# Assessing Mapping Progress in the Human Genome Project

David R. Cox, Eric D. Green, Eric S. Lander, Daniel Cohen, Richard M. Myers

An important goal of the Human Genome Project in the United States is to construct a physical map of the human genome, consisting of unique genomic landmarks at an average spacing of 100 kilobases (kb). When completed (which is projected for the year 1998), the map will consist of 30,000 ordered sites, distributed relatively evenly throughout the genome. Such a map should provide the scientific community with an invaluable resource for the localization and isolation of any human DNA sequence of interest.

Although the goal is clear, no generally accepted measures have been adopted by either the scientific community or the funding agencies for assessing progress toward the completion of such a map. To some extent, this is because scientists around the world are using a diverse array of mapping methods—including meiotic (*1*), radiation hybrid (*2*), in situ hybridization (*3*), sequence-tagged site (STS)–content (*4*), and clone-based fingerprint (*5*) mapping techniques—that might not seem amenable to direct comparison. In important ways, however, the lack of universal standards for measuring mapping progress is hindering efficient completion of the map. Without some ongoing global assessment of progress, how can one distinguish those genomic regions requiring additional work and resources from those that are essentially complete? How can anyone determine which projects are proceeding in the most cost-effective manner?

We believe that universal measures of mapping progress are both desirable and possible. Although each mapping method has its own special characteristics, they share a fundamental similarity: each involves the ordering of unique sites in the genome. This common feature provides a basis for assessing mapping progress. Below, we outline four measures that should be applied to any mapping project to describe its progress.

D. R. Cox and R. M. Myers are in the Department of Genetics, Stanford University, Stanford, CA 94305, USA. E. D. Green is at the National Center for Human Genome Research, National Institutes of Health, Bethesda, MD 20892, USA. E. S. Lander is at the Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA, and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. D. Cohen is at the Centre d'Etude du Polymorphisme Humain, Paris 75010, France.

1) *Define the "objects" used to construct the map, calling special attention to those objects that can be shared among different maps.* Most genomic mapping projects involve ordering two classes of objects relative to one another (*6*): (i) breakpoints, so-called because they represent subdivisions of the genome, which are defined by a specific experimental resource and (ii) markers, consisting of unique sites in the genome that are intended to be independent of any particular experimental resource (Table 1).

Although both types of objects are essential for map construction, the map itself should be defined in terms of markers, especially those based on DNA sequence, rather than breakpoints. One reason for this is that markers are permanent and easily shared; they can be readily stored as DNA sequence information and distributed in this fashion (*7*). By contrast, breakpoints are defined by experimental resources that tend to be transient and cumbersome to distribute. The most important reason to use sequence-based markers is that they can be easily screened against any DNA source and thus can be used to integrate maps constructed by diverse methods and investigators. Such integration, based on common sets of markers, is crucial to the assembly and assessment of maps.

2) *Define the number of occupied "bins" in the map. Describe the observed distribution of the number of markers per bin and compare it to the distribution expected for a randomly spaced collection of markers.* The breakpoints in an experimental resource divide the genome into "bins," corresponding to the regions between consecutive breakpoints. Markers that cannot be distinguished by the experimental resource occupy the same bin. For example, in meiotic linkage mapping, markers that show no recombination within a given set of pedigrees occupy a single bin. In STS-content mapping, those STSs residing on the same subset of yeast artificial chromosomes (YACs) in a library reside in the same bin.

The size of a typical bin determines the maximal resolution of a given experimental resource. The number of occupied bins puts an upper bound on the number of ordered markers in the map, as markers in the same bin cannot be ordered with respect to one another. Although some investigators report only the total number of markers used to construct the map, it is the number of occupied bins that provides the measure most relevant to mapping progress.

The distribution of the number of markers per bin provides information about whether the markers are evenly, or at least randomly, spread throughout the genome (which is the desired goal) or are clustered. For any given mapping method, the bins tend to be of roughly equal size. In this case, one can compare the observed distribution of the number of markers per bin with the expected distribution for a collection of randomly-spaced markers (*8*).

3) *Identify those occupied bins that are ordered relative to one another and provide an estimate of the confidence in the ordering.* A key measure of mapping progress is the number of occupied bins that have been ordered relative to one another. The assignment of markers to bins can proceed at a steady pace throughout a mapping project. By contrast, the ordering of bins is only possible as a project matures and is a good indication of the degree of completion.

The analysis of mapping data typically begins with the identification of "clusters" of nearby occupied bins. For example, in linkage mapping and radiation hybrid mapping, the clusters are called "linkage groups" and consist of a set of markers connected by pairwise lod scores exceeding a given threshold. For STS-content mapping, the clusters are called "contigs" and consist of sets of STSs connected to one another through their presence on a common clone. (With YAC clones, special care is needed because of the high frequency of chimeric inserts; it may be wise to believe

**Table 1.** Categorization of map objects.

| Mapping method | Experimental resource | Breakpoints | Markers |
|---|---|---|---|
| Meiotic | Pedigrees | Recombination sites | DNA polymorphisms |
| Radiation hybrid | Hybrid cell lines | Radiation-induced chromosome breaks | STSs |
| In situ hybridization | Chromosomes | Cytological landmarks | DNA probes |
| STS content | Library of clones | End points of clones | STSs |
| Clone-based fingerprinting | Library of clones | End points of clones | Genomic restriction sites |

only those STS connections supported by at least two independent clones.) In each case, the identification of clusters is an inherently statistical process, with an uncertainty that should be specified.

Once clusters are identified, the investigator needs to (i) order the bins within each cluster, and (ii) order and orient the clusters relative to one another. The ordering of bins within a cluster again involves a statistical procedure for comparing alternative orders. For linkage mapping and radiation hybrid mapping, there are widely accepted procedures (involving likelihood ratios) for assessing the relative confidence in alternative orders. In contrast, there is currently no comparable approach for in situ hybridization mapping, clone-based fingerprint mapping, or STS-content mapping. We strongly encourage research to develop approaches for defining the confidence of bin order for these mapping methods.

In their initial stages, most mapping methods do not provide a direct way to order and orient clusters along the chromosome (the exception being in situ hybridization mapping). One solution is to wait until all bins on a chromosome fall into a single cluster. This is practical for mapping methods capable of detecting clustering over relatively large distances (as for linkage mapping, which can readily detect linkage over 5 to 10% of a chromosome), but not for methods in which only local adjacency can be detected (as in clone-based fingerprint mapping). For the latter methods, the best solution for ordering and orienting clusters is to use previously ordered markers to provide a "scaffold" on which to hang the new map. Two clusters can be ordered with respect to one another if they each contain such a scaffold marker; a cluster can be oriented on the chromosome if it contains two ordered scaffold markers. For maps built in this way, investigators should always distinguish between the order information provided by the preexisting scaffold and the order information actually provided by the new map.

In summary, the order information in a map should be described in terms of (i) the criteria used to define clusters; (ii) the number of clusters obtained; (iii) the number of uniquely ordered occupied bins within each cluster, together with a measure of the confidence in the ordering; (iv) the number of uniquely ordered and oriented clusters within the map; and (v) the order information imported from a preexisting scaffold versus that provided by the mapping data itself.

4) *Estimate the distance between adjacent ordered bins, and provide the basis for the estimates. Describe the observed distribution of distances between adjacent ordered bins and compare it to the distribution expected at random.* It is important to estimate the distance in kilobases between adjacent ordered markers in a map, although the precision of such estimates varies with the mapping method. Meiotic mapping, radiation hybrid mapping, and in situ hybridization mapping each involve estimating the distance between markers in a specific unit of measure (centimorgans for meiotic mapping, centiRays for radiation hybrid mapping, and fractional length for in situ hybridization mapping); these estimates often have a large standard deviation, especially for nearby markers (9). Crude estimates of physical distance can be obtained by assuming a constant relationship between the relevant unit of measure and physical distance in kilobases. For example, if a radiation hybrid map of 500 centiRays spanned a distance of 15,000 kb (as determined by an alternative mapping method), an investigator would derive a conversion factor of 30 kb per centiRay. The assumption of strict proportionality may not be correct in all cases, but the estimated distances nonetheless provide useful benchmarks.

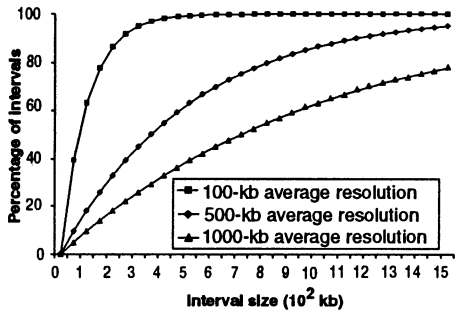


**Fig. 1.** Expected cumulative distribution $c^*(x)$ of interval sizes for maps consisting of randomly distributed markers with an average spacing of 100, 500, and 1000 kb.

For clone-based fingerprint mapping based on measuring all restriction fragments, one can directly estimate the length of each clone and the length of the overlap region. For other fingerprint mapping approaches (such as those that detect only those restriction fragments containing an Alu repeat) and for STS-content mapping, one can directly measure clone lengths and can indirectly estimate the extent of overlap (based on the proportion of restriction fragments or STSs shared in common). YAC clones, however, pose a special problem because the high frequency of both chimeric inserts and internal deletions means that measured clone length provides neither an upper nor lower bound on actual genomic distance. An alternative approach is to estimate distance based on information from other mapping methods and to assume that STSs are randomly distributed within this interval.

Having estimated the size of the intervals between adjacent markers, one should calculate the proportion $c_1(x)$ of intervals of size $\leq x$ kb and the proportion $c_2(x)$ of the genome residing within distance $x$ kb of a marker. The observed cumulative probability distributions can be compared to those expected for randomly distributed markers (*10*), to provide a measure of mapping progress. Figure 1 illustrates the expected cumulative distribution of interval sizes for various average marker spacings.

*Conclusion.* The promise of the Human Genome Project is that large-scale mapping efforts will result in a valuable, high-resolution map in a more timely and cost-effective fashion than otherwise possible. Well-defined, universally applied measures of mapping progress are essential to realizing this promise. Above, we have described a set of four measures that can be applied, regardless of the mapping method used. We believe that widespread acceptance of such a set of guidelines will accelerate progress toward completion of a map of the human genome with 100-kb average resolution.

## REFERENCES AND NOTES


1. G. Gyapay *et al.*, *Nature Genet.* **7**, 246 (1994).
2. D. R. Cox *et al.*, *Science* **250**, 245 (1990).
3. G. van den Engh, R. Sachs, B. J. Trask, *ibid.* **257**, 1410 (1992).
4. E. D. Green and M. V. Olson, *ibid.* **250**, 94 (1990).
5. E. Branscomb *et al.*, *Genomics* **8**, 315 (1990).
6. M. V. Olson and P. Green, *Cold Spring Harbor Symp. Quant. Biol.* **58**, 349 (1993).
7. M. V. Olson *et al.*, *Science* **245**, 1434 (1989). Clone-based fingerprint mapping is the one case in which the markers are not directly accessible for use or conversion into DNA sequence.
8. For any given mapping method, bin sizes tend to follow a common probability density function $f(x)$. Consider a map consisting of randomly spaced markers with an average spacing of $d$ kb. The expected proportion of bins containing $\leq k$ markers is

$$\int_0^\infty P(x/d, k)\, f(x)\, dx$$

where $P(a,k)$ is $Prob(X_a \leq k)$, with $X_a$ having a Poisson distribution with mean $a$, and $f(x)$ is the probability density function of bin lengths. For meiotic and radiation hybrid mapping, bin lengths are approximately exponentially distributed and the formula reduces to a simple geometric distribution; see, for example, W. F. Dietrich *et al.*, *Nature Genet.* **7**, 220 (1994).
9. For the distance estimate associated with crossing a single breakpoint in meiotic and radiation hybrid mapping, the standard deviation is approximately equal to the mean.
10. Consider a map consisting of randomly spaced markers with an average spacing of $d$ kb. The expected proportion $c_1^*(x)$ of intervals having size $\leq x$ kb is given by $c_1^*(x) = 1 - exp(-x/d)$. The expected proportion $c_2^*(x)$ of the genome residing within $x$ kb of a marker is $c_2^*(x) = 1 - exp(-2x/d)$. To measure progress toward a map with randomly spaced markers at 100 kb, one could compare the observed $c_1(x)$ and $c_2(x)$ to the expected distributions corresponding to $d$ = 100 kb.