

# Mapping Heredity: Using Probabilistic Models and Algorithms to Map Genes and Genomes

*Eric S. Lander*

**T**he human genome is a vast biochemical jungle in which scientists have begun hunting for the genetic basis of inherited diseases. Even a one-letter error in the  $3 \times 10^9$  base pairs (bp) of deoxyribonucleic acid (DNA) inherited from either parent may be sufficient to cause a disease. Thus, to detect inherited diseases, one must be able to detect mistakes present at just over one part in  $10^{10}$ . The task is sometimes likened to finding a needle in a haystack, but this analogy actually understates the problem: the typical 2-gram needle in a 6,000-kilogram haystack represents a 3,000-fold larger target. In certain respects, the gene hunter's task is harder still, because it may be difficult to recognize the target even if one stumbles upon it. Although molecular biologists refer to the human genome as if it were well defined in mathematicians' terms, it is recognized that, except

---

*Adapted with permission from "Calculating the Secrets of Life". Copyright 1995 by the National Academy of Sciences. Courtesy of National Academy Press, Washington, D.C. Part II of this article will appear in the August 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.*

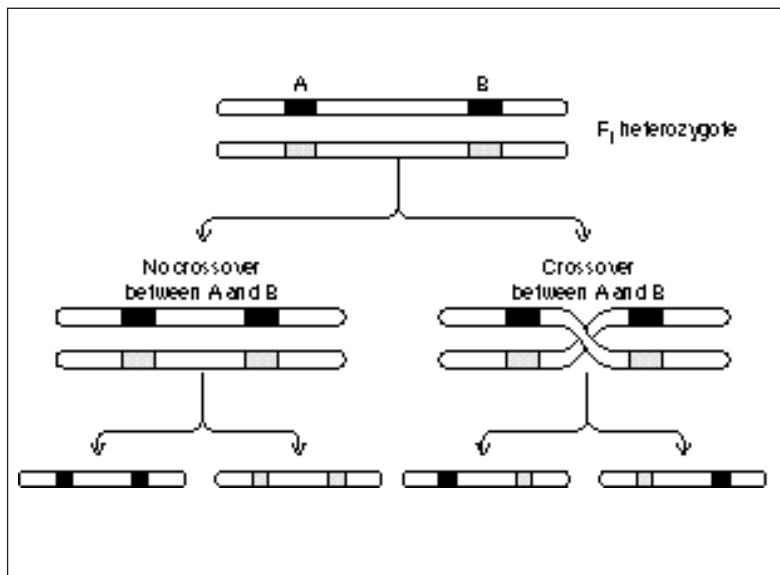
for identical twins, no two humans have identical DNA sequences. Two genomes chosen from the human population are about 99.8 percent identical, affirming our common heritage as a species. But the 0.2 percent variation translates into some six million sequence differences. Common sites of sequence variations are called *DNA polymorphisms*. Most polymorphisms are thought to be nonfunctional variations, arising by mutation, having no deleterious consequence, and increasing (and decreasing) in frequency by stochastic drift. The presence of considerable DNA polymorphism in the population has sobering consequences for disease hunting. Even if it were straightforward to determine the entire DNA sequence of individuals, one could not find the gene for cystic fibrosis (CF) simply by comparing the sequences of a CF patient and an unaffected person: there would be too many polymorphisms.

How then does a geneticist find the genes responsible for cystic fibrosis, diabetes, or heart disease? The answer is to proceed hierarchi-

---

*How does a geneticist find the genes responsible for cystic fibrosis, diabetes, or heart disease?*

---



**Figure 1**  
Schematic drawing of genetic recombination in an  $F_1$  heterozygote with distinct alleles at two loci (marked as A and B) on a chromosome. When no recombination occurs between A and B in meiosis, chromosomes carrying the original pair of alleles result. When recombination occurs, the resulting chromosomes carry a new combination of alleles.

cally. The first step is to use a technique called *genetic mapping* to narrow down the location of the gene to about 1/1,000 of the human genome. The second step is to use a technique called *physical mapping* to clone the DNA from this region and to use molecular biological tools to identify all the genes. The third step is to identify candidate genes (based on the pattern of gene expression in different tissues and at different times) and to look for functional sequence differences in DNA (for example, mutations that introduce stop codons or that change crucial amino acids in a protein sequence) in affected patients. We focus on genetic and physical mapping because they essentially involve mathematical analysis.

## Genetic Mapping

### The Concept of Genetic Maps

Genetic mapping is based on the perhaps counterintuitive notion that it is possible to find where a gene is without knowing what it is. Specifically, it is possible to identify the location of an unknown disease-causing gene by correlating the inheritance pattern of the disease in families with the inheritance pattern of known genetic markers. It is useful to return to Mendel's Laws of Inheritance:

- *First Law.* For any gene, each parent transmits one allele chosen at random to its offspring.

- *Second Law.* For any two genes, the alleles transmitted by a parent are independent (that is, there is no correlation in the alleles transmitted).

Although Mendel's First Law has held up well over the past 130 years, the Second Law turned out to be false in general. Two genes on different chromosomes show no correlation in their inheritance pattern, but genes on the same chromosome typically show correlation.

Consider the simple backcross in Figure 1, showing the inheritance of two genes A and B on the same chromosome. The  $F_1$  individual carries one chromosome with alleles  $a_1$  and  $b_1$  at the two genes and another chromosome with alleles  $a_2$  and  $b_2$ . Often, one or the other chromosome is transmitted completely intact to the offspring. If this always happened, the inheritance pattern at the two genes would be completely dependent:  $a_1$  would always be co-inherited with  $b_1$ . But the situation is more interesting. *Crossing over* can occur at random points along the chromosomes, involving an even swap of DNA material. If a crossover occurs between genes A and B, it results in *recombination* between the genes, producing a chromosome carrying a new combination of alleles:  $a_1b_2$  or  $a_2b_1$ . In fact, multiple crossovers can occur along a chromosome; recombination between two loci will result whenever an odd number of crossovers occur.

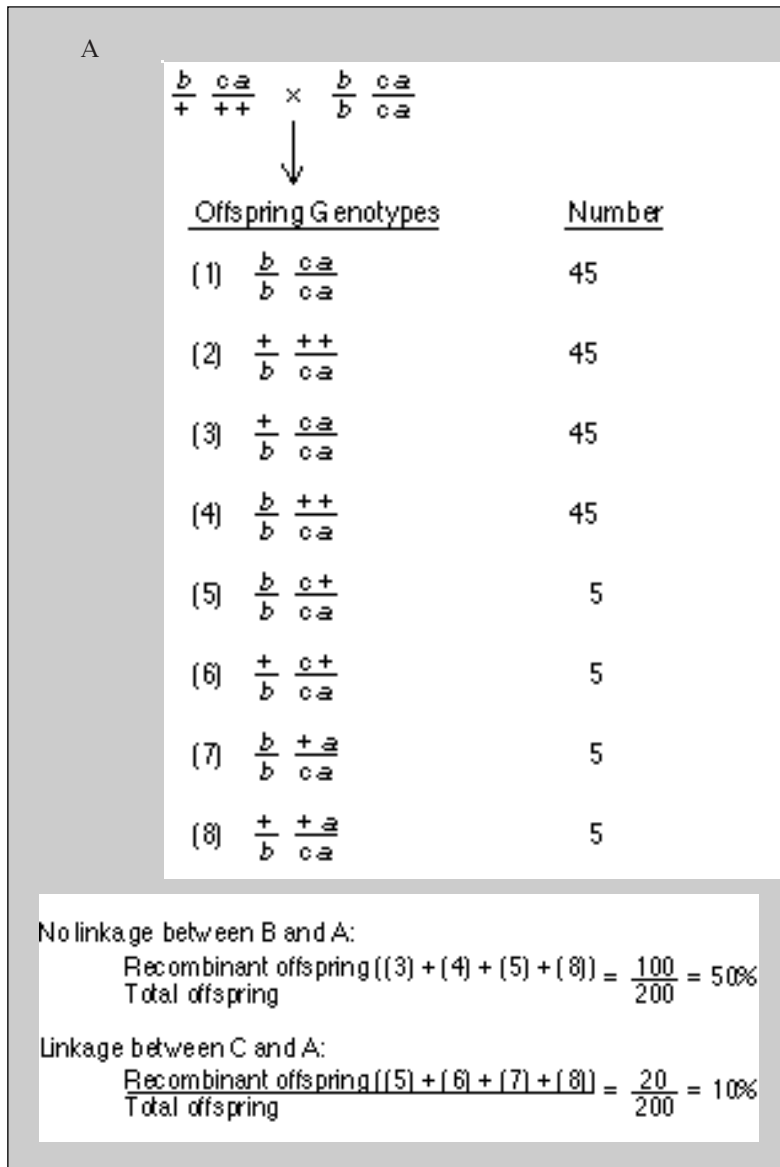
Genetic mapping is based on the recognition that the *recombination frequency*  $\theta$  between two genes (or loci) provides a measure of the distance between them. If two genes are close together,  $\theta$  will be small. If the recombination frequency is clearly less than 0.50, the genes are said to be *linked*.

The *genetic distance*  $d_{A,B}$  between two genes A and B is defined as the expected number of crossovers between the genes. If one assumes that crossovers are distributed independently with respect to one another (this assumption is not quite right but is adequate for many purposes), genetic distances can easily be converted into recombination frequency, for the number of crossovers between genes A and B will then be Poisson distributed with mean  $d = d_{A,B}$ ; and so the probability of an odd number of crossovers can be shown to be

$$\theta = (1 - e^{-2d}) / 2.$$

For small distances, the formula is  $\theta \approx d$ , which reflects the fact that the possibility of more than one crossover can be neglected. For large distances  $d$ , the recombination frequency  $\theta$  approaches 0.50, that is, independent assortment.

Genetic mapping is an essential first step in characterizing a new mutation. Consider first the

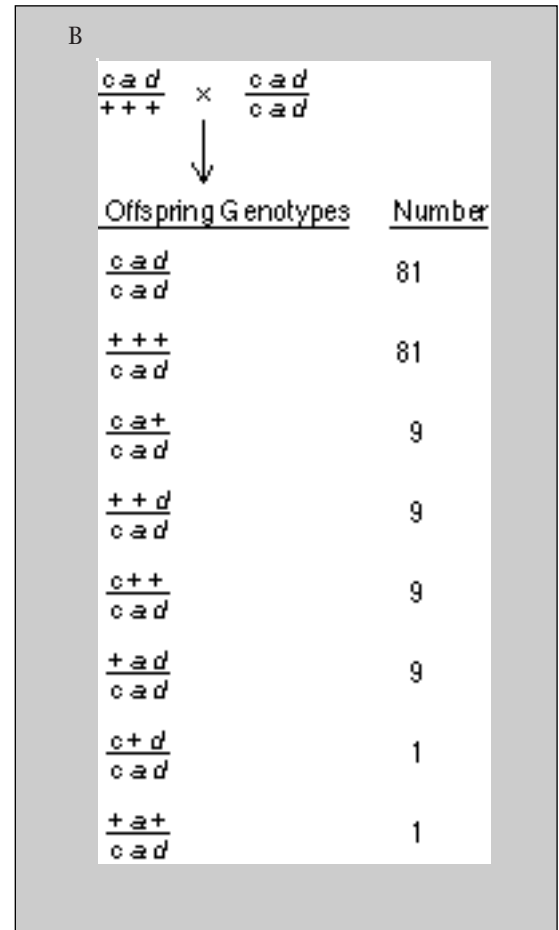


**Figure 2a**  
 Examples of three-point crosses. (A) Locus A is unlinked to locus B but is linked to locus C at a recombination fraction of 10 percent.

situation of (1) a laboratory organism in which experimental matings can be set up at will and (2) traits that are monogenic and fully penetrant (that is, the phenotype is completely determined by the genotype at a single gene). For example, a *Drosophila* geneticist might find a dominantly acting mutation at a locus X, causing flies to have an extra set of wings (in fact, such mutations exist). He would set up crosses with strains carrying different *genetic markers* (that is, variants in other genes of known location) in order to find the regions showing correlated inheritance. Figure 2a shows the result of a backcross of this type. The gene A is clearly not linked to locus B

but is tightly linked to locus C. The proportion of recombinant chromosomes provides a straightforward statistical estimator of the recombination frequency. In this case, the recombination frequency between A and B is about  $20/200 = 10$  percent. The gene A can be positioned more precisely by using a three-point cross shown in Figure 2b, in which two nearby genetic markers are segregating. Here, it is clear that A maps about midway between genes C and D (see figure caption).

For experimental organisms and simple traits, genetic mapping provides a straightforward way to locate the trait-causing gene to a small inter-



**Figure 2b**  
 (B) Locus A is located between loci C and D, at about 10 percent recombination fraction from each. The first two types of progeny involve chromosomes with no recombination; the next four involve a single recombination, and the last two involve double recombination (between C-A and A-D). The double recombination class is always least frequent, a property that allows one to determine the order of three linked loci from a cross in which they are all segregating.

val. *Drosophila* geneticists rarely need to appeal to statistical or mathematical concepts. For geneticists studying human families or complex traits, however, the situation is quite different.

### Challenges of Genetic Mapping: Human Families and Complex Traits

Medical geneticists studying diseases face two major problems: (1) for human diseases, one cannot arrange matings at will but must rather retrospectively interpret existing families; and (2) for both human diseases and animal models of these diseases, the trait may not be simply related to the genotype at a single gene. Owing to these complications, genetic mapping of disease genes often requires sophisticated mathematical analysis.

The first problem is the inability to arrange matings. To offset this limitation, human geneticists need to have a huge collection of frequent, naturally occurring genetic markers so that the inheritance pattern of each chromosomal region can be followed just as if one had deliberately set up a cross incorporating specific genetic markers. In 1980 David Botstein set off a revolution by recognizing that the naturally occurring DNA polymorphisms in the human population filled the need [4]. By 1994, over 4,000 DNA polymorphisms had been identified and mapped relative to one another.

Even with a dense genetic map of DNA polymorphisms, human genetic mapping confronts several special problems of incomplete information: (1) For individuals homozygous ( $a_1/a_1$ ) at a gene, one cannot distinguish between the two homologous chromosomes at this location. (2) For individuals heterozygous ( $a_1/a_2$ ) at a gene, one cannot tell which allele is on the paternal chromosome and which is on the maternal chromosome unless one can study the individual's parents. (3) Information for deceased individuals (or for those who choose not to participate in a genetic study) is completely missing from the pedigree.

Another problem is that many traits and diseases do not follow simple Mendelian rules of inheritance. This problem has several aspects:

- Incomplete penetrance. For some “disease genes”, the probability that an individual inheriting the disease gene will have the disease phenotype may be less than 1. This probability is called the *penetrance* of the disease genotype. Penetrance may depend on other unknown genes, age, environmental exposure, or random chance. For example, a gene called BRCA1 on chromosome 17 predisposes to early onset of breast cancer in some women, but the penetrance is estimated to be about 60 percent by age 50 and 85 percent by age 80.

- Phenocopy. Some diseases can be due to non-genetic causes. For example, colon cancer can be caused by mutations in the APC gene on human chromosome 5, but most cases of colon cancer are thought to be nongenetic in origin (and are often attributed to diet).
- Genetic heterogeneity. Some diseases may be caused by mutations in any one of several different genes.
- Polygenic inheritance. Some diseases may involve the interaction of mutations at several different genes simultaneously.

### Maximum Likelihood Estimation

To handle the problem of incomplete information, geneticists have adopted the statistical approach of *maximum likelihood estimation* (MLE).

The geneticist would ideally like to have complete genotypic data  $\mathbf{X}$ —for example, the genotype for every family member, including the precise parental chromosome from which each allele was inherited. Given complete information, it is usually easy to estimate the required parameters: for example, the recombination frequency can be estimated by counting recombinant chromosomes, and the penetrance can be estimated by finding the proportion of individuals with a disease-predisposing genotype who manifest the disease. Unfortunately, one typically has only incomplete data  $\mathbf{Y}$ , from which it is difficult to estimate  $\theta$  directly.

The maximum likelihood estimate  $\hat{\theta}$  is the value that makes the observed data  $\mathbf{Y}$  most likely to have occurred, that is, the value that maximizes the likelihood function  $L(\theta) = \text{prob}(Y|\theta)$ . Using Bayes' Theorem, one can calculate  $L(\theta)$ .

To determine whether  $\hat{\theta}$  is significantly different from a null value  $\theta_0$  (for example, to see whether an estimated recombination frequency is significantly less than 50 percent), one examines the likelihood ratio  $Z = L(\hat{\theta})/L(\theta_0)$ . If  $Z$  exceeds some appropriate threshold  $T$ , a statistically significant effect has been found.

In principle, virtually any genetic problem can be treated by this approach. In practice, two important issues arise:

**Efficient Algorithms.** The number of terms in the Bayes sum scales as roughly  $O(e^{cmn})$ , where  $m$  is the number of people in the family,  $n$  is the number of genetic markers studied, and  $c$  is a constant. Except in the case of the smallest problems, it is infeasible to enumerate all the terms in the sum. Thus, it is a challenge even to calculate the likelihood  $L(\theta)$  at a single point, let alone to find the value  $\hat{\theta}$  that maximizes the function. Considerable mathematical attention has been devoted to finding efficient ways to calculate  $L(\theta)$ .

Recently, mathematical geneticists have explored ways to approximate  $L(\theta)$  by sampling

from the sum. Modern techniques such as Gibbs sampling and importance sampling [12, 14, 15, 13, 21] have been introduced in the past few years. These methods exploit the fact that each piece of missing data depends only on local information in the pedigree. Finding good ways to compute the likelihood function remains a problem from the standpoint of genetics and an excellent test bed for new statistical estimation techniques.

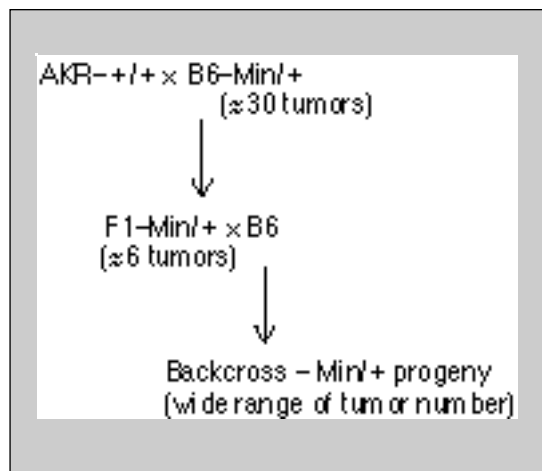
**Statistical Significance.** In many genetic situations, one may search for a disease gene by estimating  $\hat{\theta}$  at many locations along the genome. When doing multiple comparisons, the threshold for statistical significance must be higher than the threshold for a single comparison. Surprisingly, the answer to this threshold question turns out to depend on relatively recent results from the theory of large deviations of diffusion processes. We elaborate on this idea in the next section, using an example from recent work in our laboratory on susceptibility to colon cancer.

### Excursion: Susceptibility to Colon Cancer in Mice and the Large Deviation Theory of Diffusion Processes

Colon cancer is one of the most prevalent malignancies in western societies, with an estimated 145,000 new cases and 60,000 deaths per year in the United States alone. Although environmental factors such as diet can markedly influence the incidence of the disease, genetic factors are known to play a key role. Some families show striking clusters of colon cancer. Among such colon cancer families, there is a distinctive subtype called familial adenomatous polyposis (FAP), which is characterized by the fact that affected individuals develop a large number of intestinal growths called polyps that can become tumors. Genetic mapping studies [2, 17] showed that FAP was genetically linked to a region on the long arm of human chromosome 5; subsequently, physical mapping studies led to the isolation of the responsible gene, named APC [10, 11, 19].

One way to study the role of APC in tumorigenesis is to turn to biochemistry. Another way is to turn back to genetics for further insight. One observation about FAP families is that individuals inheriting precisely the same APC mutation may be affected to very different degrees.

By the usual scientific serendipity, animal studies turned out to hold an important clue. In 1990 William Dove's laboratory at the University of Wisconsin was performing mutagenesis experiments and identified a mouse that spontaneously developed colon tumors [18]. The dominantly acting mutation responsible for the trait was named Min (multiple intestinal neoplasia).



**Figure 3**

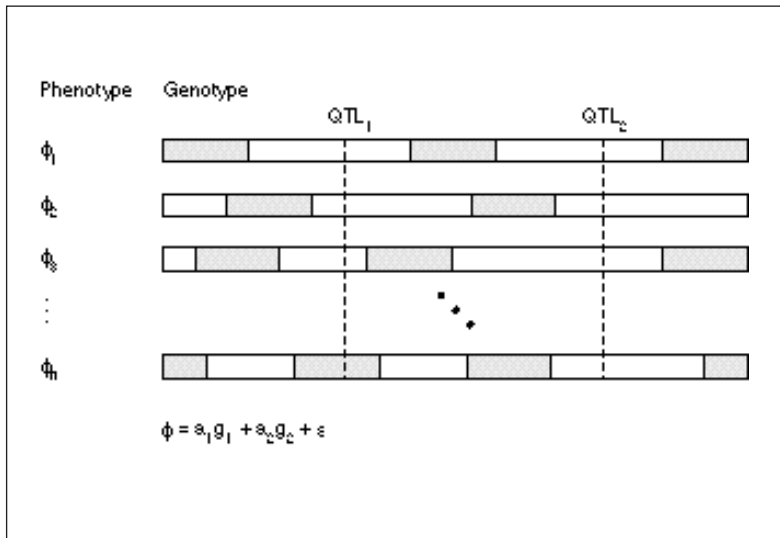
Distribution of colon tumors caused by the Min mutation. Mice from the B6 strain carrying the genotype Min/+ develop about 30 tumors on average. When these mice are crossed to the AKR strain, the resulting  $F_1$  progeny develop only about 6 tumors. When  $F_1$  progeny are crossed back to the B6 strain, the resulting backcross progeny show a wide distribution in tumor number. This figure shows the design of cross.

After considerable genetic mapping and cloning, Dove and his colleagues showed that Min was in fact a mutation in the mouse version of the APC gene [20].

The Min mouse thus provided a model of human colon cancer and, in particular, a way to look for other genes that might suppress the development of colon tumors. When Dove and colleagues crossed this mouse to another mouse strain called AKR, they got a surprising result: progeny develop many fewer colon tumors.

A backcross was arranged in which the  $F_1$  progeny were mated back to the more susceptible strain (Figure 3). For any modifier locus, 50 percent of the progeny should inherit one copy of the suppressing allele from the AKR strain (that is, have genotype AB) and 50 percent should be homozygous for the nonsuppressing allele (that is, have genotype BB). Each animal inheriting the Min mutation was scored for its phenotype by dissecting the intestine and counting the number of tumors and for its genotype by typing the mice for a dense map of DNA polymorphisms that had been constructed in our laboratory [6].

The data for animal  $i$  can be thought of as a phenotype  $\phi_i$  and a continuous function  $g_r(x)$  indicating the genotype, which is either AB or BB at each position along the chromosome (Figure 4).



**Figure 4**  
 Schematic representation of data for genetic analysis of quantitative traits in a backcross. Every offspring ( $i = 1, 2, \dots, n$ ) has a phenotype that is a continuous variable  $\Phi_i$  and a genotype at every position in the genome. The genotype  $g_i(x)$  at position  $x$  has two possible states in a backcross (homozygous or heterozygous, encoded as 0 or 1 and represented by black or white in the figure). The figure illustrates the case where the phenotype might depend on two quantitative trait loci (QTL<sub>1</sub> and QTL<sub>2</sub>), according to a linear model  $\phi = a_1 g_1 + a_2 g_2 + \varepsilon$ , where  $g_1$  is the genotype at QTL<sub>i</sub>, and the  $a$ 's are constants and  $\varepsilon$  is a normal random variable.

At every position  $x$  along the chromosome, the animals can be divided into two sets according to their genotype:

$$AB(x) = \{\text{animal } i \mid g_i(x) = AA\}$$

$$\text{and } BB(x) = \{\text{animal } i \mid g_i(x) = BA\}.$$

If a major modifier gene occurs at location  $x$ , then the animals in  $AB(x^*)$  should have many fewer tumors than the animals in  $BB(x)$ . One could thus perform a  $t$ -test at every position along the chromosome to find a region where the  $t$ -statistic  $Z$  exceeds some critical threshold  $T$ .

How high a threshold is needed to ensure statistical significance, if one scans the entire genome? We will focus on a single chromosome. If there is no modifying gene along the chromosome, the  $t$ -statistic  $Z(x)$  at any given point  $x$  should be normally distributed with mean 0. It is thus easy to determine the appropriate significance level for the single test at  $x$ . But we need to know about the distribution of  $\max Z(x)$ , where the maximum is taken over the entire chromosome.

It is not hard to show that the statistics  $Z(x)$  in our genetic example follow an Ornstein-Uhlenbeck process with  $\beta = 2$ . Using recent math-

ematical results [7, 16], one can thus show that, for large  $t$ ,

$$\text{Prob} \{\max_{0 \leq x \leq G} Z(x) \geq t\} \sim 2Gt^2(1 - \Phi(t)),$$

where  $\Phi(t)$  is the standard normal distribution function and  $G$  is the length of the chromosome measured in expected numbers of crossovers (a unit called the morgan). In short, the probability of exceeding threshold  $t$  somewhere along a genome of length  $LG$  is larger by a factor of  $2Gt^2$  than the probability of exceeding it at a single point.

Returning to the problem of colon cancer, we applied this analysis to the entire mouse genome (the genetic length  $G \approx 16$  morgans). By genetic mapping, we found a striking region on mouse chromosome 4 for which  $Z_{\max} = 4.5$ . The nominal significance level of the statistic is  $p = 3.5 \times 10^{-6}$ . After correcting for searching over an entire genome (by multiplying by  $2G(Z_{\max})^2$ ), the genome-wide significance level is  $p \approx 0.002$ . This suggests that there is indeed a modifying gene in this region of chromosome 4.

On the strength of this analysis, several additional crosses were arranged to confirm this result. With more than 300 animals analyzed, the results are now unambiguous: the corrected significance level is now  $< 10^{-10}$ , and it appears that a single copy of the suppressing form of the gene can decrease tumor numbers at least twofold. Physical mapping is now underway to clone the gene in order to learn its role of reducing colon cancer in genetically predisposed mice. With luck, it may suggest ways to do the same in humans.

## References

- [1] R. ARATIA, E. S. LANDER, S. TAVARÉ, 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. RIGAUULT, 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. USA* **87** (1990), 1213–1217.

[10] J. GRODEN, A. THLIVERIES, 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. FINNIEAR, 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] I. NISHISHO, Y. MIYOSHI, Y. MIKI, H. ANDO, A. HORII, K. KOYAMA, J. UTSUNOMIYA, S. BABA, P. HEDGE, A. MARKHAM, A. J. KRUSH, G. PETERSON, S. R. HAMILTON, M. C. NILBERT, S. K. SU, K. W. KINZLER, and B. VOGELSTEIN, *Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients*, *Science* **253** (1991), 665–669.

[20] 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.

[21] E. THOMPSON and E. WIJSMAN, *The Gibbs sampler on extended pedigrees: Monte Carlo methods for the genetic analysis of complex traits*, Technical Report **193** (1990), Department of Statistics, University of Washington, Seattle.