Of Genes and Genomes

DAVID BOTSTEIN

Chair, Department of Genetics, Stanford University School of Medicine, Stanford, California 94305-5120, USA

Nancy Wexler's presentation gives a clear impression of what motivates researchers to find out what genes do, especially in humans. I'll give you the nuts and bolts of how we learn about genes, how we get and interpret information about our genes, and how the requisite technology led to the Human Genome Project. My view (and I think, now, the general view) is that the Project was the only possible response to the kind of advances in biology that Joseph Goldstein presents in this volume. And I will address my remarks so that they can be appreciated by professors and those that are not—always a difficult task. I will try to boil the subject down to its essence.

FIGURE 1 shows you all that you really need to know about DNA: that it is an information carrier and it makes decisions. DNA contains the instructions for which proteins are to be made and the amino acid sequence of these proteins; in other words, it encodes the information that is required to make the proteins. Proteins do the work in biology; to a first approximation—and everything I say here is a first approximation—proteins do everything.

There is, of course, a genetic code. It was figured out in the 1960s, and it tells how the sequence of letters in the DNA language (the so-called "bases," or "nucleotide bases") is translated into the sequence of amino acids in the proteins. The big insight of our time, arising out of the breaking of the genetic code, is that all living organisms use the same code. This fact makes it possible to express human genes in other

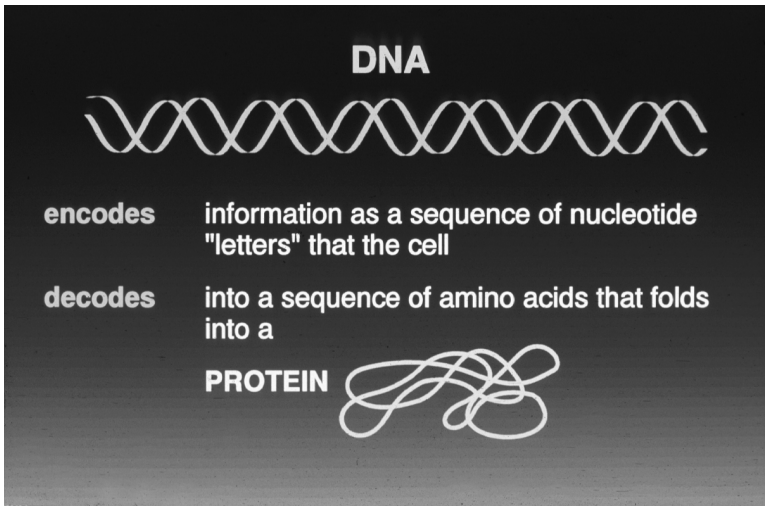

FIGURE 1

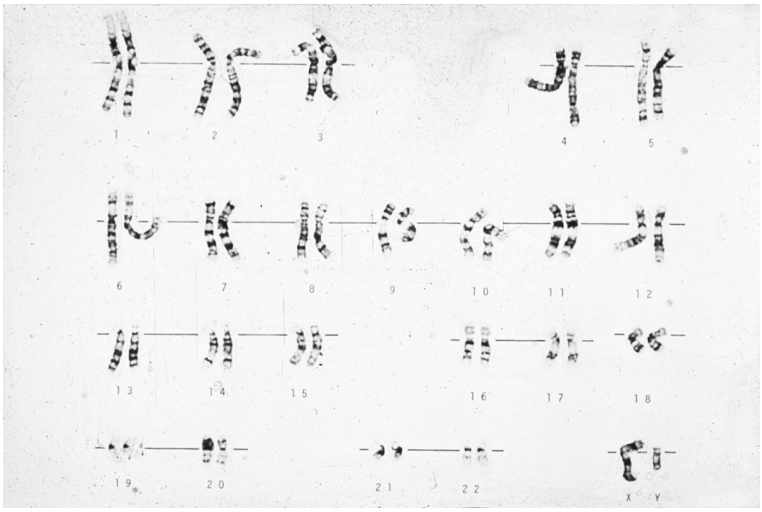

FIGURE 2. Chromosomes of normal human male.

organisms—not just in the mouse but, as we shall see, in truly any living being, plants or animals, even in bacteria and yeast.

FIGURE 2 shows the 23 pairs of chromosomes of a normal human male. All the genetic information about this particular individual is in these chromosomes. Our job as geneticists is to figure out what the information is. Everything that makes one of us different from the other—ignoring, for now, environmental influences—is in the genes that are on our chromosomes.

If humans reproduce in the manner of other organisms, and if our genes follow Mendelian rules (as, indeed, they do), then the following logic can be applied (FIG. 3). If I can distinguish, at any position on a particular chromosome, A1 from A2 by some technical means, I should be able to follow A1 and A2 from a father into his children; and I would find that Mendel's first law applies, namely, that either A1 or A2, but not both, is contributed to an offspring. And which one of the two is passed on to a given child is a purely random event. Likewise, if I can tell A3 from A4 in the mother, I will find that either A3 or A4, but not both, is transmitted to the offspring.

Similarly for B (another locus on another chromosome), Mendel's first law, and therefore the same logic, apply. Also, by looking simultaneously at the results for A and B, I could see that there is no relationship between what happens in A and what happens in B. That is Mendel's second law: of independent assortment.

The point is that if I had yet another locus, C, which happened to be nearby B, then I could show that, whereas A and B are independent and A and C are independent, B and C are not independent—that, in fact, wherever I find B1, I also find C1. The degree to which that last generalization is true has something to do with the distance between A, B, and C; in fact, I can make that relationship a measure of the distance between B and C. And if I do so—and if I limit myself to very short

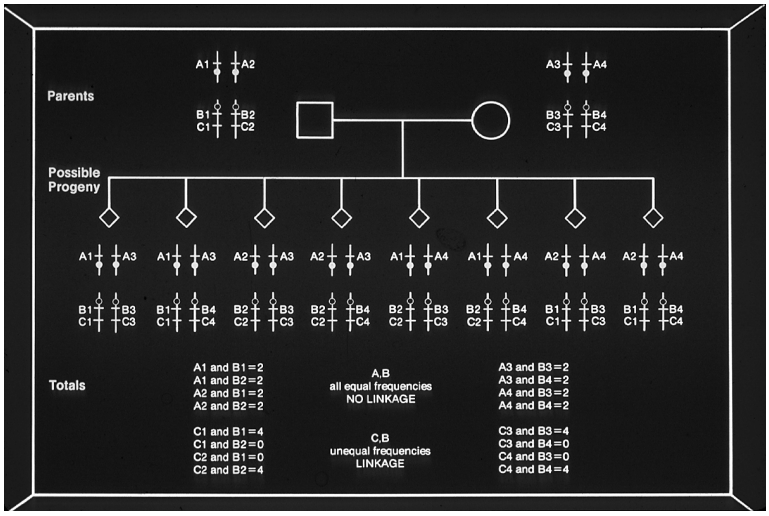

FIGURE 3

distances—I can get a close correlation between B and C (or, in genetic terms, "linkage"). Then, if there is a disease gene that happens to be near B and C (for example, the gene for Huntington's disease), I can follow B and C instead of the Huntington's disease gene (which at that point, I know nothing about) and see that B and C are carried, so to speak, with Huntington's disease. And that's where Nancy Wexler was in 1983.

That was the gist of the workshop that Nancy Wexler referred to in her paper: how we could use linkage to locate and identify genes that cause disease. Unfortunately, the level of genetic education at that time was such that nobody understood the concept of linkage. Nancy's sister, Alice Wexler, has a very amusing description of the workshop in her book, *Mapping Fate*,[1] which is entirely in accord with my recollection. Nobody seemed to understand what we were saying, and we were very frustrated because to us, knowing Mendelism (the principles of genetics), the concept was absolutely elementary.

(The problem, by the way, is one that we should not forget in Dartmouth Medical School's 200th year, namely, that medical education is lousy at dealing with principles. It's great at facts, it's great at practices, it's great at rumors, but it is *not* great with principles. Virtually all medical students shy away from any discussion of principles, especially if it involves a numerical or calculational basis—and genetics, unfortunately, has both. I will say more about this major curricular problem at the end of my this presentation.)

The concept of genetic linkage led to the idea that one wanted to find markers (or pieces of DNA) that fulfill the role of A and B and C, so one could separate A1 from A2, or from A3, or from A4. The short segment of DNA shown in FIGURE 4 was the first useful such marker, isolated by Wyman and White in 1979. The figure shows data from nine randomly selected individuals who happened to give blood at the Uni-

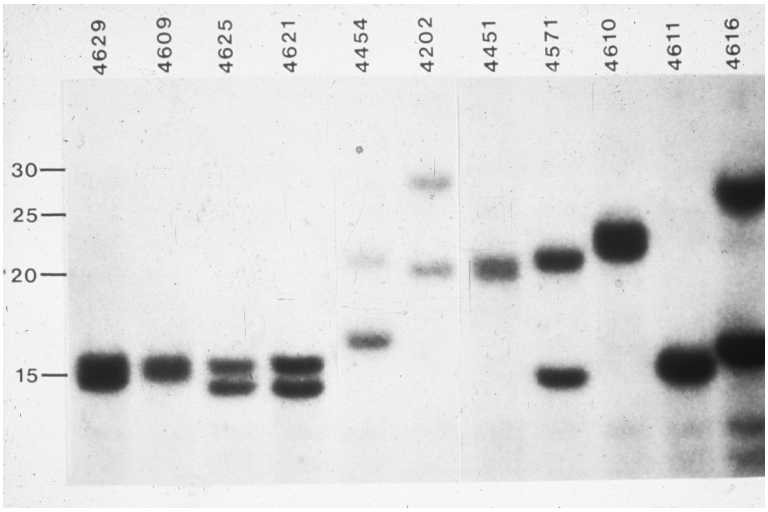

FIGURE 4

versity of Massachusetts in Worcester. (I do not know about informed consent in those days, but fortunately that was not an issue.) The important point is that they were just nine individuals, each different from the others.

It became clear, then, that we wanted to have DNA markers—not just for a single gene in one place (not just, that is, to solve Nancy Wexler's problem of Huntington's disease), but a whole map of markers all over the genome so we could find genes for all diseases. (I need to digress here for just a little bit of vocabulary about markers: a *single copy DNA probe* is a piece of DNA that occurs only once in the genome, which means that we each have two copies of it—one on that specific chromosome from our mom and one on that chromosome from our dad. *RFLP* [restriction fragment length polymorphism] is a technique that allows localization of a gene to a certain chromosome—if you followed the O.J. Simpson trial, you heard this term; and *sequence tag site* [STS] is an even more refined marker than RFLPs, which allows a more precise localization of genes on chromosomes.)

Lots of human disease genes have been mapped by the marker (or linkage) method; I believe the count now is over one thousand. Of course, in 1979 we were told that we were crazy in even proposing to map all of the disease-causing genes. But ultimately, the idea penetrated, thanks in large part to Nancy, and because she found something so quickly. Despite Nancy Wexler's success, however, the concept did not penetrate the establishment to the point where financial support was forthcoming. In fact, the lack of funding forced us to go outside the normal channels and produce a new organization, the Human Genome Project, in order to accomplish what I would have thought the National Institute of General Medical Sciences would have wanted to do as a matter of course. Not at all. They resisted it. To this day, I don't understand why.

In tackling the project, we needed to take into account the size of the human genome. There was the major question of how we would remember the order of things

on the map. The 23 chromosomes from our parents contain 3×10^9 base pairs and those from the other parent another 3×10^9 base pairs. There are four bases, so each carries two bits of information; in other words, approximately 12 billion bits of information needed to be stored. In those days, the capacity of a computer was measured in 8-bit units called bytes; but we had need for 750 million bytes (750 megabytes). Joseph Goldstein spoke of the invention of the microchip and the development of compact disks (CDs); today, virtually all medical students have laptop computers with memories that can accommodate the entire human genome.

In addition, of course, we needed lots of polymorphic markers (RFLPs). The original goal, which was regarded as much too ambitious, was one RFLP for every two recombination units (approximately 2 million bases); we now have roughly 20 times that number of markers—approximately one for every 100,000 bases. Thus, it did not turn out to be such a big deal after all. Of course, the sequence in which the bases (strictly speaking, "nucleotide bases") follow one another in any given gene was the ultimate goal; we wanted to know not only the location of the genes on the chromosomes, but also the sequence of the bases (adenine [A], guanine [G], cytosine [C], and thymine [T]—the letters that form the alphabet of the genetic code). We wanted to know the sequence in which these bases are linked, not only in the Huntington's gene and the gene for cystic fibrosis, but, in fact, in all genes.

Eventually, a committee of the National Research Council (NRC), of which I was a member and which was headed by Bruce Alberts, came out with a grand compromise between the people who were for the project and those who were against it. The compromise allotted about a quarter of the total funds towards experiments with model organisms. At that time, I was attacked on this point for feathering my own nest, because I work with such a model organism, yeast. But actually, two things were realized by the committee at this time: one (to which I have already alluded) was the need for a better way of cataloguing the vast amount of information that was to be gained; and the other was the fact that a lot of sequencing in model organisms would have to be a prerequisite for understanding the human genome. Those were scientific insights, and as a direct consequence of the discussion in the committee, the money was made available to realize both of those things.

In regard to the last point, it is worth mentioning that the deliberations of the NRC committee were the only instance that I know of where science policy was made purely on the basis of scientific needs. The credit for this has to go to Bruce Alberts, who emphasized that we needed to take advantage of a major scientific opportunity, and that we should concentrate on what was needed to exploit that opportunity.

In order to give you an idea of the infrastructure that would be required, let me try to illustrate the size of the human genome. FIGURE 5a is a picture of the earth above the Midwest, somewhat comparable to the size of the human genome. As you increase the resolution you see a part of the Midwest, including all of Lake Michigan (FIG. 5b), comparable to the order of resolution of the chromosomes pictured in FIGURE 2. Going further down, you see the Chicago area (FIG. 5c), a specific lakeside marina (FIG. 5d), and finally a person lying on a blanket in a park within the marina (FIG. 5e).

Here is another way of making the analogy: If one compares the human genome (23 pairs of chromosomes) to a map of the United States, one can consider the information on a single chromosome as analogous to the map of a single state. That de-

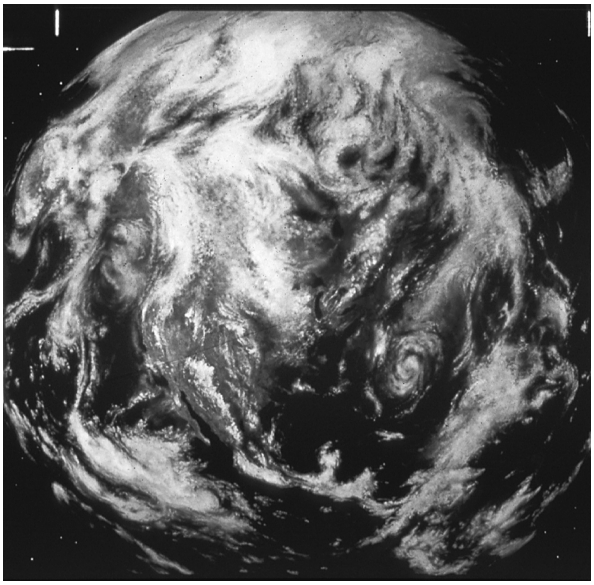

FIGURE 5a. The Earth at 10,000 kilometers above Lake Michigan.

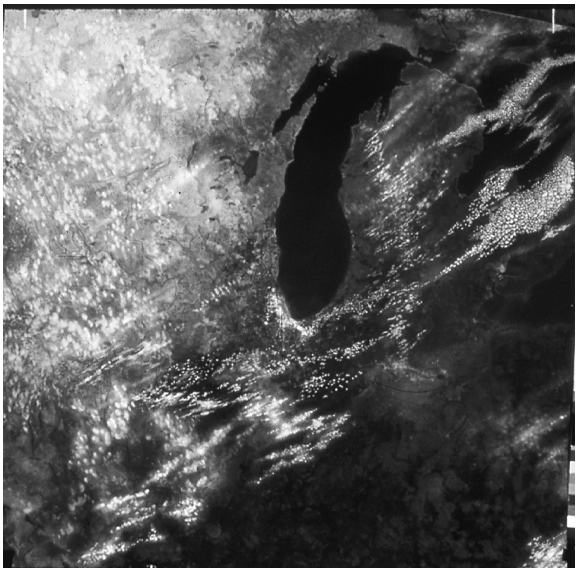

FIGURE 5b. Lake Michigan at 1000 kilometers.

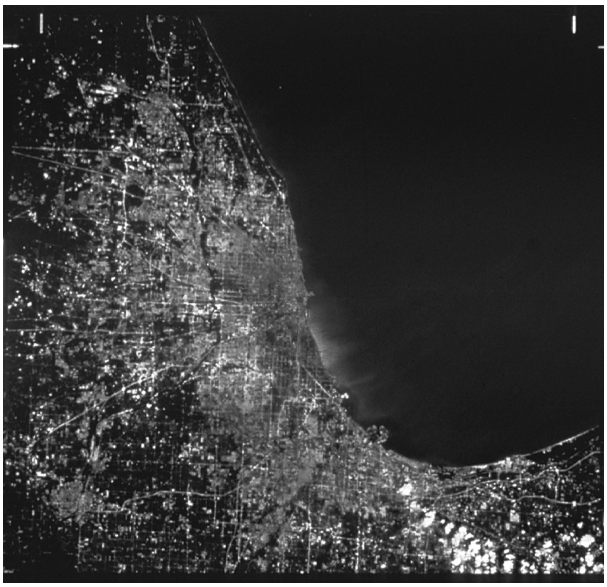

FIGURE 5c. The lower end of Lake Michigan at 100 kilometers showing the Chicago metropolitan area.

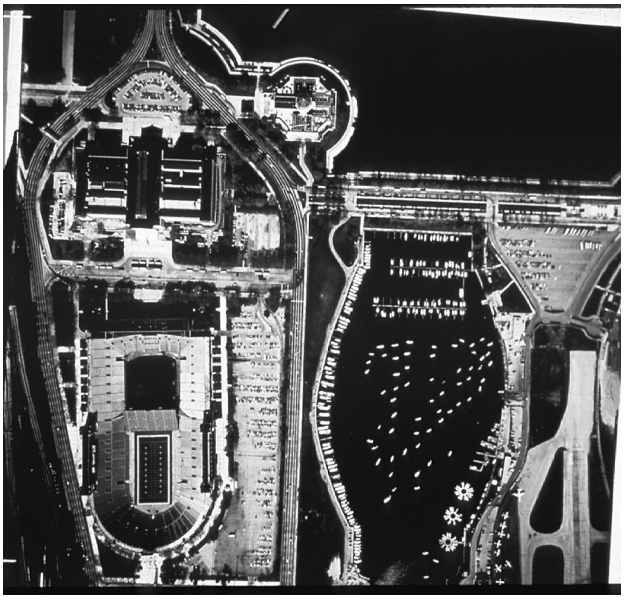

FIGURE 5d. Marina on Lake Shore Drive in Chicago seen at 1 kilometer.

FIGURE 5e. Man on a blanket in park in marina on Lake Shore Drive in Chicago seen at 1 meter. This and FIGURES 5a–d are from *Powers of Ten* by Philip and Phylis Morrison.

gree of resolution will allow us only to say that a given gene is located on a certain chromosome, like saying we are in Illinois. Obviously, a map of considerably higher resolution is needed in order to determine exactly where on a given chromosome a specific gene is located. That further resolution was provided by finding short sequences of DNA (markers) that are unique to specific locations on a single chromosome, or, to continue the analogy, unique to specific locations in Illinois. The original markers, RFLPs, provided approximate locations on chromosomes, like precincts in Chicago. The resolution was then further improved by STSs (sequence tag sites), allowing us to determine a "street address" for each gene.

And finally, let me put the cost of the Human Genome Project in perspective: Just think how much money must have been expended to get detailed maps of the United States, right down to street addresses. It was a lot. When viewed in that context, the proposed expenditure for the Human Genome Project is really modest.

The invention of the polymerase chain reaction (PCR) eliminated the need to remember probes (the DNA STSs) and store them—a problem of infrastructure that I alluded to earlier. The technique of PCR is completely automatable. It enables one to select a small piece of anybody's DNA and replicate it in such a way that one always gets out exactly the same piece of DNA. With the invention of PCR, all you had to do was to remember a short sequence (say, 50 letters) from the gene in question (e.g., for Huntington's disease, or cystic fibrosis, or neurofibromatosis), and thereafter you could always isolate that particular gene.

Another mapping technique—another form of marker—is to get probes from complementary DNA (cDNA), the subset of DNA that is actually expressed in cells. But that is a further refinement, which is not critical towards understanding the principles that I am trying to explain here.

Now, remember that after you do linkage mapping, you are still at a low level of resolution. Mapping locates the gene only in Illinois, so to speak; it's a U2 spy-plane level of resolution and not a walking-around level of resolution—orders of magnitude too low. To come down on the specific disease-causing gene, you need to have more than linkage mapping; you need a "street map," what is called in genetics a *physical map*. That problem has been solved by David Cox and Rick Myers at Stanford (previously at the University of California in San Francisco), who figured out a way to do physical mapping very efficiently, nearly automatically. Their method is called *radiation hybrid mapping*. I will not describe the details here; suffice it to say that it is an ingenious method that makes it possible to localize—within just a few hours—a given segment of DNA situated somewhere in no less than two million base pairs.

I'm not going to belabor these descriptions of techniques for identifying disease-causing genes. The point I would like to leave you with is that we are now beginning to learn how to use the genomic information in order to look into the future. There is a big opportunity to try to understand what goes on in cancer—because we have reasons to believe that current technology will allow us to discover many genes that have mutated to cause aberrations in growth and differentiation.

In the more general sense, I'd like to leave you with several lessons. One is that principles—for example, the basic principles of how organisms are organized and how one studies these processes—are more valuable than individual facts. Progress is held up if medical students and administrators do not pay enough attention to the basic sciences and to underlying principles as opposed to facts and detailed technologies. The technology that was very important when I was a student is no longer in use today. I spent much of my time as a graduate student aligning a centrifuge that is now a museum piece, if it exists at all. I worked with the same Texas Instruments calculator that Dr. Goldstein spoke of. I learned how to program it and make it go, with its puny memory of 32 bits. The programming stood me in good stead, but the machine is now not even a museum piece; I'm sure it's a metal ingot someplace! So it goes with technology.

However, Mendelism is still with us. The principles of biochemistry are still with us. The rules of inference (of what is a proper control); and statistics (what is the variance, what is the reproducibility, what are the figures of merit)—those things are still with us and will continue to be with us; and those are the things that our students are <u>not</u> learning. That is one issue I'd like to see you concerned about.

My second point is that the conservation, and hence availability, of all the newly-gained information is going to put a great premium on separating the wheat from the chaff. We will have to do very careful and rigorous experiments to figure out what each of these proteins (which are encoded by our roughly 100,000 genes) do in whatever organism can best accomplish the task, be that yeast, or bacteria, or mice. A great unification has been given to us by nature, so that a given gene in yeast may have been preserved in humans and will be responsible for the same function in humans as it has been in yeast (or other micro- or macroorganisms). We had, in fact,

no right to such a slow rate of protein evolution; but having seen it, we should take advantage of it. Yeast proteins, for example, can be used to find human genes, and vice versa. Our organism-centric focus must disappear; all these silly barriers—between prokaryotes and lower eukaryotes and higher eukaryotes—all this sort of stuff is going to have to go.

Researchers focus too much on their own work in their own little biological niche. For example, papers in the journal *Cell* make almost no reference to things other than the very narrow area the author is working in. If the author is a good scientist and a little bit of a scholar, the report might go a little further, but anything beyond that, any discussion of principle, is, by and large, not to be found.

So at the end of the day, basic science is important. The explosion of information is going to take all the factoids that you'll remember now, and render them obsolete, if only by swamping them with so many more factoids. We will have to learn how to think about 10,000 facts simultaneously. Remember: rely on much more information technology and much less Tris buffer, and stay close to the basic principles and to the fundamentals.

REFERENCE

1. WEXLER, A. 1995. Mapping Fate. Random House. New York.