Mapping the Human Genome: Current Status

J. CLAIBORNE STEPHENS,* MARK L. CAVANAUGH, MARGARET I. GRADIE, MARTIN L. MADOR, KENNETH K. KIDD

The human genome has already been the subject of extensive research activity even though the Human Genome Project is only just officially starting. This review and the accompanying wall chart attempt to provide an integrated, quantitative, and detailed summary of the status of knowledge on the human genome in mid-1990. The analysis has highlighted the rudimentary nature of many of the information links needed for the task. While this overview could not be fully comprehensive and required simplifying assumptions, the results have provided estimates of relative progress on a region-by-region basis throughout the genome.

A MAP IS A REPRESENTATION OF THE RELATIONSHIPS AMONG landmarks organized according to a defined coordinate system. Genetic maps have been constructed from many different types of data and have used different metrics, from the first genetic linkage map in 1913 (*1*) to today's detailed molecular maps. Intensified mapping activity over the last decade has generated a substantial amount of data relating to genome organization including comparisons between organisms (*2*), functional groups of genes or dispersed gene families, and chromosome regions that have been associated with pathologies (*3*). The initial phase of the official mapping effort has been focused on the description of order and spatial relationships among genetic landmarks [such as polymorphisms, genes, and DNA sequences (*4*)] in order to generate a dense linkage map, a variety of physical maps, and the beginnings of a composite DNA sequence.

Linkage maps. Genetic linkage maps are based on the coinheritance of allele combinations across multiple polymorphic loci. Parental combinations are usually transmitted if the loci are molecularly close, but recombination at meiosis will generate nonparental combinations more frequently if the loci are farther apart. The primary source of linkage data is the observation of gametic allele combinations. The allelic constitution of gametes for human linkage studies has conventionally been determined indirectly by family studies and statistical inference, but direct molecular analysis of gametes and single chromosomes has recently become possible (*5*). Whereas distances between loci in kilobases of DNA are additive across successive intervals along a chromosome, this is not true for distances measured as recombination frequencies. Therefore, linkage maps use centimorgan units (cM), a measure based on the frequency of recombination, but identical to it only in the limit of small distances (*6*). A map unit of 1 cM corresponds, at that limit, to an observation of recombination in 1% of the gametes in the sample (*7*). The predicted total length for the sex-averaged linkage map is 3300 cM (*8*).

Physical maps. Physical maps can be cytogenetically or molecularly based. Cytogenetically based physical maps order loci with respect to the visible banding pattern or relative position along the chromosomes, primarily by means of data from somatic cell hybrids and in situ hybridization (*9*). Molecularly based physical maps directly characterize large tracts of DNA by establishing molecular landmarks, such as restriction endonuclease sites and sequence tagged sites (STSs) (*10*). These maps are usually constructed from data generated by pulsed-field gel electrophoresis (PFGE) (*11*) or by related techniques that size and order large restriction fragments of genomic DNA. Another molecular strategy is to characterize cloned DNA [in the form of yeast artificial chromosomes (YACs) (*12*), cosmids, or shorter phage vectors] sufficiently to establish overlapping assemblages of clones, known as contigs.

Cytogenetic maps have a scale and coordinate system corresponding to the chromosome banding pattern. The large-scale restriction maps have a scale on the order of kilobases of DNA (kb) and have not been related to specific chromosome bands. Conversion and comparison between physical maps with such different types of scales is a high priority, but common reference points will have to be mapped and the conversion factors determined empirically. Sequence-tagged sites have been proposed as the common reference points that could be used to coordinate information from different mapping strategies (*10*).

The highest level of resolution for a molecularly based physical map is the DNA sequence, which gives the linear order of nucleotides for each of the 24 distinct human chromosomes. Leaving aside for the moment the question of sequence polymorphism, a complete reference sequence will contain roughly 3×10^9 bp of DNA (*13*).

The order of loci in physical maps and linkage maps will be the same, but there exists no simple means to convert physical distances into recombination frequencies. Recombination frequencies per megabase of DNA vary considerably by sex and by chromosomal region. The emerging pattern of regional variation shows that telomeric regions have proportionately more recombination in male meiosis and that centromeric regions have higher frequencies of recombination in female meiosis. Overall, there are higher frequencies of recombination in female meiosis than male meiosis (usually resulting in longer maps for females), but the ratios of sex-specific map lengths differ among the chromosomes (*14*).

J. C. Stephens is in the Laboratory of Viral Carcinogenesis, National Cancer Institute, Frederick, MD, 21701. M. L. Cavanaugh can be contacted through K. K. Kidd. M. I. Gradie is in the Department of Anthropology, Yale University, New Haven, CT 06520. M. L. Mador and K. K. Kidd are in the Department of Human Genetics, Yale University School of Medicine, New Haven, CT 06510. The authors were affiliated with the Yale–Howard Hughes Medical Institute Human Gene Mapping Library prior to its closing.

*To whom correspondence should be addressed.

Short-term (5-year) goals have been established for a few specific types of maps. For physical mapping, the short-term goals are (i) to assemble STS maps of all human chromosomes with the goal of having markers spaced at approximately 100,000-bp intervals, and (ii) to generate overlapping sets of cloned DNA or closely spaced, unambiguously ordered markers with continuity over lengths of 2×10^6 bp for large parts of the human genome (*4*).

The short-term goal for the linkage map is to enhance its resolution for each chromosome to a spacing of 2 to 5 cM (*4*). The >2000 DNA polymorphisms already identified and cataloged would allow construction of a linkage map with average interval size between adjacent loci of about 2 cM. However, published maps have not yet incorporated enough of these to achieve a resolution of more than approximately 5 to 10 cM. Furthermore, since as the distribution of known polymorphic markers is far from uniform, many large gaps will remain until new markers are added.

Quantifying Progress in Human Genome Mapping

The great variability in levels of resolution and the different goals for the various mapping strategies made a single coherent description of the progress of genome mapping quite challenging. The wall chart in this issue (*15*) along with the results presented below attempt to provide such a summary. Our objectives were to develop a generally applicable method for quantifying mapping activities on a regional basis throughout the genome and to relate data among different types of maps.

Sources of data. Three primary sources of data were utilized for this study: the Human Gene Mapping Library (HGML) Database (*16*), the GenBank sequence database (*17*), and published linkage maps (*18–47*). The HGML provided the cytogenetic map locations of loci. Broadly defined, a locus may be any genomic location ranging in size from single base pairs to entire gene clusters. For the most part, loci in the HGML are functional genes, pseudogenes, fragile sites, or anonymous pieces of DNA that have been mapped. Probes, on the other hand, are generally cloned pieces of DNA or primer pairs from polymerase chain reaction (PCR) analyses that are used to define or delimit a locus; there are often multiple probes for a given locus. Polymorphisms entered into the HGML are DNA variations that have been detected primarily by restriction endonuclease analysis, or more recently by PCR-based analyses. The HGML database included all map locations from the Tenth International Human Gene Mapping Workshop (HGM10) and subsequent updates based on articles published or in press through 31 July 1990.

A continuing collaborative effort with GenBank (*48*) has provided us with the links needed to relate DNA sequence to cytogenetically mapped loci. Most sequence lengths and nucleotide frequencies were current as of GenBank release 64.0.

Preparation of illustrative linkage maps for the wall chart. The knowledge necessary for the construction of comprehensive linkage maps is in a state of rapid transition. Few comprehensive linkage maps existed only 4 years ago (*49*). The CEPH collaboration (*50*) has had a catalytic effect on this field. There is now the expectation that for most chromosomes consensus maps at better than 10-cM resolution will be published within the next year or two. Many chromosomes already have multiple linkage maps published by different researchers, but these maps often have few markers in common, making complete integration impossible without additional data. An added complication is the different analytic methods being used to generate maps; multipoint methods simultaneously evaluate several loci while pairwise methods consider loci only two at a time. These different methods involve different assumptions and can yield different results, as noted below. Thus, an attempt at a comprehensive synthesis of the human linkage map is premature. We have instead chosen to present illustrative maps based on the most comprehensive of the published maps for each chromosome. The maps actually used to prepare these illustrative maps are from the articles noted in Table 1. Many other published linkage studies dealing with smaller numbers of loci exist (*51*), but were not used in constructing the maps on the wall chart.

The map for each chromosome was prepared by selecting loci to give an average spacing of roughly 10 to 20 cM. The selection of loci was conservative and included only markers with unambiguous orders. Markers from separate studies were combined when warranted, but only when no ambiguity of order would be introduced. For example, separate maps of the arms of a chromosome could be combined if a centromeric locus is shared. Combinations largely involved blocks of genes, as for the example above, but occasionally a locus was intercalated into the illustrative map when flanked by markers shared by both maps. In such cases, the flanking markers were far apart and the evidence was compelling that the intercalated marker is between the other two. In choosing which markers to illustrate in the maps, preference was given to genes over anonymous DNA segments, to loci with higher heterozygosities, and to loci more frequently studied. The resulting maps are heuristic in nature and were prepared primarily as a framework to portray the distribution of polymorphic loci in relation to the linkage intervals.

Genomic coordinate systems. Perhaps the most familiar coordinate system for genomic mapping data is the cytogenetic banding pattern. Conventional summaries (*52*) of cytogenetic mapping data depict all genes and other loci in relation to the cytological banding patterns of each chromosome, with vertical bars or brackets indicating the interval—encompassing one or more bands—to which a locus has been mapped. The bands are themselves genomic intervals, and can be thought of as the fundamental intervals in the cytogenetic coordinate system. Because of overlaps among map locations, cytogenetic summaries become quite cumbersome when a large number of loci are mapped with varying degrees of precision.

Other coordinate systems exist, such as the centimorgan scale for linkage and percentage of chromosome length for in situ hybridization (*53*), but conversion among different systems is very difficult, at best. A unifying approach for summarizing the distribution of genes and other mapped loci over different types of maps with diverse coordinate systems would be highly desirable. For all coordinate systems currently in use, the inherent uncertainty in the mapping procedure yields an interval that contains the actual map location, rather than a point. Thus, overlap among map locations are likely to be a continuing problem.

Our solution to this problem was to develop a generally applicable scheme (described below) for quantifying the regional distribution of mapping activity. We have applied the resulting algorithm to the cytogenetic coordinate system. As an approximation to the relative lengths of all chromosome bands, we have used the lengths of the 860 bands (measured at a resolution of 0.1 mm) in the high-resolution International System for Human Cytogenetic Nomenclature 1985 (ISCN) depiction (*54*). The ISCN provides an internationally recognized standard nomenclature and banding pattern with which researchers can unambiguously indicate mapping coordinates for a given locus.

Allocation algorithm. Our logic for quantifying mapping activity as a distribution was to break down each map location into its fundamental intervals, and allocate each locus proportionally to those intervals. As an example of this procedure, consider a locus mapped to chromosome 1 without regional localization; it has some

probability of being in any band on the entire chromosome. On the other hand, a locus mapped to a single band can be said to be in that band with probability = 1, barring mapping errors. Thus, b_i/l_j is an estimate of the probability that locus j is in band i; where b_i is the length of band i, l_j is the length of the map location of locus j, and i varies over each band in locus j's map location. We will take these probabilities as the proportional allocations of locus j into each band of its map location. When this is done for all mapped loci whose map location includes band i, the total number of loci allocated to band i is

$$\sum_j \frac{b_i}{l_j} \tag{1}$$

In this sum b_i is a constant, so that the remaining term, $\Sigma 1/l_j$, is a type of density (loci/length) which, when multiplied by the band's length b_i, yields the number of loci allocated to band i. These probabilities are calculated on the basis of the simplifying assumption that the only factor determining locus distribution is length of each chromosome band.

Estimating progress towards closure or completeness. As a working definition, cytogenetic mapping of genes can be considered complete when all genes have been identified and have map resolutions of one band. Obviously, this definition does not consider order within bands, but it does allow us to estimate completeness by comparing the number of genes allocated to each band with the number expected to be there.

Let L be the length of the genome, B the number of bands each of length b_i, N the true number of genes, and C the number of mapped genes. Our expectation is that each band contains a number of genes (n_i) that is proportional to that band's relative cytogenetic length [$n_i = N(b_i/L)$] following the same assumption made for the allocation algorithm. Hence, completeness for band i can be calculated as the ratio of current allocation to expectation:

$$\left(\sum_j \frac{b_i}{l_j}\right) \Big/ N\left(\frac{b_i}{L}\right) = \left(\sum_j \frac{1}{l_j}\right) \Big/ \left(\frac{N}{L}\right) \tag{2}$$

where, again, the sum is over all genes whose map location includes band i. This is the ratio of gene density within band i to the global gene density (N/L). This rationale can be applied to larger genomic regions, such as individual chromosomes. However, for regions larger than the coordinate system's fundamental intervals, precision of mapping must be taken into account; otherwise, we get trivial estimates such as C/N as an estimate of completeness of gene mapping for the entire genome. The sum $\Sigma 1/l_j$ is an estimate of the total precision for genes mapped to a given genomic region. If this sum is taken over all C mapped genes, it is an estimate taken over the entire genome; if it is taken over the C_k genes mapped to chromosome k, the sum is an estimate specific to chromosome k. This sum increases as additional genes are mapped and as map precision increases. At completion, the sums are taken over N and N_k, respectively, in which case we have the approximations

$$\sum_j^N \frac{1}{l_j} = \sum_i^B \frac{n_i}{b_i} \approx \sum_i^B N\left(\frac{b_i}{L}\right) \Big/ b_i = \frac{NB}{L} \tag{3}$$

Table 1. Data used for the wall chart and analyses are displayed. The length column represents the length of each chromosome as a percent of the total length of the genome measured from the ISCN ideograms. Most of the HGML data in these analyses were used in a straightforward manner and are raw counts, but some aspects require clarification. (i) Map locations. Most loci are mapped to a single cytogenetic interval, which has an easily defined length. However, 112 loci are mapped to discontinuous intervals, and some localizations span two different chromosomes. Conceptually these loci pose no real problems, since their more complex map locations are still definable as lengths of the human genome. (ii) Sequence overlaps. Most GenBank entries are unique DNA sequences, each for a specific region of the genome. However, as the sequencing effort intensifies, there will be a tendency for separate database entries to have overlapping sequences, generating redundancy. Currently there is no automatic means for identifying and flagging these overlaps. We have compared all GenBank sequence entries that were linked to loci with overlapping map locations and eliminated 90 kb of redundant sequence. The remainder represents about 80% of the DNA sequence data that has been cataloged. An additional 1,545,255 bp have yet to be associated with specific chromosomal locations. (iii) Probes. Since we wanted to determine the distribution of unique probes with unambiguous map locations, we have filtered out 550 probes known to be subclones of larger probes, and 628 probes that show homology to multiple genomic locations.

Chromosome	Length	Loci				Probes	Sequence (bp)	Linkage references
		All loci	Genes	Sequenced genes	Polymorphic loci			
1	8.3	311	192	82	146	677	388,576	(*20–22*)
2	7.9	196	116	50	90	522	538,478	(*20, 23–25*)
3	6.4	786	75	29	130	872	112,721	(*18, 20*)
4	6.1	242	73	34	138	461	199,261	(*26*)
5	5.8	192	74	28	112	382	157,066	(*19, 20, 27–29*)
6	5.5	207	110	55	86	620	451,606	(*18, 20, 30*)
7	5.1	555	121	50	189	965	285,589	(*31*)
8	4.5	172	58	25	55	332	190,517	(*20*)
9	4.4	110	65	24	47	206	151,098	(*18–20, 32*)
10	4.4	156	62	28	88	253	153,856	(*33, 34*)
11	4.4	624	140	55	189	1,191	336,252	(*18, 20,35*)
12	4.1	155	103	45	56	402	276,461	(*18–20*)
13	3.6	122	29	12	53	265	76,751	(*18–20, 36, 37*)
14	3.5	98	56	33	51	493	243,892	(*18*)
15	3.3	126	52	20	49	163	118,313	(*38*)
16	2.8	335	59	25	122	457	155,443	(*20*)
17	2.7	451	99	47	150	662	312,904	(*39, 40*)
18	2.5	55	23	10	32	143	100,080	(*18–20*)
19	2.3	194	82	38	59	462	243,674	(*41*)
20	2.1	64	37	17	22	141	106,462	(*19, 20, 42*)
21	1.8	202	34	7	60	308	32,675	(*43, 44*)
22	1.9	238	57	22	99	396	101,325	(*45*)
X	4.7	730	179	31	235	1,245	317,945	(*18, 46, 47*)
Y	2.0	231	13	5	17	234	15,105	(*47*)
Totals	100	6,552	1,909	772	2,275	11,852	5,066,049	

as the expectation for the entire genome, and

$$\sum_{j}^{N_k} \frac{1}{l_j} = \sum_{i}^{B_k} \frac{n_i}{b_i} \approx \sum_{i}^{B_k} N\left(\frac{b_i}{L}\right) \Big/ b_i = \frac{NB_k}{L} \quad (4)$$

as the expectation for chromosome k. In the second equation, N_k is the number of genes on chromosome k and B_k is the number of its bands. The first equality in each equation follows from our assertion that completion is the mapping of every gene to a specific band, and the approximation follows from our assumption that genes are distributed proportionately according to band length. We may now calculate completeness as the ratio

$$\left(\sum_{j}^{c} \frac{1}{l_j}\right) \Big/ \left(\frac{NB}{L}\right) \quad (5)$$

for the entire genome, and

$$\left(\sum_{j}^{c_k} \frac{1}{l_j}\right) \Big/ \left(\frac{NB_k}{L}\right) \quad (6)$$

for chromosome k. For sequencing, the same logic was applied: we replaced the number of genes allocated to a region with the number of base pairs allocated, and allowed N to be the true number of base pairs in the genome.

Since there is no meaningful limit to the number of probes or polymorphisms in each band, there can be no clear definition of completeness. A contig consisting of overlapping probes that spanned an entire band would constitute completeness for one type of physical map, but a large volume of such data is not yet readily available. Consequently, although we have used color bars to depict gene mapping and sequence completion on the well chart, the color bars for probes and polymorphisms only depict estimates of numbers per band.

Map location refinement by means of locus order. Most cytogenetic localizations arise directly from physical mapping procedures. However, indirect information such as locus order from linkage maps can be used to verify or refine these localizations. For instance, if three loci are in a known order, the cytogenetic localization of the middle locus is bounded distally by the distal limit of the regional assignment for the distal locus and proximally by the proximal limit of the proximal locus. This logic can be applied sequentially along the chromosome to each locus for which order information is known. The most precisely localized marker will exert considerable constraint on the regional localizations of a large number of other loci. A clear example is the newly described locus D10S96 (*34*): the marker only has a chromosomal localization (chromosome 10). Yet, its unambiguous position on the linkage map between D10S5 and CDC2, both of which are localized to 10q21.1, refines the localization of D10S96 to 10q21.1.

Since the order of each linkage map limits the possible map refinements, the refinement procedure can be applied iteratively when a collection of maps exists for a single chromosome. When loci are shared among these maps, the shared loci may receive different refinements that can be detected and resolved on each pass. Iteration terminates when there is no further refinement. We have used more than 100 linkage maps containing over 950 distinct loci (*55*). Two hundred and seventy-four of these loci were selected for illustration in the linkage maps on the wall chart (*51*). For most chromosomes there were no further refinements after four iterations.

Allocation of polymorphic loci to linkage intervals. The refined localizations were used to portray the number of cataloged polymorphic loci in each interval of the illustrative linkage maps. Polymorphic loci, other than those used in linkage studies, cannot be allocated directly to linkage intervals unless correspondence has been established between linkage intervals and cytogenetic intervals. The cytogenetic interval corresponding to a specific linkage interval is taken as the union of the refined cytogenetic intervals of the pair of loci determining that linkage interval (*56*). The sum of polymorphic loci allocated to each band within this cytogenetic interval is our estimate of the number of polymorphic loci within the corresponding linkage map interval.

Although each linkage interval is treated independently in this allocation, the fact that adjacent intervals share a common polymorphic locus means that all polymorphic loci allocated to the cytogenetic interval of this common locus are counted in both linkage intervals. Because of our refinement procedure, the cytogenetic overlap between adjacent linkage intervals is precisely the cytogenetic map location of the shared locus. A consequence of this unavoidable overlap is that the allocated numbers of polymorphic loci are not additive across the linkage map. They are however, for each single interval, valid maximal estimates.

Analyses of Human Genome Mapping Activity

The mapping data used in our overview are summarized in Table 1. Three overlapping subsets of the 6552 loci—genes, sequenced loci, and polymorphic loci—have been analyzed individually because of their relevance to functional characterization, sequencing, and linkage mapping activities, respectively. The allocation algorithm was applied to all loci, genes, sequenced genes, polymorphic loci, probes, and DNA sequence.

In the wall chart, distributions were shown in relation to the chromosome banding pattern. The distributions for genes and mapped DNA sequences were shown as progress towards completion for each band, whereas progress in probe identification was shown as the number of probes within each band. Numbers of polymorphic loci were shown only in relation to the linkage map as described above.

For convenience, each mapping activity has been depicted in Fig. 1 as a histogram with a numeric scale by recasting the results of our allocation to the cytogenetic bands in terms of 400 intervals of roughly equal size. In Fig. 1 polymorphic loci are represented on a cytogenetic length scale, rather than on the centimorgan scale used on the wall chart. In both the wall chart and Fig. 1 the considerable overlap among the various mapping activities is immediately obvious. For instance, localized spikes of activity on chromosomes 6, 11, and 14 (due in part to the HLA cluster, the β hemoglobin and WAGR regions, and the immunoglobulin heavy chain region, respectively) were found for all types of mapping activity. To quantify this impression, we calculated the correlation coefficient between each pair of mapping activities. All calculated coefficients would have been highly significant if the distributions of mapping activities were Gaussian. As this does not appear to be the case, variation in magnitude (0.35 to 0.83) could only be used to suggest trends. There was a trend for regions with more sequenced genes to have larger numbers of base pairs allocated. Mapping of all loci and polymorphic loci were positively correlated, but each was poorly correlated with sequenced genes and with base pairs of sequence. The poor association between (polymorphic) loci and sequencing may reflect a preference toward sequencing functional genes. Lastly, the probe distribution was relatively highly associated with all mapping activities, which confirms that probes are central to current mapping strategies.

The genome-wide estimate for complete identification of genes (Eq. 5) was 0.52%. The simple estimate of 1.9% (C/N, where C = 1909 and N = 100,000) was almost fourfold higher, but does not account for mapping resolution. The peaks of activity in Fig. 1 suggest substantial variability in regional progress toward completion. We have used Eq. 6 to estimate progress towards complete

identification of genes in all of the bands for each chromosome (Table 2). Chromosomes 22, X, 11, and 17 were each estimated to be over 1% complete, whereas chromosomes 15, 3, 5, and 8 were farthest from completion. This ranking takes into account not only the number of genes mapped to each chromosome and their mapping resolutions, but also the relative goals for each chromosome; a chromosome with few bands is easier to complete than a chromosome of similar size that has many bands.

Ten bands were more than 10% complete, as measured by Eq. 2 (Table 3). These bands are the well-characterized regions associated with β-hemoglobin, immune function (HLA, immunoglobulin chains), and several diseases. Most of these bands were relatively highly ranked (>1%) for DNA sequence completion as well. The largest band (Yq12) was estimated to be the farthest from gene mapping completion. In fact, this band consists of highly polymorphic heterochromatin, and is thought unlikely to harbor many genes. As more information is gathered, it will be possible to look for systematic differences among the different types of bands. The other nine of the ten least complete bands are all on chromosome 13. There are relatively few genes mapped to 13 (Table 1), but they are highly localized, leaving many bands virtually empty.

Completion of the linkage map can be estimated from the summary linkage map presented in HGM10 (*18*). In this summary map, 5.7% of the total centimorgan length spanned by the maps was contained in intervals ≤2 cM, 29.2% was in intervals ≤5 cM, while 58.3% was contained in intervals ≤10 cM. These figures are very crude estimates of the completion at the various levels of resolution and may be misleading for a variety of reasons. First, very few linkage maps have included telomeric loci, which means that the total centimorgan length is an underestimate of the true length. Second, this particular set of maps was not created for the purpose of estimating completion, and may not be ideal for this purpose. Third, the various statistical methods for calculating centimorgan length often yield very different results and the distances in this set of maps (*18*) are consistently smaller than in other maps of the same loci. A more rigorous estimate of completion would have to take these factors into account, and should attempt to estimate ranges of values for each component interval.

Table 2. Chromosomes (C) ranked by percent completeness (% Comp) as estimated by Eq. 6.

C	% Comp	C	% Comp	C	% Comp	C	% Comp
22	1.28	16	0.77	Y	0.40	4	0.28
X	1.18	21	0.72	2	0.39	9	0.27
11	1.12	12	0.71	13	0.34	8	0.24
17	1.10	1	0.56	18	0.33	5	0.21
14	0.80	6	0.56	20	0.32	3	0.21
19	0.78	7	0.54	10	0.29	15	0.19

A different perspective on the mapping effort comes from an examination of the distribution of mapping resolutions. Currently, 170 genes have been mapped to a single ISCN band; in other words, these genes are "completely" mapped to this level of cytogenetic resolution. There are 72 fragile sites and 340 anonymous DNA segments that have also been mapped to a resolution of one band. A more quantitative analysis of mapping resolutions is to

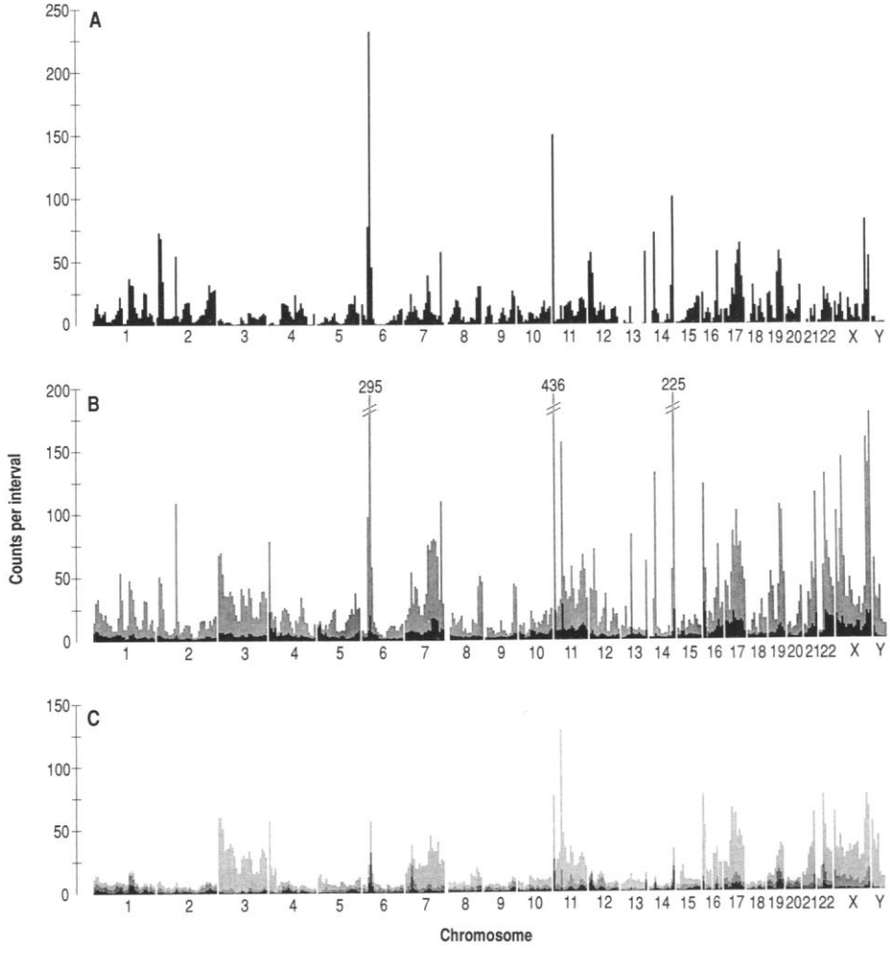

Fig. 1. Density distributions for various types of mapping data. The genome was divided in 400 intervals of approximately equal sizes (0.25% of the genome) based on length of the ISCN bands. For each chromosome, an integral number of intervals was used so that chromosomes ends would coincide with interval boundaries. Counts per interval for each of the following are shown: (**A**) base pairs (in 1000s); (**B**) probes (gray) and polymorphic loci (black); (**C**) all loci (light gray), genes (gray), and sequenced loci (black).

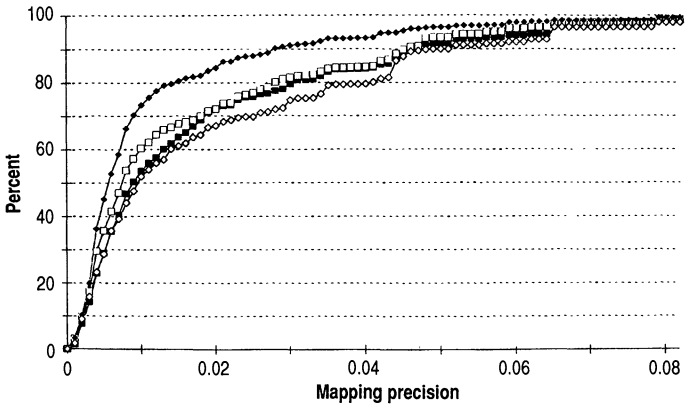

Fig. 2. Cumulative distribution functions of cytogenetic mapping precisions. Mapping precision is measured as percent of genome length. Closed squares, all loci; open squares, genes; open diamonds, polymorphic loci; closed diamonds, sequenced genes.

examine the lengths of all map localizations (Fig. 2). Most loci are mapped to within 1% of the genome; all are mapped to a resolution less than 8.3% (the length of chromosome 1). The average lengths of map locations are: 1.7% of the genome length for all loci, 1.6% for genes, 1.1% for sequenced genes, and 2.0% for polymorphic loci. Localization of sequenced genes had the highest precision, which suggests that they are generally under intense study. This trend may be bolstered by high-resolution PCR-based mapping techniques that make use of the DNA sequence data. The lower precision of localization of polymorphic loci reflects the fact that many such loci are anonymous DNA segments that have only been localized by linkage, which typically yields less precise map locations. As loci become more precisely mapped along the chromosome, the ability to order loci unambiguously will improve, as will the ability to draw relationships between the cytogenetic map and other types of maps such as the linkage map.

Our analyses have only used the chromosomal banding pattern as a coordinate system. In reality, biological differences underlie the observed banding pattern. Many explanations have been proposed, the most common of which invokes the observed difference in replication timing between light and dark bands (*57*). Although we cannot test this hypothesis with our data, we can demonstrate how the data can be used to test two other hypotheses. The first is that functional genes tend to be clustered in light bands. If the bands that cannot be simply categorized as light or dark (for example, centromeric bands) are ignored, the genome consists of 413 light bands and 377 dark bands representing 51.2% and 48.8% of the genome length, respectively. When these lengths are treated as the binomial probabilities that a gene falls in a light band versus a dark band, the 99% confidence interval for the proportion of genes expected in light bands is 48.1 to 54.2%. However, 59.1% (1052.1 of the 1780.7 total genes allocated to light or dark bands) was allocated to light bands, a significant excess. This difference is even more impressive when one considers that the allocation algorithm could obscure this tendency—all genes with map localizations that include multiple bands are allocated proportionally among these bands. A more direct test is to examine genes that have been mapped to a single band. Of the 170 such genes, 115 (68%) were in light bands. Sixty-four of 72 fragile sites (89%) and 266 of 340 anonymous DNA segments (78%) that have been mapped to single bands were also in light bands. Most of the single-copy fraction of the genome is thought to be nonfunctional. Most anonymous DNA segments are single copy and if they are, therefore, nonfunctional, then the perceived clustering of genes in light bands may be due to systematic mapping biases rather than actual clustering of the genes. Possibilities include greater ability to resolve loci within light bands, sampling bias for the loci being mapped (for example, exclusion of repetitive elements or preference for highly expressed genes), or greater condensation of DNA in light bands, and hence more DNA—an explanation contrary to prevailing expectations. Investigation of transcriptional activity does not show a bias between light and dark bands (*57*).

The second hypothesis is that GC content (the percentage of DNA composed of guanine and cytosine base pairs) is higher in light bands than in dark bands (*57, 58*). The loci that have been sequenced and mapped to single bands provide a relevant data set. The average GC content for sequences linked to loci mapped to a single light band was 0.517; that for sequences linked to loci mapped to a single dark band was 0.507. There was a great amount of variability in the estimates for both band types; GC content for 16 light bands ranged from 0.434 to 0.610, and for 18 dark bands ranged from 0.397 to 0.594. These data suggest no difference in GC content between light bands and dark bands as two classes, but leave open the question of major differences in GC content among individual bands of either class. The data used here represent only a fraction of the information that will eventually be available. Improvements in the resolution of map locations that have been sequenced and acquisition of sequencing and mapping data for additional loci will enable researchers to elucidate such basic aspects of genome organization.

Table 3. Individual bands with greater than 10% mapping completeness and the official symbols for selected genes found in each band.

Band	Completion		Representative loci
	Mapping (%)	Sequencing (%)	
14q32.33	18.3	3.5	IGH@*
12p13.2	15.39	1.74	PRB1, PRB2, PRB3, PRB4
Xp22.32	14.98	0.99	PABX, STS, XG, KAL, CDPX
6p21.3†	12.88	3.1	HLA@*, CYP21, HFE, C2, C4A, TNFA, TNFB,
17q21.32	12.18	2.22	GP2B, GP3A
22q12.2	10.99	0.12	FRA22B
11p15.5	11.68	2.64	HBB@*, HRAS, MAFD1, INS
Xq27.2	10.3	0.08	FRAXD

*@ represents unofficial symbols (as of HGM10) used for gene clusters. †The three sub-bands of 6p21.3 share over 70 markers, none of which have been mapped to individual sub-bands.

Discussion and Conclusion

The methodology developed for this analysis has produced a multifaceted overview of current mapping activity and provides a preliminary evaluation of current progress toward mapping the human genome. Through this exercise, some of the limitations of the methodology and the barriers to presenting a comprehensive overview inherent in both the data and the manner in which it is recorded have become apparent.

Pertinent data for the human genome project come from a diverse spectrum of disciplines, many of which have immediate goals only incidentally related to gene mapping (for example, medicine, biochemistry, physical anthropology, and molecular biology). The compilation and coordination of data from these various disciplines is daunting, and suffers from the lack of common channels of communication for the relevant data items. Two major barriers to the expedient integration of data are the lack of a flexible, universal

nomenclature, and the variety of mapping resolutions associated with these data items. Even though a rigorous nomenclature exists for genes and mapped segments of DNA (*59–61*), this nomenclature has only limited usage outside the human gene mapping community and its scope is still too limited to accommodate more general "loci," such as contigs, YACs, and other objects arising in physical mapping. In addition, official names and symbols are often assigned only after initial publication under very different names. A more generalized system is needed to allow the broader scientific community easy access to all information on a locus when that information is published in diverse journals under a multiplicity of names. Such a system would facilitate identification of links between different data, such as DNA sequence and map location, which until now have been largely established retrospectively through a laborious manual evaluation of data.

Variability in the level of resolution of a data item is a very common problem. For example, there are currently several entries for amylase-2 DNA sequences in GenBank; however, the official HGM10 nomenclature recognizes two specific loci, AMY2A and AMY2B. This situation would be partially solved by recognizing these two adjacent genes as a cluster with a single location in the genome, thereby allowing a link between the two types of data. Problems also arise in relating other locus attributes, such as polymorphism, to the individual members of gene clusters. The hierarchical nomenclature needed to resolve such issues for gene clusters is also needed for contigs and their component clones.

To date, efforts to compare maps, even when the links have been clear, have been limited to individual chromosomes or chromosomal regions. Under the rationale that all maps should share the same relative order of loci, we have used locus order as a means for refining cytogenetic map locations across the entire genome. If used indiscriminately, this method has the potential to propagate errors from one incorrectly but seemingly precisely positioned locus through the nearby loci. However, as larger numbers of loci are localized with increasing precision, these errors will be detectable as differences in locus order. In an application to one set of linkage maps involving 578 ordered loci (*18*), only 15 inconsistencies between linkage order and cytogenetic order were found, suggesting that most of the regional localizations are accurate and consistent with order derived from linkage. The inconsistencies clearly indicate loci and regions requiring additional evaluation. One of the inconsistencies detected occurred in the middle of the short arm of chromosome 1; it had previously been noted (*62*). There will inevitably be conflict between maps constructed with different techniques, or even maps constructed independently with the same technique. Clearly, continuous coordination of the different mapping results is one of the best defenses against mapping errors.

We cannot construct a comprehensive, unified, linkage map at this time. The various published linkage maps for a chromosome often differed considerably in the estimates of distance between loci; in some cases they differed in the order of loci and the sets of loci used in different maps often had minimal overlap. It is not our objective to review these differences in detail, but two points can be made quite strongly. First, different methods of analysis of the same data can give quite different map distances. One example is the map of chromosome 4 where one map (*26*) based on multipoint mapping methods of primarily CEPH data is almost twice (1.94 times) the length of another map (*18*) based on a different method of analyzing essentially the same data. Without discussing the merits of one analytic approach over the other, it is clear that human map distances are estimates that vary depending on analytic assumptions. The second point is that distance estimates also vary among different data sets because of the inherent sampling error. Compared to the numbers of meiotic products sampled to assemble maps in such experimental organisms as *Drosophila melanogaster*, human linkage maps are based on very small numbers, and the distance estimates have a correspondingly large statistical uncertainty.

Our use of the ISCN as a coordinate system, without corrections for nonuniform distribution of genes and DNA between chromosome bands, constrains the interpretation of our analyses. Band lengths measured from the ISCN (1985) were used because a better coordinate system is not yet available. A molecularly based coordinate system could provide the basis for the organization of data from the different mapping activities. For example, a map based on physical reference points uniformly spaced every 7.5 Mb would allow generation of histograms with roughly the same resolution as those in Fig. 1. One goal of physical mapping is an STS map with intervals approximately 0.1 Mb in length. Even the smallest cytogenetic band is several times larger than this and the average band contains close to 4 Mb; thus, there is a sizable difference in scale between the best molecular resolution represented by cytogenetic coordinates and the objective of an STS map. The estimation and allocation algorithms, and the completeness calculations presented above, could all be cast in terms of intervals defined in an STS map.

The enormity of the task at hand becomes apparent when considering the types of mapping information not included in our analyses, yet considered to be crucial to the broader scientific goals of the human genome project. Our analyses have not incorporated information from interspecies comparative maps, nor have we presented (other than as numbers of identified polymorphic loci) any measure of the kinds of and levels of normal variation that are expected in human nuclear DNA sequence. Although estimates for the human genome are that 0.3 to 0.5% of base pairs are polymorphic, no studies representative of the genome as a whole have yet been undertaken (*63*). Thus, no global estimate is really justified and region specific estimates need to be empirical.

Ultimately we would like to know the informational content of the genome, and the order and relationship (physical and functional) among the various regions. The Human Genome Project can be seen as providing the necessary underpinning for this ultimate goal. The impact of these data should go beyond the immediate concerns of the genome mapping community, allowing scientists in the various fields that have contributed data to answer questions relevant to their own disciplines. This will only be possible if the data are accessible in a form that can readily be utilized by all researchers. With the increase in data generated by mapping efforts, our concept of genome organization will change, with ramifications for a common coordinate system, for nomenclature, and for database management. These issues deserve careful consideration, since the way we organize and record data will limit the questions we can pose.

REFERENCES AND NOTES

1. A. H. Sturtevant, *J. Exp. Zool.* **14**, 43 (1913).
2. S. J. O'Brien, Ed., *Genetic Maps: Locus Maps of Complex Genomes* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1990); J. H. Nadeau, *Linkage and Synteny Homologies Between Mouse and Man* (Jackson Laboratory, Bar Harbor, ME, 1990).
3. V. A. McKusick, *Mendelian Inheritance in Man: Catalogs of Autosomal Dominant, Autosomal Recessive, and X-Linked Phenotypes* (Johns Hopkins Univ. Press, Baltimore, MD, 1990).
4. U.S. Department of Health and Human Services and U.S. Department of Energy, *The U.S. Genome Project: The First Five Years FY 1991–1995* (National Technical Information Service, Springfield, VA, 1990).
5. H. Li *et al.*, *Nature* **335**, 414 (1988); A. J. Jeffreys *et al.*, *Cell* **60**, 473 (1990); G. Ruano *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 6296 (1990).
6. J. Ott, *Analysis of Human Genetic Linkage* (Johns Hopkins Univ. Press, Baltimore, MD, 1985).
7. Recombination between any two loci has an upper limit of 50%—independent inheritance—but map distances assembled by adding smaller intervals can reach more than 200 cM for larger chromosomes. See (*6*) for a discussion of the several mapping functions that relate recombination fraction to map distance.
8. N. E. Morton *et al.*, *Hum. Genet.* **62**, 266 (1982).
9. F. H. Ruddle and R. P. Creagan, *Annu. Rev. Genet.* **9**, 407 (1975); S. J. Goss and

H. Harris, *Nature* **255**, 680 (1975); P. Lichter *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 9664 (1988).
10. M. Olson, L. Hood, C. Cantor, D. Botstein, *Science* **245**, 1434 (1989).
11. D. C. Schwartz and C. R. Cantor, *Cell* **37**, 67 (1984).
12. D. T. Burke, G. F. Carle, M. V. Olson, *Science* **236**, 806 (1987).
13. National Research Council, *Mapping and Sequencing the Human Genome* (National Academy Press, Washington, DC, 1988).
14. J. Wu *et al.*, *Am. J. Hum. Genet.* **46**, 624 (1990); P. O'Connell *et al.*, *Genomics* **1**, 93 (1987); see also (*19*), (*20*), and (*33*).
15. The wall chart enclosed with this issue of *Science*, titled "The Human Genome Map 1990" is a graphical representation of many of the ideas presented in this article. Additional copies are available directly from *Science*.
16. More detailed information about the contents of the HGML can be found in the following papers: J. C. Stephens *et al.*, *The Human Gene Mapping Library: Present Status and Future Directions*, in *Computers and DNA*, G. I. Bell and T. G. Marr, Eds. (Addison-Wesley, New York, 1990), p. 69; R. K. Track *et al.*, *Banbury Report 32: DNA Technology and Forensic Science* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1989), p. 335; J. C. Stephens *et al.*, *The Human Gene Mapping Library: A Resource for Studies of Molecular Evolution, Population Genetics, and Comparative Mapping*, in *Molecular Evolution: UCLA Symposia on Molecular and Cellular Biology*, M. T. Clegg and S. J. O'Brien, Eds. (Wiley, New York, 1990), p. 273. The HGML closed permanently on 29 August 1990. The data accumulated in the HGML are available to nonprofit organizations so that they will not be lost to the scientific community. Most specifically, the HGML data became the initial contents of the new Genome DataBase at Johns Hopkins University. Requests for an electronic copy of the HGML database (more than 50 megabytes) should be directed to K. K. Kidd, Department of Human Genetics, Yale University Medical School, 333 Cedar Street, New Haven, CT 06510.
17. C. Burks *et al.*, *Methods Enzymol.* **183**, 3 (1990).
18. B. Keats *et al.*, *Cytogenet. Cell Genet.* **51**, 459 (1989).
19. R. White *et al.*, in *Genetic Maps: Locus Maps of Complex Genomes*, S. J. O'Brien, Ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1990), p. 5.134.
20. H. Donis-Keller and C. Helms, in *ibid.*, p. 5.158.
21. C. Dracopoli *et al.*, *Am. J. Hum. Genet.* **43**, 462 (1988).
22. P. O'Connell *et al.*, *Genomics* **4**, 12 (1989).
23. S. Povey and C. T. Falk, *Cytogenet. Cell Genet.* **51**, 91 (1989).
24. P. O'Connell *et al.*, *Genomics* **5**, 738 (1989).
25. A. J. Pakstis *et al.*, *Cytogenet. Cell Genet.* **51**, 1057 (1989).
26. D. R. Cox *et al.*, *ibid.*, p. 121.
27. L. A. Giuffra *et al.*, *ibid.* **49**, 313 (1988).
28. L. A. Giuffra *et al.*, *ibid.* **51**, 1004 (1989).
29. J. L. Kennedy *et al.*, *Schizophr. Bull.* **15**, 383 (1989).
30. H. Blanche *et al.*, *Cytogenet. Cell Genet.* **51**, 963 (1989).
31. G. M. Lathrop *et al.*, *Genomics* **5**, 866 (1989).
32. S. Chamberlain *et al.*, *Cytogenet. Cell Genet.* **51**, 975 (1989).
33. R. L. White *et al.*, *Genomics* **6**, 393 (1990).
34. J. Wu *et al.*, *Genomics*, in press.
35. P. Charmley *et al.*, *ibid.* **6**, 316 (1990).
36. J. L. Haines *et al.*, *Genet. Epidemiol.* **5**, 375 (1988).
37. A. Bowcock *et al.*, *Cytogenet. Cell Genet.* **51**, 966 (1989).
38. Y. Nakamura *et al.*, *Genomics* **3**, 342 (1988).
39. Y. Nakamura *et al.*, *ibid.* **2**, 302 (1988).
40. E. C. Wright *et al.*, *ibid.* **7**, 103 (1990).
41. Y. Nakamura *et al.*, *ibid.* **3**, 67 (1988).
42. M. L. Summar *et al.*, *Mol. Endocrinol.* **4**, 947 (1990).
43. R. E. Tanzi *et al.*, *Genomics* **3**, 129 (1988).
44. A. C. Warren *et al.*, *ibid.* **4**, 579 (1989).
45. G. A. Rouleau *et al.*, *ibid.*, p. 1.
46. F. Rouyer *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* **51**, 221 (1986).
47. D. C. Page *et al.*, *Genomics* **1**, 243 (1987).
48. J. C. Stephens *et al.*, *Cytogenet. Cell Genet.* **51**, 1085 (1989); J. C. Stephens *et al.*, in preparation.
49. K. K. Kidd, *J. Psychiatr. Res.* **21**, 551 (1987).
50. J. Dausset *et al.*, *Genomics* **6**, 575 (1990).
51. A complete list of the linkage maps surveyed is available from the authors.
52. I. H. Cohen *et al.*, in *Genetic Maps: Locus Maps of Complex Genomes*, S. J. O'Brien, Ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1990).
53. P. Lichter *et al.*, *Science* **247**, 64 (1990).
54. D. G. Harnden and H. P. Klinger, Eds., *An International System for Human Chromosome Nomenclature (1985): Report of the Standing Committee on Human Cytogenetic Nomenclature* (Karger, New York, 1985).
55. The data used to prepare the illustrative linkage maps represent only a portion of the maps used in refinement. Presentation of the details of the refinement procedure and resulting refinements for these loci are beyond the scope of this review, and are being prepared for publication elsewhere.
56. Strictly speaking, the cytogenetic interval corresponding to the linkage interval is defined by the refined top limit of the upper locus and the refined bottom limit of the lower locus. Although the refinement procedure is not a prerequisite for the allocation of polymorphic loci, our goal was to provide the best possible estimates for the linkage intervals.
57. G. Holmquist *et al.*, *Cell* **31**, 121 (1982).
58. G. Bernardi *et al.*, *Science* **228**, 953 (1985); G. Bernardi, *Annu. Rev. Genet.* **23**, 637 (1989).
59. T. Shows *et al.*, *Cytogenet. Cell Genet.* **46**, 11 (1987).
60. P. McAlpine *et al.*, *ibid.* **51**, 13 (1989).
61. K. Kidd *et al.*, *ibid.*, p. 622.
62. J. Kidd *et al.*, *Genomics* **6**, 89 (1990).
63. L. Cavalli-Sforza, *Am. J. Hum. Genet.* **46**, 649 (1990).
64. The authors thank L. Manuelidis for in-depth discussions of higher order chromosome structure, and P. Nadkarni for developing the sequence overlap detection algorithm applied to GenBank sequences. M. Seashore provided feedback on the representative genes for the wall chart. B. Jasny provided valuable editorial commentary. H. Cann engaged in many helpful discussions about the linkage maps. The efforts of P. Gilna, M. Skolnick, E. Hildebrand, and M. Conneally for their review of the wall chart, are greatly appreciated. We would like to extend a special thanks to the staff of the Yale-HHMI Human Gene Mapping Library: H. Chan, M. Chipperfield, I. Cohen, A. Ferrara, V. Ferriouolo, C. Gawron, S. Knoble, C. Partridge, F. Ricciuti, and R. Track; and to the HGML student assistants A. Boni, C. Boyajian, A. Dienard, and M. Simmons. Thanks also to W. Alles of the HGM10 workshop staff.

CORRECTION TO HUMAN GENOME MAP

Please note that the linkage map scale bar in the legend of the enclosed Human Genome Map (pages 262*a*–262*p*) should read 40 cM rather than 10 cM.