

# Mapping Heredity: Using Probabilistic Models and Algorithms to Map Genes and Genomes (Part II)

*Eric S. Lander*

## Physical Mapping

### Assembling Physical Maps by “Fingerprinting Random Clones”

Genetic mapping is only the first step toward positional cloning of a gene. Once a gene has been determined to lie between two genetic markers, the geneticist must produce a physical map—consisting of overlapping clones spanning the chromosomal region between the two flanking markers. Traditionally, physical maps have been produced by the process of *chromosomal walking*: one starts with clone  $C_1$  containing one of the genetic markers, uses  $C_1$  as a probe to find an overlapping clone  $C_2$ , uses  $C_2$  as a probe to find  $C_3$ , and so on until the region has been spanned (Figure 1). Chromosomal walking is an inherently serial procedure, and each step may take several weeks (due to the laboratory steps involved in making and using a probe).

This tedious process could be eliminated if one simply constructed a complete physical map of overlapping clones spanning the *entire* genome. The idea is more practical than it may seem at first glance. Whereas chromosomal walking proceeds serially, a physical map of an entire genome can be constructed in parallel. The idea is to describe each clone  $C$  by an easily determined *fingerprint*  $F(C)$ —which can be thought of as a set of “attributes” of  $C$ . If two clones have substantial overlap, their fingerprints should be similar. Conversely, if two clones have very similar fingerprints, they are likely to overlap. In principle, one should be able to construct a physical map by fingerprinting a large collection of clones and using computer analysis to compare the fingerprints and recognize the overlaps. The

choice of a fingerprinting method depends principally on laboratory considerations; certain types of clones are more amenable to certain types of analysis.

For very large genomes such as the human genome with three billion bp, it is necessary to work with large subclones of length  $> 100,000$  bp. For such large subclones, a fingerprinting strategy called STS content has gained favor in recent years. The method is based on sequence tagged sites (STS), which are very short unique sequences taken from the genome. The fingerprint of a clone is the list of STSs contained within it; the data form an incidence matrix of clones by STSs (Figure 2). Clones containing even a single unique STS in common should overlap. As an aside, the determination of which clones contain a given STS is typically made using a combinatorial pooling scheme that avoids having to test each STS against each clone (Green and Olson, [9]). Using this approach, Foote et al. [8] and Chumakov et al. [5] constructed the first

---

*Adapted with permission from “Calculating the Secrets of Life”. Copyright 1995 by the National Academy of Sciences. Courtesy of the National Academy Press, Washington, D.C. Part one of this article appeared in the July issue of Notices.*

*Eric S. Lander is a member of the Whitehead Institute for Biomedical Research, professor of biology at the Massachusetts Institute of Technology, and director of the Whitehead Institute/MIT Center for Genome Research.*

complete maps of human chromosomes (Y and 21, respectively).

There are two key mathematical issues pertinent to the construction of a physical map:

(1) Algorithms for map assembly. Given the fingerprinting data, what algorithm should be used for constructing a physical map? This question is closely related to graph theory: given information about adjacency among clones inferred from their fingerprints, one must reconstruct the underlying geometry of the physical map.

(2) Statistics of coverage. How many clones must be studied to yield a map covering virtually the entire genome? This question belongs to probability theory: assuming that subclones are distributed randomly across the genome, one needs to know the distribution of gaps—uncovered regions or undetected overlaps—in the map.

Mathematical analysis is thus essential to the design and execution of physical mapping projects (Arratia et al., [1]; Lander and Waterman, [21]). To illustrate this, we discuss the considerations involved in making a physical map of the entire human genome.

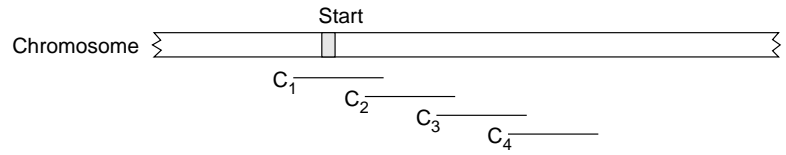
### Excursion: Designing a Strategy to Map the Human Genome

Under the auspices of the Human Genome Project, our laboratory is engaged in constructing complete physical maps of the mouse and human genomes, each about  $3 \times 10^9$  bp in length. The task is daunting—requiring analysis of tens of thousands of clones, each carrying extremely large DNA fragments.

Currently, the best clones for making a human physical map are *yeast artificial chromosomes* (YACs). A good YAC library might contain inserts of about 1,000,000 bp in length. Even with such large inserts, it would take 3,000 YACs to cover the human genome if they were laid end-to-end. Of course, clones taken from an actual library will be arrayed randomly, and so considerably more clones are required to ensure coverage.

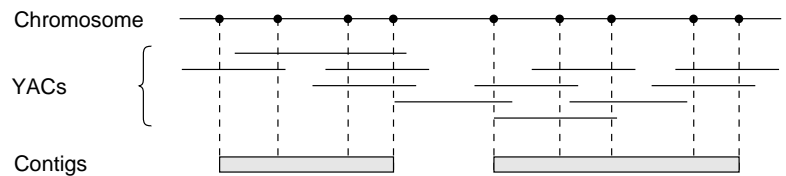
The best fingerprint for studying YACs is STS content. STSs are unique sites in the genome that can be easily assayed by polymerase chain reaction (PCR). STSs are created by sequencing short (200 bp) fragments. Each STS is screened simultaneously against the entire YAC library to identify the clones that contain it. Because STSs are screened in parallel, it is most efficient to work with a fixed YAC library and to test STSs sequentially.

For mathematical analysis of physical mapping, the YACs and STSs can be abstracted to a set  $I$  of intervals (which may vary in size) and a set  $P$  of points distributed randomly along a line segment. An interval is said to be anchored



**Figure 1**

Schematic diagram illustrating chromosome walking. One starts by isolating a clone  $C_1$  containing the initial starting point.  $C_1$  is then used as a probe to isolate overlapping clones, such as  $C_2$ . The process is iterated to obtain successive steps in the walk. Although at each step one isolates clones extending in either direction, only those clones extending the walk to the right are shown in the diagram.



**Figure 2**

Schematic diagram illustrating the principle of STS content mapping. Various unique points in the genome, called STSs, are tested against a collection of random large-insert clones, such as YACs, to determine which STSs are contained in which YACs. Based on the resulting adjacency matrix, one attempts to reconstruct the order of the STSs in the genome. “Contigs,” consisting of groups of STSs connected by YACs, are assembled based on the adjacency data. In the figure, the STSs can be grouped into two contigs.

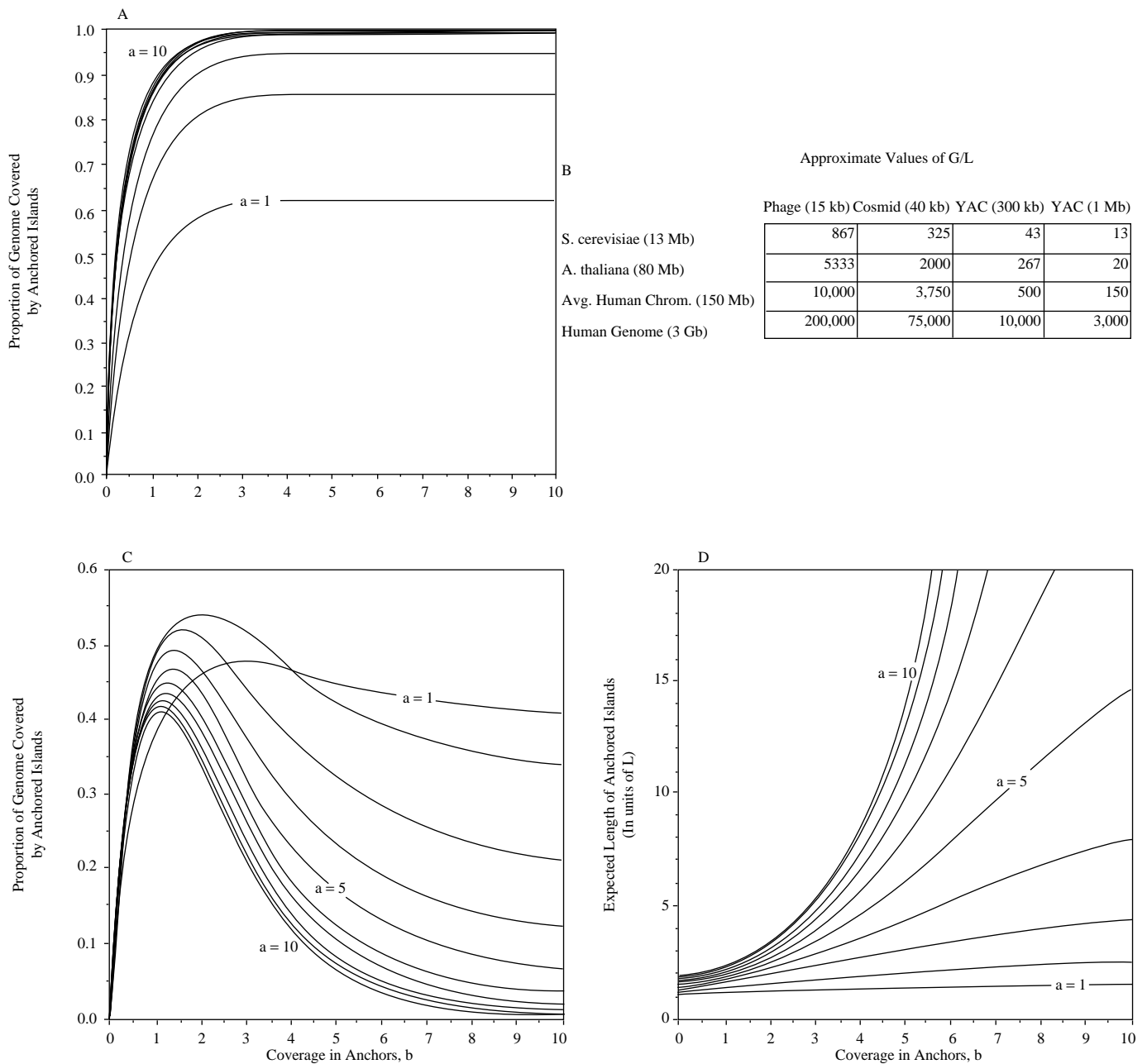
if it contains at least one point  $p \in P$ . Two anchored intervals  $I_1$  and  $I_2$  are said to be connected if there is a point  $p \in P$  contained in their intersection. Note that two intervals may overlap but fail to be connected. If we take the transitive closure of the connectivity relation, the resulting equivalence classes of anchored intervals are called *anchored contigs*.

The key question is: How many intervals and how many points should be analyzed to construct a reasonably complete physical map—that is, one in which the vast majority of the genome is contained in a modest number of large contigs? We define the following notation:

$G$ , the length of the genome in base pairs;

$L$ , the length of a random clone in base pairs, a random variable;

$\mathbf{L}$ , the expected length of a random clone,  $\mathbf{L} = E(L)$ ;



**Figure 3**

Expected coverage properties for STS content mapping, as a function of the coverage  $a$  in YACs and  $b$  in STSs. Calculations assume YACs of constant length  $L$  and a genome of length  $G$ . The graphs show (A) the expected proportion of the genome covered by anchored “contigs” (also called islands); (C) the expected number of anchored contigs, and (D) the expected length of an anchored contig. Graphs show the situation for  $a = 1, 2, \dots, 10$  (only the cases  $a = 1, 5, 10$  are explicitly marked). Results are expressed in units of  $G/L$ . Table B lists the value of  $G/L$  for certain representative genomes and cloning vectors, including two different sizes of YACs. Reprinted by permission from Arratia et al. (1991).

Copyright ©1991 by *Genomics*.

$N$ , the number of clones to be used;  
 $M$ , the number of STSs to be used;  
 $a = LN/G$ , the expected number of clones covering a random STS; and

$b = LM/G$ , the expected number of STSs contained in a random clone.

Clone lengths  $L$  will be assumed to be independent, identically distributed random variables, with the probability density function of the

normalized length  $l = L/L$  denoted by  $f(l)$  and the inverse cumulative distribution function denoted  $F(l) = \text{Prob}(l/L > x)$ . It is also useful to define the auxiliary function

$$J(x) = \exp \left\{ -a \int_x^\infty F(l) dl \right\},$$

which can be interpreted as the probability that two points separated by distance  $x$  are not covered by a common clone.

The problem belongs to the area of *coverage problems*, which treat processes of covering a space with random sets of a given sort. Often, mathematical authors focus on the goal of attaining *complete* coverage. Such results are not really appropriate from a biological standpoint—because they depend sensitively on the distribution of covering sets being absolutely random, an assumption that is biologically implausible. Instead, it is more sensible to focus on central behavior—that is, the goal of covering *most* of the space.

STS content mapping poses a slightly unusual coverage problem, because the definition of coverage involves joining together random intervals with random points. It is nonetheless possible to analyze many features of the stochastic process in order to derive many prescriptive results. Arratia and colleagues [1] proved the following result, which describes the basic coverage properties:

**Proposition:** With the notation as above,

(1) the expected number of anchored contigs is  $Np_1$ , where

$$p_1 = \int_0^\infty be^{-bu}J(u)F(u)du,$$

(2) the expected length of an anchored contig is  $\lambda E(L)$ , where

$$\lambda = \left\{ 1 + \int_0^\infty (b^2u - 2b)e^{-bu}J(u)du \right\} / ap_1,$$

(3) the expected proportion  $r_0$  of the genome not covered by anchored contigs is:

$$r_0 = \int_0^\infty \int_0^\infty b^2 e^{-b(u+v)} \frac{J(u)J(v)}{J(u+v)} dudv.$$

Figure 3, taken from Arratia et al. [1], plots these functions for the case of clones of constant size. From these graphs, experimentalists can plan their experimental approach. For our project with the human physical map, the typical clone size is about  $1 \times 10^6$  bp. Based on the tradeoffs between screening more YACs and using more STSs, we selected  $a = 6$  and  $b = 3$ —corresponding to about 18,000 YACs and about 9,000 STSs. This selection should ensure that about  $r_0 \approx 99$  percent of the genome is covered, with about 850 anchored contigs having average length of about 3.5 Mb.

Having explored the question of experimental design, it is worth briefly discussing the issues involved in data analysis. Although the process of STS content mapping may consume several person-years of laboratory work, the final data consist simply of a large ( $18,000 \times 9,000$ ) adjacency matrix  $A = (a_{ij})$ , with  $a_{ij} = 1$  or 0 in position  $i, j$  according to whether YAC<sub>*i*</sub> contains STS<sub>*j*</sub>. Based on this information, how

do we determine the correct order of the STSs in the genome?

In principle, a proposed order of the STSs is consistent with the observed data if and only if permuting the columns of the adjacency matrix  $A$  according to this order causes  $A$  to have the *consecutive ones property*—that is, in each row, the ones occur in a single consecutive block. This property follows from the fact that each YAC should consist of a single connected interval taken from the genome (Figure 2). The consecutive ones property has been extensively studied in computer science. Booth and Leuker [3] devised an elegant linear-time algorithm for solving the problem in a very strong sense: Given a (0,1)-matrix  $A$  with  $n$  rows and  $m$  nonzero entries, the algorithm needs running time of only  $O(m + n)$  to determine whether there is *any* column permutation causing the matrix to have the consecutive ones property and, if so, to produce a simple representation of *all* such column permutations.

In practice, there is a serious problem with this approach: it assumes that the data are absolutely error-free. Laboratory work is never flawless. Certainly it is not when the task involves filling in 162 million entries in an adjacency matrix. If even a few errors are present, the Booth-Leuker algorithm is almost certain to report that there is no consistent order! In fact, there are likely to be many errors, including

- false negatives: one may fail to identify some proportion of the YACs containing an STS;
- false positives: some proportion of the YACs detected as containing an STS may not actually do so; and
- chimeric YACs: some proportion of the YACs may not represent a single contiguous region, but two unrelated regions that have been joined together in a single clone.

Moreover, the occurrence of false negatives and positives may not be random but systematic (owing to deletions of clones or contamination of samples). In short, algorithms must be robust to errors in the data. Producing such algorithms is an interesting challenge that draws on methods from graph theory, operations research, and statistics. As of this writing, the best approach has not yet been determined.

In essence, mapping problems—like many problems in computational biology—involve indirect inference of the structure of a biological entity, such as a chromosome, based on whatever data can be effectively gathered in the laboratory. It is not surprising that mapping problems draw on statistics, probability, and combinatorics. Although the field of mapping dates nearly to the beginning of the twentieth century, the area remains rich with new challenges—because new laboratory methods con-

stantly push back the frontiers of the maps and features that can be mapped in DNA.

## References

- [1] R. ARRATIA, E. S. LANDER, S. TAVAR, and M. S. WATERMAN, *Genomic mapping by anchoring random clones: A mathematical analysis*, *Genomics* **11** (1991), 806–827.
- [2] W. F. BODMER, C. J. BAILEY, J. BODMER, H. J. R. BUSSEY, A. ELLIS, P. GORMAN, F. C. LUCIBELLO, V. A. MURDAY, S. H. RIDER, P. SCAMBLER, D. SHEER, E. SOLOMON, and N. K. SPURR, *Localization of the gene for familial adenomatous polyposis on chromosome 5*, *Nature* **328** (1987), 614–616.
- [3] K. S. BOOTH and G. S. LUEKER, *Testing the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms*, *J. Comput. System Sci.* **13** (1975), 335–379.
- [4] D. BOTSTEIN, R. L. WHITE, M. SKOLNICK, and R. W. DAVIS, *Construction of a genetic linkage map in man using restriction fragment length polymorphisms*, *Amer. J. Human Genetics* **32** (1980), 314–331.
- [5] I. CHUMAKOV, P. RIGAUT, S. GUILLOU, P. OUGEN, A. BILLAUT, G. GUASCONI, P. GERVY, I. LE GALL, P. SOULARUE, and L. GRINAS, *Continuum of overlapping clones spanning the entire human chromosome 21q*, *Nature* **359** (1992), 380–387.
- [6] W. DIETRICH, H. KATZ, S. E. LINCOLN, H.-S. SHIN, J. FRIEDMAN, N. C. DRACOPOLI, and E. S. LANDER, *A genetic map of the mouse suitable for typing intraspecific crosses*, *Genetics* **131** (1992), 423–447.
- [7] E. FEINGOLD, P. O. BROWN, and D. SIEGMUND, *Gaussian models for genetic linkage analysis using complete high resolution maps of identity by descent*, *Amer. J. Human Genetics* **53** (1993), 234–251.
- [8] S. FOOTE, D. VOLLRATH, A. HILTON, and D. C. PAGE, *The human Y chromosome: Overlapping DNA clones spanning the euchromatic region*, *Science* **258** (1992), 60–66.
- [9] E. GREEN and M. V. OLSON, *Systematic screening of yeast artificial chromosome libraries using the polymerase chain reaction*, *Proc. Nat. Acad. Sci. U.S.A.* **87** (1990), 1213–1217.
- [10] J. GRODEN, A. THLIVERIS, W. SAMOWITZ, M. CARLSON, L. GELBERT, H. ALBERTSEN, G. JOSLYN, J. STEVENS, L. SPIRIO, M. ROBERTSON, L. SERGEANT, K. KRAPCHO, E. WOLFF, R. BURT, J. P. HUGHES, J. WARRINGTON, J. MCPHERSON, J. WASMUTH, D. LE PASLIER, H. ABDERRAHIM, D. COHEN, M. LEPPERT, and R. WHITE, *Identification and characterization of the familial adenomatous polyposis coli gene*, *Cell* **66** (1991), 589–600.
- [11] K. W. KINZLER, M. C. NILBERT, L. K. SU, B. VOGELSTEIN, T. M. BRYAN, D. B. LEVY, K. J. SMITH, A. C. PREISINGER, P. HEDGE, D. MCKECHNIE, R. FINNIAR, A. MARKHAM, J. GROFFEN, M. S. BOGUSKI, S. F. ALTSCHUL, A. HORII, H. ANDO, Y. MIYOSHI, Y. MIKI, I. NISHISHO, and Y. NAKAMURA, *Identification of FAP locus genes from chromosome 5q21*, *Science* **253** (1991), 661–665.
- [12] A. KONG, *Efficient methods for computing linkage of recessive diseases in inbred pedigrees*, *Genetics and Epidemiology* **8** (1991), 81–103.
- [13] A. KONG, N. COX, M. FRIGGE, and M. IRWIN, *Sequential imputation for multipoint linkage analysis*, *Genetics and Epidemiology* **10** (1993), 483–488.
- [14] A. KONG, M. FRIGGE, N. COX, and W. H. WONG, *Linkage analysis with adjustments for covariates: A method combining peeling with Gibbs sampling*, *Cytogenetics and Cell Genetics* **59** (1992a), 208–210.
- [15] A. KONG, M. FRIGGE, M. IRWIN, and N. COX, *Importance sampling. I. Computing multimodel p values in linkage analysis*, *Amer. J. Human Genetics* **51** (1992b), 1413–1429.
- [16] E. S. LANDER and D. BOTSTEIN, *Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps*, *Genetics* **121** (1989), 185–199.
- [17] M. LEPPERT, M. DOBBS, P. SCAMBLER, P. O. CONNELL, Y. NAKAMURA, D. STAUFFER, S. WOODWARD, R. BURT, J. HUGHES, E. GARDNER, M. LATHROP, J. WASMUTH, J. M. LALOUEL, and R. WHITE, *The gene for familial polyposis coli maps to the long arm of chromosome 5*, *Science* **238** (1987), 1411–1413.
- [18] A. R. MOSER, H. C. PITOT, and W. F. DOVE, *A dominant mutation that predisposes to multiple intestinal neoplasia in the mouse*, *Science* **247** (1990), 322–324.
- [19] L. K. SU, K. W. KINZLER, B. VOGELSTEIN, A. C. PREISINGER, A. R. MOSER, C. LUONGO, K. A. GOULD, and W. F. DOVE, *A germline mutation of the murine homolog of the APC gene causes multiple intestinal neoplasia*, *Science* **256** (1992), 668–670.
- [20] E. THOMPSON and E. WIJSMAN, *The Gibbs sampler on extended pedigrees: Monte Carlo methods for the genetic analysis of complex traits*, Technical Report no. 193, Department of Statistics, Univ. of Washington, Seattle, 1990.
- [21] E. S. LANDER and M. S. WATERMAN, *Genomic mapping by fingerprinting random clones: A mathematical analysis*, *Genomics* **2**, 231–239, 1988.