Current trends in mapping human genes

VICTOR A. MCKUSICK

Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

ABSTRACT **The human is estimated to have at least 50,000 expressed genes (gene loci). Some information is available concerning about 5000 of these gene loci and about 1900 have been mapped, i.e., assigned to specific chromosomes (and in most instances particular chromosome regions). Progress has been achieved by a combination of physical mapping (e.g., study of somatic cell hybrids and chromosomal in situ hybridization) and genetic mapping (e.g., genetic linkage studies). New methods for both physical and genetic mapping are expanding the armamentarium. The usefulness of the mapping information is already evident; the spin-off from the Human Genome Project (HGP) begins immediately. The complete nucleotide sequence is the ultimate map of the human genome. Sequencing, although already under way for limited segments of the genome, will await further progress in gene mapping, and in particular creation of contig maps for each chromosome. Meanwhile the technology of sequencing and sequence information handling will be developed. It is argued that the HGP is a new form of coordinated, interdisciplinary science; that its primary objective must be seen as the creation of a tool for biomedical research—a source book that will be the basis of study of variation and function for a long time; that the impact on scientist training will be salutary by relieving graduate students of useless drudgery and by training scientists competent in both molecular genetics and computational science; and that the funding of the HGP will have an insignificant negative effect on science funding generally, and indeed may have a beneficial effect through economy of scale and a focusing of attention on the excitement of biology and medical science. —McKusick, V. A. Current trends in mapping human genes. *FASEB J.* 5: 12-20; 1991.**

Key Words: chromosome • pedigree pattern • family linkage • gene mapping • hybridization • genomics

THE ESSENCE OF MENDEL'S DISCOVERY (1865) is that inheritance is particulate. The chromosomes were first described 12 years later by Walther Fleming of Kiel. Meiosis, or the reduction division, as it was then called, was described in the 1880s. In part as the basis for rationalizing the reduction division of the chromosomes in gametogenesis, the notion was put forward that the chromosomes carry factors that determine development: the Roux-deVries-Weissmann hypothesis. Before the rediscovery of Mendelism in 1900, E. B. Wilson, in the first edition of his landmark monograph *The Cell in Development and Inheritance* (1896), stated as follows (pp. 182-185):

> ". . . the chromosome is a congeries or colony of invisible self-propagating vital units . . ., each of which has the power of determining the development of a particular quality. Weismann conceives these units . . . [to be] associated in linear groups to form the . . . chromosomes."

Thus, the chromosome theory of heredity preceded the rediscovery of Mendelism. The chromosome theory of Mendelism was advanced by Sutton and Boveri in 1902-1904.

The phenomenon of linkage was discovered in the first decade of this century in the domestic fowl by Bateson and Punnett, who introduced the terms coupling and repulsion. The linear arrangement of Mendel's particulate elements of heredity along the chromosome and the estimation of the intervals separating two such elements, by then called genes (Johansson's term, about 1909), were developed by Thomas Hunt Morgan on the basis of studies of *Drosophila* beginning about 1911. Undergraduate Alfred H. Sturtevant was important in developing the concept of linkage mapping.

Also in 1911, Morgan's colleague at Columbia, E. B. Wilson, for the first time assigned a specific gene to a specific chromosome in a mammal: the colorblindness gene to the human X chromosome. He wrote as follows in his 1911 paper (1):

> "In the case of color-blindness, for example, all the facts seem to follow under this assumption [that the gene is on the X chromosome] if the male be digametic (as Guyer's observations show to be the case in man). For, in fertilization this character will pass with the affected X chromosome from the male into the female, and from the female into half her offspring of both sexes (Diagram, Fig. 5). Color-blindness, being a recessive character, should therefore appear in neither daughters nor granddaughters, but in half the grandsons, as seems actually to be the case"

Friedrich Horner, a Zürich ophthalmologist, had described the typical pedigree pattern of colorblindness in 1876 (2). As Horner pointed out, this pedigree pattern was known also for hemophilia and later it was recognized for a few other disorders, which by the same reasoning as that applied by Wilson to colorblindness, must also be coded by genes on the human X chromosome.

Because of the distinctive pedigree pattern, about 36 sex-linked traits or disorders were described in the human before the first X-linked trait was discovered in the mouse, about 1950. On the other hand, demonstration of autosomal linkages proceeded much faster in the mouse because experimental matings such as the highly informative double backcross mating could be done experimentally in that species. The first linkage in a mammal was that demonstrated between albinism and pink eye in 1915 by J. B. S. Haldane and his sister Naomi, working with A. D. Sprunt (3). The authors gave the following excuse for publishing a preliminary report:

---

[1]Abbreviations: RFLP, restriction fragment length polymorphism; VNTR, variable length tandem repeats; PIC, polymorphism information content; PCR, polymerase chain reaction; YAC, yeast artificial chromosome; HGM, human gene mapping; GDB, genome database; CEPH, Centre d'Etudes du Polymorphisme Humain.

0892-6638/91/0005-0012/$01.50. © FASEB

"Owing to the war it has been necessary to publish prematurely, as unfortunately one of us (A. D. S.) has already been killed in France."

Although the interval separating the hemophilia and colorblindness loci on the X chromosome was estimated by Haldane and colleagues in two separate analyses in 1937 and 1947, no autosomal linkage was demonstrated, let alone quantitated, until 1951—about the same time that the first X-linked trait was demonstrated in the mouse. Jan Mohr, in his doctoral thesis in Copenhagen, found linkage of Lutheran blood group with secretor factor. His studies also suggested linkage of these two loci to that for myotonic dystrophy, a finding that was subsequently confirmed. Indeed, the three loci were shown to be on chromosome 19, and it was mainly work in the department of Professor Jan Mohr, by then the long-time director of the Institute of Medical Genetics at Copenhagen, that provided the critical evidence of the chromosomal location of this linkage group.

In his studies in the early 1950s, Jan Mohr made use of the sib-pair method of Penrose based on the principle that if two loci are linked, sibs will fail to show random association of traits determined by genes at those loci. The method of estimating the likelihood of linkage accounting for findings in particular pedigrees was developed by workers such as C. A. B. Smith and Newton Morton in the 1940s and 1950s. This work was the basis of the now familiar lod score—the logarithm of the odds of linkage as opposed to nonlinkage (4).

Between 1951 and 1968, which is the next watershed year, a number of further autosomal linkages were described, such as ABO vs. nail-patella syndrome and Rh vs. elliptocytosis-1. But for all these autosomal linkages, the precise chromosome carrying the genes was not known. In 1968, Roger Donahue (5), then a candidate for a Ph.D. degree in human genetics at Johns Hopkins University, demonstrated in his own family linkage of the Duffy blood group locus to chromosome 1 as distinctively marked in him and a number of his relatives by a so-called heteromorphism. The relatively easy study of human chromosomes and the introduction of cytogenetics to the clinic had come in the previous 10 years. Donahue's unusual chromosome in preparations made in 1968 in the just prebanding era had the appearance of an uncoiled area subjacent to the centromere. (With the advent of banding methods about 1970, especially centromeric banding, it was clear that the anomaly represented an unusually long heterochromatic segment.) Donahue realized that this might be segregating as a dominant trait, as it was demonstrated by one of the two chromosomes 1, and could therefore be used as a linkage marker. Furthermore, he had the gumption to do a linkage study, which was not easy because of his far-flung family for collecting blood samples and because of the labor involved in the testing of markers.

In the Donahue pedigree, the lod score for linkage of Duffy blood group to the long heterochromatic segment of chromosome 1 was far below the figure of 3.0 (1000 to 1 odds on linkage) generally taken as evidence for linkage nowadays. However, the observation was quickly confirmed by others. Many studies of linkage of loci on chromosome 1, where now more than 190 genes have been located (see later), indicate that the Duffy blood group is on the proximal short arm, i.e., on the opposite side of the centromere from the long heterochromatic segment with which it shows linkage. There may be suppression of crossing-over in the pericentric region favoring demonstration of the linkage.

By 1968, when the first autosomal assignment was made, about 68 genes had been assigned to the X chromosome on the basis of characteristic pedigree patterns. (The regional location of the 68 on the X chromosome was known for none of them.) The growth of information on chromosomal assignments is diagramed in **Fig. 1.** As of September 10, 1990, 1868 expressed genes had been assigned to specific chromosomes (**Table 1**), and in most instances to specific bands of those chromosomes. This has been possible because of four commingling methodologic streams: *1*) family linkage study, *2*) chromosome study, *3*) somatic cell genetic study, and *4*) molecular genetic study (**Fig. 2**).

Beginning about 1970, mapping took off at an accelerated pace, particularly through the use of somatic cell hybridization. The sorting out of the chromosomes in the subclones derived from rodent-human hybrid cells was comparable to the assortment of human chromosomes in meiosis. Somatic cell hybridization was thus what Pontecorvo referred to as a parasexual method and what Haldane called a substitute for sex. The development of chromosome banding, also about 1970, provided a highly important, indeed essential, methodologic element by permitting the unique identification of each human chromosome and the differentiation of human chromosomes from rodent chromosomes in the hybrid cells.

About 1980, a further acceleration in gene mapping occurred with the introduction of molecular genetic techniques. Recombinant DNA technology provided probes that could be used in three ways: first, in connection with DNAs from somatic hybrid cells (obviating the necessity to have gene expression in the cultured cells, as one could directly "go for the gene"); second, for in situ hybridization to chromosome spreads; and third, as DNA markers—e.g., restriction fragment length polymorphisms (RFLPs) and variable length tandem repeats (VNTRs)—in family linkage studies (6). Chromosomes sorted by means of the fluorescence-activated cell sorter could also be probed with the new armamentarium. (Chromosome sorting by this physical method serves the same role as meiosis in vivo and somatic cell hybrid clones in vitro. In situ hybridization and somatic cell hybridization are methods of physical mapping; family link-

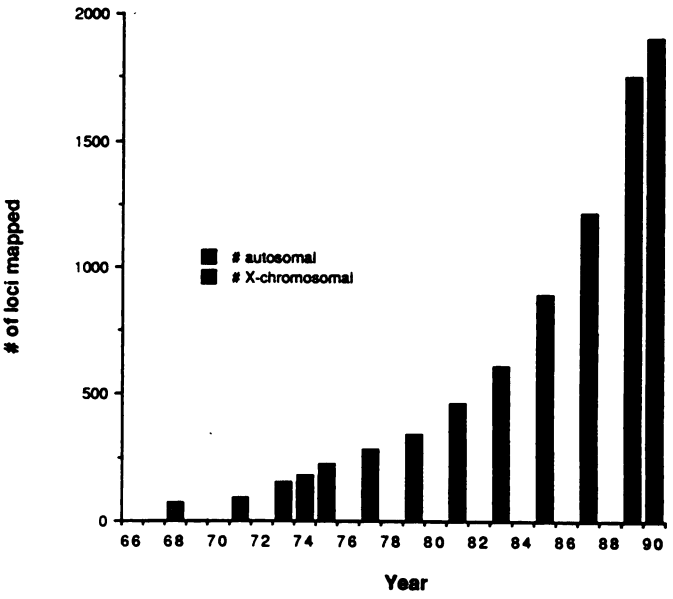

Figure 1. Growth of information on chromosome assignments.

TABLE 1. *Number of gene assignments by chromosome*[a]

Chromosome	No. of genes	Chromosome	No. of genes
1	194	13	24
2	110	14	60
3	68	15	51
4	74	16	68
5	72	17	107
6	104	18	23
7	107	19	91
8	53	20	34
9	59	21	36
10	58	22	58
11	132	X	187
12	100	Total	1868

[a]Number of expressed genes assigned to each chromosome (Human Gene Mapping Workshop 10.5, Oxford, England, September 10, 1990).

age studies give information on the genetic map, but when the markers used have been mapped to specific chromosomal sites, physical location can be inferred.)

STATUS OF THE HUMAN GENE MAP AS OF SEPTEMBER 10, 1990

At the Human Gene Mapping Workshop in Oxford early in September, 1990, the available information on the human gene map was collated, with results that are tabulated in Table 1 and **Fig. 3.** A total of almost 1900 genes have been mapped to specific chromosomal locations. In addition, more than 4500 DNA segments, so-called anonymous DNA segments because their function, if any, is not known (indeed most are known not to be expressed), have been mapped to specific chromosomal sites. About half of these segments have been shown to be polymorphic and many of them are sufficiently variable to make them useful as linkage markers.

The human is estimated to have about 50,000 genes, almost certainly not more than 100,000. On the basis of indirect arguments, estimates of this order were arrived at in the past; in more recent times, studies of the density of genes on the chromosomes and information on the range of sizes of genes support these estimates.

Many more genes have been assigned to some chromosomes than to others, if one examines either the absolute (Fig. 3*A*) or the relative numbers, derived by dividing the absolute number by the length of the chromosome expressed as a percentage of the length of the haploid set of autosomes (Fig. 3*B*). The preponderance of loci on the X chromosome derives from the relative ease of assignment by pedigree pattern; this factor was aided by the availability of a selection system, based on the X-linked HPRT locus, that could be used in mapping by somatic cell hybridization. The large number on chromosome 17 is related in large part to the availability of the selection system based on the thymidine kinase locus on that chromosome. Chromosomes 11 and 16 are rather well mapped, perhaps particularly because of the location there of the β-globin and α-globin loci, respectively. Chromosome 16 also had a good linkage marker early on in the form of the haptoglobin locus. Chromosome 6 had the advantage of HLA, a superb highly polymorphic marker system for genetic linkage. Chromosome 7 underwent intensive study when it was discovered that the cystic fibrosis locus maps to 7q. Chromosome 21 has had attention for the obvious reason of interest in Down syndrome. The strong performance of chromosome 22 may be related to the fact that the λ light-chain genes are there, and the BCR/ABL fusion gene that results from the 9;22 translocation of chronic myeloid leukemia has stimulated much study.

The reason for the large number of genes assigned to chromosome 19 (the highest relative number second only to the X) may be merely that it got started early with the secretor/Lutheran/myotonic dystrophy linkage group, and the relatively large number on chromosome 1 may have the same explanation. Is it possible that some chromosomes are genically less densely populated than others? Chromosomes 13 and 18 stand out for a low number of mapped genes. It is noteworthy that except for the smallest autosome, no. 21, only nos. 13 and 18 show trisomy that is compatible with live birth. This fact suggests also that there are relatively few genes or at least few genes of critical importance on these chromosomes.

In parallel with the mapping in humans, gene mapping has been proceeding apace in the mouse, which is the most useful mammalian genetic model for the human because of the extensive amount of genetic information available, and the relative ease of experimental manipulation. At the same time (June, 1976), mapping in both mouse and the human achieved the point that at least one gene had been assigned to every chromosome. Although in the mouse, as mentioned earlier, linkage mapping proceeded much faster until the surrogate methods such as somatic cell hybridization were developed, the human has outpaced the mouse in recent times.

Caenorhabditis elegans, a nematode, is also a useful model (7). As in humans, there is extensive anatomic and physiologic information, and relatively speaking, the genetic and developmental information is even more extensive than in humans. The developmental lineage of each of the 959 cells of the organism is known; all the synapses of the nervous system are described; and more than 1000 genes have been

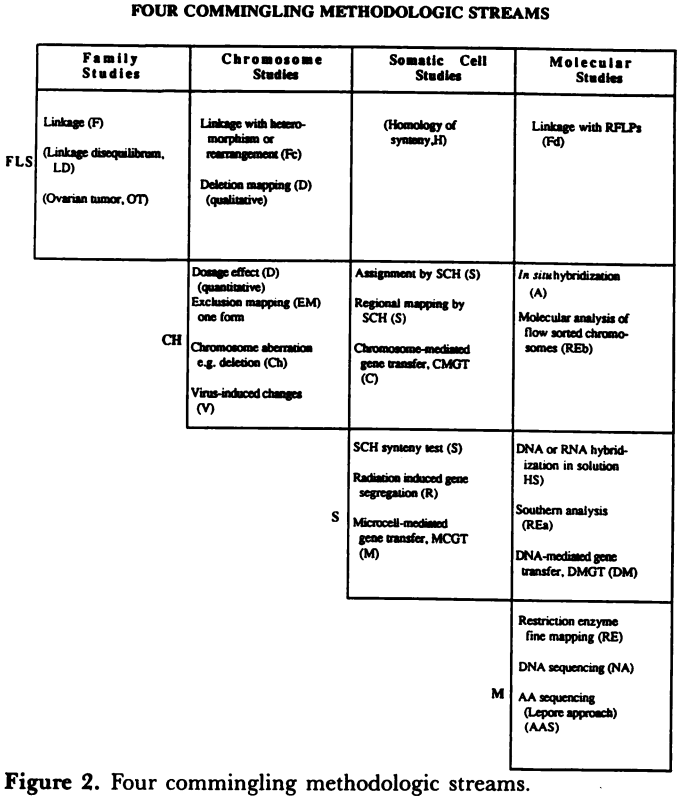

Figure 2. Four commingling methodologic streams.

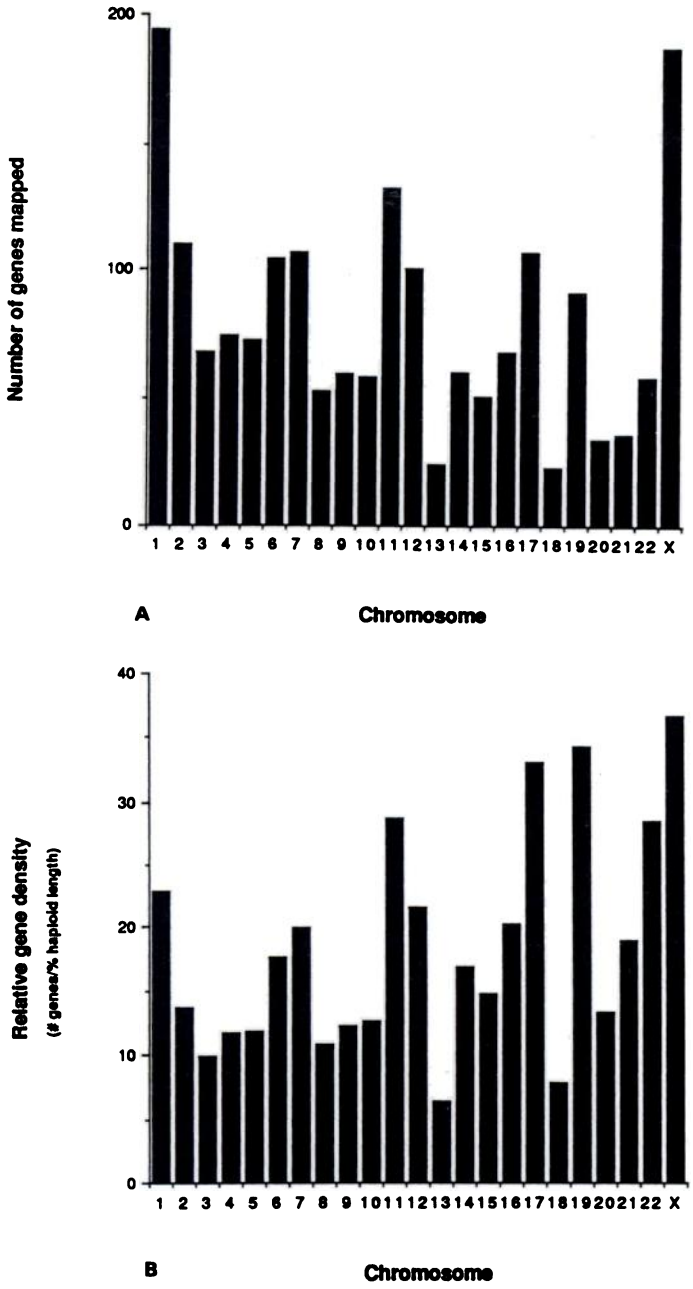

Figure 3. *A*) Number of gene assignments by chromosome. *B*) Relative density of genes assigned to each chromosome (see text).

mapped to one or another of the four chromosome pairs. Contig maps are approaching completion. Now what remains is nucleotide sequencing.

THE SOCIOLOGY OF HUMAN GENE MAPPING

Information on the human gene map has been collated in a series of human gene mapping (HGM) workshops. The first of these was convened at Yale by Frank Ruddle in 1973, with the financial support of The March of Dimes, represented by Dr. Daniel Bergsma, Vice President for Professional Education. Subsequent workshops were held in The Netherlands (D. Bootsma, organizer), Baltimore (V. A. McKusick, organizer), Winnipeg (J. Hamerton, organizer), Edinburgh (J. Evans, organizer), Oslo (K. Berg, organizer), Los Angeles (R. Sparkes, organizer), Helsinki (A. de la Chapelle, organizer), Paris (J. Frézal, organizer), and again New Haven (F. Ruddle and K. Kidd, co-organizers). The workshop in early September, 1990, was held in Oxford, England, under the direction of Sir Walter Bodmer and Ian Craig. For these HGM workshops a committee with co-chairs assumed responsibility for collating the information on the gene map of each chromosome. A standing committee of the HGM workshops, cochaired by Phyllis J. McAlpine and Thomas B. Shows, renders opinions on matters of nomenclature, particularly choice of gene symbols, to assure consistency. Committees at the workshops are responsible for other overall considerations, namely, a committee on clinical disorders that have been mapped, and a committee on neoplasia that reports on specific chromosomal changes associated with specific neoplasms, as well as the mapping of oncogenes and genes at cloned breakpoints in the chromosome rearrangements in neoplasia, and recently, loss of heterozygosity and coamplification of genes in neoplasms as indicators of map location.

Beginning with the most recent Oxford workshop, the data on the human gene map are entered directly into a database that will be continuously updated. This database, maintained in Baltimore at Johns Hopkins University with the support of the Howard Hughes Medical Institute, is called GDB (for genome database). The chairs of each of the committees will be responsible for ongoing editing, which they can perform from remote sites. Furthermore, the database will be generally accessible by the scientific community worldwide. Its development is under the direction of Dr. Peter Pearson, formerly chairman of the Department of Human Genetics in Leiden University, and a leading cytogeneticist and gene mapper. The development of the systems underlying GDB has been the responsibility of Richard Lucier.

Fully integrated with GDB and distributed worldwide along with it (and indeed a main reason for the development of GDB in Baltimore) is OMIM, the on-line version of *Mendelian Inheritance in Man* (8), the encyclopedic catalog of genes that I have maintained on computer since late 1963 and have published as a computer-based book since 1966 (9th edition, 1990). My colleagues and I have provided the scientific and medical community access to the on-line version for more than 3 years. Although the print version has appeared at

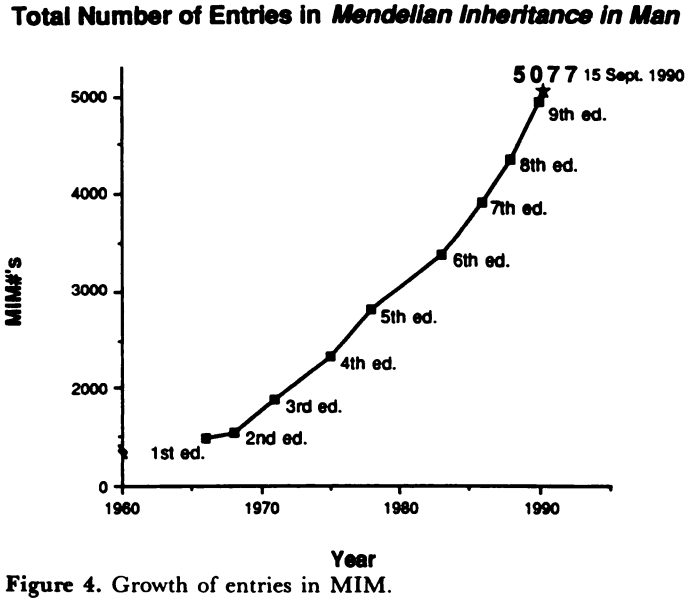

Figure 4. Growth of entries in MIM.

2-year intervals, in recent times the rapid evolution of the field gives the on-line version the extra value of timeliness. Another advantage of the on-line version is the ability to search the database electronically. The size of the database is reflected by the number of entries, more than 5,000 (**Fig. 4**), each representing a single gene locus, as well as by the counts of individual authors (more than 56,000) and references (more than 36,000). The database is clearly too large to be handled efficiently by print indices.

Within the United States, OMIM and (more recently) GDB have been easily accessible through Telenet and Internet; access and even transcription charges have been waived through the generosity of the Howard Hughes Medical Institute.[2] To facilitate distribution elsewhere in the world, arrangements have been made for distribution nodes in the United Kingdom, Germany, Sweden, and Japan.

Figure 4 indicates the growth in number of entries in MIM over the 25 years of its existence. The principle is that no more than one entry per gene locus is created. All the allelic mutations at a given locus are listed under a single entry. Thus, the total gives some indication of the proportion of all genes about which we have information. That proportion is now about 10%. Until about 1980, the only method for identifying genes was the occurrence of Mendelian variation; hence the title, *Mendelian Inheritance in Man.* In recent times, of course, genes have been cloned, mapped, and sequenced without any Mendelian variation in phenotype having been associated with them. Such genes have deservedly been incorporated as entries in MIM. These have contributed greatly to the recently accelerated pace of accessions.

METHODS OF MAPPING, OLD AND NEW

As indicated by **Table 2,** somatic cell hybridization in all its variations has been by far the most productive method of human gene mapping. The method second in position is in situ hybridization which attained this place by the fall of 1987. Its rapid ascendancy is in some ways surprising because the method was made to work reliably for single-copy genes only as recently as 1981. Presently, whenever a gene is cloned, one has not adequately characterized it until one has mapped it, first, by determining its chromosomal location by hybridizing a gene probe to a panel of DNAs from somatic cell hybrids, and second, by corroborating that assignment and regionalizing it by in situ hybridization. If a newly cloned gene is found to have an associated RFLP, there is a third method for mapping the gene: linkage against DNA markers that have previously been mapped in the Centre d'Etudes du Polymorphisme Humain (CEPH) families. The CEPH in Paris, developed by Nobel Laureate Jean Dausset, has a collection of cell lines from more than 40 three-generation families in which all four grandparents and eight or more grandchildren, as well as their parents, have been available for sample collection. The DNAs from the families have been subjected to linkage studies to create a reference map for each chromosome. Nonisotopic labeling for in situ hybridization (9, 10) has the advantage that one does not need to wait for development of the autoradiographs or have the nuisance and expense involved in the handling of radioisotopes. In addition, one can expect to study the relative position of two or more genes by using fluorescent markers of different colors.

TABLE 2. *Method of mapping*[a]

Method	No. of loci mapped
Somatic cell hybridization	1148
In situ hybridization	687
Family linkage study	466
Dosage effect	159
Restriction enzyme fine analysis	176
Chromosome aberrations	123
Homology of synteny	110
Radiation induced gene segregation	18
Others	143
Total	3047

[a]Number of autosomal loci mapped by several methods, September 15, 1990.

There are, however, many genes, including clinically important disorders such as cystic fibrosis, Huntington disease, and polycystic kidney disease, that could not be mapped by first cloning the gene because the nature of the mutant gene product was not known. In these cases, mapping of the phenotype by family linkage studies has been necessary. The availability of DNA markers such as RFLPs, distributed more or less uniformly over the genome at intervals of 5 to 10 centimorgans (cM) to constitute a linkage reference map, means that the chances are good that linkage can be demonstrated between a given rare dominant (or even recessive) and one of the markers. The informativity of a particular RFLP depends on its degree of heterozygosity. On the suggestion of Botstein et al. (6), the informativity is measured by the PIC (polymorphism information content) value. PIC is the sum of the proportion of all parental matings that have at least one heterozygous parent. The disorders listed in **Table 3** were all mapped by genetic linkage to markers.

The application of the polymerase chain reaction (PCR) to single sperm, in the hands of Arnheim and his colleagues (11, 12), permits the direct ascertainment of linkage and estimation of recombination frequencies. It is like determining directly the genetic composition of the stick diagrams used in textbooks to explain linkage and recombination and presenting the data derived from family studies. The method has the potential advantage that one can study large numbers of meioses from a single doubly heterozygous male. Ordinarily, in human families, when one must depend on analysis of the phenotype of his offspring for identification of recombination, one has an opportunity to study no more than 8 or 10, or in truly exceptional circumstances, 16 or 18 meioses from a single male. Because of the large number of meioses that can be studied by the Arnheim method, information on close linkage can be determined. Furthermore, the possibility of differences in the frequency of recombination in the same DNA segment in different males can be studied, as well as the effect of age in the individual male.

Linkage disequilibrium can be used as a clue to close linkage. In a sense, homozygosity mapping of recessive genes (13) is based on this principle. In a population such as the Amish or the Finns where founder effect makes it likely that all cases of a given, ordinarily rare recessive are descendents from a common ancestor, one expects that genes closely situated to the disease-producing allele will stand a good chance of being transmitted to affected descendants and not to the

[2]For establishing access, contact GDB/OMIM User Support, Welch Medical Library, Johns Hopkins University, 1830 E. Monument St., Baltimore, MD 21205, USA. Tel.: (301) 955-7058; FAX: (301) 955-0054.

TABLE 3. *Diseases mapped by RFLPs*[a]

Charcot-Marie-Tooth disease	1q,17,Xq13
Usher syndrome	1q
van der Woude lip-pit syndrome	1q
Aniridia	2p, 11p
Waardenburg syndrome	2q
von Hippel-Lindau syndrome	3p
Huntington disease	4p
Facioscapulohumeral muscular dystrophy	4q
Spinal muscular atrophy, several types	5q
Adenomatous polyposis of colon	5q
Hemochromatosis	6p
Juvenile myoclonic epilepsy	6p
Spinocerebellar ataxia (one form)	6p
Craniosynostosis	7p
Greig craniopolysyndactyly syndrome	7p
Cystic fibrosis	7q
Langer-Giedion syndrome	8q
Friedreich ataxia	9q
Torsion dystonia	9q
Tuberous sclerosis	9q, 11q
Nail-patella syndrome	9q
Multiple endocrine neoplasia, type II	10q
Wilms tumor-1	11p
Multiple endocrine neoplasia, type I	11q
Ataxia-telangiectasia	11q
Retinoblastoma	13q
Wilson disease	13q
Marfan syndrome	15q
Polycystic kidney disease	16p
Batten disease	16p
Cataract, Marner type	16q
Neurofibromatosis	17q
Myotonic dystrophy	19q
Malignant hyperthermia	19q
Alzheimer disease (one form)	21q
Acoustic neuroma, bilateral	22q
Duchenne muscular dystrophy	Xp
Retinitis pigmentosa (two forms)	Xp
Wiskott-Aldrich syndrome	Xp
Alport syndrome	Xq

[a] List of mapped Mendelian disorders for which the biochemical basis was not previously known (mapping as a first step toward basic understanding).

unaffected. The closer the genes the greater the chance that both the marker and the disease allele will remain on the same chromosome in most affected persons. This is the argument underlying the use of haplotypes[3] to identify the presence of thalassemia or other disease genes. In the Amish, because of the defined genealogies, DNA panels that reflect an identifiable number of meioses might be used in the same way the reference families in CEPH are used. (Homozygosity mapping bears similarities in principle to mapping in mice by use of recombinant inbred strains) (14).

Goss and Harris in the 1970s developed a method for determining the interval between genes on the basis of the chance that they would be separated when a standard dose of radiation was administered to the cell. The radiated cell was rescued by hybridization with a rodent cell. This method of radiation-induced gene segregation or radiation mapping has been developed recently by David Cox and others (16, 17). The product might be called the "zap map," a term Cox does not like.

In recent times we have learned how to use enzymes that cut the DNA only rarely, at intervals of a few hundred kilobases or more. In general, the larger the number of nucleotides represented by the recognition site for the restriction enzyme, the rarer is the cut site and the larger the fragments that are produced. In addition, the methods for separating large fragments, of which the Schwartz and Cantor method of pulsed field gel electrophoresis was a pioneer, provide the material that can be used for hybridization of gene probes. Given that gene A has been assigned to a particular location in the genome by some other method, if gene B hybridizes to the same large fragment, one has determined where gene B is located also, as well as placed an upper limit on the interval separating genes A and B. Restriction mapping using frequently cutting endonucleases can define the relationship of A and B in more detail. Yeast artificial chromosome (YAC) cloning also provides large DNA fragments to which cloned genes can be hybridized for mapping purposes.

THE ROLE OF GENE MAPPING IN HUMAN BIOLOGY AND MEDICINE

As Sir Walter Bodmer has pointed out, what seemed like a rather recondite activity when the human gene mapping workshops were initiated in 1973 has achieved a central role in both human biology and scientific medicine. Charles Scriver of Montreal suggested that gene mapping is providing a neo-Vesalian basis for medicine. Beginning in the 1950s, the availability of relatively easy methods for microscopic study of the human chromosomes gave the clinical geneticist "his organ" and abetted the development of human genetics as a clinical speciality. Continually improving methods of chromosome study, and particularly the methods for mapping genes on chromosomes, have given all of medicine a new paradigm. Specialists in all medical areas approach the study of their most puzzling diseases by first mapping the genes responsible for them. Thus, just as Vesalius's anatomical text of 1543 formed the basis for the physiology of William Harvey (1628) and the morbid anatomy of Morgagni (1761), gene mapping is having a widely pervasive influence on medicine.

In examining the significance of mapping information in clinical medicine, it may be useful to substitute the anatomical metaphor (the anatomy of the human genome) for the cartographic metaphor (the human gene map). The anatomic metaphor prompts one to think in terms of the morbid anatomy, comparative anatomy and evolution, functional anatomy, developmental anatomy, and applied anatomy of the human genome.

The approximately 1900 expressed genes that have been mapped to specific chromosomes, and in most instances to specific chromosome regions or bands (Table 1), code for blood groups, enzymes, hormones, clotting factors, growth factors, receptors (e.g., for hormones and growth factors), cytokines, oncogenes, structural proteins (e.g., collagens and elastin), etc. They also code for a large number of disease genes for which the biochemical basis is not known or was not known before the mapping (Table 3).

In all, the chromosomal location of over 500 genetic disorders (the morbid anatomy of the human genome) has been

[3] *Haplotype* refers to the combination of specific alleles at several closely linked loci, e.g., the haplotypes at the major histocompatibility complex for HLA types, such as A3, B27, C2, DR2. The approach was adapted to tracing β-thalassemias and sickle cell anemia in the early 1980s and more recently has been used in the study of inborn errors of metabolism such as PKU and cystic fibrosis.

determined. For some disorders, the mapping has been done by locating the wild-type gene such as that for phenylalanine hydroxylase which is deficient in phenylketonuria (PKU) and maps to 12q. In other disorders, the mendelizing clinical phenotype has been mapped by family linkage studies using anchor markers such as RFLPs; examples are Huntington disease, polycystic kidney disease, and Marfan syndrome. In yet other disorders, mapping has been done by both approaches; for example, elliptocytosis, type I, was mapped to the distal region of the short arm of chromosome 1 by linkage to Rh and other markers situated there and also by mapping of the gene for protein 4.1 (which is mutant in that disorder) to the same region.

To date, the applied anatomy of the human genome has related particularly to those disorders for which the basic biochemical defect was not yet known (**Fig. 5**). Because of this ignorance, it was impossible to devise a fully specific diagnostic test or to design rational therapy. Once the chromosomal location of a disease-producing gene is known, together with its proximity to other genes and especially DNA markers, one can do diagnosis (prenatal, presymptomatic, and carrier) by the linkage principle. Furthermore, one can expect to determine the fundamental nature of the genetic lesion by walking or jumping in on the gene, a process often labeled, with questionable appropriateness, reverse genetics. Then, knowing the nature of the wild-type gene, one can work out the pathogenetic steps that connect gene to phene, mutation to clinical disorder. Secondary prevention and therapy through modification of those steps can be developed. In many instances gene therapy will probably find gene mapping information useful background.

In connection with the morbid anatomy of the human genome, mapping, more than any other single factor perhaps, has been responsible for establishing the chromosome basis of cancer: by the demonstration of specific, microscopically evident chromosomal aberrations in association with specific neoplasia, by the mapping of oncogenes and antioncogenes (recessive tumor suppressor genes), and by correlation of the two sets of observations. Somatic cell genetic disease as the basis not only of all neoplasia but also of some congenital malformations and probably of autoimmune diseases joins the three cardinal categories of disease as to genetic factors: single gene disorders, multifactorial disorders, and chromosomal disorders.

Mitochondrial genetic disease is a fifth category. The mitochondrial chromosome, the 25th chromosome, was completely sequenced by 1981 in the laboratory of Fred Sanger at Cambridge, and its genes were mapped in the 6 years that followed. Deletions and point mutations in the mitochondrial chromosome were then identified as the basis of Leber optic atrophy, myotonic epilepsy/ragged red fiber disease, Pearson pancreas-bone marrow syndrome, oncocytomas, and in cultured cells, chloramphenicol resistance. This sequence of discovery is the opposite, for the most part, of that pursued to date in the delineation of the nuclear genome where the progression has been from study of diseases, then to their mapping, and finally to the sequencing of the genes. The mitochondrial chromosome, with its mere 16,569 base pairs, can be considered a paradigm for what the human genome project (see later) hopes to achieve for the nuclear genome (which is approximately 200,000 times larger in terms of nucleotides). Indeed, complete sequencing may help greatly in identifying all the nuclear genes just as it did in the case of the mitochondrial genes. Presently we have information, as catalogued in *Mendelian Inheritance in Man* (8), on only about 5000 of the 50,000 to 100,000 genes, or gene loci (Fig. 4). Walter Gilbert suggested that complete sequencing may be the best way to find the rest. Thereafter, one can determine the function of the genes and the disorders caused by mutations therein, as has been done for the mitochondrial chromosome.

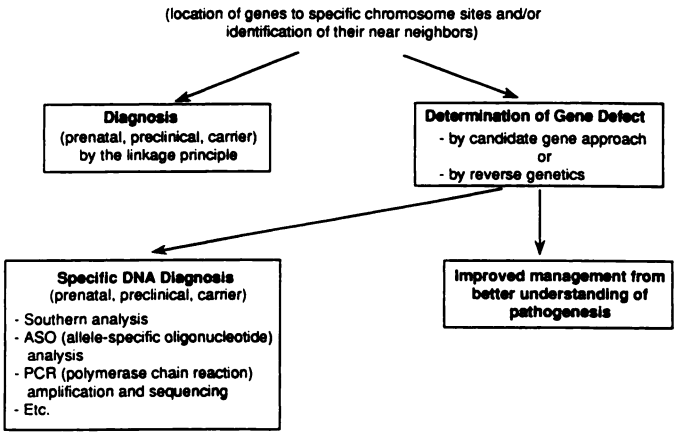

Figure 5. Clinical application of gene mapping.

OTHER TYPES OF MAPS

Gene mapping subsumes both genetic linkage maps, which are derived from meiotic recombination frequencies and measured in centimorgans, and physical maps, based on various experimental methods of which several have been described earlier. As discussed elsewhere, the large fragment clones that are produced by YAC cloning permit the creation of contig maps for each chromosome, i.e., maps of overlapping, and therefore contiguous, DNA segments. Smaller segments, cosmid clones, have been used for the same purpose but are less efficient. As already indicated, these mapped segments can be used for mapping genes of unknown location by hybridization. They are also the raw material for nucleotide sequencing. The contig map is the penultimate physical map; the nucleotide sequence is the ultimate map.

The correlation of the genetic map (based on the location of cloned genes through family linkage studies) and the physical map (as represented by the contig map, for example) is likely to be aided by the use of STSs (sequence-tagged sites) (19). They may obviate, to a considerable extent, the need to store and distribute cloned DNA segments for study by investigators in many laboratories. Knowing the STS identifier of a particular segment, the researcher can clone that segment and not require it as a clone from a repository. One will store information, not DNA, related to each part of the genome. STSs are a sort of Esperanto for the scientists using diverse genetic and physical methods.

Finding all the genes could be helped by a concerted effort to create the cDNA map—the map of probes made by reverse transcription of messenger RNAs that have tissue and developmental stage specificity. The cDNA (or exon) map would provide candidate genes for diseases mapped by phenotype. Because by definition the exons are where the action is, sequencing could logically and efficiently start with them.

THE HUMAN GENOME PROJECT

The human genome initiative aims to map and sequence completely the human genome. The sequence is the ultimate

map. The NRC/NAS committee, which concluded that the Human Genome Project was worthwhile and could be done in a reasonable time frame and budget, suggested "map first, sequence later" (20). No absolute stepwise dichotomy was in the mind of the committee, but this order was considered advisable because both the gene map and the contig map are necessary for most efficient sequencing, and the technology for sequencing requires improvement.

Given adequate funding (an estimated $200 million a year for the worldwide effort, in 1988 dollars), the human genome project should be completed in 15 years. Spin-off in applications to medicine, in particular, occurs continuously during that period. It is well to be aware of the end-product of the project: It will produce a reference map (and sequence)—a source book for human biology and medicine for centuries to come. Two broad areas will occupy scientists for many years: variation and function. It is urged by some that studies of variation in populations worldwide should begin immediately. Important as these studies are, it seems that the extent of variation is not such that confusion will arise in assembling the reference map. The journalists' question is irrelevant, Whose genome will be mapped and sequenced? Although the source of DNA used in sequencing should be recorded, the final map and sequence will be a composite of information from many individuals. The extent of variation will need to be studied later. It and function are all of biology and genetics; it is unreasonable to expect the human genome project to do it all and it would be imprudent for the effort of the human genome project to be diffuse. It will be important to "keep our eyes on the ball" if the main objective of a complete map and sequence is to be achieved in a timely manner and within the budget limits set. It is projected that doing the entire job will represent economy of scale and having the map/sequence information will facilitate greatly the elucidation of genetic disease. Cystic fibrosis is an example: identification of the gene was an expensive project (how expensive is difficult to estimate precisely) and would undoubtedly have been much less expensive (how much less is also hard to say) if the complete sequence were available.

COUNTER-ARGUMENTS AND COUNTER-COUNTER-ARGUMENTS CONCERNING THE HUMAN GENOME PROJECT

Among scientists, the main arguments against the Human Genome Project seem to be four: that it is bad science; that it is big science (with the implication that it is not the way science is most effectively done); that it creates an improper milieu for doctoral training in science; and that it is taking money away from other worthy (in the opinion of some, more worthy) projects.[4]

The purpose of the HGP is to create a tool for science—the source book referred to earlier. A good deal of applied science and engineering will go into the HGP, but it may not be appropriate to criticize the project through a comparison with biological science as it has been pursued traditionally.

The HGP is not so much big science as it is coordinated, interdisciplinary science. Especially in the sequencing part of the project, both the generation of the data and particularly their interpretation will require the recruitment of experts from disciplines that have had little involvement in biology to date. It seems clear that handling the information, validating, storing, and retrieving it, searching it for patterns indicative of functional domains, and identifying the coding portions, will require much computer-assisted expertise. It represents a formidable challenge to the information scientist.

To the argument that an institution where genomics (a generic term for mapping and sequencing) is being done is a poor site for graduate training and that the HGP will have a major adverse effect on graduate programs in biology, reality may be quite the contrary. With the full map and sequence, graduate students will be spared the drudgery of cloning and sequencing particular genes before they can get down to the much more interesting and intellectually demanding work of studying variation, function, regulation, and so on. The genomics laboratories will be superb settings for training a new breed of scientist—one who is prepared to capitalize on both the molecular genetics revolution and the computation revolution. These will be the leaders in biology in the 21st century.

It appears that the HGP is being made a scapegoat by those frustrated by the tight funding for research. If the HGP went away completely, the funding situation would not change perceptibly. Indeed, the excitement stimulated by discussion of the HGP has had and can continue to have a beneficial effect on science funding generally. FJ

REFERENCES

1. Wilson, E. B. (1911) The sex chromosomes. *Arch. Mikrosk. Anat. Enwicklungsmech.* **77**, 249-271
2. Horner, J. F. (1876) Die Erblichkeit des Daltonismus. *Amtl. Ber. Ueber die Verwaltung des Medizinalwesens des Kt. Zurich v. Jahre 1876.* **31**, 208-211
3. Haldane, J. B. S., Sprunt, A. D., and Haldane, N. M. (1915) Reduplication in mice. (Preliminary communication.) *J. Genet.* **5**, 133-135
4. Ott, J. (1985) *Analysis of Human Genetic Linkage.* The Johns Hopkins Univ. Press, Baltimore
5. Donahue, R. P., Bias, W. B., Renwick, J. H., and McKusick, V. A. (1968) Probable assignment of the Duffy blood group locus to chromosome 1 in man. *Proc. Natl. Acad. Sci. USA* **61**, 949-955
6. Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980) Construction of a genetic linkage map in man using restriction length polymorphisms. *Am. J. Hum. Genet.* **32**, 314-331
7. Roberts, L. (1990) The worm project. *Science* **248**, 1310-1313
8. McKusick, V. A. (1990) Mendelian inheritance in man: catalogs of autosomal dominant, autosomal recessive, and X-linked phenotypes, 9th ed. Johns Hopkins University Press, Baltimore
9. Fan, Y.-s., Davis, L. M., and Shows, T. B. (1990) Mapping small DNA sequences by fluorescence in situ hybridization directly on banded metaphase chromosomes. *Proc. Natl. Acad. Sci. USA* **87**, 6223-6227
10. Lawrence, J. B., Villnave, C. A., and Singer, R. H. (1988) Sensitive high resolution chromatin and chromosome mapping in situ: presence and orientation of closely integrated copies of EBV in a lymphoma cell line. *Cell* **52**, 51-56
11. Boehnke, M., Arnheim, N., Li, H., and Collins, F. S. (1989) Fine-structure genetic mapping of human chromosomes using the polymerase chain reaction on single sperm: experimental design considerations. *Am. J. Hum. Genet.* **45**, 21-32
12. Cui, X., Li, H., Goradia, T. M., Lange, K., Kazazian, H. H., Jr., Galas, D., and Arnheim, N. (1989) Single-sperm typing: determination of genetic distance between the G-gamma-globin and parathyroid hormone loci by using the polymerase chain reaction and allele-specific oligomers. *Proc. Natl. Acad. Sci. USA* **86**, 9389-9393

[4] I am indebted to Leroy E. Hood for this listing of counterarguments and, in part, for the formulation of counter-counterarguments.

13. Lander, E. S., and Botstein, D. (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science.* **235,** 1567-1570
14. Taylor, B. A. (1989) Recombinant inbred strains. In *Genetic Variants and Strains of the Laboratory Mouse,* 2nd ed. (Lyon, M., and Searle, A. G., eds) Oxford University Press, New York
15. Goss, S. J., and Harris, H. (1975) New method for mapping genes in human chromosomes. *Nature (London)* **255,** 680-684
16. Cox, D. R., Pritchard, C. A., Uglum, E., Gashere, D., Kobori, J., and Meyers, R. M. (1989) Segregation of the Huntington disease region of human chromosome 4 in a somatic cell hybrid. *Genomics* **4,** 397-407
17. Goodfellow, P. J., Povey, S., Nevanlinna, H. A., and Goodfellow, P. N. (1990) Generation of a panel of somatic cell hybrids containing unselected fragments of human chromosome 10 by X-ray irradiation and cell fusion: application to isolating the MEN2A region in hybrid cells. *Somat. Cell Mol. Genet.* **16,** 163-171
18. McKusick, V. A. The morbid anatomy of the human genome: a review of gene mapping in clinical medicine. *Medicine* **65,** 1-33, 1986; **66,** 1-63, 237-296, 1987; **67,** 1-19, 1988
19. Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989) A common language for physical mapping of the human genome. *Science* **245,** 1434-1435
20. National Research Council (1988) Committee on Mapping and Sequencing the Human Genome: mapping and sequencing the human genome. National Academy Press, Washington, D.C.