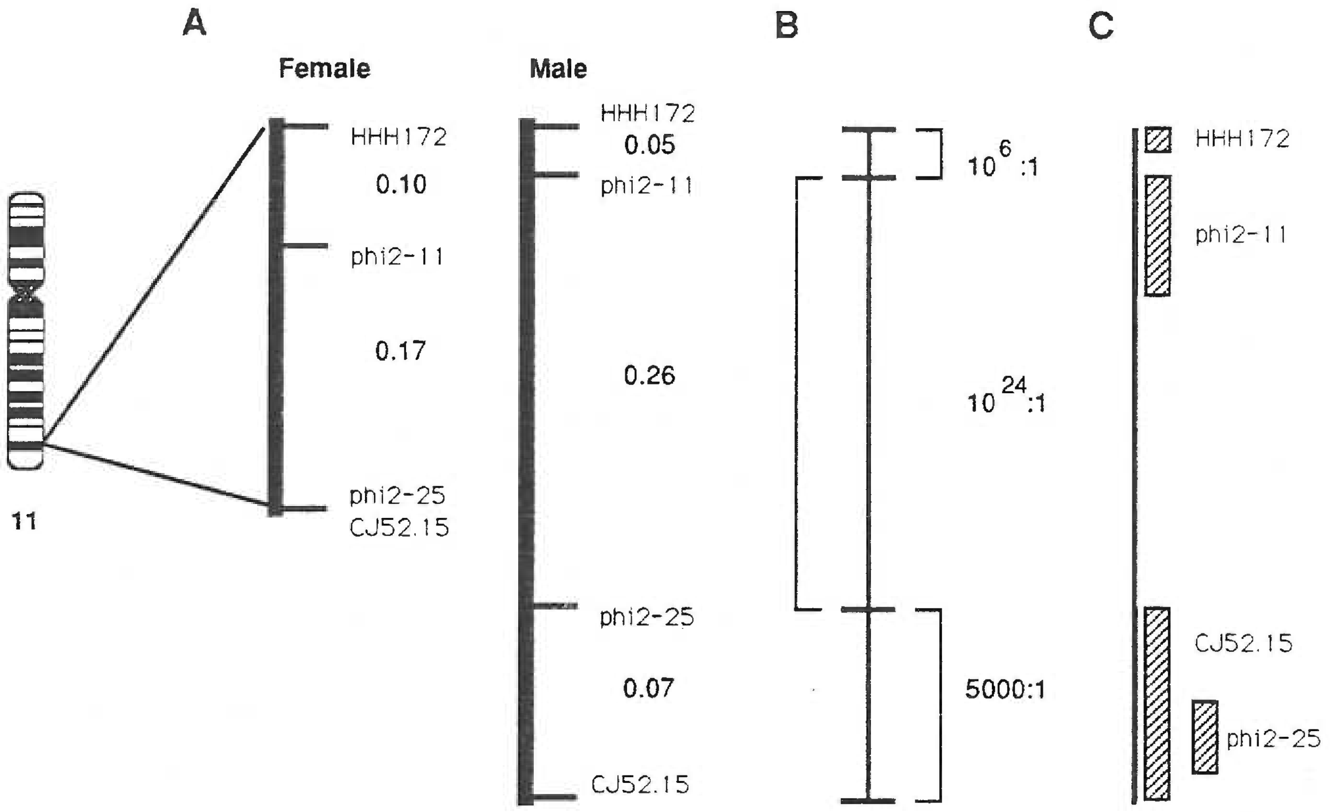


Figure 5. This figure , modeled after (54), shows A) male and female genetic maps for four markers, B) odds against inversion of location of two adjacent markers, and C) the LOD-1 confidence intervals for each marker's location.

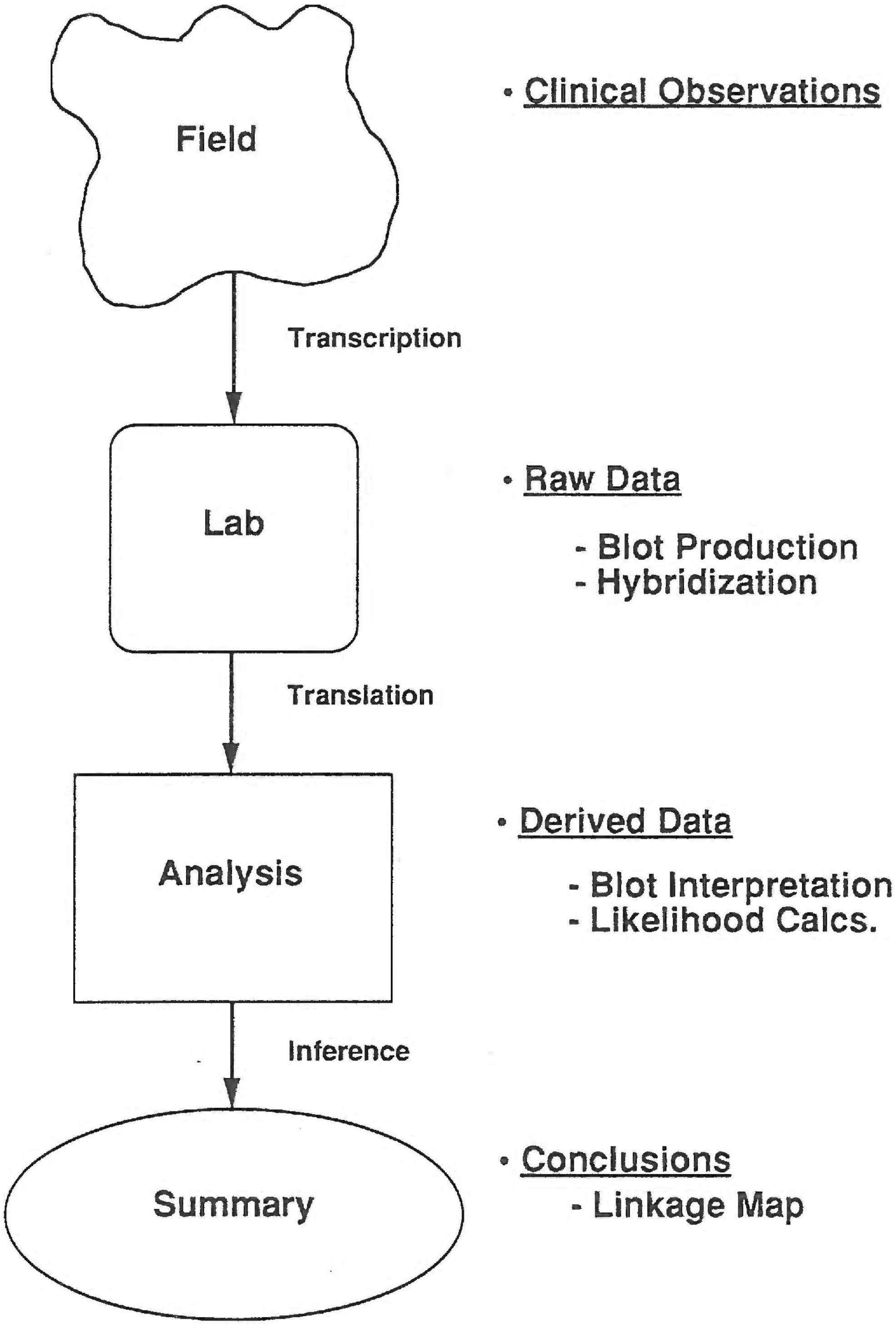
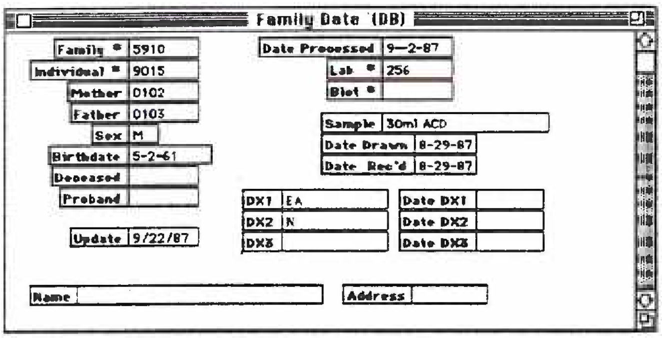


Figure 6. This figure represents a systems approach to linkage studies. Data flows from the field to the lab where it is translated for analysis. Inferences are made from the results of the analysis, and summarised as a basis for conclusions.



Family Data (DB)

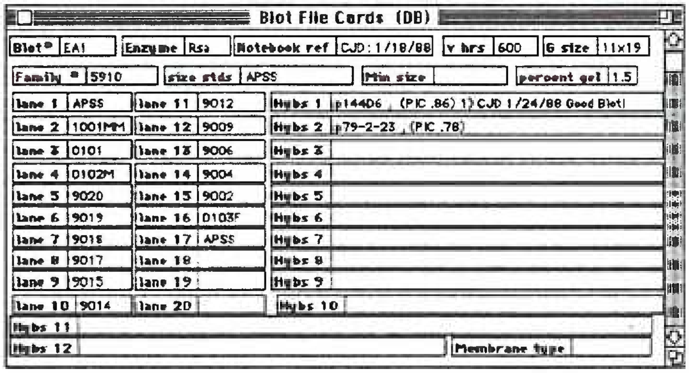
Family #	5910	Date Processed	9—2-87
Individual #	9015	Lab #	256
Mother	0102	Blot #	
Father	0103	Sample	30ml ACD
Sex	M	Date Drawn	8-29-87
Birthdate	5-2-61	Date Rec'd	8-29-87
Deceased			
Proband			

DX1	EA	Date DX1	
DX2	N	Date DX2	
DX3		Date DX3	

Update 9/22/87

Name

Address



Blot File Cards (DB)

Blot # EA1 | Enzyme Rsa | Notebook ref CJD: 1/18/88 | v hrs 600 | G size 11x19

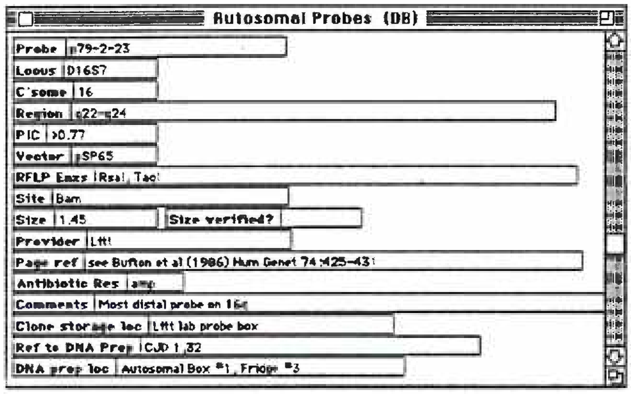
Family # 5910 | size stds APSS | Min size | percent gel 1.5

lane 1	APSS	lane 11	9012	Hybs 1	p144D6 , (PIC .86) 1) CJD 1/24/88 Good Blot!
lane 2	1001MM	lane 12	9009	Hybs 2	p79-2-23 , (PIC .78)
lane 3	0101	lane 13	9006	Hybs 3	
lane 4	0102M	lane 14	9004	Hybs 4	
lane 5	9020	lane 15	9002	Hybs 5	
lane 6	9019	lane 16	0103F	Hybs 6	
lane 7	9018	lane 17	APSS	Hybs 7	
lane 8	9017	lane 18		Hybs 8	
lane 9	9015	lane 19		Hybs 9	
lane 10	9014	lane 20		Hybs 10	

Hybs 11

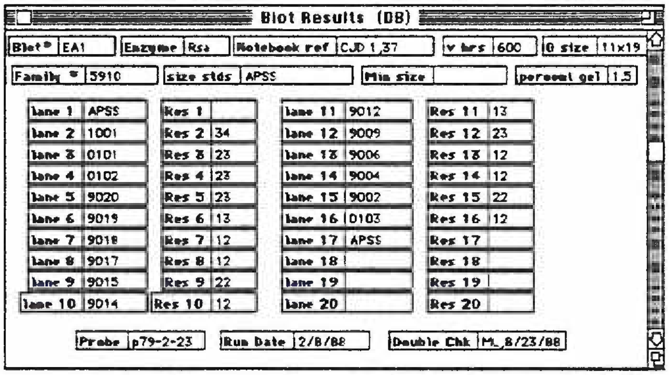
Hybs 12

Membrane type



Autosomal Probes (DB)

Probe	p79-2-23
Locus	D16S7
C'some	16
Region	p22-p24
PIC	>0.77
Vector	pSP65
RFLP Enzs	Rsal, Taq
Site	Bam
Size	1.45
Size verified?	
Provider	Lttt
Page ref	see Bufton et al (1986) Hum Genet 74:425-431
Antibiotic Res	amp
Comments	Most distal probe on 16p
Clone storage loc	Lttt lab probe box
Ref to DNA Prep	CJD 1,32
DNA prep loc	Autosomal Box #1, Fridge #3



Blot Results (DB)

Blot # EA1 | Enzyme Rsa | Notebook ref CJD 1,37 | v hrs 600 | G size 11x19

Family # 5910 | size stds APSS | Min size | percent gel 1.5

lane 1	APSS	Res 1		lane 11	9012	Res 11	13
lane 2	1001	Res 2	34	lane 12	9009	Res 12	23
lane 3	0101	Res 3	23	lane 13	9006	Res 13	12
lane 4	0102	Res 4	23	lane 14	9004	Res 14	12
lane 5	9020	Res 5	23	lane 15	9002	Res 15	22
lane 6	9019	Res 6	13	lane 16	0103	Res 16	12
lane 7	9018	Res 7	12	lane 17	APSS	Res 17	
lane 8	9017	Res 8	12	lane 18		Res 18	
lane 9	9015	Res 9	22	lane 19		Res 19	
lane 10	9014	Res 10	12	lane 20		Res 20	

Probe p79-2-23 | Run Date 2/8/88 | Double Chk ML 8/23/88

Figure 7. The four Macintosh screens shown here are the Microsoft Works representations of the PMMS files : FAMILY file, BLOT file, PROBE file, and RESULT file.

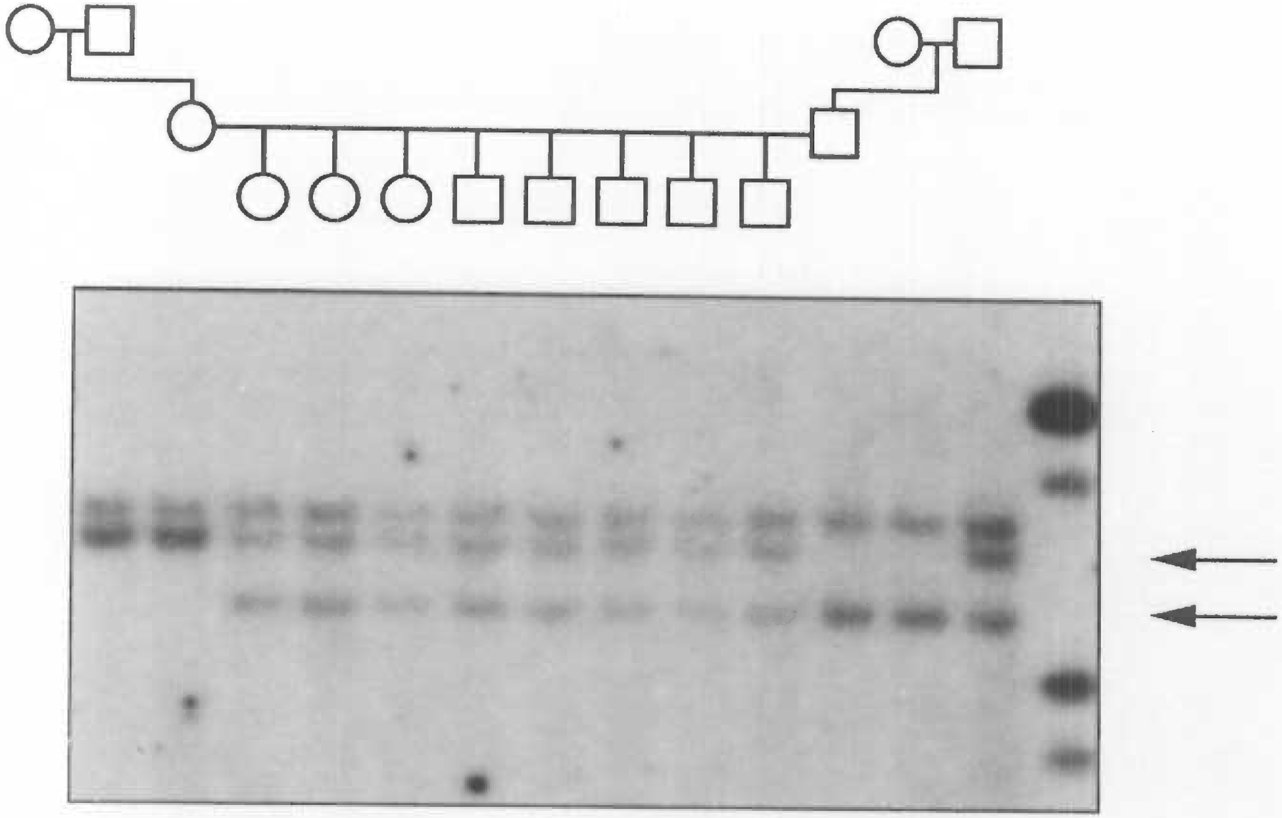


Figure 8. Southern blot of DNA from members of Utah reference family K1346 digested with RsaI, and probed with cosmid 3-27, demonstrating Mendelian inheritance of this RFLP. Arrows indicate polymorphic bands. (Last Lane APSS=Size Standards)

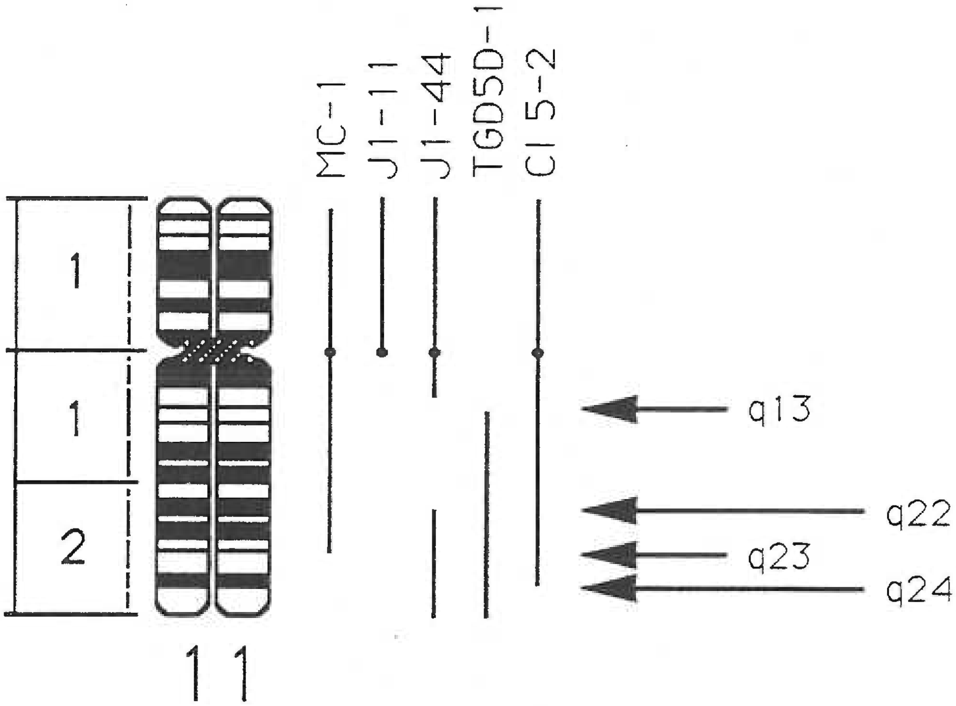


Figure 9. This is the somatic cell mapping panel used to localize probes to sub-regions of chromosome 11.

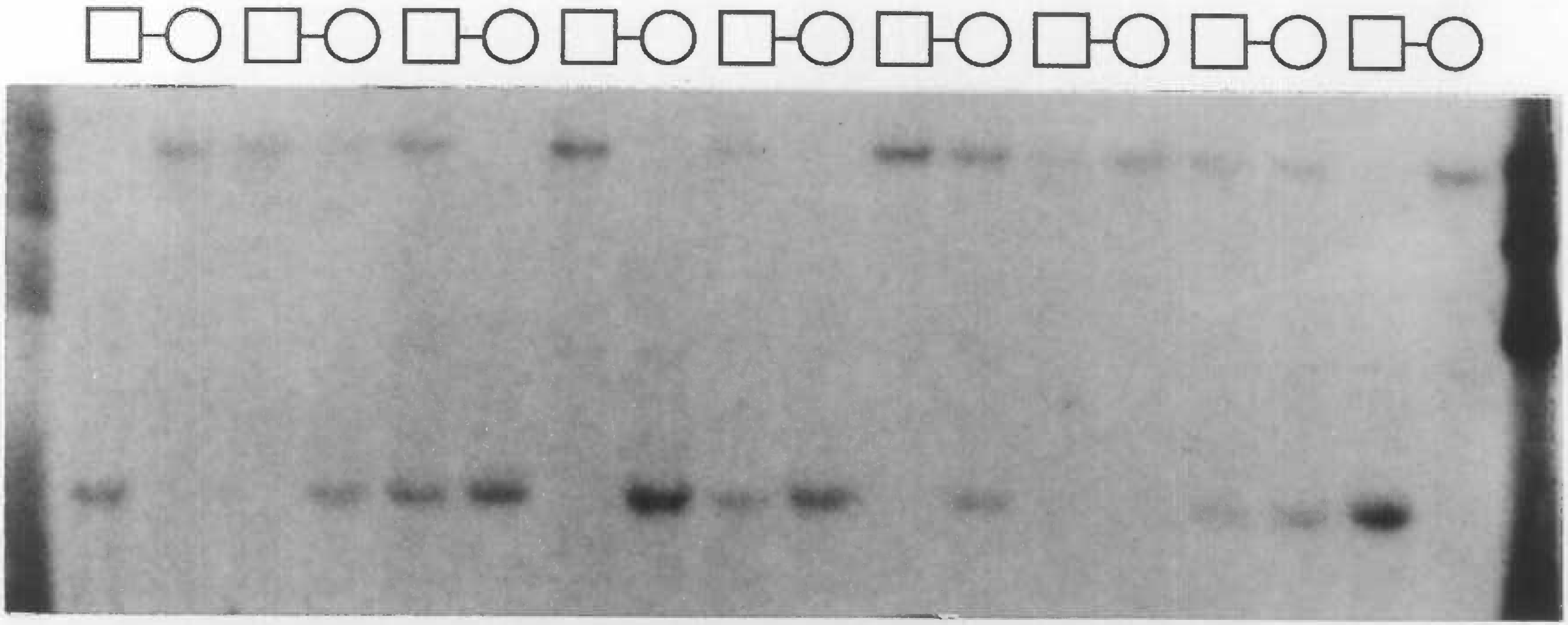


Figure 10. Southern blot of DNA form parents of CEPH reference families, digested with TaqI, and probed with cosmid 20-13. (Lanes 1 and 20 APSS=Size Standards).

1 Chapter 1: Episodic Ataxia Exclusion Mapping

1.1 Abstract

Determination of a genetic location for episodic ataxia (EA) in three study families has been approached using the technique of exclusion mapping. A mapped gene locus for spinocerebellar ataxia (SCA) on the short arm of chromosome 6, and 3 regions of chromosome 17, have been excluded as locations for EA genes in three study families.

1.2 Introduction & Overview

1.2.1 Rationale & Purpose

Since Botstein et al. first described a strategy for mapping disease genes using RFLPs and linkage analysis (31), the technique has been applied with success in a number of studies: Huntington's disease (41), cystic fibrosis (42), neurofibromatosis type 1 (43), chronic granulomatous disease (44), and others (reviewed in 11) including a form of spinocerebellar ataxia (SCA)(14). The mapping of a disease gene is significant in providing a tool for prenatal or early diagnosis of the disease (32). Mapping a gene is also an important first step in obtaining a clone of that gene (49). Linkage analysis also serves to exclude the location of a gene from the region of a genetic marker. We have approached mapping autosomal dominant Episodic Ataxia (EA) by linkage analysis. Discovery of linkage would indicate the presence of a major gene in the etiology of EA (20,21). Currently the chromosomal location of EA is unknown. The only previous linkage study tested 24 markers for linkage with EA and found 6 regions of exclusion (3).

EA is an autosomal dominant neurological disease which is characterized by periodic attacks, usually precipitated by physical or emotional stress, and asymptomatic interictal periods (i.e. between episodes). The biochemical basis of EA is unknown. A putative gene for spinocerebellar ataxia, which causes impairment of neurological function with some phenotypes overlapping with EA, has been mapped to the short arm of chromosome 6, and could be allelic with EA. A linkage map of markers exists for chromosome 6p (16) allowing us to perform multipoint linkage analysis of EA in this region to test this hypothesis.

In addition to the region of SCA, we have tested a series of probes representing loci on chromosome 17 for linkage with EA in study families. Chromosome 17 was chosen because of the existence of a good linkage map of this chromosome (17), which has allowed us to perform multipoint mapping of EA in 3 regions of this chromosome.

1.2.2 Clinical EA Picture

Episodic Ataxia is a relatively benign neurologic condition, which shows a pattern of autosomal dominant inheritance, and has been well characterized in a number of families (1,2). Individuals affected with EA are usually asymptomatic, but have occasional attacks, separated by intervals ranging from hours to years, and lasting from minutes to hours. Attacks are often precipitated by: physical stress in the form of exercise, sudden postural changes when excited, fatigue, sickness (e.g. fever or injury), and inconsistently by alcohol or caffeine, and by mental stress in the form of heightened emotions. There does not appear to be any aura, or warning of the attacks.

Attacks consist of a sudden onset of severe limb ataxia (defective limb coordination), making the individual unable to control their arms and legs for any useful purpose, and dysarthria (difficulty with speech). Attacks are described as ranging from mild to severe and often include dizziness, light headedness or drugged feeling, nausea, headache, gait ataxia, oscillopsia (movement of the visual field), and nystagmus (involuntary cyclical movement of eyes). Examination of individuals during attacks reveals ataxia, nystagmus, dysarthria, diaphoresis (profuse sweating), pallor and ptosis (drooping of eyelids). Attacks of EA are distinctive, and not easily confused with other syndromes.

Diagnosis is by history and neurological examination, with special attention paid to the patient's gait and balance. In many of the affected members of study families there is an interictal finding of nystagmus, which has been noted before the age of onset, and can aid in the diagnosis.

1.3 Materials and Methods

1.3.1 Family Ascertainment

EA study families were ascertained at the OHSU genetics clinic by Drs. Nutt and Gancher of the Oregon Health Sciences University Department of Neurology. Diagnosis is by clinical examination by Drs. Nutt or Gancher. Figure 1 shows the three EA families used in this study: 5910, 3149, and 9778.

Demographic and pedigree data for members of all families has been collected and stored along with affection status for all individuals. DNA samples have been extracted from blood specimens drawn from each available family member.

These families are well suited to linkage studies in that the mode of inheritance of EA is known and observed in all families, the families are large with multiple affected members in multiple branches, the diagnoses are clear, and there appears to be complete penetrance of EA after a usually early age of onset (i.e. post-puberty).

1.3.2 Determination of Phenotypes at Marker Loci

DNA was prepared from 50ml blood samples, with a typical 10-30 microgram DNA yield per milliliter of blood. Crude nuclear fractions, pelleted from blood samples after cell lysis with a Triton-X 100 containing buffer, were stored at -70 degrees centigrade, where they have been shown to be stable for periods of more than 7 years (127). DNA was then purified from aliquots of the crude nuclear pellets by treatment with Proteinase K/sodium dodecyl sulfate, phenol and chloroform extractions, and two successive ethanol precipitations, first in the presence of 0.3M sodium acetate and then 2.5M ammonium acetate.

DNAs, stored at a final concentration of 105 micrograms per milliliter, were digested with a five to ten fold excess of restriction enzyme under conditions recommended by the manufacturer. Test digests were created by removing 10 microliters of the restriction reaction mixture, which is mixed with 0.5 micrograms of bacteriophage lambda DNA in a separate tube, and allowed to incubate in parallel with the samples. At the end of the incubation period test digests were monitored by agarose gel electrophoresis. If the resulting fragment patterns were consistent with a complete digestion of the lambda DNA, and the background smear of human fragments appeared to cover the expected size range, the main digests were assumed to be complete. If an incomplete digest was indicated, more enzyme was added to the main digest, and a second test digest was made and monitored. Complete restriction enzyme digestions are essential, as incomplete digests can cause appearance of false "polymorphic" bands revealed by probes on Southern blots, and cause errors in typing. This can be monitored to some extent since all of the probes we are using have alleles demonstrated by fragments of known sizes and we can make note of bands with higher than expected molecular weights, or inconsistent inheritance patterns.

Gels for creation of Southern blots were run on 11 by 13 centimeter trays poured with twenty-well slot formers. 2.5 micrograms of restricted DNA samples were loaded per lane, size fractionated on the agarose gel by electrophoresis, and transferred to positively charged nylon membranes as per manufacturer's recommended methods. The membranes with bound DNA were then exposed to brief UV light treatment to cross-link the DNA fragments (40).

Probe inserts, or whole probes, were radioactively labeled with alpha dCTP ^{32}P to specific activities of 1-2 X 10^9 dpm per microgram using random primers and the Klenow fragment of DNA polymerase I (45). Probes labeled in this manner were allowed to hybridize to appropriate blots in hybridization solution (50% formamide, 10% dextran sulfate, 0.2 mg/ml denatured total human DNA, 5X SSC, 1X Denhart's solution (0.02% Ficol, 0.2% PVP, 0.02% BSA), 0.2M sodium phosphate pH 6.6) for 18 hours at 42 degrees centigrade. The blots are then washed sequentially in 2XSSC/0.1%SDS, and 0.1XSSC/0.1%SDS at room temperature, and twice in 0.1XSSC/0.1%SDS at 55-68 degrees centigrade (depending on the probe being used), and then exposed to X-ray film (usually with an intensifying screen at -70 degrees centigrade) for 1 to 5 days to produce autoradiographs. Blots may then be stripped of probe by mild alkaline treatment, and re-used.

1.3.3 Probe Selection

Table 1 lists the probes used in this study. These probes were chosen because of their location on chromosomes 6p and 17. Many of them have been placed on linkage maps, making them useful for multipoint linkage analysis (14,16).

1.3.4 Linkage Analysis

Both two-point and multipoint linkage analyses were facilitated by PMMS (27), and performed using programs in the LINKAGE system (24).

1.4 Results

Table 2 shows the results of two-point linkage analyses in the three study families using probes from Table 1. No markers showed significant linkage with EA. The pCH6 marker excluded EA from a 22 cM region of the short arm of chromosome 6 which overlaps the assigned location of the SCA gene. And when pCH6 and pHHH157 are run in multipoint analysis, the region of exclusion of EA is extended by 8 cM toward the centromere. No genetic maps which include the other markers we typed on chromosome 6 are available, preventing us from extending our region of exclusion with multipoint analyses.

Figure 2 shows the map for the markers I have run on chromosome 17. Sex averaged map distances were used in the multipoint analyses. Figure 3 shows the multipoint linkage analysis results. Figure 4 summarizes the results of the chromosome 17 multipoint and two-point analysis. These data exclude EA from regions of chromosome 17 where informative markers were found, totalling approximately 40% of the chromosome.

1.5 Discussion

We have succeeded in excluding the location of genes for EA from the region of the short arm of chromosome 6 known to be linked with SCA (14,15), and thereby shown that EA cannot be allelic with SCA. We have also excluded three regions of chromosome 17 as a location for EA using multipoint linkage analysis.

Our laboratory has been testing other regions of the genome for linkage to EA concurrently with this work, and has excluded more than 30% of the genome. Recently, a suggestive linkage was found for EA in family 5910 on the long arm of chromosome 11. Efforts to increase the significance of this linkage, to continue to exclude other regions of the genome and collect more study families, are currently in progress in our laboratory.

The development of genetic maps of many regions of the human genome, and especially chromosome 11q, combined with suitable EA families, gives us an excellent chance of successfully mapping its genetic location, or alternatively excluding the EA gene from these regions of the genome. Simulations of linkage analysis between EA in family 5910 and a fully informative marker showing no recombination have generated LOD scores greater than the commonly accepted significance level of three.

Mapping EA becomes especially significant when one considers that the episodic nature of EA is unique among the disorders which are the focus of much current genetic research (e.g. degenerative and cancer causing disorders). Understanding of EA's pathogenesis may therefore involve a different class of gene products or regulation mechanisms than has been previously described. Additionally, the initial indications that EA is an uncommon disease may be incorrect. With only 12 previously reported kindreds, Gancher and Nutt have reported 7 kindreds in Oregon, a state with only 2.5 million people. The incidence, and therefore the benefits of improved diagnosis and eventual treatment, may be greater than has been generally recognized.

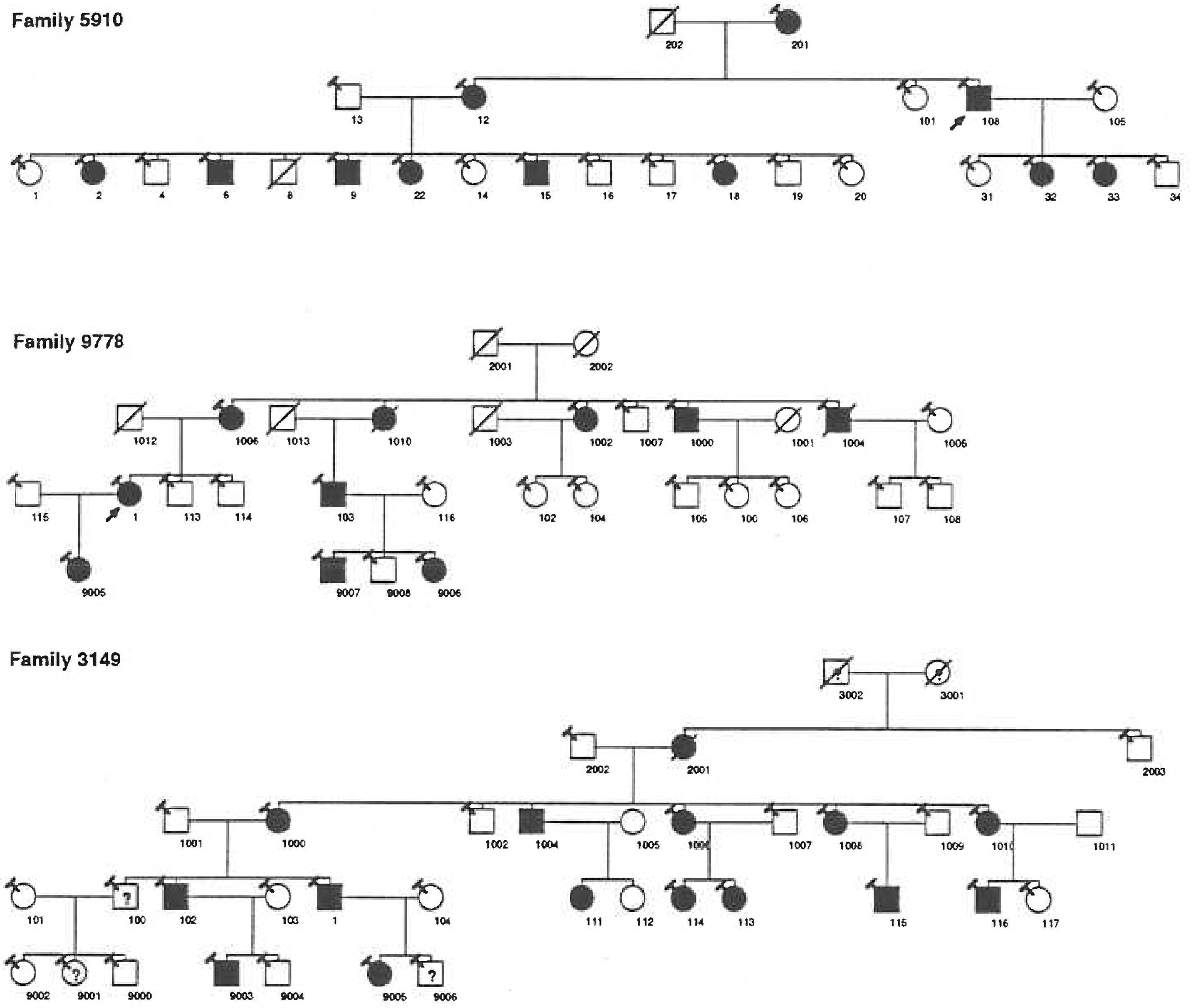
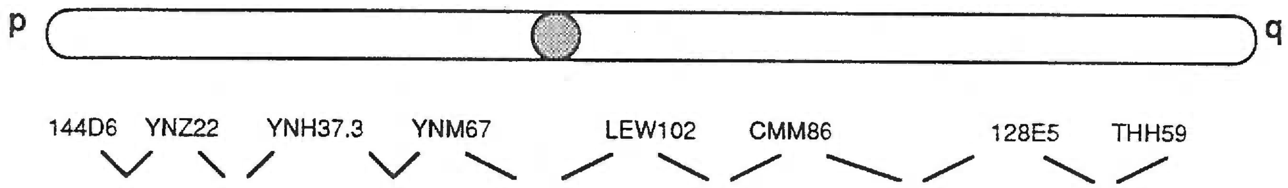


Figure 1. Pedigrees of our three EA Study families. Probands are indicated by arrows. Individuals sampled for DNA are marked ⋀ . Affecteds are dark, unknown diagnoses are marked with ?.

Estimated Genetic Map of Chromosome 17 (Distances in Morgans)



Males	.15	.02	.50	.24	.04	.35	.11
Females	.03	.001	.69	.58	.13	.47	.10
Averaged	.09	.021	.595	.41	.085	.41	.105

Figure 2. This depicts the loci used in multipoint analysis of chromosome 17 markers with EA, and the map distances between them estimated from a linkage map of chromosome 17. (Not to Scale)

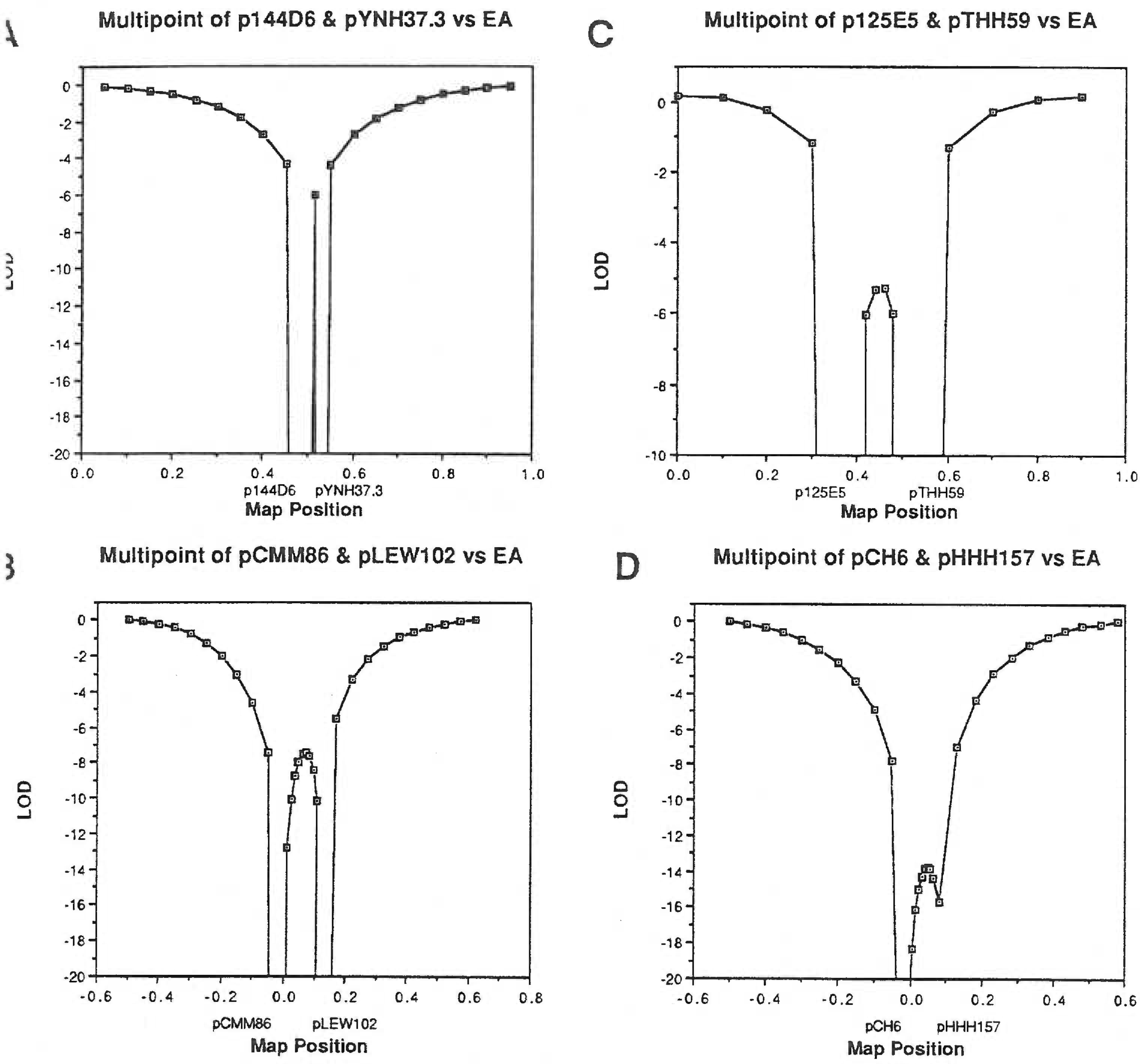


Figure 3. This figure shows multipoint linkage analysis results for EA on chromosomes 6p and 17. Panel A) shows the results for markers on 17p, B) shows markers at the 17 centromere, C) shows markers on 17 q, and D) shows markers on 6p. All map positions are in Morgans relative to an arbitrary zero point.

Table 1

Probes and Enzymes Used in Linkage Analysis

Probe	Locus	Enzyme	Informativeness	
p144D6........	D17S34	Taq I	.86	PIC
pYNZ22.........	D17S30	Msp I	.86	PIC
pYNH37.3.....	D17S28	Msp I	78%	Het
pHHH202......	D17S33	Rsa I	.38	PIC
p131A8........	D17S78	Msp I	.73	PIC
pCMM86........	D17S74	Taq I	90%	Het
pYNM67.........	D17S29	Taq I	.33	PIC
pLEW102......	D17S41	Taq I	.30	PIC
pTHH59.........	D17S4	Pvu II	71%	Het
pRMU3...........	D17S24	Pvu II	85%	Het
pEFD52.........	D17S26	Msp I	86%	Het
p128E5.........	D17S77	Msp I	.85	PIC
pEFD75.1......	D6S25	Msp I	65%	Het
pEFD70.2......	D6S26	Pst I	55%	Het
pCH6.............	D6S10	Taq I	.43	PIC
pCRI1065....	D6S21	Rsa I	.74	PIC
pYNZ132......	D6S40	Msp I	69%	Het
pYNB3.6.......	D6S30	Msp I	.37	PIC
pHHH157.....	D6S29	BamHI	.36	PIC
pHHH171.....	D6S38	MspI	35%	Het
pTHH5..........	D6S39	TaqI	55%	Het
p7H4............	D6S7	BamHI	.24	PIC

NOTE: Het = Heterozygosity
PIC = Polymorphic Infromation Content

Table 2

Pairwise LOD Scores for EA and Chromosome 6 & 17 Markers

EA vs.	Recombination Fraction							
Markers	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.40
p144D6								
5910	-inf.	-3.070	-1.749	-1.062	-0.632	-0.352	-0.168	0.007
9778	-inf.	-0.403	0.072	0.346	0.327	0.317	0.264	0.113
3149	-inf.	-1.376	-0.828	-0.533	-0.344	-0.215	-0.126	-0.027
	-inf.	-4.849	-2.505	-1.249	-0.649	-0.250	-0.030	0.093
pYNZ22								
5910	-inf.	-3.369	-2.039	-1.334	-0.884	-0.574	-0.352	-0.085
9778	0.221	0.187	0.154	0.122	0.093	0.066	0.044	0.011
	-inf.	-3.182	-1.885	-1.212	-0.791	-0.508	-0.308	-0.074
pYNH37.3								
5910	-inf.	-4.348	-2.702	-1.812	-1.233	-0.829	-0.536	-0.168
	-inf.	-4.348	-2.702	-1.812	-1.233	-0.829	-0.536	-0.168
pHHH202								
9778	-inf.	-2.457	-1.604	-1.124	-0.795	-0.550	-0.362	-0.111
3149	-2.797	-0.446	-0.199	-0.083	-0.021	0.010	0.021	0.012
	-inf.	-2.903	-1.803	-1.207	-0.816	-0.540	-0.341	-0.099
p131A8								
9778	-2.538	0.636	0.766	0.753	0.674	0.556	0.416	0.141
3149	-inf.	-1.154	-0.633	-0.366	-0.205	-0.104	-0.042	0.006
	-inf.	-0.518	0.133	0.387	0.469	0.452	0.374	0.147
pCMM86								
5910	-inf.	-4.607	-2.918	-1.985	-1.367	-0.926	-0.601	-0.186
9778	-inf.	-1.666	-0.920	-0.548	-0.329	-0.193	-0.107	-0.022
3149	-2.672	-0.681	-0.400	-0.246	-0.147	-0.080	-0.036	0.007
	-inf.	-6.954	-4.238	-2.779	-1.843	-1.199	-0.744	-0.201

Table 2 (Continued)

Pairwise LOD Scores for EA and Chromosome 6 & 17 Markers

	Recombination Fraction							
EA vs. Markers	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.40
pYNM67								
5910	-inf.	-0.721	-0.443	-0.292	-0.194	-0.125	-0.076	-0.018
	-inf.	-0.721	-0.443	-0.292	-0.194	-0.125	-0.076	-0.018
pLEW102								
5910	-inf.	-3.813	-2.236	-1.417	-0.909	-0.573	-0.344	-0.082
9778	-2.865	-0.811	-0.373	-0.179	-0.080	-0.030	-0.007	0.002
3149	-2.575	-0.596	-0.329	-0.192	-0.110	-0.060	-0.029	-0.004
	-inf.	-5.220	-2.938	-1.788	-1.099	-0.663	-0.380	-0.084
pTHH59								
5910	-inf.	-2.68	-1.552	-0.945	-0.571	-0.340	-0.181	-0.008
	-inf.	-2.68	-1.552	-0.945	-0.571	-0.340	-0.181	-0.008
pRMU3								
5910	0.010	0.170	0.110	0.080	0.060	0.040	0.020	0.010
	0.010	0.170	0.110	0.080	0.060	0.040	0.020	0.010
pEFD52	Uninformative							
p128E5								
5910	-inf.	-2.040	-1.080	-0.47-	-0.160	0.020	0.130	0.016
	-inf.	-2.040	-1.080	-0.47-	-0.160	0.020	0.130	0.016
pEFD75.1								
5910	-inf.	-1.235	-0.716	-0.448	-0.283	-0.175	-0.102	-0.023
3149	-inf.	-2.270	-1.445	-0.995	-0.700	-0.490	-0.333	-0.122
	-inf.	-3.505	-2.161	-1.443	-0.983	-0.665	-0.435	-0.145

Table 2 (Continued)

Pairwise LOD Scores for EA and Chromosome 6 & 17 Markers

EA vs.	Recombination Fraction							
Markers	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.40
pEFD70.2								
5910	-inf.	-2.136	-1.323	-0.888	-0.608	-0.414	-0.274	-0.096
9778	-inf.	-2.329	-1.275	-0.748	-0.443	-0.258	-0.146	-0.038
3149	-inf.	-4.162	-2.734	-1.932	-1.389	-0.991	-0.685	-0.258
	-inf.	-8.627	-5.332	-3.568	-2.440	-1.663	-1.105	-0.392
pCH6								
5910	-inf.	-4.019	-2.507	-1.664	-1.108	-0.716	-0.435	-0.102
9778	-inf.	-2.514	-1.600	-1.080	-0.731	-0.483	-0.303	-0.084
3149	0.159	0.127	0.098	0.073	0.051	0.033	0.019	0.002
	-inf.	-6.406	-4.009	-2.671	-1.788	-1.166	-0.719	-0.184
pCRI1065								
5910	-inf.	-0.721	-0.444	-0.293	-0.194	-0.125	-0.076	-0.018
9778	-inf.	-3.112	-1.827	-1.165	-0.757	-0.484	-0.294	-0.072
3149	-inf.	-0.715	-0.242	-0.024	0.086	0.136	0.145	0.091
	-inf.	-4.548	-2.513	-1.482	-0.865	-0.473	-0.225	0.001
pYNZ132								
5910	-inf.	-0.770	-0.055	0.275	0.439	0.506	0.503	0.343
9778	-inf.	-2.292	-1.385	-0.891	-0.575	-0.360	-0.212	-0.047
3149	-inf.	-1.137	-0.658	-0.421	-0.278	-0.185	-0.121	-0.042
	-inf.	-4.199	-2.098	-1.037	-0.414	-0.039	0.170	0.254
pYNB3.6	Uninformative							

Table 2 (Continued)

Pairwise LOD Scores for EA and Chromosome 6 & 17 Markers

EA vs.	Recombination Fraction							
Markers	0.00	0.05	0.10	0.15	0.20	0.25	0.30	0.40
pHHH157								
5910	-inf.	-2.878	-1.767	-1.167	-0.782	-0.518	-0.334	-0.113
9778	-inf.	-0.888	-0.377	-0.122	0.021	0.098	0.127	0.093
3149	-inf.	0.171	0.141	0.117	0.084	0.060	0.039	0.010
	-inf.	-3.595	-2.003	-1.172	-0.677	-0.360	-0.168	-0.010
pHHH171								
5910	-inf.	0.443	0.768	0.830	0.777	0.654	0.481	0.058
9778	-inf.	-0.998	-0.689	-0.510	-0.385	-0.290	-0.214	-0.095
3149	-inf.	0.171	0.141	0.111	0.084	0.060	0.039	0.010
	-inf.	-0.384	0.220	0.431	0.476	0.424	0.306	-0.027
pTHH5								
5910	-inf.	-1.770	-0.754	-0.248	0.041	0.204	0.281	0.245
9778	-inf.	-0.257	-0.086	-0.029	-0.009	-0.001	0.004	0.013
3149	-2.878	-0.567	-0.313	-0.184	-0.107	-0.059	-0.030	-0.004
	-inf.	-2.594	-1.153	-0.461	-0.075	0.144	0.255	0.254
pMCOB5								
5910	-inf.	-4.328	-2.663	-1.755	-1.163	-0.750	-0.455	-0.107
	-inf.	-4.328	-2.663	-1.755	-1.163	-0.750	-0.455	-0.107

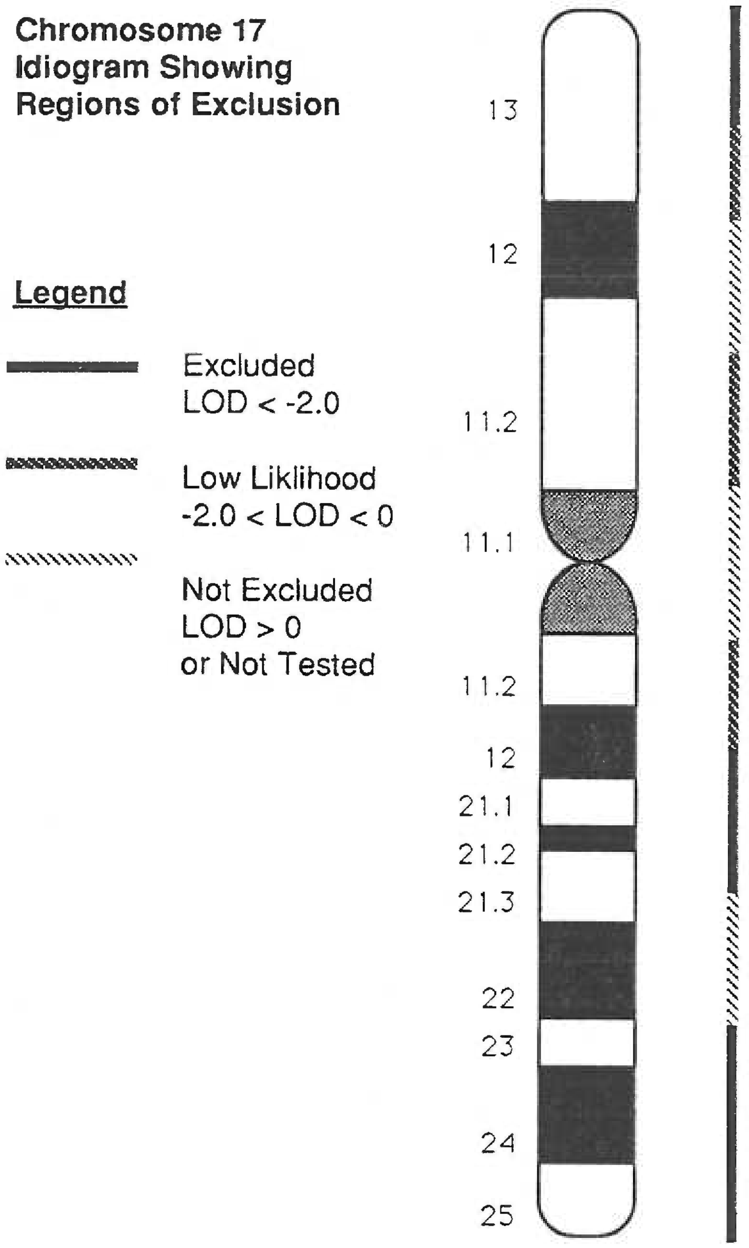


Figure 4. This idiogram of chromosome 17 represents regions excluded as locations for EA in our study families.

2 Chapter 2: Development of RFLPs on 11q

2.1 Abstract

Cosmids from a chromosome 11q library have been screened for their ability to detect restriction fragment length polymorphisms (RFLPs). Six cosmids have been found which allow determination of 8 polymorphic genetic loci. These loci have been typed in CEPH reference families, and added to regions of two genetic linkage maps of chromosome 11q. Localization of markers by linkage analysis is consistent with physical mapping results from somatic cell hybrids. The addition of markers to the genetic map of chromosome 11q will help in the study of the genomics of this important chromosomal region.

2.2 Introduction & Overview

Chromosome 11q represents approximately 2.5% of the genetic length of the human genome (22). There are currently 85 polymorphic marker loci and 62 genes listed as mapped to 11q in the Yale human gene mapping library (HGML)(105). This represents 3.5% of all mapped autosomal genes listed in the HGML, but only 2.3% of all autosomal marker loci. Many important diseases, for which no gene has yet been identified, map to 11q including ataxia telangiectasia (62), tuberous sclerosis (63), multiple endocrine neoplasia type 1 (60), and some forms of atopy (64). A number of interesting genes are located on 11q including the dopamine D2 receptor (104), which is a candidate gene for many neuropsychiatric disorders, as well as members of the immunoglobin supergene family (106). Chromosome 11q holds numerous regions of interest in the human genome, and there is a need for more genetic markers to help in the development of a detailed genetic map of those regions.

We have characterized eight new RFLPs, discovered by screening twenty cosmids from a library derived from cell line TG 5D1-1, which contains the chromosomal fragment 11q13.1-qter as its only human component, representing about 0.9% of the physical length of the human genome (122). Cesium chloride-banded DNA was prepared from each of the cosmids, and whole cosmids prehybridized with a vast excess of total human DNA (90) were used as probes on southern blots containing DNA from six or more individuals. These DNAs were digested with at least three separate restriction enzymes (TaqI, MspI and RsaI) as an initial screening for RFLPs, since approximately 40% of the more than 1500 probes in the HGML are polymorphic with one of these three enzymes.

All polymorphisms found were checked for Mendelian inheritance in families. The cosmid probes were mapped on a somatic cell hybrid mapping panel to provide a physical localization on 11q. Loci detected by the same probe were checked for linkage disequilibrium. All loci were used to type informative CEPH families. Linkage analysis was performed between these new markers and other probes in the CEPH version 3 database and probes reported in a primary linkage map of 11q (54).

2.3 Materials & Methods

2.3.1 Cosmid Library

The source for all probes was twenty cosmids randomly selected from a chromosome 11q library constructed by G.A. Evans and K.A. Lewis (122). The vector is sCos1 (122) with inserts cloned into a unique BamHI cloning site, which is closely flanked by NotI sites.

2.3.2 RFLP Screening

Cesium chloride-banded DNA preparations from each of the cosmids were produced. Aliquots of whole cosmid preparations were radioactively labeled by random hexamer priming (45), prehybridized with a vast excess of total human DNA (90), and used as probes on southern blots containing DNA from six or more individuals, digested with at least three separate restriction enzymes (TaqI, MspI and RsaI). Cosmids 1-16, 4-7, and 12-16, which map to distal regions of 11q, have been screened for RFLPs in at least six unrelated individuals, whose DNA has been digested with at least six separate restriction enzymes in addition to TaqI, MspI and RsaI.

Southern blots were prepared and hybridized as described elsewhere (90,123), with the additional step of crosslinking the DNA to the nylon membrane by irradiation of the damp membrane, immediately after blotting, with 1600 joules/m^2 of UV light in a Stratalinker (Stratagene) (40).

Blots were prehybridized with total human DNA at 42 degrees centigrade for at least two hours before hybridization with probes. After hybridization, blots were washed in 2XSSC/0.1%SDS, and 0.1XSSC/0.1%SDS for 15 minutes each at room temperature, and twice for 15 minutes in 0.1XSSC/0.1%SDS at 65-68 degrees centigrade.

2.3.3 Restriction Fragment Isolation

Gel-fragments which revealed RFLPs were isolated from restriction digests of cosmids by excision of bands from low-melting agarose gels. The gel-fragments are washed with distilled water twice for 30 minutes each time, mixed with an equal volume of distilled water, melted at 70 degrees centigrade, and stored at -20 degrees centigrade. Gel-fragments so prepared were radioactively labeled in the same manner as whole cosmid DNA.

2.3.4 Densitometry

To determine which bands on Southern blots represent alleles, images of autoradiographs were scanned using a HP Scan-jet plus, and analyzed on a Macintosh computer using the Image software package (123).

2.3.5 Linkage Disequilibrium

RFLPs detected by the same probe were analyzed for linkage disequilibrium between alleles of the two systems using the chi-square test. Haplotypes of unrelated individuals typed for both markers, and homozygous in at least one system to allow unambiguous haplotype determination, were counted and the observed frequencies were compared to expected haplotype frequencies. Expected haplotype frequencies were calculated based on observed allele frequencies assumed to be in Hardy-Weinberg equilibrium.

2.3.6 Somatic Cell Hybrids

A panel of 5 somatic cell hybrid cell lines containing portions of chromosome 11q was used to physically map cosmids to regions of 11q (Figure 1). Cell lines MC-1, J1-11, J1-44, and TGD5D1-1 are described in reference 61. Cell line Cl 5-2, provided by B.S. Emanuel, contains chromosome fragment 11pter-q24 as its only chromosome 11 component (71). This panel defines 6 chromosomal regions of 11q, including a region in 11q13 of unknown size due to uncertainties in the end-points of the J1-44 and TGD5D1-1 cell hybrids.

2.3.7 Linkage Analysis

To determine their location on genetic maps, all 8 RFLPs were typed in informative CEPH families; this dataset was then used for two-point analysis between the new markers and two sets of markers for which a CEPH family dataset had been produced: the CEPH V3 dataset (25 markers), and an 11q CEPH family dataset provided by Lathrop (31 markers) (54).

For each new marker, a set of linked markers (with LOD scores greater than 2) was used to perform interval mapping (128) between the new marker and intervals defined by the pair wise combination of linked markers, separated by their published two-point distances. These interval maps demonstrate the likelihoods, reported as LOD scores, that the new marker resides between, or to either side of, a pair of linked markers.

2.4 Results

2.4.1 Polymorphic Loci

Of the 20 cosmid clones screened, 6 cosmids revealed 8 RFLPs (Table 1). Five of these cosmids were physically mapped to regions of 11q using a somatic cell hybrid mapping panel (Table 2). Cosmid 20-13 could not be mapped using somatic cell hybrids due to a high background hybridization with mouse DNA.

Cosmid 24-10, and probe 3-27.b detect two-allele systems which revealed no individuals homozygous for the less frequent minor allele in the parents of the 36 CEPH families screened. In the case of 24-10, the sizes and relative intensities of the major and minor allelic fragments are unchanged when the enzyme to DNA ratio is increased to 4 times the standard 5 fold enzyme excess, indicating that the bands are not the result of partial digestions. To verify the identity of the major allelic fragment in cosmid 24-10, densitometry was performed on autoradiographs of blots probed with this marker (Figure 2), and the results support the assignment of the major allele. This assignment is also supported by the identification of two children homozygous for the minor allele, which were found in two separate families where both parents were heterozygotes for 24-10.

Many of these RFLPs can be detected by restriction fragments isolated from agarose gels (Table 3), which often give more consistent results, with a higher signal to noise ratio on autoradiography, than those obtained using whole cosmids as probes. The gel fragment revealing locus 3-27.b has been subcloned into a plasmid vector (BamHI site of pTZ18).

2.4.2 Population Studies

Two probes, 3-27 and 1-16, each reveal two RFLP loci. Haplotype frequencies were calculated for both loci of each probe by inferring haplotypes from phenotypes of unrelated individuals. Chi square analyses of observed versus expected haplotype frequencies (Table 4) indicate that the systems revealed by 3-27 are in strong linkage disequilibrium, with an overabundance of the A2B1 and A2B2 haplotypes, and deficiencies of the A1B2 and A3B2 haplotypes. The systems revealed by 1-16 are in linkage equilibrium, with haplotype frequencies showing no significant deviations from the expected values.

All markers were typed on blots containing DNA from parents of CEPH families to determine which families were informative for the marker systems. Frequencies of alleles were determined in this set of unrelated individuals. Percent heterozygosity and polymorphism information content (PIC) (34) have been calculated for all probes. These results, and a compound PIC for 1-16 based on observed haplotype frequencies, are reported in Table 1.

2.4.3 Linkage Analyses

Probes were typed in informative CEPH families to produce a dataset which has been translated into the CEPH data format and sent to CEPH for inclusion in their Version 4 database.

This dataset was then used to map these new markers with respect to two sets of markers which have been typed in CEPH families: 25 markers from the CEPH V3 dataset which have been localized to 11q, and 31 markers from a dataset provided by Lathrop which he used to construct a linkage map of 11q (54). These two sets of markers, which have only four markers in common, are listed in Table 5, and tables of the two-point LOD scores calculated for all pair-wise combinations of markers in each set have been published (54,94). Two-point linkage analyses between markers in each of these sets and the 8 new markers were performed in 24 CEPH families typed with the set of new markers. Tables 6 and 7 list the two-point analyses which gave LOD scores of 0.8 or greater.

Interval mapping of these new markers (Table 8) has determined 7 intervals in which markers are likely to reside, and has excluded markers from 2 intervals. These localizations are based on a likelihood of marker order which is 100 times (i.e. two LOD units) greater than that of alternative orders. Marker 3-27 lies in a 28 cM region between p3C7 and CJ52.99, and has also been excluded from a 7 cM region within this interval, localizing it to the region between PYGM and Cj52.99. Localization of 20-13 is also enhanced by exclusion from a sub-interval. Figure 3 shows results from interval mapping, and two-point analysis with maximum LOD-1 confidence intervals, which allows rough localization of all new markers on genetic maps. Note that two markers used in interval mapping, but not shown in figure 3: phi8-10, and CJ52.102, are 4 cM from HBI18, and 3 cM CJ52.15 respectively.

2.5 Discussion

The rarity of individuals homozygous for the minor allele in marker 24-10 is disturbing. We would have expected to see approximately 10 individuals homozygous for the minor allele of 24-10 in the panel of 36 CEPH parents, based on the allele frequencies calculated in this population, and we have seen none (p<.001). Densitometry data for 24-10 is consistent with our assignment of the location of the major allele. Furthermore, by typing two families (total of 13 children) in which both parents are heterozygous for 24-10, we have been able to find two individuals homozygous for the rare allele, supporting our assignments of the correct allele sizes. Speculation as to the cause of this deviation from Hardy-Weinberg expectations includes the possibility that the minor allele of 24-10 could be in partial linkage disequilibrium with a recessive allele not compatible with life, but which may be maintained in the population by heterozygote advantage or other mechanism. If so, the individuals we found that were homozygous for the rare allele must represent cross-over events between the this allele and the postulated recessive lethal allele. None of the CEPH families screened with 3-27.b have two parents heterozygous for the system. However, by typing a series of random individuals for marker 3-27.b we have found an individual homozygous for the minor allele in this system, supporting our assignment of alleles.

Using 24 CEPH reference families, two-point linkage analysis between these new markers and two sets of markers already typed in CEPH families has provided evidence of linkage between several of the markers. Evidence of linkage between these new markers and previously mapped markers, including the markers previously developed in this lab by Maslen (61), is consistent with previous chromosomal localizations and somatic cell hybrid mapping results.

Cosmid 1-16 has been mapped by high-resolution *in situ* techniques to 11q23.3 (58). This assignment is consistent with linkages observed between 1-16 and other markers localized by high-resolution *in situ* mapping (e.g. THY1 with 1-16: theta=.029, LOD=7.29; APOA1 with 1-16: theta=.0001, LOD=3.9). Localizations of cosmids 3-27 and 24-10 by linkage analysis is consistent with the 11q13-11q22 somatic cell hybrid panel localizations for these markers. Linkage analysis localizations of cosmids 4-7 and 12-16 are also consistent with their 11q23-qter and 11q24 somatic cell hybrid panel localizations. These data support the localizations of the new markers by linkage analysis.

The eight new markers described here are well distributed along 11q, and will be useful in studies of the genomics of this region of the genome. The usefulness of these markers is extended by their addition to the CEPH V4 database, which will allow their inclusion in future genetic maps of 11q.

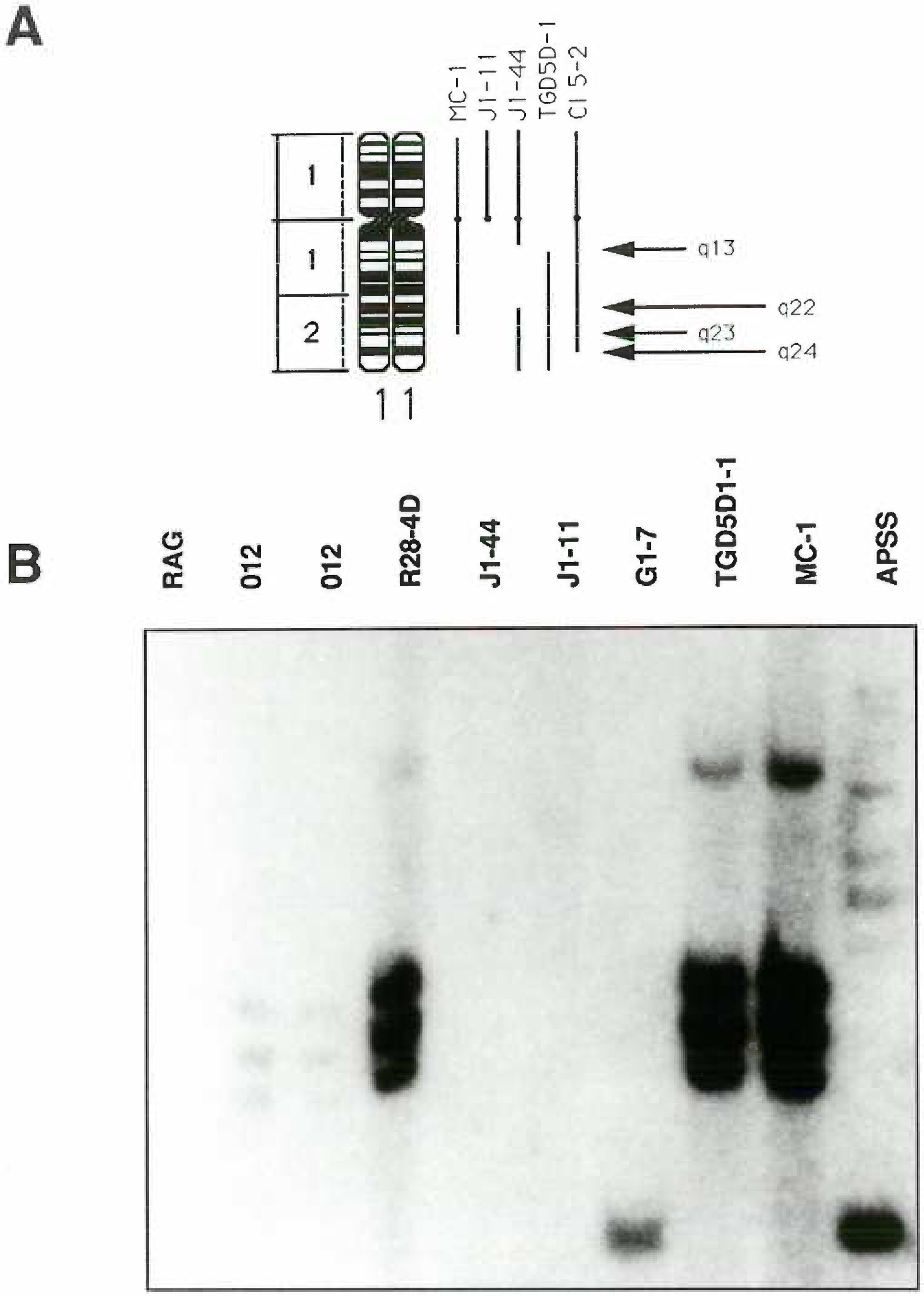


Figure 1. Panel A shows the somatic cell hybrid mapping panel used to localize our markers on 11q. Panel B show a result for marker 3-27. (APSS = Size Standards, G1-7 is not useful , R28-4D is identical to TGD5D1-1, 012 = Total Human DNA, RAG = Mouse DNA)

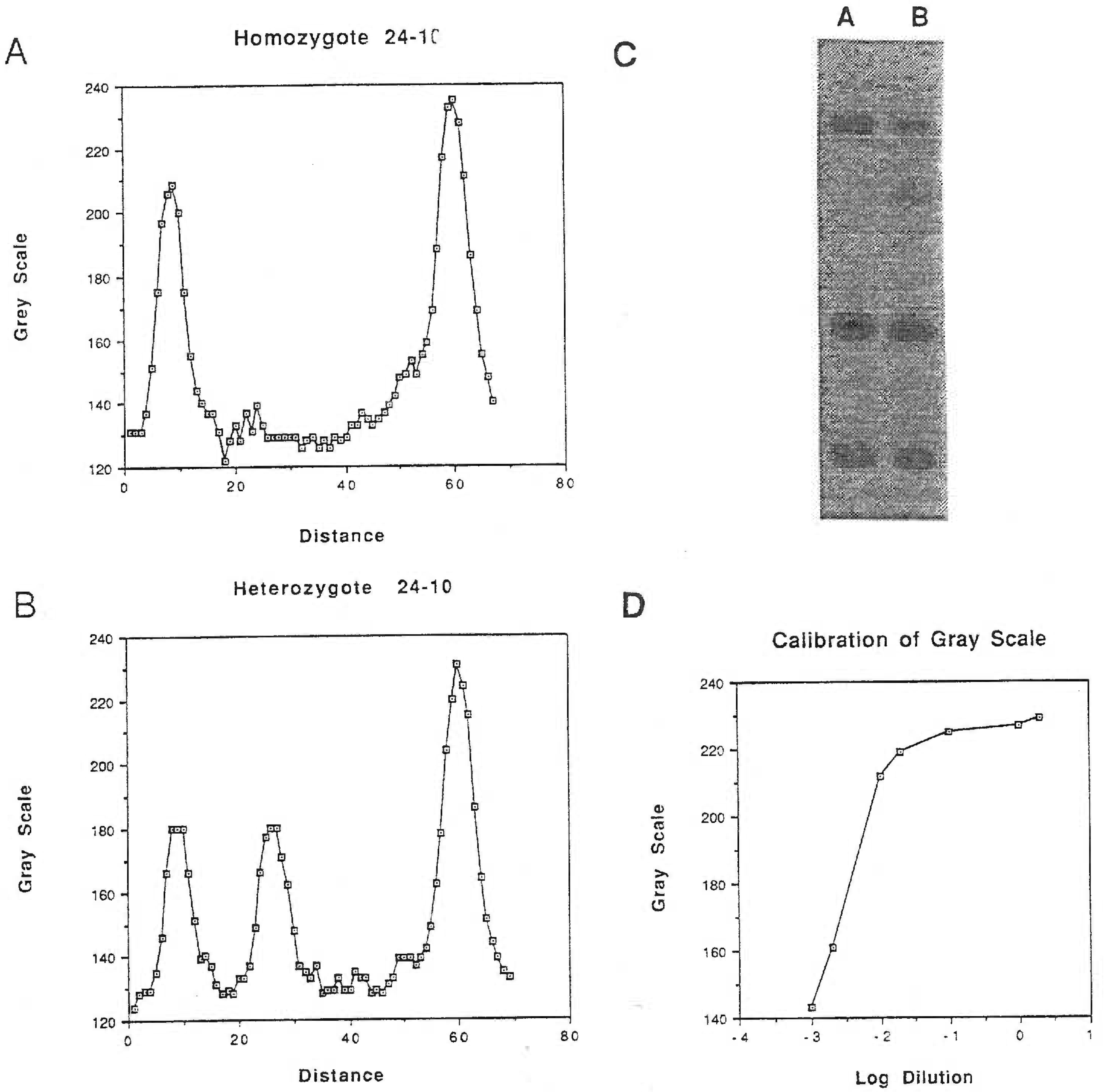


Figure 2. Panel A and B show densitometry scans of two lanes of a MspI southern blot (Panel C) probed with cosmid 24-10. Calibration of gray scale vs amount of radioactivity (Panel D) was performed by scanning an autoradiograph of filters which were spotted with serial dilutions of ^{32}P oligolabled pRB322 and TCA precipitated, to confirm linearity of relationship. Note that the intensity of the first band in panel A is split between the first two bands in panel B, while the intensity of the non-polymorphic band at 60 distance units remains constant.

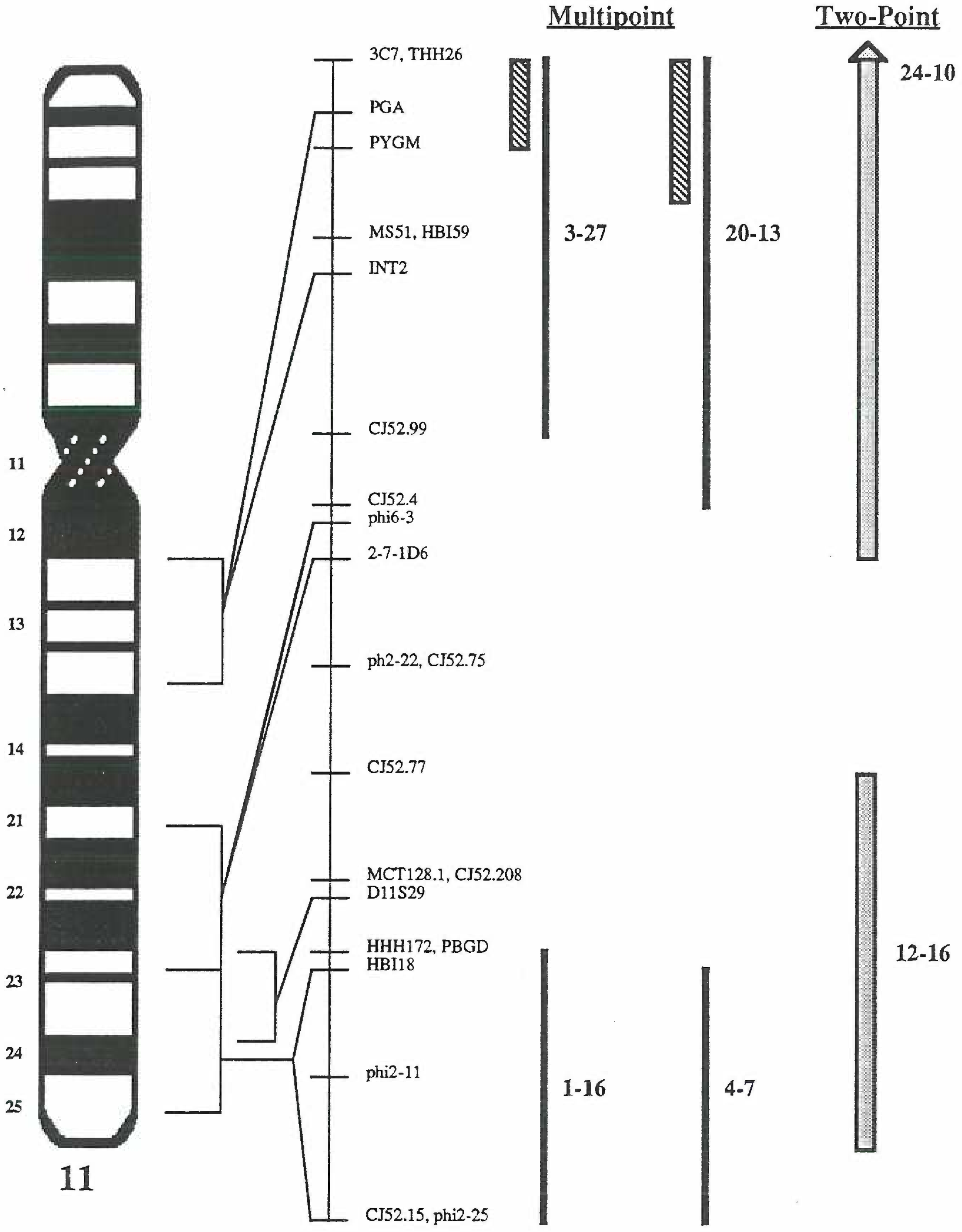


Figure 3. This figure shows results of multipoint interval mapping placing four markers, and two-point linkage analysis, with maximum LOD -1 confidence intervals, placing the remaining 2 markers. Note: confidence interval for 24-10 extends above map. (Figure after reference 54)

Legend

▬	100:1 or greater odds marker is inside the interval
▨	100:1 or greater odds marker is outside the interval
░	Maximum LOD -1 confidence interval

Table 1

					Hetero.		Number	
Probe	Enzyme		Size *	Freq.	%	PIC	Chrom.	Location
24-10	MspI	Constant:	8.5 kb					11q13
		A1	14.1	0.64	46%	.35	150	- or -
		A2	11.6	0.36				11q22-11q23
3-27	RsaI	Constant:	4.5 kb					11q13-q22
			2.8					
		A1	2.9	0.14	42%	.53	128	
		A2	2.6	0.5				
		A3	2.4	0.35				
3-27.b	TaqI	Constant:	1.1 kb					
		B1	3.9	0.14	26%	.21	158	
		B2	3.0	0.86				
1-16	TaqI	A1	13.3 kb	0.19	31%	.26	116	11q23-q24
		A2	11.1	0.81				
1-16.b	TaqI	B1	3.4	0.56	57%	.37	86	
		B2	3.3	0.44				
				Compound PIC= .56				
12-16	EcoRI	Constant:	11.0 kb					11q24
			8.7					
			7.5					
			5.4					
			3.7					
		A1	4.3	0.75	38%	.30	134	
		A2	2.9	0.25				
20-13	TaqI	Constant:	1.7 kb					no cell panel data
		A1	14.9	0.62	47%	.35	108	
		A2	6.7	0.37				
4-7	TaqI	Constant:	3.9 kb					11q23-qter
			3.3					
			2.8					
		A1	9.7	0.64	46%	.35	56	
		A2	8.0	0.36				

* Note: Other Constant bands may be present, only most intense noted.

Table 2

Probe Hybridization Patterns for 11q Mapping Panel

Probe	MC-1	J1-11	J1-44	TG 5D1-1	Cl 5-2	Location
24-10	+	-	+	+	+	11q13 or 11q22-q23
3-27	+	-	-	+	+	11q13-q22
1-16	-	-	+	+	+	11q23-q24
12-16	-	-	+	+	+	11q24
4-7	-	-	+	+	-	11q23-qter

Table 3

Restriction Fragments Revealing Polymorphism

Probe	Enzyme	Fragment Size
1-16	TaqI	10.5 kb
1-16.b	TaqI	3.7 kb
3-27	RsaI	2.8 kb
3-27.b	TaqI	4.0 kb
4-7	TaqI	9.6 kb
25-3	TaqI	4.8 kb

Table 4

Linkage Disequilibrium between Loci Detected by Single Probe

Part A.

Haplotype Counts

Probe	Haplotype	Observed	Expected
1-16	A1B1	6	7.87
	A1B2	6	6.19
	A2B1	30	33.57
	A2B2	32	26.37
3-27	A1B1	0	1.72
	A1B2	1	10.58
	A2B1	8	5.74
	A2B2	59	35.26
	A3B1	0	4.02
	A3B2	14	24.68

Part B.

Chi Square Analysis

Locus	X^2	N	P	DF
1-16	2.03	74	0.10	1
3-27	35.91	82	<0.0001	2

N = Number of Chromosomes Counted

Table 5

11q Markers from 11q Map & CEPHV3 Database

Probe	Locus	Enzyme	Number of Alleles
11q Map Markers:			
pTHH26	D11S149	PvuII	2
p3C7	D11S288	MspI	2
L7	D11S29	TaqI	2
pMCMP1	PYGM	MspI	>4
pMCT128.1	D11S144	MspI	2
pPGA101	PGA	Bg1II	2
phi6-3	D11S85	MspI	2
p2-7-1D6	D11S84	TaqI	2
pHBI59	D11S146	MspI	2
phi2-25	D11S83	MspI	2
pHHH172	D11S350	MspI	2
pSS6	INT2	TaqI	2
pPGBC9	CD3D	TaqI	2
phi9-11	D11S98	MspI	2
pPstPstPNE	PBGD	MspI	2
CJ52.12	D11S382	TaqI	2
CJ52.15	D11S383	TaqI	2
CJ52.193	D11S384	TaqI	2
pCJ52.75M1	D11S385	MspI	2
phi2-22	D11S35	TaqI	2
phi8-10	D11S286	BamHI	2
CJ52.5T1	D11S386	PstI	2
pCJ52.102T1	D11S387	TaqI	3
phi2-11	D11S34	MspI	2
APOA1	APOA1	TaqI	2
pHBI18P2	D11S147	PstI	2
CJ52.208M2	D11S351	MspI	2
CJ52.4	D11S388	MspI	2
pMS51	D11S97	TaqI	5
pCJ52.99M2	D11S389	MspI	2
pCj52.77M1	D11S...	MspI	2

Table 5 (Continued)

11q Markers from 11q Map & CEPHV3 Database

Probe	Locus	Enzyme	Number of Alleles
CEPH V3 Database Markers:			
p3C7	D11S288	MspI	2
CRI-L605	D11S127	MspI	9
S6		BamHI	2
CRI-L962	D11S128	MspI	9
CRI-R365	D11S129	Bg1II	8
CRI-R975	D11S131	TaqI	2
CRI-L424	D11S132	EcoRI	2
CRI-L451	D11S133	EcoRI	2
CRI-L834	D11S134	MspI	2
CRI-L1382	D11S136	MspI	3
pINT-800	CAT	TaqI	2
CRI-R83	D11S137	HindIII	2
CRI-R83	D11S137	TaqI	2
CRI-R83	D11S137	MspI	2
CRI-L937	D11S140	HincII	2
CRI-L762	D11S141	TaqI	5
CRI-V928	D11S142	MspI	2
THY1	THY1	MspI	2
phi2-25	D11S83	MspI	2
L7	D11S29	TaqI	2
E79	D11S37	HindIII	2
p32-1	D11S16	MspI	3
SS6(int-2)	SS6	TaqI	2
pLC11A	D11Z1	Sau3AI	3
H31	D11S348	HindIII	2

Table 6

Two-Point LOD Scores of 11q Cosmids with 11q Map Markers

Probe	24-10		3-27		3-27b		1-16		1-16b		12-16		20-13		4-7	
	Theta	LOD	Theta	LOD	Theta	LOD	Theta	LOD	Theta	LOD	Theta	LOD	Theta	LOD	Theta	LOD
pTHH26	0.07	2.3	0.88	2.1												
p3C7	0.12	2.1	0.16	2.1									0.14	1.9		
L7																
PYGM			0.14	3.0									0.82	2.1		
pMCT128.1							0.07	6.4	0.13	2.7					0.23	0.8
pHBI59			0.15	1.1									0.14	2.6		
ph2-25							0.26	0.9	0.10	2.3						
pHHH172							0.001	3.6								
INT2			0.07	4.6									0.067	2.0		
PBGD							0.04	4.5	0.14	0.8	0.06	2.6				
CJ52.12							0.04	4.2	0.001	3.6					0.001	1.5
CJ52.15							0.001	1.8							0.08	4.4
ph2-22	0.17	1.2	0.24	1.0	0.001	1.5										
ph8-10							0.04	3.9	0.001	4.2			0.12	3.7		
pCJ52.102T1											0.24	0.8			0.05	3.9
ph2-11							0.001	2.7	0.08	3.3						
APOA1							0.001	3.9							0.11	1.0
pHBI18P2							0.001	4.5	0.001	4.5					0.21	1.9
CJ52.4			0.24	1.6	0.001	2.1							0.16	4.5		
pMS51	0.33	0.9	0.19	4.8	0.001	1.5					0.29	0.8	0.26	2.3		
pCJ52.99M2	0.13	1.6	0.06	3.1	0.001	1.5							0.001	1.5		

Table 7

Two-Point LOD Scores of 11q Cosmids with CEPH V3 Database

Probe	24-10		3-27		1-16		1-16b		12-16		20-13		4-7	
	Theta	LOD	Theta	LOD	Theta	LOD	Theta	LOD	Theta	LOD	Theta	LOD	Theta	LOD
p3C7	0.11	1.3									0.14	2.1		
CRI-L962									0.14	2.0			0.28	1.3
CRI-R975			0.44						0.15	1.4			0.14	2.5
CRI-L424					0.15	0.8								
CRI-L834			0.20	1.4										
CRI-L1382									0.06	3.4			0.11	2.0
CRI-R83.1	0.16	1.2							0.001	2.1	0.11	3.3		
CRI-R83.2											0.09	3.2		
CRI-R83.3											0.07	5.0		
CRI-L937			0.17	2.4										
CRI-L762	0.001	2.4							0.15	2.9				
CRI-V928			0.08	1.7										
THY1					0.02	7.2	0.001	5.1						
L7					0.12	5.0	0.17	1.5					0.27	1.1
p32-1			0.18	2.5										
SS6(int-2)			0.15	3.8										
pLC11A	0.001	2.4	0.16	1.4					0.11	2.4				

Table 8

Interval Maps for 11q Markers

Test Locus	Marker 1	Marker 2	Fixed Theta	LOD Score for Order T-1-2	1-T-2	1-2-T
3-27	pTHH26	CJ52.99	0.28	4.84	6.67	4.38
3-27	p3C7	CJ52.99	0.28	5.27	7.49	5.34
3-27	p3C7	PYGM	0.07	7.58	6.34	8.33
1-16	PBGD	CJ52.15	0.24	6.38	8.74	5.90
1-16	CJ52.15	HBI18	0.25	5.83	8.01	4.89
1-16.b	phi2-25	phi8-10	0.04	4.62	7.64	5.42
20-13	p3C7	CJ52.4	0.25	5.59	7.87	6.62
20-13	p3C7	MS51	0.07	8.31	6.23	9.796
4-7	HBI18	CJ52.102	0.30	2.64	5.00	2.64

3 Chapter 3: Pedigree/Marker Management System

3.1 Introduction & Overview

3.1.1 PMMS

The purpose of this chapter is to document the design and use of the Pedigree Marker Management System (PMMS). The PMMS is designed to help manage pedigree and marker information used in linkage studies.

PMMS is a tool for information management in a unique situation; data for PMMS is gathered on an Apple Macintosh, and PMMS data is analyzed by a linkage analysis package on an IBM personal computer. In this situation, PMMS can be considered as an interface between these two areas, as well as a database management system in its own right.

The main goal of the PMMS is to organize and combine data gathered in the field and the laboratory into datasets ready for analysis by the LINKAGE program suite (24). Results, in the form of LOD scores, from the LINKAGE programs are then used to draw conclusions about linkage of two genetic loci, usually a genetic marker (i.e. probe locus) and a putative disease gene.

3.1.2 PMMS Data Base

The PMMS database organizes information from four major areas:

- Family Information
- Probe Information
- Blot Information
- Result Information

Each of these areas will be described in detail in the following documentation. Included in these descriptions are file definitions, entry standards, menu utilization, and organizational design specifications.

3.1.3 PMMS Interfaces

The entire PMMS is menu driven (i.e. all options are selectable from a menu), and easy to use. Complete descriptions, with examples, of the User Interface are contained in the PMMS Interface documentation.

Since PMMS deals with information that is gathered and analyzed on separate systems, there are many translation and import options available. The interface documentation also contains descriptions and examples of data transfer in and out of PMMS.

3.2 System Requirements

The PMMS is designed to operate under the dBASE III plus software package (or equivalent, e.g. dBXL). The minimum system requirements for implementation on a IBM PC/XT compatible are: 512K RAM, 10MB Hard Disk, 5 1/4 in. floppy disk, and a 80 column printer. If you wish to use the Macintosh based portion of PMMS you will need that computer, the Microsoft WORKS program, and a method of transferring text files between the PC and the Macintosh.

An important point about PMMS is that all the programs are written in the dBASE command language. When dBASE compatible database management systems (DBMS) become available on Apple Macintosh or UNIX based systems, PMMS will be portable to those environments. (Note: dBASE for the Mac is expected this Spring). When this, coupled with the fact that we now have efficient two-way communication and file transfer between the IBM and Apple worlds (via Mac-II 5 1/4 in. disk drive), it is apparent that the PMMS goes a long way towards integrating our two data environments, and allows flexibility in distributing computing tasks between them.

3.3 PMMS Data Base Design

3.3.1 System Description

The purpose of this section of the documentation is to describe the PMMS in terms of what information it can hold, and how it makes that information available to the user in terms of queries and reports.

The main database files used by PMMS include the FAMILY, PROBES, LOCI, BLOTS, and RESULTS files. These files are organized in the PMMS file system along with other index, import/export, program and support files. The File Definition section contains a description and layout for each file. The Data Standards section describes the organization of the file, and its location in the PMMS file system. (An outline of the PMMS file system can be found in Appendix A).

3.3.2 Overview of Files

3.3.2.1 FAMILY File

3.3.2.1.1 FAMILY File Definition

The FAMILY file contains information on individuals in families. For each individual there is one record in the file which contains the individual's family ID number, and individual ID number. Any given individual in the PMMS may be identified by the combination of these two numbers, which are unique for each individual. The record also contains basic demographics on the individual: Sex, Date of Birth, Affection Status, etc. A very important piece of information in the family file is Mother ID and Father ID. These are the individual ID numbers of the person's mother and father, and they act as pointers for the computer to determine family relationships.

A given family file may contain one or many families, each distinguishable by its family number. Typically one would create a family file for all the families in a given study. For example the FAMILY file "CEPHFAM" might be the file for storage of all CEPH study families.

The user is given complete access to the family file and may add, delete and edit any family information. The family file definitions are as follows:

FIELD NAME	DESCRIPTION	TYPE	LENGTH
FAM_NO	Family ID Number	Char	5
IND_NO	Individual ID Number	Char	5
MOM_NO	Mother's ID Number	Char	5
DAD_NO	Father's ID Number	Char	5
SEX	Individuals Sex (M,F or blank)	Char	1
DOB	Date of Birth	Char	8
AFFECTED	Affection Status	Char	2
UPDATE	Date of Last Information Update	Char	8
DECEASED	Date of Death (Blank if living)	Char	8
PROBAND	Individual is Proband (Y or N)	Char	1
NAME	First 20 Characters of Name	Char	20

3.3.2.1.2 Data Standards

All FAMILY files are indexed by FAM_NO and IND_NO, the index has the same name as the FAMILY file. All family files and indexes are stored in the FAMILY sub-directory of the PMMS file system. Import to a family file follows the usual procedure described in Appendix B.

Family file manipulation is performed via the Family file menu, which is described in Appendix D.

3.3.2.2 PROBES File

3.3.2.2.1 PROBES File Definition

The PROBES file contains information on DNA probes that are available in the Lab. It is mainly a list kept for convenience in helping decide which probes might be most informative to run next. Reports from this file can also help identify areas of the genome where our probes are few, and keep us on the alert for availability of new probes in these areas. Probe reports are also a useful tool in the process of deciding what blots to produce.

Each record in the file contains information for one probe. The user is given complete access to the probe file and may add, delete, and edit any probe information. Here is a list of the PROBE file definitions:

FIELD NAME	DESCRIPTION	TYPE	LENGTH
PROBE	Name of Probe	Char	10
LOCUS	Name of Locus (D Number or Gene Name)	Char	10
CHROMOSOME	Number of Chromosome	Char	3
REGION	Chromosome Localization	Char	15
VECTOR	Vector in which Probe Resides	Char	7

SITE	Cloning Site in Vector	Char	10
SIZE	Size of Probe Insert	Char	7
PROVIDER	Name of Probe Provider	Char	15
NOTEBOOK	Notebook Reference of Probe Receipt	Char	10
DNA_PREP	Notebook Reference of DNA Preparation	Char	10
RFLP_ENZY	Enzymes Showing RFLPs (Separated by commas)	Char	20
PIC	PIC for First Listed Enzyme (Usually Highest)	Char	4
ANTIBIOTIC	Antibiotic Resistance of Vector	Char	4
CLONE_LOC	Location of Clone (Freezer/Box)	Char	10
DNA_LOC	Location of Probe DNA (Refrig/Box)	Char	10

3.3.2.2.2 Data Standards

The PROBES file is indexed by PROBE, and has the same name as the file. The PROBES file and index are stored in the base directory of the PMMS file system. Import to the PROBES file follows the usual procedure described in Appendix B.

Probe file manipulation is performed via the Probe file menu, which is described in Appendix D.

3.3.2.3 BLOTS File

3.3.2.3.1 BLOTS File Definition

The BLOTS file contains information on Southern blots that have been made, or are to be made, in the lab. For each blot there is one record in the file. The record contains the basic parameters of the blot: what samples are on it, what enzyme was used to cut the DNA, percent of the gel, volt hours, etc.

This file is useful for two purposes. It can be used to check which blots an individual or family appears on, and what enzymes we have run for that individual or family. This file also serves as an "order desk" for blots to be made, records with a blank notebook reference field represent gels that have been designed, but not made yet. Reports from this file are used to help plan the blots we need to make.

The user is given complete access to the blot file and may add, delete and edit any blot information. Here is a list of the blot file definitions:

FIELD NAME	DESCRIPTION	TYPE	LENGTH
BLOT	Blot Number (e.g. EA12)	Char	10
ENZYME	Restriction Enzyme used on DNA	Char	6
NOTEBOOK	Notebook Reference on Blot Production	Char	15
PCNT_GEL	Percentage of Gel	Char	5

GEL_SIZE	Dimensions of Gel (in cm)	Char	8
VOLT_HRS	Voltage Hours Gel was Run	Char	4
UGMS_LANE	Micrograms DNA Sample Loaded per Lane	Char	5
FAM_1	Family ID of Sample in Lane 1	Char	5
IND_1	Individual ID of Sample in Lane 1	Char	5

..... (Fields for Lanes 2 thru 19) ....

FAM_20	Family ID of Sample in Lane 20	Char	5
IND_20	Individual ID of Sample in Lane 20	Char	5

3.3.2.3.2 Data Standards

The BLOTS file is indexed by BLOT number, and the index has the same name as the BLOTS file. The BLOTS file and index are stored in the base directory of the PMMS file system. Import to a BLOTS file follows the usual procedure described in Appendix B.

Blot file manipulation is performed via the Blot file menu, which is described in Appendix D.

3.3.2.4 LOCI File

3.3.2.4.1 LOCI File Definition

The LOCI file contains information on the alleles that comprise a given genetic locus. Each record represents a possible allele for a locus and contains a code, used by the linkage analysis program, to describe that allele. The purpose of the LOCI file is to translate the codes entered in the results file for co-dominant markers, and the family file for sex and affection status, into codes that can be interpreted by the LINKAGE program package.

The user is given complete access to the LOCI file and may add, delete and edit any locus information. Here is a list of the locus file definitions:

FIELD NAME	DESCRIPTION	TYPE	LENGTH
LOCI	Name of Locus (e.g. SEX, D17S34, etc.)	Char	10
RESULT	The RESULTS File Entry	Char	4
CODE	The LINKAGE Interpretable Code	Char	14

3.3.2.4.2 Data Standards

The LOCI file index is ordered by the LOCI plus RESULT fields, and has the same name as the data file. The LOCI file and index are stored in the base directory of the PMMS file system.

Loci file manipulation is performed via the Loci file menu, which is described in Appendix D.

3.3.2.5 Results File

3.3.2.5.1 RESULTS File Definition

The RESULTS file contains the data on observed phenotypes from Southern Blots hybridized with a given probe and exposed to autoradiography. The format of the file is one record for a given individual for a given probe.

The user is given complete access to the results file and may add, delete, and edit any result information.

FIELD NAME	DESCRIPTION	TYPE	LENGTH
PROBE	Name of Locus (e.g. SEX, pl44d6, etc.)	Char	10
FAM_NO	Family ID Number	Char	5
IND_NO	Individual ID Number	Char	5

RESULT	Blot Lane Result Code	Char	4
BLOT	Blot Number	Char	10

3.3.2.5.2 Data Standards

The RESULTS file is indexed by PROBE and Family ID plus Individual ID, allowing quick access to all the results for a given probe. The RESULTS file and indices are stored in the root level of the PMMS file system.

Importing data to the RESULTS file follows a slightly different procedure than for other imports. This is due to the fact that an input record contains results for all individuals on a given blot, and represents one output record for each individual.

The Macintosh Microsoft-WORKS input results files have different organizations depending on the designer of the form (this is unfortunate and will be rectified soon). Thus for a given import session some program variables need to be set in the PMMSRIMP.PRG program file to define which field is the first valid result, which field is the first valid ID, and the format for interpreting the Family ID for individuals on the blot. Also unfortunate (and also to be rectified) is the use of commas in the WORKS files, these need to be edited out before translation into IBM format (see Appendix C for example).

Result file manipulation is performed via the Result file menu, which is described in Appendix D.

3.3.3 Overview of Reports

3.3.3.1 REPORT

3.3.3.1.1 Report Definition

Specifics on reports are still being defined. Here is an initial list of reports:

- Family Report: Shows all members of a given family with generation and parent information, plus demographics.
- Individual Report: Shows all information for an individual, including all results at all loci.
- Probe Report: Shows all individuals run for a given probe.
- Basic List reports: Show all data for each database.
- Etc.

3.3.3.1.2 Report Standards

This section describes the style and function of each report.

3.4 PMMS Data Base Interface

3.4.1 Menu Driven User Interface

All PMMS functions are accessible from menu options presented to the user for selection. Options are selected by number, and there is an organizational hierarchy for access. This hierarchy is presented in appendix D to assist you in navigation through the PMMS system.

3.4.2 File Import

The PMMS file import utilities are of great importance in this release of PMMS since no real data entry will be occurring on the IBM side of the PMMS system. Instead all PMMS data is to be entered into Microsoft WORKS on the Macintosh, and translated to an IBM readable text format (all fields separated by commas, all records on a separate line), which is the entry point to the IBM portion of PMMS.

To facilitate the easy import of diverse files we have implemented a generic import scheme which is used for all the files in the database. As a guide to which field of an import file goes where in a PMMS file, we have created a Translation Format file. This file has only one record, that record has 55 fields (an arbitrary limit for the maximum number of fields transportable from WORKS). For each of the input file's fields that the user wants to have transferred to the PMMS file, the corresponding field in the translation file contains the field number of the destination field in the PMMS file.

For example, suppose that we were importing PROBE file records, and that the WORKS file field for "Chromosome Location" was the fifth field in a database record. Knowing that the CHROMOSOME field is number 3 in the PMMS PROBE file, we would build our translation file with a 3 in field "F5". The translation file also contains the name of the PMMS file (e.g. PROBE), and the number of fields in the import record (e.g. 20 for the "Probes On Hand" file). Here is a description of the translation file:

FIELD NAME	DESCRIPTION	TYPE	LENGTH
FILE	Name of PMMS File to which Import will go	Char	8
FIELDS	Number of Fields in the Import File	Numb	2
F1	Number of PMMS field for the 1st Import Field	Numb	2
F2	Number of PMMS field for the 2nd Import Field	Numb	2
...	....	...	..
...	....	...	..
F55	Number of PMMS field for the 55th Import Field	Numb	2

Each translation session is prefaced with questions on what the name of the Import file is, and what Translation file is to be used in the Import process. Currently there is no utility for the creation of translation files in the PMMS system. Translation files for all the current WORKS files have been prepared, and are listed in Appendix B. To create a new Translation format file one enters the DBXL program, selects the XFORM database (in the TRANSFER sub-directory), and copies the structure to an appropriately named new database (also in the TRANSFER sub-directory). Then one appends a single record containing the field translation information to the newly created file. When prompted for the name of the translation format file when performing the transfer, the user enters the name of the newly created file and the transfer is processed. See Appendix C for a complete example using the results file import.

3.4.3 File Export

The file export portion of the PMMS is really the most important part of the system, since the main purpose of PMMS is to organize laboratory and field data for linkage studies. File export allows creation of a data file suitable for use with the LINKAGE programs to report LOD scores for linkage analysis. The format for pedigree and phenotype data required by LINKAGE is reviewed in the LINKAGE ANALYSIS PROGRAMS USER'S GUIDE, which is provided with the LINKAGE software (24). Family and result data from

PMMS are combined to create the output file. The output file is analyzed in conjunction with a loci description file that is created under LINKAGE to produce a LOD score table.

When users select the Export File option from the PMMS main menu, they are prompted for the file name and ID number of the family they wish to export. Then they select loci for which phenotype data exists for the family. Finally, they enter the name of the output file and processing begins. The final output file is placed in the DATA sub-directory of the PMMS file system. See Appendix C for a complete example.

3.5 APPENDICES

3.5.1 Appendix A: PMMS File System

The PMMS file system has been designed to be relocatable, and its root level may be set in a dBase memory variable file which contains two variables: DR for the drive, and BP for the base or root level file path (e.g. \DBXL\PMMS) for the PMMS programs and files. These exist in the file PMMS.PRM and may be changed using dBase. Currently the root level of PMMS is set to 'C:\DBXL\PMMS'. Below this root level there are five sub-directories:

o FORMATS: This sub-directory contains all of the formats for all files associated with the PMMS.

o FAMILY: This sub-directory contains all of the family database files.

o IMPORT: This sub-directory is where PMMS looks for any files to be imported into the database.

o DATA: This sub-directory is where all exported LINKAGE ready files are written.

o TRANSFER: This sub-directory contains all the transfer file formats.

A batch file called "PMMSBACK" has been created that copies all files and directories involved with PMMS to floppy disk (Drive A:) for backup purposes. It is suggested that this batch file be executed after any major changes to the database, or at a point when the user decides they would rather not re-enter the most current data if a file was somehow lost.

A batch file called "PMMSBEG" that sets paths and starts the PMMS has been created and is copied into the computer's root file system (e.g.C:\) by the INSTPMMS program.

3.5.2 Appendix B: Import File Definitions

This appendix lists the Macintosh Microsoft WORKS files for which translation files have been created under PMMS, and describes their use.

The "EA Family Data" WORKS file has a translation file named "XFAM".

3.5.3 Appendix C: Interface to LINKAGE - An Example

3.5.3.1 Purpose

The purpose of this appendix is to take the user through the use of the PMMS. The example we will use will be the process of transferring blot result data from the Macintosh to the PC, importing it to PMMS, and exporting it to LINKAGE. This is a fairly involved process, but don't worry, I will describe every step and make it as simple as possible.

3.5.3.2 Import From Macintosh

The first step in exporting data from the Macintosh Microsoft WORKS program is to open the file to export from, in this case the "Results File", on the Mac. Once open the user should select the records for export, using selection criteria, and select the "SAVE AS" option from the FILE menu. The user should enter a new name for the transfer file, and click both the EXPORT FILE and SAVE SELECTED RECORDS ONLY options. This will save a text file containing only the records we want, with fields separated by tabs, and one record on each line.

The next step is to open this file with the WORKS word processor. What you will see is the previously selected records, wrapped across the screen. We want to change two things (in this order): 1) remove all commas, and 2) change tabs to commas. So select REPLACE from the SEARCH menu, enter a comma in the FIND field, and leave the REPLACE WITH field blank, select replace all, and wait (this can take several minutes, be patient). Now using the cursor select the first tab between the first two field names on the top line of the file (select by dragging over the tab, which will make a darkened tab that looks like a bar), now select the REPLACE option and you will see the FIND field is darkened (the find is now set to tab, although you can not see it), select the REPLACE WITH field and type a comma. Select replace all, and wait again. When this process is finished delete the first line, which may be wrapped around (it only contained the names of the fields), and save the file as an export file (under the SAVE AS option of the FILE menu). You should consider naming the file with a PC like name: eight characters, a ".", and a three character extension (e.g. CEPHRES1.TXT). NOTE: the three character extension must always be "TXT" for import files.

We now have the exact file we want, except that it is on the Mac, and we need it on the PC. We will use a Mac II with a 5 1/4 inch MS-DOS disk drive for the next step. From the Finder, open the Apple File Exchange (AFE) utility which is in the System folder, in the Utilities 2 folder, in the Apple File Exchange folder. AFE is simple to operate, when it starts, open the folder containing the file for export on the left half of the screen, and select the file (darkens by clicking on it). Put a PC formatted disk with enough room for your transfer file in the Mac II 5

1/4 inch disk drive, and close the door. You will see the directory of the PC disk in the right hand window. Now de-select the MacWrite to DCA option from the Mac to MS-DOS menu, and click the translate button to begin the copy. When done, you may remove the PC disk and quit the program. The remainder of the work will now be on the PC, all PC commands should be ended with a press of the Enter key (large key on right of keyboard with the bent arrow).

Take the PC disk and put it in the PC's 5 1/4 inch drive (either A or B drive). Copy the file from the floppy to the PMMS IMPORT sub-directory (i.e. "COPY A:CEPHRES1.TXT C:\DBXL\PMMS\IMPORT"). Now start the PMMS by changing to the C drive's root directory (i.e. "CD C:\" at the C prompt) and typing "PMMS".

You will now see the PMMS program start up, and will eventually be presented with the PMMS main menu. From here we want to select the Result File Management option, and then the Results Import option (Note that you do not hit the Enter key after making a selection from a PMMS menu). You will be prompted for the name of the import file (with the option to list all files in the IMPORT sub-directory). Use the file just copied to the import sub-directory (TRANSFER.TXT), you should only type the first part of the name, PMMS assumes a ".TXT" extension. The file will then be opened and PMMS will report on how many records have been read in. The computer will then transfer the new records to the results file, PMMS will give a running count of each result transferred. Remember, a result record corresponds to data for a single individual (i.e. one gel lane), so the number of results added will be much larger than the number of records read.

To check the results you added, you can select the Browse option from the results menu. If everything looks good we are ready to proceed to the export of the data to the analysis software. Exit the result file management by selecting the Return to Main Menu option.

3.5.3.3 Export to LINKAGE

To export data for a family to a file usable by the LINKAGE analysis programs select the Transfer Utility menu option from the PMMS Main Menu. You will be asked to enter a family file name (use "?" as the first character in the answer for a list). Enter the name of the family file that contains the family you wish to run linkage analysis on.

Next you will be asked to identify the specific family in the family file by entering the family number for that family, again the "?" will give you a list of all the families in the file. NOTE: Some family files may contain a single family. Now that we know the set of individuals for which we want to export data, we can select which data we want to export.

The last question is for which loci are results to be translated. This is answered in a dialog with PMMS where you are shown two windows: one is a list of probes (from the LOCI file) and the record numbers associated with them, the second is a list of the probes you have selected for export (initially empty - note that sex and affection status are automatically transferred). The process of selecting probes is as follows: you can see more of the list in the Probes window by pressing the PGDN key (the 3 on the keypad on the right side of the keyboard). NOTE: the PMMS program is VERY SLOW here, press keys once and WAIT for the result. To select a probe, type the record number of the probe you want and press return (the large key marked "ENTER"), it will be added to the selection list. To change a probe simply move to the line that probe is listed on (using the UP and DOWN ARROW keys, also on the keypad) and typing the new number followed by return. If you type a wrong key, or a record number outside the range of probes, the computer will beep. To remove a probe from the selection move to the line and press DEL (the "." on the keypad). When finished with the selection press return to exit the selection window, and again to confirm your selections (if you with to abort the export at any time you may press "Q").

The final step is to enter the name of the file you want the linkage formatted data to be placed in. Just enter the first part of the name, PMMS adds a '.TXT' extension, and places the file in the "C:\DBXL\PMMS\DATA" sub-directory. The export module keeps you informed of its progress showing you the output it creates (it takes a bit of time), and returns to the Main Menu when it's done.

3.5.3.4 Running LINKAGE

For all the popularity of the LINKAGE program suite, I have yet to find a decent written explanation of how to really use the programs. Much trial and error has gone into the following brief guide to using the MLINK (multi-point & two-point linkage analysis) program, which is only a small subset of the suite. The LINKAGE ANALYSIS PROGRAMS USER'S GUIDE is required reading for background information, but I would be derelict in my duties if I simply referred you to it, since it does not contain a step by step approach to doing anything.

The first step in our analysis is to copy the file from the PMMS export directory to the LINKAGE disk area. First quit the PMMS system (last option on the Main Menu). When returned to the C> prompt change directories to the root directory (type "CD C:\"), and type "CD C:\LINKAGE\DATA". You will be whisked to the LINKAGE "DATA" directory. Set the path to execute programs in the LINKAGE "EXE" directory by typing "PATH=C:\LINKAGE\EXE". Now to copy the file, say we named it "CEPHFAM.TXT", type "COPY C:\DBXL\PMMS\DATA\CEPHFAM.TXT C:\LINKAGE\DATA". There are three steps to perform before we begin our analysis: editing the export file, running PEDPOINT, and running PREPLINK.

Editing the export file can be done with the DOS EDLIN editor. To begin simply type "C:\DOS\EDLIN CEPHFAM.TXT". Now use the editor to remove any branches of the family not of interest to our analysis. Directions for using EDLIN may be found in any DOS manual, or use the editor of your choice.

Running PEDPOINT converts the export file into a LINKAGE format file by re-coding family and individual numbers, and creating family pointers. At the C> prompt type "PEDPOINT", enter the name of the input file (e.g. CEPHFAM.TXT) and the name of the output file (e.g. CEPHFAM.PED), don't use the same name for your input and output files. You will be asked if the file contains pedigree numbers, the answer is no. Answer no to all the remaining questions, and the program will process the output file. If any errors occur (almost always), you may need to re-edit the file and try again. The error messages are mostly self explanatory, checking that all referenced people are present, and that "0000" is entered for unknown individuals is the most common process.

Running PREPLINK is fairly straightforward, it creates a parameter file describing the loci in the data file. It has many other important functions as well, which are too involved to be presented here.

Now that all our support files are in place we may begin the analysis. LINKAGE V4.6 has been set up to make this step quite simple. A control program (LCP) creates a batch file that organizes all the files and runs the analysis. To create this batch file type "LCP" at the C> prompt. You will be presented with a list of files to be specified: the first is the name of the batch or COMMAND file - use something to remind you of the analysis at hand (e.g. CEPHRUN1.BAT). The second is the output or LOG file which will contain all the results of the run - use the same name as for the batch file with a ".OUT" extension (e.g. CEPHRUN1.OUT). The third file is a diagnostic report on the analysis and can be left as is

(STREAM.OUT). The fourth is your family or PEDIGREE file (e.g. CEPHFAM.PED). The fifth is your PARAMETER file (CEPHFAM.DAT). The sixth option is to be left "NO", and the last two options as blank. When the screen has been completed press PGDN.

Next you will be asked to pick an analysis program, select "MLINK" using the arrow keys, and press PGDN. Select "Specific Evaluations", press PGDN, select "No Sex Difference" press PGDN. Now you are presented with the final screen for the run, the first entry is locus order. This is the locus numbers of the loci in the PEDIGREE file you which to analyze for linkage, usually locus one separated by a space from one of the marker loci (e.g. "1 3"), and press return. Next is the initial recombination fraction, use ".05", press return. Next is which recombination to vary, for two-point analysis this is always "1", for multipoint enter "1" to vary recombination between the first and second entered loci, "2" to vary recombination between the second and third entered loci, etc. The increment value should be ".05", and the stop value should be ".45". This will give you a LOD score table starting at theta equal to .05 and going in .05 increments to theta equal to 0.5. Press PGDN when the screen is complete, and the commands for this analysis will be placed in the COMMAND file (e.g. CEPHRUN1.BAT). If you wish to produce LOD scores for other loci, repeat this process beginning at the selection of the MLINK program. When done press the CTRL and "Z" keys together (CTRL first and hold it down, then "Z") to exit. Don't press CTRL and A or all the entry you have done in the LCP program will be lost (i.e. Aborted).

When you return to the C> prompt simply type the name (no extension) of the command file and watch linkage take off! Your results will be left in the LOG file (e.g. CEPHRUN1) you selected. You may print them by typing "COPY CEPHRUN1.OUT LPT1:", or display them on the screen with "TYPE CEPHRUN1.OUT".

3.5.4 Appendix D: Menu Access to PMMS

All of the PMMS menus are basically identical, so a single general description of the menus will be given. There are two levels of menus in PMMS: the Main Menu, and Sub-Menus. The single main menu has 6 options, and each PMMS sub-menu has 7 options; we will look at each option in detail, and describe its processing.

Selecting menu options is easy; when PMMS starts you will be presented with the main menu, simply press a key for the number of the option you wish to select (e.g. press '1' to select Blot file maintenance). Pressing any other key besides those available on the menu will have no effect. The first four options of the main menu are for working with the various data files (i.e. FAMILY, PROBE, etc.). Selection of these options takes you to the sub-menu for the data file you selected. Note that both Probe and Loci management are accessed by option 2. Also note that the first question presented after selecting the FAMILY maintenance option is to enter the name of the FAMILY file with which you wish to work. If you have not yet imported any family data just use the default FAMILY database. The fifth option is for translating PMMS data into LINKAGE format (see Appendix C). The sixth option is to exit the PMMS program. When this option is selected you will be asked if you are sure you wish to exit the system. A "Y" or yes response, or just pressing the enter key, will stop PMMS and return you to DOS, leaving you in the PMMS sub-directory. A 'N' or no response will return you to the main menu.

The PMMS sub-menus have seven options to allow maintenance of data in each of the five PMMS files. The user may Add, Edit, Delete, Browse, Report, or Import data, and exit the sub-menu with these seven sub-menu options. Let's look at each option in detail.

The add data option, when selected, presents the user with a blank entry screen with a highlighted input field for every data element in the file. The entry screen may be filled in with data for a new data file record. Each field may be moved through with the right and left arrow keys, the insert key toggles insert and overwrite modes. The delete key deletes the character selected by the cursor, and backspace deletes the character to the left of the cursor. The home key moves the cursor left one word, the end key moves the cursor right one word. Using the control key with the left and right arrows moves the cursor to the beginning or end of an input field. The standard procedure for adding a record is to fill in the entry screen a field at a time, reaching the end of an entry field will move you to the next one, or the up and down arrows will move you back and forth between fields. When the last field is filled, and the screen looks correct, pressing return with the cursor on the last entry field, or holding down the control key and pressing the PgUp key, will save the record to the data file, and return you to the sub-menu. If at any time you wish to abort the addition of the record, pressing the Esc (i.e. Escape) key will return you to the sub-menu without saving the new record to the data file.

ADD RECORD

SUMMARY:	Up and Down Arrows	.....	Move Between Fields
	Left and Right Arrows	..	Move Within Fields
	Ins Key	....	Toggle Insert & Overwrite
	Home & End Keys	....	Left and Right One Word
	Return & CTRL-PgUp	....	Save New Record
	Esc	....	Don't Save New Record

The edit data option first presents the user with a screen on which to enter the 'key' for the record to be edited. The key is the field or fields by which the data file is indexed, and acts to identify a record. A lone question mark entered as the first character in this field will make a small window that scrolls a list of the keys for the data file (one key is displayed for each record in the data file) across the screen, allowing users to spot the key they want to access. The question mark is the default entry, and by just pressing "return" the user will see such a display. When the window fills with keys the message 'MORE..' is highlighted at the bottom of the window, by pressing enter another window of data is displayed. Pressing the space bar at the 'MORE..' message scrolls the keys through the window one at a time, and pressing the Esc key aborts the key display, returning the user to the key entry screen. A lone exclamation mark entered as the first character in the field causes a window to be displayed asking the user to enter a record number for the desired record. This record number is then used as a key to select the record from the data file. Note that the record number is displayed in the key window.

Once the key is entered, the data file is searched, and the first match is displayed, along with a message asking if the displayed record is the one the user wishes to edit. If the user responds with a 'Y' for yes (the default), then the record is re-displayed and the user may edit it, using the same keys as described in the add record section. If the user responds with a 'N' for no, the next record is displayed and the user is again asked if it is correct. By responding with a 'Q' for quit, the user is returned to the sub-menu.

EDIT RECORD

SUMMARY:	? in Key Field	....	Display Key Window
	Space Bar	....	Single Scroll in Key Window
	Return	....	Continue in Key Window
	! in Key Field	....	Display Record Number Window

The delete data option starts identically to the edit data option, in that the user selects a record by key. Then the user is asked to confirm the deletion of this record by typing 'YES'. Any other entry will abort the deletion. Note that when a record is deleted under PMMS it is not visible in the system, but it still exists in the file. To either remove it permanently (say if your files are getting too big) or restore it to the file you must use dBase III to PACK or RESTORE the deleted records.

The browse option is a very useful option for scanning data in a file. When browse is selected the user is presented with a screen depicting the first record in the data file. By pressing PgUp and PgDn keys the user can scroll through the database one record at a time. By pressing the '?' key the user can get a window of all the keys, as described in the edit data option. Also in the key window is a record number. By pressing the exclamation point key the user is presented with an entry field for a record number. The record for the number entered by the user is then displayed. If the number entered is larger than the total records in the file, the last record is displayed.

BROWSE RECORD

SUMMARY:	PgUp	....	Display Next Record
	PgDn	....	Display Previous Record
	?	....	Display Key Window
	!	....	Display Record Number Window

The report option is designed to display another sub-menu of reports. However in this release of PMMS, due to problems with printer incompatibilities, we have implemented only a very basic report option. The user should have their printer turned on when selecting this option.

The import data option is reviewed in Appendix C. Note that the import data sub-option is not available for the Loci file.

The return to main menu option allows the user to exit the sub-menu and return to the main menu.

4 Discussion

4.1 Overview

In this discussion I will recount discoveries and obstacles encountered in the evolution of this thesis, and provide a perspective from which the significance of the work may be viewed. I will begin with a brief summary of the line my work has followed, by reviewing the process of developing my results in terms of the successes and failures I have encountered. I will then give a review of what I did in each chapter of this thesis.

4.2 Summary

This work began as a linkage study of episodic ataxia (EA) in a single family. It has grown to include an EA linkage study in three families, which have been used to successfully exclude the location of EA from 3 regions of chromosome 17 representing 40% of its length, and from the short arm of chromosome 6 in the region of a potentially allelic neurological disorder: spinocerebellar ataxia (SCA).

Concurrent work by others in our laboratory provided a suggestive linkage of EA to a marker on the long arm of chromosome 11 (11q) in study family 5910. And although more than 30 markers from 11q, concentrated in the region of the initial linkage, have been tested for linkage with EA in all three study families, no more informative linked marker has been found. Three 11q markers less significantly linked to EA in family 5910 have been found (CRI-L962: theta=0.18, LOD=1.40, phi8-10: theta=0.001, LOD=1.21, cTBZ6: theta=.085, LOD=2.11).

An important tool in solving problems in gene mapping of this sort is a linkage map of the region of interest. Such a map would include both genetic distances between the markers we have used in linkage analyses, which can be used to increase the significance of a linkage by multipoint analysis, and their relationship to contiguous cloned DNA segments, which can be used to find new closer markers, and ultimately the gene itself.

Our efforts to increase the significance of the observed results in the EA study by using linked markers on 11q in a multipoint linkage study have been hampered by the lack of a linkage map which includes all of these markers. Of the four linked markers, phi8-10 and phi2-25 appear on one 11q map based on CEPH family linkages (54), cTBZ6 has been mapped in CEPH families but is not on any map, and CRI-L962 is on the CRI map of 11q (also created using CEPH families (19)) which contains none of the other markers. It is theoretically possible to combine the CEPH based datasets created for each of these markers and map their relative locations. We will attempt to do this in the near future.

This situation has prompted the use of another approach to increasing our EA linkage significance: development of genetic markers on 11q in an attempt to find more closely linked markers. Eight RFLP markers spanning 11q were found by screening 20 cosmids from a chromosome 11q specific cosmid library. Four of the these new marker loci (1-16, 1-16.b, 4-7, and

12-16) are in the region of interest for EA. Unfortunately, none of these RFLPs, or the four RFLPs from other regions, have been informative enough to provide a closer linkage for EA in study families. Efforts to find more 11q markers are currently underway in our laboratory.

4.3 Episodic Ataxia

4.3.1 Episodic Ataxia Linkage Study

We have succeeded in excluding the location of a gene for EA from the region of the short arm of chromosome 6 known to be linked with SCA using two markers in multipoint analysis: pCH6 and pHHH157. This demonstrates that EA is not allelic with SCA in our study families. Three regions of chromosome 17, representing approximately 40% of that chromosome, have also been excluded as locations for EA in our families using multipoint linkage analysis.

Concurrently with this work, the process of testing many markers distributed throughout the genome for linkage with EA has been pursued in our laboratory. Testing of over 100 genetic markers has excluded EA from more than 30% of the genome. A suggestive linkage was found by Marilyn Jones between EA in family 5910 and phi2-25 (theta=0.063, LOD=2.68) on the long arm of chromosome 11. Efforts to increase the significance of this linkage by developing better genetic maps of this region, and discovering additional markers in the region, are currently in progress in our laboratory, and are the subject of the next section.

Scanning the genome for EA has been made possible by the development of many genetic markers by a diverse group of laboratories. The determination of which markers to test is dictated by a number of factors including: how informative they are (based on PIC and heterozygosity), how well they are localized, if they have been genetically mapped, are they revealed with common restriction enzymes or will they require production of special blots, etc. To keep informed of available markers and the development of new markers many resources are employed.

The Human Gene Mapping Library (HGML) is a valuable resource in this regard, maintaining a list of all known genetic markers in a database which is accessible over the telenet network. The database of markers is cross-referenced to literature citations and addresses of the source laboratories. We obtain markers for use in our studies through collaborations with these laboratories. As a convenience, often the source laboratory will make the marker available through the American Type Culture Collection (ATCC) which provides both clones and DNA for a number of genetic markers. The ATCC order numbers are posted on the HGML, and orders to ATCC for markers can be placed electronically. The ATCC is currently working to develop genome screening kits, consisting of a series of well spaced mapped markers for each chromosome. This will be a great help in studies of this type, making it unnecessary for each lab to invent a mapping strategy as we have had to.

The development of genetic maps of many regions of the human genome, and especially chromosome 11q, combined with suitable EA families, gives us an excellent chance of successfully mapping EA's genetic location.

4.3.2 Lessons

There is an important lesson which I have learned in this study: Things do not always work out as you plan.

Of the 30 markers we have run in our families from the region of 11q 26 have been uninformative or only slightly informative. Almost all the markers run in the region of our most significant linkage (phi2-25 in family 5910: LOD=2.68 at theta=0.068) have not been able to confirm or exclude this region for linkage of EA. The most informative marker, and many of the other markers run, are only informative in the larger of the two sibships in family 5910. This is extremely frustrating since the information added by a second informative sibship would increase the significance of any linkage results.

An additional hurdle we face in a study of this type is based on the statistical nature of linkage studies. As more markers are tested for linkage with EA, we increase the statistical chances of seeing random co-segregation of EA with marker alleles. The acceptance of a LOD score of 3 as proof of linkage is based on a statistical argument which considers the a priori chance of observing a genetic linkage. The significance of a LOD score of 3 is thus decreased when many loci are tested for linkage, and the a priori chance of co-segregation increases. When the number of markers tested for linkage approaches 100, LOD score of 5 or more may be required for proof of linkage (30).

This fact underscores our previously identified need for more polymorphic markers in the region of 11q. To help fill this need I have developed more RFLPs in this region, as described in chapter 2, and summarized in the next section. All of the RFLPs I have developed, with the exception of 20-13 which gives a weak indication of linkage with EA (LOD < 1), are uninformative in both branches of family 5910. My only consolation in this fact is that it is consistent with our observations so far of other markers in the region.

4.4 11q RFLP Development

4.4.1 Characterization of 11q RFLPs

In an effort to improve the genetic map of chromosome 11q, eight RFLP loci revealed by six cosmids isolated from a set of 20 cosmids from a chromosome 11q library have been characterized.

G.A. Evans and K.A. Lewis used the TG 5D1-1 Friend cell line, which contains chromosome 11q13.1-qter as its only human component, as a source of DNA for the manufacture of this cosmid library (122). The resulting library of clones with human DNA inserts represents approximately 0.9% of the human genome. We were sent 20 randomly picked cosmids from this library.

To get usable DNA for RFLP screening from these cosmids it was necessary to cesium chloride-band the DNA preparations. It is important to note that using whole cosmids as probes on southern blots of digests of total human DNA is not an exact science. A certain number of poor results were obtained for each usable result in the process of optimizing the parameters of the experiments. Reasonable results were eventually obtained for each of the 20 cosmids screened by repetitive iterations of parameters including wash temperatures, probe preparation, and hybridization technique.

A problem affecting one of the 8 RFLP loci I isolated should be mentioned:

The initial absence of individuals homozygous for the rare allele in marker 24-10 was disturbing. In system 24-10 we would have expected to see approximately 10 individuals homozygous for the minor allele in the panel of 36 CEPH parents based on Hardy-Weinberg expectations calculated from the allele frequencies in this population, and we had seen none (p<.001). The major allele detected by 24-10 is constant in size on blots containing DNA digested with a 20 fold excess of MspI, indicating the band is not caused by partial digestion. A gel fragment which revealed the RFLP was isolated from a restriction digest of cosmid 24-10. However, the gel fragment gave identical phenotypes as the whole cosmid.

Densitometry data for 24-10 allowed us to determine the location of the major allele. By typing two families in which both parents were heterozygous for 24-10 (with a total of 13 children), I was finally able to find a homozygote for the minor allele, providing unequivocal evidence of the correct allele sizes.

A ninth RFLP revealed by cosmid 25-3 was eventually dropped from our dataset because no individual homozygous for the rare allele was found in the 36 CEPH parents, or in any of the CEPH families typed (including three families with both parents heterozygous). Speculation as to why this RFLP does not conform to Hardy-Weinberg expectations has included the possibility that the rare allele is in complete linkage disequilibrium with a recessive condition not compatible with life, but which is maintained in the population by heterozygote advantage or other mechanism. It is also possible that the absence of the major allele is masked by a constant band of the same size. The probability of this explanation is decreased by the fact that a gel fragment revealing the RFLP gives identical phenotypes as the whole cosmid.

Restriction fragments which reveal many of the other RFLPs have been identified, and are in the process of being cloned.

Using 24 CEPH reference families, two-point linkage analysis between these new markers and two sets of mapped markers has provided evidence of linkage between each new marker and one or more of the previously mapped markers. Evidence of linkage between these new markers and previously mapped markers which localizes their genetic location is consistent with previous chromosomal localizations and somatic cell hybrid mapping results for all markers. The six cosmids have been mapped by two-point linkage analysis with LOD scores proving linkage (i.e. LOD > 3), with the exception of 24-10 which has a maximum LOD of 2.4 for localization. The locations of the new markers are well distributed along 11q, making them useful for increasing the number of markers spanning 11q. The three cosmids mapping to the region closest to our EA linked markers have been screened for additional RFLPs. No additional RFLPs have been found with these cosmids in at least six unrelated individuals, whose DNA was digested with six seperate restriction enzymes.

Cosmid 1-16 has been mapped by high-resolution in situ techniques to 11q23.3 (58). This assignment is consistent with linkages observed between 1-16 and other markers that have been high-resolution mapped (e.g. THY1 with 1-16: theta=.029, LOD=7.29; APOA1 with 1-16: theta=.0001, LOD=3.9). Glen Evans has been informed of my RFLP localizations, and is in the process of performing high-resolution in situ florescent hybridization with these cosmids. This work should eventually yield data on the relation between physical and genetic distance on the long arm of chromosome 11.

4.4.2 Future Directions

Publication of the first CEPH consortium map (53) heralds the realization of the promise of collaborative genetic map building (107). The map was produced from data developed and analyzed at several separate laboratories, using a number of techniques. This map represents the first comparative review of the various mapping strategies and tools. This, coupled with the fact that CEPH has made the collaborative dataset available to the scientific community, means that the time for standardizing map building approaches and the datasets they involve is at hand.

I plan to lend my skills to the fulfillment of this standardization by doing post-doctoral research with Dr. Mark Lathrop at the CEPH institute. I plan to work on the development of new genetic markers, and to address the issues of genomic informatics.

4.5 PMMS

4.5.1 Laboratory Workstations

The PMMS was developed to process linkage study data in the framework defined by studies of genetic disease (i.e. systems approach). Chapter 3 is a manual of the PMMS system describing what it does, the structures of the datasets it uses, and how it is operated. I will now talk about why it is useful.

The concept of the electronic lab book has not yet been realized. This theoretical computer/book would allow easy recording and annotation of any laboratory results, organization of results for reporting purposes, interfaces for the analysis of results, and storage and annotation of the results of these analyses. Although systems of this type are available, their cost is prohibitive and they are targeted for specific applications.

In our laboratory we have employed a general purpose laboratory workstation, which is shared between lab workers to reduce costs. General purpose software running on this workstation has been tailored for gathering and organizing data related to linkage studies. Analysis of this data is provided by the LINKAGE program suite. The PMMS facilitates the transfer of this data between the workstation and the analysis software. I plan to extend PMMS to support the transfer of the linkage results back to the workstation.

Having our linkage data managed by electronic methods provides versatility in the reporting and analysis, and centralized control of its integrity.

I have also used PMMS to integrate data from disparate sources for linkage analysis. I have written programs to convert PMMS datasets into the CEPH data format. This has allowed the results of my 11q markers run in CEPH families to be analyzed with data from the CEPH V3 database, and to be sent to CEPH for inclusion in the CEPH V4 database. I have also written programs to convert a dataset of CEPH family typing results generated with 24 markers on 11q received from Dr. Mark Lathrop into the PMMS format. This dataset had been used to construct a multipoint linkage map of these markers (54). Using PMMS to combine these datasets with my dataset of new markers has allowed analysis of the localization my 11q markers with respect to other mapped markers.

PMMS is really a prototype for systems that use distributed database technology to integrate and standardize genetic data between laboratories. The insights I have gained in developing PMMS will be applied in my future contributions to the development of these new genetics systems.

REFERENCES

1. ST Gancher, JG Nutt (1986) Autosomal Dominant Episodic Ataxia: A Heterogeneous Syndrome. Movement Disorders 4:239-253.

2. D Margolin, JG Nutt, EW Lovrien (1982) Familial Periodic Ataxia. Trans. Am. Neurol. Assoc. 106:1-5.

3. EW Lovrien. Personal Communication.

4. NL Zasorin, RW Baloh, LB Myers (1983) Acetazolamide-responsive Episodic Ataxia Syndrome. Neurology 33:1212-1214.

5. R Mayeux, S Fahn (1982) Paroxysmal Dystonic Choreoathetosis in a Patient with Familial Ataxia. Neurology 32:1184-1186.

6. HL Parker (1946) Periodic Ataxia. Coll. Papers of Mayo Clinic 38:642-645.

7. JC White (1969) Familial Periodic Nystagmus, Vertigo, and Ataxia. Arch. Neurol. 20:276-280.

8. DH Van Dyke, RC Griggs, MJ Murphy, MN Goldstein (1975) Hereditary Myokymia and Periodic Ataxia. J. Neurol. Sci. 25:109-118.

9. RC Griggs, RT Moxley, RA LaFrance, J McQuillen (1978) Hereditary Paroxysmal Ataxia: Response to Acetazolamide. Neurology 28:1259-1264.

10. TW Farmer, VM Mustian (1963) Vestibulocerebellar Ataxia. Arch. Neurol. 3:471-480.

11. SH Orkin (1986) Reverse Genetics and Human Disease. Cell 47:845-850.

12. FH Ruddle (1981) A new era in mammalian gene mapping: somatic cell genetics and recombinant DNA methodologies. Nature 294:115-120.

13. NE Morton (1955) Sequential Tests for the Detection of Linkage. Am. J. Hum. Genet. 7:227-318.

14. JL Haines, JA Trofatter (1986) Multipoint Linkage Analysis of Spinocerebellar Ataxia and Markers on Chromosome 6. Genet. Epidemiol. 3:339-406.

15. NE Morton, J-M Lalouel, JF Jackson, RD Currier, S Yee (1980) Linkage Studies in Spinocerebellar Ataxia. Am. J. Med. Genet. 6:251-257.

16. M Leppert, P O'Connell, Y Nakamura, R Leach, M Lathrop, P Cartwright, J-M Lalouel, R White (1987) Extension to a Primary Genetic Linkage Map of Chromosome 6p. Human Gene Mapping 9: Ninth International Workshop on Human Gene Mapping. Cytogenet Cell Genet 46:727.

17. Y Nakamura, M Lathrop, P O'Connell, M Leppert, D Barker, E Wright, M Skolnick, S Kondoleon, M Litt, J-M Lalouel, R White (1987) A Mapped Set of Markers for Human Chromosome 17. Genomics 2:302-309.

18. P O'Connell, GM Lathrop, Y Nakamura, ML Leppert, RH Ardinger, JL Murray, J-M Lalouel, R White (1989) Twenty-Eight Loci form a Continuous Linkage Map of Markers for Human Chromosome 1. Genomics 4:12-20.

19. H Donis-Keller, et. al. (1987) A Genetic Linkage Map of the Human Genome. Cell 51:319-337.

20. RC Elston, J Stewart (1971) A General Model for the Analysis of Pedigree Data. Hum. Hered. 21:523-542.

21. JK Haseman, RC Elston (1972) The Investigation of Linkage Between a Trait and a Marker Locus. Behav. Genet. 2:3-19.

22. J Ott (1985) Analysis of Human Genetic Linkage. The Johns Hopkins University Press:Baltimore & London.

23. J-M Lalouel, T Elsner, GM Lathrop, P Callahan, A Oliphant, RL White, D Cohen, J Dausset (1985) The CEPH System of Management of RFLP Data. Eighth International Workshop on Human Gene Mapping. Cytogenet. Cell Genet. 40:676.

24. GM Lathrop, J-M Lalouel (1984) Easy Calculation of LOD Scores and Genetic Risk on Small Computers. Am. J. Hum Genet. 36:460-465.

25. GM Lathrop, J-M Lalouel, C Julier, J Ott (1984) Strategies for Multilocus Linkage Analysis in Humans. PNAS USA 81:3443-3446.

26. NE Morton, CJ MacLean, R Lew, S Yee (1986) Multipoint Linkage Analysis. Am. J. Hum. Genet. 38:868-883.

27. C Dubay (1988) A Database for Linkage Studies That Uses the Macintosh Computer for Data Entry. Abstract. Am. J. Hum Genet. 43:A143. Supplement.

28. DD Kosambi (1944) The Estimation of Map Distances from Recombination Values. Ann. Eugen. 12:172-175.

29. JH Edwards (1982) The Use of Computers. Cytogenet. Cell Genet. 32:43-51.

30. KK Kidd, J Ott (1983) Power and Sample Size in Linkage Studies. Abstract. Human Gene Mapping 7: Seventh International Workshop on Gene Mapping. Cytogenet Cell Genet 37:510.

31. D Botstein, RL White, M Skolnick, RW Davis (1980) Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. Am. J. Hum. Genet. 32:314-331.

32. H Donis-Keller, DF Barker, RG Knowlton, JW Schumm, JC Braman, P Green (1986) Highly Polymorphic RFLP Probes as Diagnostic Tools. CSH Symp. on Quant. Bio. 51:317-324.

33. WRA Brown, AP Bird (1986) Long Range Restriction Site Mapping of Mammalian Genomic DNA. Nature 322:477-481.

34. ES Lander, D Botstein (1986) Mapping Complex Genetic Traits in Humans: New Methods Using a Complete RFLP Linkage Map. CSH Symp. on Quant. Bio. 51:49-62.

35. R White, J-M Laouel (1988) Chromosome Mapping with DNA Markers. Sci. Americ. Feb:40-48.

36. Y Nakamura, M Lepert, P O'Connell, R Wolff, M Culver, C Martin, E Fujimoto, M Hoff, E Kumlin, R White (1987) Variable Number of Tandem Repeat (VNTR) Markers for Human Gene Mapping. Science 235:1616-1622.

37. M Lathrop, Y Nakamura, P O'Connell, M Leppert, M Woodward, J-M Lalouel, R White (1988) A Mapped Set of Genetic Markers for Human Chromosome 9. Genomics 3:361-366.

38. NC Dracopoli, BZ Stanger, CY Ito, KM Call, SE Lincoln, ES Lander, DE Housman (1988) A Genetic Linkage Map of 27 Loci from PYND to FY on the Short Arm of Chromosome 1. Am. J. Hum. Genet. 43:462-470.

39. EM Southern (1975) Detection of Specific Sequences Among DNA Fragments Separated by Gel Electrophoresis. J Molec Biol 98:503-517.

40. GM Church, W Gilbert (1984) Genomic Sequencing. PNAS USA 81:1991-1995.

41. JF Gusella, NS Wexler, PM Conneally, SL Naylor, MA Anderson, RE Tanzi, PC Watkins, K Ottina, MR Wallace, AY Sakaguchi, AB Yong, I Shoulson, E Bonilla, JB Martin (1983) A Polymorphic DNA Marker Genetically Linked to Huntington's Disease. Nature 306:234-238.

42. LC Tsui, M Buchwald, D Barker, JC Braman, R Knowlton, JW Schumm, H Eiberg, J Mohr, D Kennedy, N Plasvic, M Zsiga, D Markiewicz, G Akots, V Brown, C Helms, T Gravius, C Parker, K Rediker, H Donis-Keller (1985) Cystic Fibrosis Locus Defined by a Genetically Linked Polymorphic DNA Marker. Science 230:1054-1056.

43. DE Goldgar, P Green, DM Parry, JJ Mulvihill (1989) Multipoint Linkage Analysis in Neurofibromatosis Type 1: An International Collaboration. Am. J. Hum. Genet. 44:6-12.

44. RL Baehner, LM Kunkel, AP Monaco, JL Haines, PM Conneally, C Palmer, N Heerema, SH Orkin (1986) DNA Linkage Analysis of X-Chromosome Linked Chronic Granulomatous Disease. PNAS USA 83:3398-3401.

45. A Feinberg, B Vogelstein (1984) A Technique for Radiolabeling DNA Restriction Endonuclease Fragments to High Specific Activity. Anal. Biochem 132:6-13.

46. J Ott (1983) Linkage Analysis and Family Classification Under Heterogeneity. Ann Hum Genet 47:311-320.

47. A Poustka, H Lehrach (1986) Jumping Libraries and Linking Libraries: the Next Generation of Molecular Tools in Mammalian Genetics. TIG July 1986 2:174-179.

48. M Koenig, EP Hoffman, CJ Bertelson, AP Monaco, C Feener, LM Kunkel (1987) Complete Cloning of the Duchenne Muscular Dystrophy (DMD) cDNA and Preliminary Organization of the DMD Gene in Normal and Affected Individuals. Cell 50:509-517.

49. JM Rommens, MC Lannuzzi, BS Kerem, ML Drumm, G Melmer, M Dean, R Rozmahel, JL Cole, D Kennedy, N Hidaka, M Zsiga, M Buchwald, JR Rioradan, LC Tusi, FS Collins (1989) Identification of the Cystic Fibrosis Gene - Chromosome Walking and Jumping. Science 245:4922, pp1059-1065.

50. JR Riordan, JM Rommens, BS Kerem, N Alon, R Rozmahel, Z Grzelczak, J Zielenski, S Lok, N Plavsic, JL Chou, ML Drumm, MC Lannuzzi, FS Collins, LC Tusi (1989) Identification of the Cystic Fibrosis Gene - Cloning and Characterization of Complementary DNA. Science 245:4922, pp1066-1072.

51. BS Kerem, JM Rommens, JA Buchanan, D Markiewicz, TK Cox, A Chakravari, M Buchwald, LC Tusi (1989) Identification of the Cystic Fibrosis Gene - Genetic Analysis. Science 245:4922, pp1073-1080.

52. SH Friend, R Bernards, S Rogelj, RA Weinberg, JM Rapaport, DM Albert, TP Dryja (1986) A Human DNA Segment With Properties of the Gene that Predisposes to Retinoblastoma and Osteosarcoma. Nature 323:643-646.

53. RL White, et al. (1990) The CEPH Consortium Primary Linkage Map of Chromosome 10. Genomics 6:393-412.

54. C Julier, Y Nakamura, M Lathrop, P O'Connell, M Leppert, M Litt, T Mohandus, J-M Lalouel, R White (1990) Detailed Map of the Long Arm of Chromosome 11. Genomics: in press.

55. P Charmley, T Foroud, S Wei, P Concannon, DE Weeks, K Lange, RA Gatti (1990) A Primary Linkage Map of the Human Chromosome 11q22-23 Region. Genomics 6:316-323.

56. C Dubay, M Litt (1990) Addition of 8 RFLP Loci to CEPH Chromosome 11 Map. Chapter 2 of this thesis.

57. GM Wahl, KA Lewis, JC Ruiz, B Rothenberg, J Zhao, GA Evans (1987) Cosmid Vectors for Rapid Genomic Walking, Restriction Mapping, and Gene Transfer. PNAS USA 84:2160-2164.

58. P Lichter, CC Tang, K Call, G Hermanson, GA Evans, D Housman, DC Ward (1990) High-Resolution Mapping of Human Chromosome 11 by in Situ Hybridization with Cosmid Clones. Science 247:64-69.

59. V McKusick (1988) Editorial. Genomics 1:1-2.

60. C Larson, B Skogseid, K Oberg, Y Nakamura, M Nordenskjold (1988) Multiple Endocrine Neoplasia Type 1 Gene Maps to Chromosome 11 and Is Lost in Insulinoma. Nature 332:85-87.

61. C Maslen, C Jones, T Glaser, E Magenis, R Sheehy, J Kellogg, M Litt (1988) Seven Polymorphic Loci Mapping to Human Chromosomal Region 11q22-qter. Genomics 2: 66-75.

62. RA Gatti, et.al. (1988) Localization of an Ataxia-telangiectasia Gene to Chromosome 11q22-q23. NATURE 338:577-580.

63. M Smith, S Smalley, R Cantor, M Pandolfo, MI Gomez, R Bauman, P Flodman, K Yoshiyama, Y Nakamura, C Julier, K Dumas, J Haines, J Troffatter, MA Spence, D Weeks, M Conneally (1989) Mapping a Gene Determining Tuberous Sclerosis to Human Chromosome 11q14-11q23. Genomics 6:105-114.

64. W Cookson, P Sharp, J Faux, J Hopkin (1989) Linkage Between Immunoglobulin E Responses Underlying Asthma and Rhinitis and Chromosome 11q. The Lancet June 10th:1292-1295.

65. M Prochazka, E Leiter, D Serreze, D Coleman (1987) Three Recessive Loci Required for Insulin-dependent Diabetes in Nonobese Diabetic Mice. Science 237:286-289.

66. A Hillyard, D Doolittle, M Davidson, T Roderick (1989) Locus Map of Mouse with Comparative Map Points of Human on Mouse. The Jackson Laboratory, Bar Harbor, Maine.

67. VA McKusick (1987) The Morbid Anatomy of the Human Genome: A Review of Gene Mapping in Clinical Medicine. Medicine 66: 237-296.

68. Rizzuto et.al. (1989) Report of the Committee on the Genetic Constitution of Chromosome 11. HGM10. Cytogenet Cell Genet 51:226.

69. J Dausset (1986) Le Centre d'Etude du Polymorphisme Humain. La Presse Medicale 15:18 October No. 36, pp 1801-1802.

70. JP Fryns, A Kleczkowska, M Buttiens, P Marien, H Van Den Berghe (1986) Distal 11q Monosomy: The Typical 11q Monosomy Syndrome is Due to Deletion of Sub Band 11q24.1. Clin. Genet. 30:255-260.

71. M Budarf, B Sellinger, C Griffin, BS Emanuel (1989) Comparative Mapping of the Constitutional and Tumor Associated 11;22 Translocations. Amer. J. Hum. Genet. 45:128-139.

72. DP Gold, H Clevers, B Alarcon, S Dunlap, J Novotny, AF Williams, C Terhorst (1987) Evolutionary Relationship Between the T3 Chains of the T-cell Receptor Complex and the Immunoglobin Supergene Family. PNAS 84:7649-7653.

73. AF Williams, AN Barclay (1988) The Immunoglobin Superfamily: Domains for Cell Surface Recognition. Annu. Rev. Immunol. 6:381-405.

74. T Hunkapiller, L Hood (1989) Diversity of the Immunoglobin Superfamily. Adv. Immunol. 44:1-63.

75. AG Searle, J Peters, MF Lyon, JG Hall, EP Evans, JH Edwards, VJ Buckle (1989) Chromosome Maps of Man and Mouse (IV). Ann. Hum Genet. 53:89.

76. NSF Ma, DS Gerard (1988) Chromosome Assignment of the Gene Loci ETS1 and THY1 in the Owl Monkey. Cytogenet. Cell Genet. 48:170-173.

77. M Lepert, P O'Connell, Y Nakamura, GM Lathrop, C Maslen, M Litt, P Cartwright, J-M Lalouel, R White (1987) A Partial Primary Genetic Linkage Map of Chromosome 11: Ninth International Workshop on Human Gene Mapping. Cytogenet. Cell Genet. 46:648.

78. RP Donahue, WB Bias, JH Renwick, VA McKusick (1968) Probable Assignment of the Duffy Blood Group Locus to Chromosome 1 in Man. PNAS 61:949-955.

79. JB Haldane (1919) The Combination of Linkage Values and the Calculation of Linkage Between the Loci of Linked Factors. J. Genet. 8:299-309.

80. JH Renwick (1971) The Mapping of Human Chromosomes. Ann. Rev. Genet. 5:81-120.

81. JBS Haldane, CAB Smith (1947) A New Estimate of the Linkage Between the Genes for Colour-blindness and Haemophilia in Man. Ann. Eugen. 14:10-31.

82. D Nathans, H Smith (1975) Restriction Endonucleases in the Analysis and Restructuring of DNA Molecules. Ann. Rev. Biochem. 44:273-293.

83. EA Thompson, K Kravitz, J Hill, M Skolnick (1978) Linkage and the Power of a Pedigree Structure. In Genetic Epidemiology, NE Morton and CS Chung eds. New York Academic Press, pp 465-479.

84. H Harris, D Hopkinson (1972) Average Heterozygosity per Locus in Man: an Estimate Based on the Incidence of Enzyme Polymorphism. Ann. Hum. Genet. 36:9-20.

85. M Rosenbash, M Campo, K Gummerson (1975) Conservation of Cytoplasmic Poly-A-containing RNA in Mouse and Rats. Nature 258:682-686.

86. R Britten, A Cetta, E Davidson (1977) The Single Copy DNA Sequence Polymorphism of the Sea Urchin Strongylocentrotus purpuratus. Cell 10:509-519.

87. RC Lewontin (1974) The Genetic Basis of Evolutionary Change. New York, Columbia Univ. Press, .

88. W Upholt (1977) Estimation of DNA Sequence Divergence from Comparison of Restriction Endonuclease Digests. Nucleic Acids Res. 4:1257-1265.

89. C Schmid, P Deininger (1975) Sequence Organization of the Human Genome. Cell 6:345-358.

90. M Litt, RL White (1985) A Highly Polymorphic Locus in Human DNA Revealed by Cosmid-derived Probes. PNAS 82:6202-6210.

91. D Hartl. A Primer of Population Genetics. Pg 11. Sutherland Sinauer As., Sutherland, MA.

92. T Grodzicker, J Williams, P Sharp, J Sambrook (1974) Physical Mapping of Temperature-sensitive Mutations of Adenovirus. Cold Spring Harbor Symp. Quant. Biol. 39:439-446.

93. JL Marx (1985) Putting the Human Genome on the Map. Science 229:150-151.

94. CEPH (1989) CEPH Version 3 Database LOD Scores and Recombination Estimates. December 1989, CEPH, Paris, FRANCE.

95. M Litt, LB Jorde (1986) Linkage Disequlibria Between Pairs of Loci within a Highly Polymorphic Region of Chromosome 2q. Am. J. Hum. Genet. 39:166-178.

96. A Chakravarti, KH Buetow, SE Antonarakis, PG Warber, CD Boehm, HH Kazazian (1984) Nonuniform Recombination within the Human Beta-Globin Gene Cluster. Am. J. Hum. Genet. 36:1239-1258.

97. E Boder, R Sedgwick (1958) Ataxia Telangiectasia. Pediatrics 21:526-554.

98. RA Gatti, M Swift, eds. (1985) Ataxia Telangiectasia: Genetics, Neuropathology, and Immunology of a Degenerative Disease of Childhood. New York, Alan R. Liss.

99. J Llerena Jr., et.al. (1989) Spontaneous and Induced Chromosome Breakage in Chorionic Villus Samples: a Cytogenetic Approach to First Trimester Prenatal Diagnosis of Ataxia- telangietasia Syndrome. Journal of Medical Genetics 26:174-178.

100. D Schindler, et.al. (1987) Screening Test for Ataxia- telangiectasia [letter]. LANCET v2 No 8572:p1398-9.

101. M Swift, et.al. (1987) Breast and Other Cancers in Families with Ataxia- telangiectasia. N Engl J Med 316:1289-94.

102. M Gomez (1988) Tuberous Sclerosis. Raven Press, New York.

103. NJ Royle, R Clarkson, Z Wong, AJ Jeffreys (1987) Preferential Localization of Hypervariable Minisatellites Near Human Telomeres. Ninth international workshop on human gene mapping. Cytogenet. Cell Genet. 46:685.

104. DK Grandy, M Litt, L Allen, JR Bunzow, M Marchionni, H Makam, L Reed, RE Magenis, O Civelli (1989) The Human Dopamine D2 Receptor Gene is Located on Chromosome 11 at q22-q23 and Identifies a TaqI RFLP. Am J Hum Genet 45:778-785.

105. RL Miller, HS Chan, MC Cavanaugh, WP Alles, ML Mador, KK Kidd (1989) HGM10. Cytogenet Cell Genet 51:9-10.

106. P Elsten, G Burns, DS Gerhard, D Pravtcheva, C Jones, D Housman, FA Ruddle, S Orkin, C Terhorst (1985) Assignment of the Gene Coding for the T3-delta Subunit of the T3-T-cell Receptor Complex to the Long Arm of Human Chromosome 11 and to Mouse Chromosome 9. PNAS 82:2920-2924.

107. J Dausset, H Cann, D Cohen, M Lathrop, J-M Lalouel, R White (1990) CEPH: Collaborative Genetic Mapping of the Human Genome. Genomics 6:575-577.

108. RV Lebo, A Chakravarti, KH Buetow, M-C Cheung, H Cann, B Cordel, H Goodman (1983) Recombination Within and Between the Human Insulin and Beta-globin Gene Loci. PNAS 80:4808-4812.

109. JB Haldane (1922) Sex Ratio and Unisexual Sterility in Hybrid Animals. J. Genet. 8:299-309.

110. JH Renwick, J Schulze (1965) Male and Female Recombination Fraction for the Nail-patella : ABO Linkage in Man. Ann. Hum. Genet. 28:379-392.

111. R White, M Leppert, T Bishop, D Barker, J Berkowitz, C Brown, P Callahan, T Holm, L Jerominski (1985) Construction of Linkage Maps with DNA Markers for Human Chromosomes. Nature 313:101-105.

112. AP Bird (1980) DNA Methylation and the Frequency of CpG in Animal DNA. Nucl. Acids Res. 9:1499-1504.

113. D Barker, M Schafer, R White (1984) Restriction Sites Containing CpG Show a Higher Frequency of Polymorphism in Human DNA. Cell 36:131-138.

114. H Youssoufian, HH Kazazian Jr., DG Phillips, S Aronis, G Tsiftis, VA Brown, SE Antonarakis (1986) Recurrent Mutations in Haemophilia A Give Evidence for CpG Mutation Hotspots. Nature 324:380-382.

115. JF Gusella, C Keys, A Varsanyi-Breiner, F-T Kao, C Jones, TT Puck, D Housman (1980) Isolation and Localization of DNA Segments from Specific Human Chromosomes. PNAS 77:2829-2833.

116. Rodent X Human Cell Panel provided by Gail Burns. Described in (126).

117. Rodent X Human Cell Panel described in (61).

118. ML Pardue, JG Gall (1970) Chromosomal Localization of Mouse Satellite DNA [Letter reviewing in-situ discovery and use]. Science 168:1356-1358.

119. DT Burke, GF Carle, MV Olson (1987) Cloning of Large Segments of Exogenous DNA into Yeast by Means of Artificial Chromosome Vectors. Science 236:808-812.

120. LM Smith, JZ Sanders, RJ Kaiser, P Hughes, C Dodd, CR Connell, C Heiner, SBH Kent, LE Hood (1986) Florescence Detection in Automated DNA Sequence Analysis. Nature 321:674-678.

121. GM Church, S Kieffer-Higgins (1988) Multiplex DNA Sequencing. Science 240:185-188.

122. GA Evans, KA Lewis (1989) Physical Mapping of Complex Genomes by Cosmid Multiplex Analysis. PNAS 86:5030-5034.

123. W Rasband (1989) "Image Processing and Analysis Software [for the Macintosh]: Image V1.22", National Institutes of Health, Research Services Branch NIMH.

124. L Bufton, GAP Burns, RE Magenis, D Tomar, D Shaw, D Brook, M Litt (1986) Four Restriction Fragment Length Polymorphisms Revealed by Probes from a Single Cosmid Map to Chromosome 19. Am J Hum Genet 38:447-460.

125. JH Nadeau (1989) Maps of Linkage and Synteny Homologies Between Mouse and Man. TIGs 5:82-86.

126. M Litt, GAP Burns, R Sheehy, RE Magenis (1986) A Highly Polymorphic Locus in Human DNA Revealed by Probes from Cosmid 1-5 Maps to Chromosome 2q35-37. Am. J. Hum. Genet. 38:288-296.

127. GI Bell, JH Karram, WJ Rutter (1981) Polymorphic Region Adjacent to the 5' End of the Insulin Gene, PNAS 78:5759-5763.

128. ES Lander, D Botstein (1986) Strategies for Studying Heterogeneous Genetic Traits in Humans by Using a Linkage Map of Restriction Fragment Length Polymorphisms. PNAS 83:7353-7357.