Am J Hum Genet 34:842–845, 1982

How Many Polymorphic Genes Will It Take to Span the Human Genome?

KENNETH LANGE[1] AND MICHAEL BOEHNKE

SUMMARY

It is desirable to know how many polymorphic marker loci will be required so that every human genetic locus can be mapped by classical pedigree methods to a specific region of a specific chromosome. Assuming a total autosomal map length of 33 morgans, it would take only about $33/(2d)$ evenly spaced markers for every locus to be within d morgans of a marker. Taking into account that the markers will fall randomly along the genome, we show that a much larger number of such loci will have to be isolated and tested before the goal of a saturated gene map is reached.

Surely, one of the dreams of the human genetics community is to find a sufficient number of polymorphic marker loci so that any new genetic locus can be mapped by family studies to a specific region of a specific chromosome. It is of interest to establish how many marker loci it will take to span the entire genome.

A simple approach to this problem is to argue that, under the best circumstances, the marker loci should be evenly spaced along the genome. To avoid the complications of edge effects, one can imagine the 22 human autosomes placed end-to-end so as to cover the circumference of a circle. For simplicity, take the circumference to have unit length. If we want any point on the circle to be within a distance d from a marker, then the markers should be a distance $2d$ apart. Provided $(2d)^{-1}$ is an integer, it will take exactly $(2d)^{-1}$ markers to cover the circle. Thus, if we accept Renwick's [1] sex-averaged map length of 33 morgans for the

Received November 5, 1981; revised January 3, 1982.

This research was supported in part by the University of California, Los Angeles; Research Career Development Award K04-HD00307 from the National Institutes of Health; and National Research Service Award 07104 from the U.S. Public Health Service.

[1] Both authors: Department of Biomathematics, University of California, Los Angeles, CA 90024.

autosomal portion of the genome, and we want every gene to be within .1 morgan of a marker, $d = .1/33$, and it takes 165 markers to span the genome.

This result is appealing because of its simplicity and because of the manageable number of markers it entails. If polymorphisms based on DNA restriction fragment lengths are as common as hoped [2–7], then it is just a matter of time until we have 165 markers. The problem with the simple argument is that it ignores the fact that the markers will occur randomly over the genome, with tight clusters and sizable gaps appearing. As a consequence, considerably more than $(2d)^{-1}$ markers will be necessary to guarantee a minimum distance d.

To get some idea of what proportion of the genome can be covered, consider a fixed point on the circle [8–10]. If there are n markers, each one will fall beyond the minimum distance from the unit with probability $1 - 2d$. Hence, all n will fall beyond the minimum distance with probability $(1 - 2d)^n$. For $n = (2d)^{-1}$, we therefore expect a proportion

$$(1 - 2d)^{\frac{1}{2d}} \simeq e^{-1} \\ = .368$$

of the genome to be uncovered. In general, for $n = k/2d$, we expect an approximate proportion e^{-k} of the genome to be uncovered. Thus, with $n = 114$ and $d = .1/33$, about half the genome is uncovered; with $n = 493$, only about 5% is uncovered. When one takes into account that there are 22 separate autosomes, the uncovered portion of the genome should be larger since the influence of a marker cannot extend beyond the autosome on which it falls. It is still possible to solve the coverage problem with the complication of edges. Using the formula in [11], the proportion of the genome uncovered is .381 for $d = .1/33$ and $n = (2d)^{-1}$ markers. As the minimum distance d gets smaller, the circle approximation naturally gets better.

Another way of approaching the covering problem is to ask how many markers n it will actually take to cover the circle so that no gap is greater than $2d$. n is now a random variable. Flatto and Konheim [12] have proved that the average value of n is

$$E(n) = 1 + \sum_{j=1}^{[(2d)^{-1}]} (-1)^{j-1} \frac{(1 - j2d)^{j-1}}{(j2d)^{j+1}},$$

where $[(2d)^{-1}]$ is the integer part of $(2d)^{-1}$. There is, in fact, a whole asymptotic theory for the moments and the limiting distribution of n as $d \to 0$ [13–18]. In table 1, we list the mean $\pm$ the standard deviation for n for selected values of d [14]. It is obvious that the simple argument provides a serious underestimate of the number of marker loci necessary to cover the genome.

To assess the magnitude of edge effects on $E(n)$, we also carried out a computer simulation of how markers fall on the genome. In this simulation, we assumed as in [11] that markers are distributed independently and uniformly with respect to

physical length along the chromosomes. A weighted average of the mitotic chromosome lengths reported at the Paris Conference [19] was used. We also assumed that map distance in morgans is proportional to physical distance. The results of the simulation are reported in table 1. Inspection of the table reveals that edge effects are not altogether trivial and only diminish the chances of complete coverage.

New techniques developed by Gusella et al. [20] permit the search for DNA restriction fragment length polymorphisms to be confined to one human chromosome at a time. These techniques exploit somatic cell hybrids between man and rodent; occasionally, the hybrid cells will have preferentially lost all but a single human chromosome or a fragment of a human chromosome. These techniques provide a much more economical approach to spanning the human genome with markers. Once a chromosome is spanned by markers, one can eliminate it and thus avoid many redundant markers.

The number n of markers required to span the human genome can then be expressed as

$$n = \sum_{i=1}^{22} n_i ,$$

where n_i is the number necessary for the ith autosome and where the n_i are now independent random variables. Cooke [14] gives appropriate formulas for the mean and variance of each n_i based on a probabilistic model of random points falling on a line segment rather than on a circle. The last column of table 1 displays our calculations applying these formulas. The results are clearly encouraging, and we anticipate even better results if chromosome fragments are systematically used.

In summary then, the original argument promotes too optimistic a view of how much work it will take to span the human genome with polymorphic markers. Adopting a divide and conquer strategy of proceeding chromosome-by-chromosome helps matters considerably. Still, human geneticists will need patience, persistence, luck, and a willingness to settle for a less than complete genetic map. Filling the last few gaps of the map will be a slow, frustrating enterprise.

TABLE 1

NO. MARKER LOCI REQUIRED TO SPAN THE HUMAN GENOME

Minimum distance to marker locus (morgans)*	Markers required for circular genome†	Markers required for 22 autosomes‡	Markers required for 22 isolated autosomes§
.1	1273 ± 238	1528 ± 349	766 ± 66
.2	570 ± 119	743 ± 182	330 ± 33
.3	354 ± 80	493 ± 131	201 ± 22
.4	252 ± 60	365 ± 103	141 ± 16

* To convert to d divide by 33 morgans.

† Expected nos. marker loci required to span a circular autosomal genome ± 1 SD.

‡ Expected nos. marker loci required to span the 22 autosomes ± 1 SD, arrived at by computer simulations of 1,000 trials each.

§ Expected nos. marker loci required to span the 22 isolated autosomes ± 1 SD.

ACKNOWLEDGMENTS

We would like to thank Richard Gatti, Steve Matthysse, and Anne Spence for their helpful suggestions and for pointing out some of the references. We also wish to thank Ray White and Mark Skolnick for calling to our attention similar work carried out by another group concurrently and independently of us. A paper by J. Williamson, C. Cannings, D. T. Bishop, and M. Skolnick is in preparation.

REFERENCES

1. Renwick JH: The mapping of human chromosomes. *Annu Rev Genet* 5:81–120, 1971
2. Botstein D, White RL, Skolnick M, Davis RW: Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331, 1980
3. Housman D, Gusella J: Use of recombinant DNA techniques for linkage studies in genetically based neurological disorders, in *Genetic Research Strategies for Psychobiology and Psychiatry*, edited by Gershon E, Matthysse S, Breakefield XO, Ciaranello RD, Pacific Grove, Calif., Boxwood Press, 1981, pp 17–24
4. White R: In search of DNA polymorphism in humans, in *Banbury Report 4: Cancer Incidence in Defined Populations*, edited by Cairns J, Lyon J, Skolnick M, Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory, 1980, pp 409–420
5. Wyman A, White R: A highly polymorphic locus in human DNA. *Proc Natl Acad Sci USA* 77:6754–6758, 1980
6. Kan YW, Dozy AM: Polymorphism of DNA sequence adjacent to human β-globin structural gene: relationship to sickle mutation. *Proc Natl Acad Sci USA* 75:5631–5635, 1978
7. Jeffreys AJ: DNA sequence variants in the $^{G}\gamma$-, $^{A}\gamma$-, δ- and β-globin genes of man. *Cell* 18:1–10, 1979
8. Stevens WL: Solution to a geometrical problem in probability. *Ann Eugen (Lond)* 9:315–320, 1939
9. Feller W: *An Introduction to Probability Theory and Its Applications*, vol 2, 2nd ed. New York, John Wiley, 1971
10. Solomon H: *Geometric Probability*. Philadelphia, Society for Industrial and Applied Mathematics, 1978
11. Elston RC, Lange K: The prior probability of autosomal linkage. *Ann Hum Genet* 38:341–350, 1975
12. Flatto L, Konheim AG: The random division of an interval and the random covering of a circle. *Siam Rev* 4:211–222, 1962
13. Flatto L: A limit theorem for random coverings of a circle. *Isr J Math* 15:167–184, 1973
14. Cooke PJ: Bounds for coverage probabilities with applications to sequential coverage problems. *J Appl Prob* 11:281–293, 1974
15. Edens E: Random covering of a circle. *Indag Math* 37:373–384, 1975
16. Flatto L, Newman DJ: Random coverings. *Acta Math* 138:241–264, 1977
17. Kaplan H: Two applications of a Poisson approximation for dependent events. *Ann Prob* 5:787–794, 1977
18. Holst L: On multiple covering of a circle with random arcs. *J Appl Prob* 17:284–290, 1980
19. Paris Conference: Standardization in human cytogenetics. *Birth Defects: Orig Art Ser* 7:7, 1971
20. Gusella JF, Keys C, Varsanyi-Breiner A, et al.: Isolation and localization of DNA segments from specific human chromosomes. *Proc Natl Acad Sci USA* 77:2829–2833, 1980