THE BIOLOGICAL CENTURY

Friday Evening Talks at the Marine Biological Laboratory

Edited by Robert B. Barlow, Jr., John E. Dowling, and Gerald Weissmann

with Pamela L. Clapp

The Marine Biological Laboratory
Woods Hole, Massachusetts

Distributed by Harvard University Press
Cambridge, Massachusetts
London, England

Genetic Maps—Fruit Flies, People, Bacteria, and Molecules: A Tribute to Morgan and Sturtevant

Joshua Lederberg

The Rockefeller University

THE MBL CENTENNIAL CELEBRATION was a wonderful reunion, at every imaginable level. My visits to Woods Hole began forty-five years ago; but I always felt like a little bit of an interloper. I never found a very good excuse to work on sea urchins or the giant squid or any of the other favorite organisms of this place. It was bad judgment to pick bacteria, which provide very little rationale for the enjoyment of the MBL: but I would visit the laboratory every now and then regardless. Unfortunately there are still very few organisms that live in the sea whose genetics have been developed. It's time we change that.

I'm going to discuss genetic maps, and I assume that the reader has a general knowledge of genes and chromosomes, but does not know so much of the details of DNA chemistry that my simplifications in that realm will be irritating. Just relax, and read a story of genetic maps, and how our understanding of the maps has developed. This essay is based on an evening lecture, not a semester course in molecular biology (see bibliographic note).

Thomas Hunt Morgan came to a chair in Zoology at Columbia University in 1904 at the behest of the renowned cytologist, E. B. Wilson. Shortly thereafter he began breeding *Drosophila* with the intention to make it the exemplar of genetic investigation. And he set up the famous fly room, shown in Figure 1. This picture was purportedly taken (according to Curt Stern's notes) secretly, without Morgan's knowledge. He was supposed to have been somewhat

Figure 1. *T. H. Morgan in the "Fly Room" at Columbia University. (Reproduced with permission from* Am. Zool. *26: 573-581, 1986.)*

phobic about the matter. The picture was taken with a camera that was concealed in one of the fly incubators.

As you can see, Morgan is almost concealed by the cream bottles that were used for growing flies at that time. It is a small exercise in personal nostalgia to recall where the fly room was. It was on the sixth floor of Schermerhorn Hall, Room 613, close to where I started my own research as a college student in 1941. Between 1908 and 1928, this room was the creche of the chromosome theory of heredity. I've started a campaign at Columbia that there should be some commemoration of this remarkable place—the Thomas Hunt Morgan fly room.

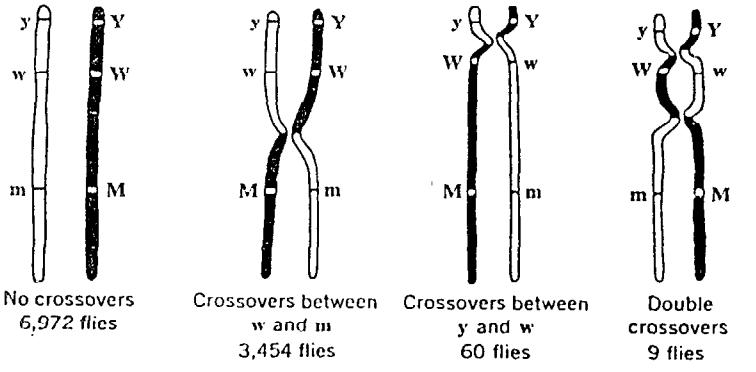
In 1911—30 years before my arrival—Alfred Sturtevant was also an undergraduate at Columbia. He had joined Morgan's laboratory together with Calvin Bridges and Herman Muller. These four people were really the steam engine for the origination of our modern theory of genetics. Here is Alfred's own statement of how he came upon the genetic map:

THE "FLY ROOM"

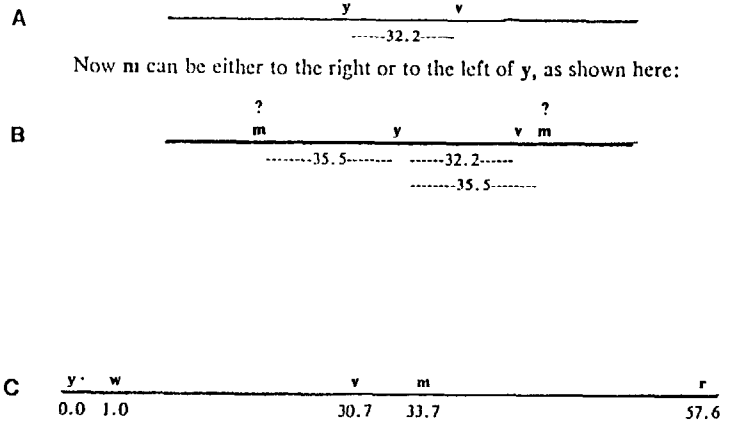
In 1909 Castle published diagrams to show the interrelations of genes affecting the color of rabbits. It seems possible now that these diagrams were intended to represent developmental interactions, but they were taken (at Columbia) as an attempt to show the spatial relations in the nucleus. In the latter part of 1911, in conversation with Morgan about this attempt—which we agreed had nothing in its favor—I suddenly realized that the variations in strength of linkage, already attributed by Morgan to differences in the spatial separation of the genes, offered the possibility of determining sequences in the linear dimension of a chromosome. I went home and spent most of the night (to the neglect of my undergraduate homework) in producing the first chromosome map, which included the sex-linked genes y, w, v, m, and r, in the order and approximately the relative spacing that they still appear on the standard maps (Sturtevant, 1913).

—Quoted from Sturtevant (1965)

Linkage is the consequence of genes being located on the same chromosome not far from one another. Such genes will co-segregate in crosses unless some other event happens. That other event is called "crossing-over," and the frequency of crossing over is a function of the distance between genes on the chromosome, and that is translated to distance on a genetic map. As Sturtevant notes above, his first maps show a number of genes located in the order and approximately the relative spacing that they still appear on a standard map (Figs. 2, 3).



Figures 2 and 3. *Examples of genetic maps. (Reproduced with permission from* Am. Zool. *26: 573-581, 1986.)*



If markers are perfectly linked they will appear at the same point on the chromosome. If they separate from one another 1% of the time, we say they are one linkage unit, one centimorgan, apart. To verify the linear order of markers, one has to take markers not just by twos, but at least by threes to resolve ambiguities as to the specific sequence. When they do all fit together reasonably well by a series of three point tests and higher, we have a corroboration of the linear map.

At that time *Drosophila* was known to have four pairs of chromosomes. A peculiarly looking chromosome called the Y was present in the male and not the female, and a few sex-linked markers, that is to say genes, were located on the X chromosome. E. B. Wilson had first hypothesized that the inheritance of color blindness could be explained by the position of that marker in the sex chromosome of the human. The sex-linked markers, besides bolstering the chromosome theory, are technically much easier to study; results can be obtained in one generation rather than two generations of breeding.

By 1915, we had the listing of the linkage groups of *Drosophila*, as shown in Figure 4. The chromosome picture hadn't changed, but now there are 85 genes. The genes fell into four linkage groups, a rational finding because four chromosome pairs were available. Only a couple of genes could be identified with the tiny chromosome number 4. One thing that is evident immediately is that the locations of the genes for factors involving different developmental outcomes make no particular sense whatsoever. Many of the chromosomes had genes for color; every one of them had genes for wing shape, and so on. Again there seems to be no developmental sense to the map. We know today that if you compare different species of *Drosophila* you will find that the genetic locations of homologous genes are terribly scrambled. You can look for correspondence between any one of the chromosomes in one species and another. You will find that, typically, the most obvious genetic difference between species is the structural change in the chromosomes where pieces of chromosomes have been moved around to many different places. So for many years the idea that the comprehensive genetic map could enhance our insight into the developmental role of the genes was substantially frustrated. Fine structure studies have been more informative.

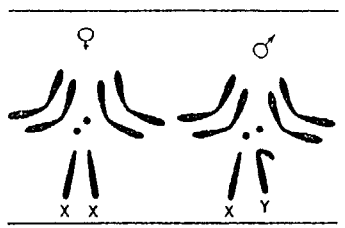
As Harlyn Halvorson mentioned in his introduction, gene mapping continued well through the '40s, having been facilitated by Theophilus Painter's serendipitous discovery, in 1933, of the giant chromosomes of the salivary glands of *Drosophila* (Fig. 5).

GROUP I

Name	Region Affected
Abnormal	Abdomen
Bar	Eye
Bifid	Venation
Bow	Wing
Cherry	Eye color
Chrome	Body color
Cleft	Venation
Club	Wing
Depressed	Wing
Dotted	Thorax
Eosin	Eye color
Facet	Ommatidia
Forked	Spines
Furrowed	Eye
Fused	Venation
Green	Body color
Jaunty	Wing
Lemon	Body color
Lethals, 13	Die
Miniature	Wing
Notch	Venation
Reduplicated	Eye color
Ruby	Legs
Rudimentary	Wings
Sable	Body color
Shifted	Venation
Short	Wing
Skee	Wing
Spoon	Wing
Spot	Body color
Tan	Antenna
Truncate	Wing
Vermilion	Eye color
White	Eye color
Yellow	Body color

GROUP IV

Name	Region Affected
Bent	Wing
Eyeless	Eye



GROUP II

Name	Region Affected
Antlered	Wing
Apterous	Wing
Arc	Wing
Balloon	Venation
Black	Body color
Blistered	Wing
Comma	Thorax mark
Confluent	Venation
Cream II	Eye color
Curved	Wing
Dachs	Legs
Extra vein	Venation
Fringed	Wing
Jaunty	Wing
Limited	Abdominal band
Little crossover	II chromosome
Morula	Ommatidia
Olive	Body color
Plexus	Venation
Purple	Eye color
Speck	Thorax mark
Strap	Wing
Streak	Pattern
Trefoil	Pattern
Truncate	Wing
Vestigial	Wing

GROUP III

Name	Region Affected
Band	Pattern
Beaded	Wing
Cream III	Eye color
Deformed	Eye
Dwarf	Size of body
Ebony	Body color
Giant	Size of body
Kidney	Eye
Low crossing over	III chromosome
Maroon	Eye color
Peach	Eye color
Pink	Eye color
Rough	Eye
Safranin	Eye color
Sepia	Eye color
Sooty	Body color
Spineless	Spines
Spread	Wing
Trident	Pattern
Truncate intensf.	Wing
Whitehead	Pattern
White ocelli	Simple eye

Figure 4. Drosophila *linkage groups. (Reproduced with permission from* Am. Zool. *26: 573-581, 1986.)*

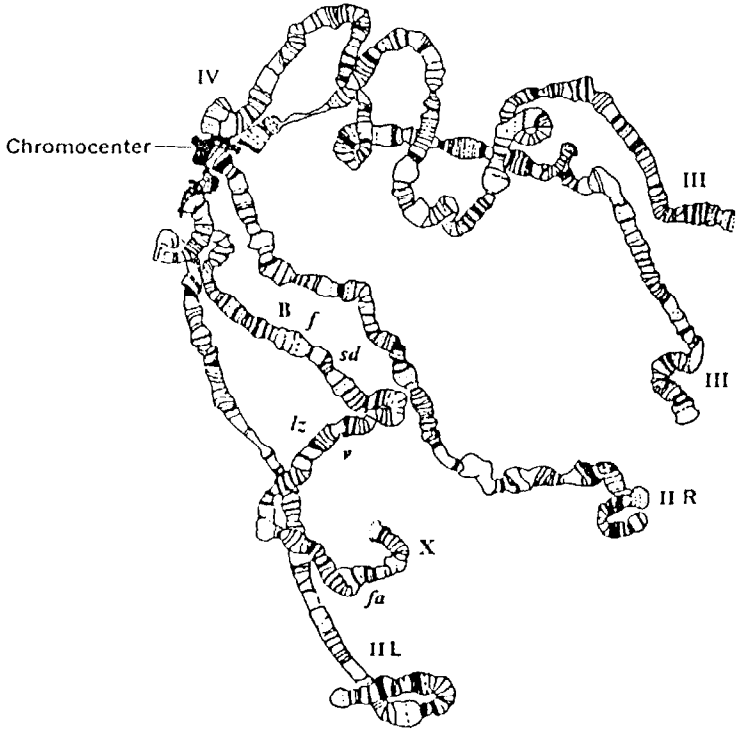


Figure 5. *Giant chromosomes of the salivary glands of* Drosophila. *(Reproduced with permission from* Am. Zool. *26: 573-581, 1986.)*

Drosophila was developed as a standard or type organism for experimental investigation in genetics because of convenience in handling. It was of a size such that you could just about see it with the naked eye, and a really clever and astute observer could pick out mutants by inspection. You didn't have to provide enormous amounts of food or other logistical care to sustain thousands—or tens of thousands—of individuals. Before the introduction of bacteria into genetics, those were regarded as astronomical numbers of organisms for genetical research. But on top of that, it now had one more treasure to offer, namely the giant chromosomes that just pop out for no reason other than that God was kind to us during the cytogenetic research on

the salivary glands of the dipteran larva. With the discovery of these large chromosomes, it suddenly became possible to study the anatomy of the chromosome in finer detail, to at least two orders of magnitude, even just using the light microscope.

Drosophila scientists were thrilled by Painter's discovery. It was so exciting that Morgan, having just received the announcement that he was to be the recipient of the Nobel Prize in physiology or medicine, decided to defer going to Stockholm because he didn't want to lose out on the opportunity to act very, very quickly, exploiting Painter's finding. He wanted to include the salivary gland chromosome mapping in the Nobel address that he gave the following summer in 1934.

The experts found that they could identify even small segments of chromosomes by their specific banding pattern. If a chromosome was broken and a piece translocated to another place, they could track that rather well. They could even eventually locate individual genes to individual bands. I guess that there were about a thousand bands all together so there is that degree of subdivision of the genome that becomes available by this technique of visualizing banding patterns.

Figure 6 shows photomicrographs of one of the salivary chromosomes. Another attribute, again God-given, is that the chromosome pairs in the salivary glands are fused and aligned in much the same fashion as happens transiently during meiosis. So if you have structural changes whereby part of one chromosome has been broken off and translocated to another one, that is faithfully replicated and magnified in the salivary chromosome. When the altered giant chromosome participates in somatic pairing with the other normal chromosome of the diploid pair, you get conformations like that shown in Figure 6, which indicates that in one of the parents this chromosome has been broken at that point. So you can really track quite small changes in the movement of the chromosome parts from one to another of that set.

The giant salivary gland chromosomes deferred a latent crisis in the development of genetics: that the usefulness of *Drosophila* as a tool, had nearly been exhausted because of the methodology available at that time.

Nevertheless, scientists were stymied about going still further into the fine chemical structure of the gene in *Drosophila*. They lacked a means of subdividing the chromosome that was sufficiently detailed so that you could look at individual genes or look into their structure by either the cytological or genetic methods available.

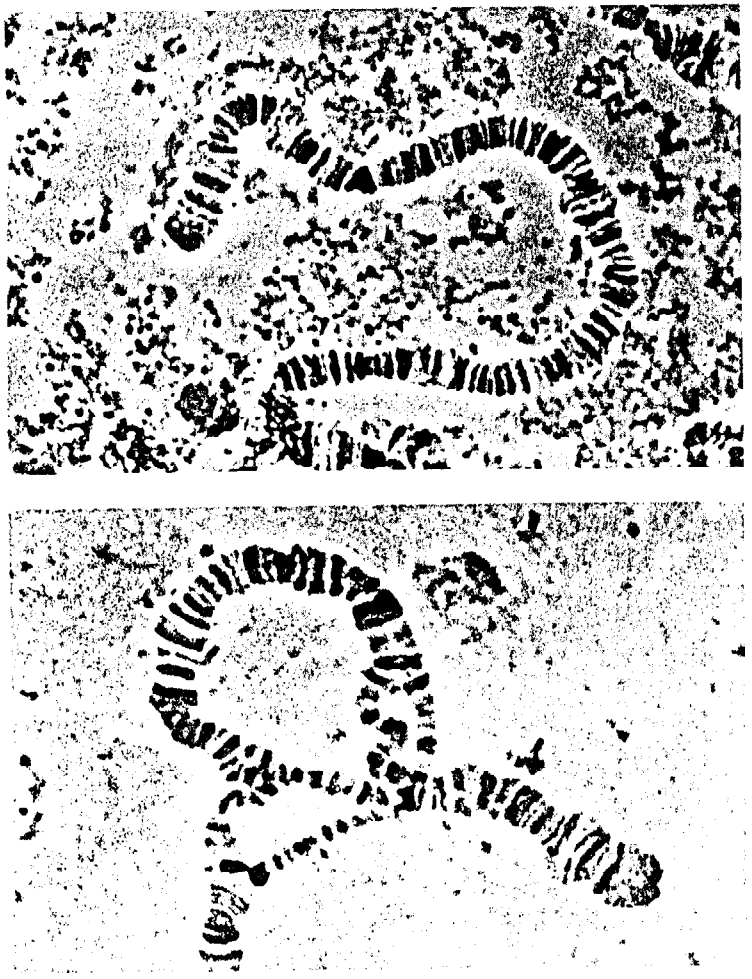


Figure 6. *A photomicrograph of one of the salivary chromosomes of* Drosophila. *(Reproduced with permission from* Am. Zool. *26: 573-581, 1986.)*

New opportunities arose by turning to other "organisms" from the world of microbiology. The classic work "PATOOMB" (Cairns *et al.*, 1966) documents the historic development of bacteriophage as a system of biological inquiry, largely through the impetus of Max Delbrück.

In 1946 at the July symposium at Cold Spring Harbor, both Delbrück, and (essentially independently) Al Hershey reported on a new phenomenon that had cropped up in their investigation of the T family of phages. The T-even phages, T2, T4, and T6, arbitrarily plucked out of the sewers of the world, turned out to be, in fact, rather closely related to one another in their genetic behavior; they are more or less interchangeable. It turns out that if you start with two virus strains in T4, one that had achieved a mutation that changed its host range and another that had achieved an alteration that changed its plaque morphology (one of the so-called R mutants), and infect a single bacterial cell with at least one each of the two parental types, you would find new combinations—genetic recombinant phages—in the crop of viruses that was generated from that cell.

In the population emerging from the mixed infection, you would get some wild type from putting the two mutants together, and you would get some that were double mutant—a virus that was altered both in host range and in its plaque morphology. When this work first appeared there was some skepticism, especially from Delbrück, about a recombinational interpretation. He had been investigating interference between phages in such depth that he viewed the results from that standpoint. He had discovered another phenomenon in which it appeared that once a single phage had penetrated a bacterium, it would tend to exclude other phages participating in growth on the same bug. So his first hypothesis to interpret the mixed infection results was not genetic recombination at all, but some effect of one virus particle, from outside the cell, on the growth of the virus of the other genetic type.

But Hershey followed his work up in some detail, and a few years later he had built a genetic map. His investigation of a variety of host range and R mutants provided the basis for the recombinational genetics that we have seen ever since. Although the T4 genetic map is described in circular terms, the actual DNA of any given particle of T4 as you see it wrapped up within its own envelope, is, in fact, a linear piece of DNA, but it is cut as a permutation from a circular map of that organism. When the DNA replicates it simply rolls over and over

again from its starting point, so you have pieces cut from a sort of rolling circle of spaghetti to wrap them up into a single bacteriophage particle. This replication proceeds continuously, resulting in a hundred or so progeny from each infected bacterial cell.

In T4, the individual genetic cross is hard to control: when you put two virus particles in, you get a few hundred progeny out, and only with difficult tricks can you snapshot the intervening process. About five successive rounds of mating occur in the pool of virus genomes that develop within the mixedly infected bacterium. I'm not going to go into the details of bacteriophage replication and recombination here, but I do want to describe the way in which Seymour Benzer, starting in about 1955, exploited the technical opportunities T4 recombination afforded "to run the genetic map into the ground."

Benzer noted a certain class of R mutations, which emphasized the laborious task of looking at individual plaques in order to type them. Instead, he found a method of selection. Certain hosts, in particular *E. coli* K-12 (a different bacterial strain than was being used for the original growth of the phages), would not sustain the growth of any of the RII mutants, but would accept the wild type. Therefore, he counted what fraction of the phage particles would give plaques on this differential host, K-12, compared to what they give on the standard host, B. He could selectively choose rare wild type phages produced as crossovers or recombinants between different RII mutants, even if the mutational sites were very close together and the frequency of recombination low.

His method was so precise that he could easily pick up recombination when it occurred at distances of only .0001%. He collected hundreds of these RII mutants, testing them against one another. From his library of hundreds of mutants, he could find mutants that behaved as if they were adjacent ultimate units of the genetic map, defining a quantum of recombination. If you took two mutants and they were very close together: either they didn't cross over at all, to a sensitivity of .0001%, or they crossed over with a frequency of at least .01%. So it appeared as if there were a certain grain in the structure of the genetic material, namely, mutations that corresponded to changes in adjacent nucleotides of the DNA sequence. You can't have two markers closer together than being adjacent nucleotides—unless they are identical. This conclusion still stands, although we must now take into account gene conversion and other complications.

Besides many stretches of closely adjacent mutants, Benzer found others that seem to cover overlapping segments. These mutants were deletions, and they proved to be an enormous technical help, and also corroborated the linearity of the sequence. Deletion mutants would fail to give recombinants with each of the single mutants within the genetic region spanned by the deletion. Deletions were invaluable in the actual construction of a genetic map of the RII gene of phage T4. If you have libraries of several hundreds of these mutants, you really don't want to have to cross every one with every other one to map them. It is enough to see whether a given mutant fails to recombine with a certain deletion to know whether it falls within this segment or outside of it.

This work was the first to show mutational events that corresponded to the physical model of the chromosome that was now being developed in terms of the DNA structure.

The total DNA sequence has now been pretty well worked out. All the mutants that Benzer worked on were nucleotides within the RII segment. The map as a whole, however, is some 165 times larger than that one segment; there are 166,000 in all of the base pairs and almost all of the genes of phage T4 have been located and identified in functional terms. Therefore, mutations having some impact in the development of the organism had been localized on the T4 map. They can be related in their connection with one another: for instance, what a wonderful thing it is that we have a complex of genes located together that all have to do with the construction of the head. As you go down the map you do begin to see locally, regionally, that genes of related function are, in fact, often physically coordinated in the structure of the chromosome. But you need a very high resolution to be able to see these relationships and there are exceptions to it. For some of the exceptions you find that the gene seems way out of place, but nevertheless can interact developmentally with something on the other side of the map, thus preserving a connection between gene location and function.

Because phage T7 is smaller than T4, it has been examined in somewhat more detail: it is only 39,936 base pairs long. It has also been run into the ground both chemically and genetically. In general terms, one can classify the characteristic genes of the organism as falling into an early, a middle, and a late class depending on when the function of those genes have to be elaborated in terms of the needs of the initial steps of infection, replication, the transcription of the phage

DNA, and finally the components needed for the assembly of the bacteriophage itself in the maturation of the phage particle.

I now turn to another system of mapping that Harlyn also referred to briefly in his introduction. I was working with Ed Tatum at Yale. In 1946, at the very same meeting at Cold Spring Harbor where Delbrück and Hershey reported on recombination phage in T4, I had the exciting opportunity of talking about our work on recombination in the bacterium *E. coli* K-12. In fact, Ed and I probably wouldn't have published a work quite that quickly if it weren't for the excitement that was being generated by the Delbrück and Hershey reports. Our report was not preplanned on the program of the meeting. I had done my first really successful experiment on crossing *E. coli* strains at about the end of May, then during all of June it had been repeated at least a dozen times. There was no doubt about the reality of the observations that *E. coli* bacteria underwent genetic recombination, but we still felt a little bit edgy about rushing to print after only six weeks of an inquiry.

But the data were there. They were quite solid and the mood was right, so Ed and I agreed that it was perfectly appropriate to publish. That really was quite an exciting summer. I devoted almost the entire next year to developing a genetic map in *E.coli* K-12, using essentially the same paradigm that Sturtevant had used in 1911. I wrote up my findings at the MBL Library during the summer of 1947; that was my doctoral thesis at Yale. The map of *E. coli* genes that we published in *Genetics* is a modest subset, but it can be superimposed on today's map.

We continued the mapping work by introducing some further markers; but by 1951 we were facing a disaster. The original markers that I indicated previously had mapped linearly; but when we had a few others like maltose, streptomycin resistance, xylose, and mannitol fermentation, they made no sense at all. In the *Genetics* paper I represented it as a branched map, but unfortunately, some people didn't read the figure legends carefully. As we insisted at the time, this diagram is purely a formal represention of recombination frequencies, and does not imply a physically branched chromosome. What it does imply is that we were in trouble. That is to say that anomalies in recombination frequencies appear in the way that the consequences of crossing were not working out simply. There was something not understood. My own mental model of conjugation in 1946 was borrowed from paramecium: the complete transfer of a whole nucleus from one cell to another. But Jacob and Wollman showed that there

is a progressive transfer of that chromosome, starting from a particular point of origin. It takes 100 minutes for the full consummation of this transfer.

The French refer to "coitus interruptus," which is how one studies the kinetics of gene transfer in bacteria. If you simply give the culture tubes in which the mating is occurring a violent shake at a given time, you physically disrupt the mating process. Then only as much DNA as has already penetrated the female cell can further participate in recombinational processes. An additional detail: a single strand of the double strand helix of the male DNA is replicated as it enters the female cell.

I now turn to today's comprehensive map of the *E. coli* genome, the compilation of which we owe so much to Barbara Bachmann.

First of all, mapping *E. coli* is done today by taking full advantage of the temporally sequenced transfer of markers. The position on the linkage map is calibrated by how many minutes it takes for the first appearance of a marker donated by the male cell to appear among the recombinants (these stretches of map are in groups of 10 min). Each minute equals about 50,000 base pairs; *i.e.*, roughly 1000/second.

The process is sufficiently fragile that hardly any crosses really go to completion to the full 100 minutes of the map length. And so you can see that there were certain complications about what was going on in conjugation in *E. coli* that confound the details of mapping by Sturtevant's method, but now almost everything has fallen into place. In bacterial conjugation we have found a different principle of mapping that we can correlate with the method of mapping by the frequency of crossing over between two markers.

Mapping can be difficult if you have a marker in a genetic region that is not otherwise well populated with genes. You can still find out where that particular gene is by using the temporal sequence method. That tells you approximately where the gene is and what to look for in terms of very closely linked markers if you want to then establish within a millimorgan exactly where it is on the map. This is a fairly arcane presentation, and you might ask, what is the use of it? It is interesting to know that we have gotten well over a thousand different genes mapped in *E. coli*. We are reaching the point where we can expect to saturate the map. This information has ended up having a surprisingly practical significance. As arcane as it seems, you can't do serious biotechnology using the recombinant DNA paradigm without exactly these kinds of maps that define where the markers are

into which you are going to splice new factors, finding the promoters to genes you want expressed, and constructing specific recombinants. So it has ended up being far from an abstruse, arcane, and useless outcome, whatever golden fleece awards might have been conferred on studies of the sex life of bacteria in the 1950s.

Another mode of genetic recombination, transduction, was first discovered by Norton Zinder in my laboratory (Zinder and Lederberg, 1952). It was his doctoral thesis at the University of Wisconsin. We found it while looking for another organism that could be crossed like *E. coli* could be crossed, and we wanted some taxonomic diversification. It seemed important to look at an organism that had a certain medical significance, whose serology was very important, and had been very well worked out as in the case of *Salmonella*. So there were both a number of practical aspects, and a certain background of microbiological knowledge, that would really make one ache to cross *Salmonella* as well as *E. coli*. We used the same paradigm as was used in *E. coli* K-12, and it just didn't work at all! It took a year or two of floundering around to realize that they were not crossing the same way *E. coli* K-12 was. Instead, there was another phenomenon.

A bacterial virus was living in a lysogenic state hidden within the chromosome of one of the strains of *Salmonella*. This phage would occasionally be liberated by the medium, would grow on the other *Salmonella* strain, then that phage crop grown on that other *Salmonella* strain would pick up some of its genes and then move them back to cells of the first strain—and give rise to a number of genetic recombinants by that route. A single phage particle can package about 1% of the total genome in the *Salmonella* system. In *E. coli* another transduction system of similar value was exploited by Yanofsky and Lennox (1959). In *E. coli* you could now relate transduction recombination data to the crossing maps that I mentioned before, but you now had the advantage of much higher resolution because only 1% of the genome is in any given phage particle. Two markers are only co-transduced—they only appear to be linked—when they are very, very close together. And you can then study the detail of crossing over that then occurs within that 1%-interval of the amount of DNA that is, in fact, carried within a given phage particle.

So Yanofsky and Lennox used that system for an enterprise rather like what Benzer had done previously in T4, down to almost the same level of resolution. Yanofsky in particular was able to find mutons that represented mutations in adjacent or nearly adjacent nucleotides.

By studying the tryptophan synthetase gene and getting thousands of mutants within that particular gene, and crossing them by transduction, he was able to relate changes in the genes according to their map position with changes in the amino acids in different positions of the protein itself. So he confirmed the linearity of the genetic map, and the colinearity of the genetic map with the amino acid sequence of the protein: there was clearly a one-to-one relationship. That is to say the information in the DNA was not only encoding for the general properties of the protein that it was responsible for, but it was determining nucleotide by nucleotide which amino acid was going to be present at a given point in the primary structure of the protein.

This set the stage and provided the biological background needed for the cracking of the genetic code, which proceeded in the early 1960s in a rather more direct chemical examination by Nirenberg and others. During my lecture I showed some examples of map structure having to do with physiological consequence. In the map, *trp* A, B, C, D, E referred to five different enzymes involved in tryptophan biosynthesis. The genes for the enzymes appear stacked one next to another on the bacterial chromosome. We get some inkling about why that should be, not so much as to how it came about (we weren't there at the creation), but what sustains it by the understanding that we are beginning to have about the control mechanisms for the expression of this gene sequence.

E. coli is very clever. It can synthesize its own tryptophan by turning on the genes for these enzymes, and it does so in minimal media—media that are not already loaded with that amino acid. But *E. coli* shows great economy in the deployment of its synthetic resources. When tryptophan is available in the outside medium, it shuts off its own internal production. It saves the need to synthesize that whole set of enzymes that is relevant to making that particular growth factor and nothing else.

There is a promoter: the DNA sequence that represents the point of binding of the RNA polymerase, the enzyme that is involved in the transcription of the DNA information into RNA. There is also an operator: a point on the chromosome that is sensitive to amino acids in the environment and is able to exert a negative influence on the further transcription of the DNA if there is trytophan bound to a repressor protein that then binds to the operator site. In addition, there is an attenuation mechanism, which at the translation stage is also sensitive to the level of tryptophan in the environment and shuts off

translation if tryptophan is present, or, more accurately, allows it to read through a stop point if tryptophan is absent from the environment.

Thus the co-location of these *trp* genes in a related biochemical sequence is connected to their co-adaptivity to the organism. It is easy to see why the organism should want to turn these enzymes on and off *en bloc* rather than have different mechanisms to deal with them separately. One way to do that is to have them co-located on the DNA sequence. The genomes of both viruses and prokaryotes contain innumerable examples of these kinds of complexes.

I'm going to go now to human chromosomes, and I'm going to jump very quickly, although the pace of events was at first inordinately delayed and then very rapid once the plug was pulled. It was not until 1956 that Tjio and Levan published the correct chromosome number in man: 2n=46. That's a staggering statement! Many of you were born by then, so this event occurred within your own memory. We were so muddled about how to look at the nucleus of the human cell, that we carried around a model that had the wrong number. What could be worse than that to confuse human biology!

However, when 23 pairs were identified, very rapid progress followed. Originally it looked much more nondescript than the four pairs of chromosomes in *Drosophila*. Torbjorn Caspersson noted that one could use certain fluorescent stains: he was looking for a way in which you could physically sort chromosomes, one from the other, to separate them. The staining provided rather peculiar banding patterns, not so different in principle from the bands in the salivary chromosomes of *Drosophila*, but not in as much detail, by an order of magnitude. But they do enable the cytogeneticist to look at a piece of chromosome and be able to identify it no matter where it appears, even if it has been the subject of a translocation, to spot it as having been a piece of X that's moved to chromosome number seven, or vice versa. And that has greatly accelerated progress in this field.

The first assignment of a gene to one of these cytological markers was in 1968 on the basis of a translocation; the Duffy marker was shown to be on chromosome number 1. At that time, nine linkage groups were known in the human, and we knew there were going to be 23 of them eventually, but it took one more finding to be able to do this with some expedition. With 23 chromosomes you have an exponential increase in the amount of labor that is needed to provide assignments of individual genes to individual chromosomes. The

answer to that came from another quarter in yet another method of mapping, namely somatic cell fusion.

There was a lot of grumbling through the 1950s, including my own, about why we didn't treat somatic cells the way we have been treating microorganisms! We should be looking for sex where it had never been seen before. Eventually there were some exciting findings by Ephrussi and by Henry Harris (1970) in particular, that showed us the technology for doing that. One of the steps in the absorption and uptake of certain viruses in the mammalian cell involves the destruction of the cell membrane. By using irradiated suspensions of the Sendai virus, Harris was able to exploit this phenomenon, get enough destruction of cell surfaces that the cells could fuse with one another, and initiate the beginning of an artificially created quasi-sexual mechanism of bringing nuclei from different sources into the same cell to see whether they would interact.

They do in a number of interesting ways. In 1965, Harris and Watkins thought to do this between a mouse cell and a human cell; they were spectacularly successful. The whole set of mouse chromosomes is rather easy to tell apart from many of the human chromosomes, and because in these mouse-human cell hybrids the human cells tend to drop out, they just get lost. So it's not too difficult to establish the clones of cells derived from these fusions, which contain the full set of mouse chromosomes, plus one, and only one, human chromosome.

Then if you have the appropriate markers for detecting a gene, you can then tell whether a given marker is on that chromosome in the human set. That methodology very quickly amplified our knowledge of the assignment of human genes to chromosomes. By 1976, each chromosome had at least one marker gene. So a combination of banding, which enables us to identify specific pieces of human chromosomes, and this wonderful tool bench, on which one can follow the addition of one human chromosome at a time against the mouse karyotype background, has led to very rapid progress.

One can also locate markers on the chromosome by direct visualization. You can make a probe for the gene that you are seeking, starting with an RNA message extracted from active cells, and using reverse transcriptase to convert to c-DNA. Cloning the DNA of interest with the recombinant DNA methods into bacteria, you can get large quantities of DNA that is specific for a given function. Today, the PCR (polymerase chain reaction) (Mullis, 1990) makes it even easier to fabricate such probes. One of the most dramatic applications

is *in situ* hybridization. If a fluorescent, or label, is attached to the probe, it can be used literally to light up the spot on the chromosome that displays homologous DNA (Lichter and Ward, 1990).

Using a combination of these techniques today, we now have extensive maps of the human genome. For many years, McKusick (1988) has been tabulating and listing about 5,000 markers; they are now beginning to be assigned to map positions. I guess there are five or six hundred that have been assigned, but maybe that was yesterday and there are a thousand today. Things are moving very very quickly.

The human chromosome set is comprised of 22 autosomes plus the X and the Y. If you examined just number 11, for instance, which is a medium-sized chromosome, you would find that it has about 120,000,000 base pairs or about 4% of the total genome. McKusick (1988) has recorded many interesting things that are now known to be located on that particular chromosome. His database emphasizes the genetic defects mapped thus far, using the variety of techniques that I have been describing. We have this level of extensive map data now for the whole chromosome set.

The ultimate map is the DNA sequence. The basic methodology consists of four experiments. To start, it is important to have some highly purified homogeneous material that is worth sequencing—for that, the recombinant DNA methodology is indispensable. Then you set up four different subsamples of that DNA for replication using one each of dideoxyadenylate or guanylate or cytidylate, or thymidylate (the A, G, C, T, of course, are the letters of the alphabet of which the DNA sequence is composed). By using the dideoxy derivative rather than the natural deoxy compounds in a small dose in a cocktail needed for synthesizing copies of the DNA to be sequenced, DNA synthesis is randomly interrupted when a dideoxy base is incorporated at that point in the enzymatic synthesis of copies of the DNA. The replication, then, goes no further. One can then sort the set of DNA molecules that have been synthesized according to the length of the piece: the long ones are on top, the short ones are on bottom. For instance, you can find out if a piece of X length appears as a band here in which case there is an A at that position because the synthesis of the DNA strand was terminated by a dideoxy nucleotide having been picked up instead of a normal deoxy nucleotide. As a control you see you don't have a band of this length in the G or the C or the T lane. And, similarly, for all the other bases. You find the sizes of the DNA fragments that end with that base. So by comparing the bands that you

see in one and only one of the lanes, you can then literally write down the sequence of the DNA in that particular circumstance so that at this particular position, you had an A, you had a G, you had a G, you had an A, you had three Ts. And so on. Fundamentally, this is how sequencing is done.

While we can do this to stretches of hundreds to thousands of nucleotides, you also carry out the procedure on different DNA samples and find overlaps so that you can match and extend the results. You can put it all together and assemble the entire map of an organism. It has been done to the extent of about 200,000 bases, maybe more than that by this afternoon, with the method mentioned.

We still lack precise information on the human mutation rate per generation (Mohrenweiser *et al.*, 1989). The best estimates range from 10^{-5} to 10^{-6} per gene per generation for mutations of consequential effect: roughly 0.1 per genome, compared to an accumulated background of 10 to 100 times that level. Therefore, our evolutionary background would comprise 10 or 100 generations worth of new spontaneous mutations. Another way of putting those numbers is that it takes an average of 10 to 100 generations to weed out every deleterious new mutation that enters the gene pool. One guesses (very crudely) that every baby is born with about 100 genes exhibiting some new nucleotide substitution, almost all inconsequential. Regardless of this, we have to be very much concerned about aggravating the accumulation of genetic damage that is inherent in our evolutionary history, and which is unavoidable from natural radiation and cellular metabolism.

I turn now to the human genome project, which has emerged out of these scientific and technical advances (adapted from Lederberg, 1988).

DNA is a structure of formidable complexity. If unraveled from a single human cell, the three billion nucleotide pairs of DNA fulfill two meters of double helix. I should double that because it is only the haploid genome that has the three billion base pairs. If, as is widely assumed, about 1% of that total length is translated into protein structures, about 100,000 gene products will have to be accounted for. The ultimate reductionism would be to build an analytical factory that would complete the reading of all three billion nucleotides of human DNA as one technical exercise. A price tag of a few billion dollars is cited, perhaps less if there is prior investment in new technology to automate the task. Is it worth the cost? Undoubtedly.

Is it the wisest use of funds at that level of expenditure? I have very grave doubts. Part of my reservations have to with the style of research that it would encourage and part with the misunderstanding of what we need to learn in mapping the genome. By now we have profound information concerning a score or so of human proteins. Each of them is at least a life's work. At a modest ten million dollars per life's work, that would amount to a trillion dollars to gain detailed information about the full set of human genes. Obviously we cannot commit this large a sum, so we must make discriminating selections of targets before committing to the task.

About 100 human proteins are now discernable as agents of important biological activity. That number will soon grow to perhaps a thousand. That is roughly the map that McKusick prepared, although at present only a small fraction of human genetic disorders can be related to specific gene products, and not every one of them is of such compelling importance that you want to spend a hundred million dollars to identify it.

The biologically active genes and gene products should be the priority list for further inquiry. I am suggesting that we have sensible means by which to decide, out the 100,000 genes that we will eventually need to learn all about, that we can pick out a hundred of them, even a thousand of them, to analyze before the rest. Even this fraction of the total will take many lifetimes of work to try to plumb them to the bottom. Getting the DNA sequence connected with them is obviously absolutely essential, but to have merely sequenced the DNA is just the tip of the iceberg. We need to understand the functional aspects of the gene and the gene product.

To learn about a gene in depth we will have to look into detail of regulation, three-dimensional structure, genetic variability, within and between species, physiological interrelationships, and therapeutic applications. To pursue such inquiries will take much more than the engineering mentality that would apply the single methodology of DNA sequencing for a single sweep through the entire human genome. To make significant advances, we will need a sensitivity to the organism and the focus of expertise, even a fascination, with the particular gene system under scrutiny. My recipe is that we not overlook exploratory research, often best done in the context of natural historical observation. The field of view may be under the microscope or at the hospital bedside, as well as in the open countryside or the oceans.

With luck, our strategies will continue to be self correcting. As we are presented with samples of maps and sequences, individual chromosomes here and there, we will be better able to assess the value of coninuing the grand sweep, at a hundredfold greater costs, compared to the more focussed and diversified inspection of a thousand blossoms wherever they may appear.

Since this address was given on August 15, 1988, rapid leaps have been made in mapping many species, including *E. coli*, yeast, *Drosophila* and the human. Accordingly, the corresponding detail, now outdated, has been deleted from the current text. For extensive synopses and posters, see *Science* **250** (October 12, 1990)—chart at p. 262: a-p; *Science* **254** (October 11, 1991) at pp. 247-262; and *Science* **258** (October 2, 1992) at pp. 87-102.

Literature Cited

Anderson, S., A.T. Bankier, B.G. Barrell, A.R. Coulson, M.H.L. Debruijn, J. Drouin, I.C. Eperon, D.P. Nierlich, B.A. Roe, F. Sanger, P.H. Schreier, A.J.H. Smith, R. Staden, and I.G. Young. 1981. Sequence and organization of the human mitochondrial genome. *Nature* **290**: 457-465.

Bernardi, G. 1989. The isochore organization of the human genome. *Ann. Rev. Genet.* **23**: 637-661.

Brewer, B.J. 1988. When polymerases collide—replication and the transcriptional organization of the *Escherichia-coli* chromosome. *Cell* **53**: 679-686.

Cairns, J., G.S. Stent, and J.D. Watson (eds.). 1966. *Phage and the Origins of Molecular Biology.* Cold Spring Harbor Laboratory of Quantitative Biology, Cold Spring Harbor, NY. 340 pp.

Harris, H. 1970. *Cell Fusion.* Clarendon Press, Oxford.

Lederberg, J. 1988. The second century of Louis Pasteur. A global agenda for biomedical research. Pp. 19-30 in *Molecular Biology and Infectious Diseases*, M. Schwartz, ed. Elsevier, Amsterdam.

Lichter, P., and D.C. Ward. 1990. Is non-isotopic *in situ* hybridization finally coming of age? *Nature* **345**: 93-95

McKusick, V.A. 1988. Mendelian inheritance in man. Catalogs of autosomal dominant, autosomal recessive, and x-linked phenotypes. Johns Hopkins University Press, Baltimore, MD.

Mohrenweiser, H.W., R.D. Larsen, and J.V. Neel. 1989. Development of molecular approaches to estimating germinal mutation-rates. 1. Detection of insertion deletion rearrangement variants in the human genome. *Mutation Res.* **212**: 241-252.

Mullis, K.B. 1990. The unusual origin of the polymerase chain-reaction. *Sci. Am.* **262**: 56.

Riley, M., and S. Krawiec. 1987. Genome organization. Pp. 967-981 in Neidhardt (1987).

Segall, A.M., and J.R. Roth. 1989. Recombination between homologies in direct and inverse orientation in the chromosome of *Salmonella*—intervals which are nonpermission for inversion formation. *Genetics* **122**: 737-747.

Singh, G., D.C. Wallace, and M.T. Lott. 1989. A mitochondrial-DNA mutation as a cause of Lebers Hereditary Optic Neuropathy. *New Eng. J. Med.* **320**: 1300-1305.

Sturtevant, A. H. 1965. *A History of Genetics.* Harper & Row, New York.

Yanofsky, C., and E.S. Lennox. 1959 Transduction and recombination study of linkage relationships among genes controlling tryptophan synthesis in *Escherichia coli*. *Virology* **8**:425-447.

Zinder N.D., and J. Lederberg. 1952. Genetic exchange in *Salmonella*. *J. Bact.* **64**: 679-699.

General Bibliographical Note:

Several monographic texts provide background detail and references for the molecular biology surveyed in this address. Particularly useful sources for the topics covered here are:

Darnell, James, Edwin Lodish, Harvey Franklin, and David Baltimore, 1990. *Molecular Cell Biology,* 2nd ed. Scientific American Books, New York.

Lederberg, J. (ed. in chief). 1992. *Encyclopedia of Microbiology.* Academic Press, San Diego, CA.

Neidhardt, F.C. (ed.). 1987. Escherichia coli *and* Salmonella typhimurium: *Cellular and Molecular Biology.* American Society for Microbiology, Washington, DC.

Neidhardt, F.C., M. Schaechter, and J.L. Ingraham. 1990. *Physiology of the Bacterial Cell. A Molecular Approach.* Sinauer Associates, Sunderland, MA.

O'Brien, S.J. (ed.). 1987. *Genetic Maps. A Compilation of Linkage and Reconstruction Maps of Genetically Studied Organisms,* 4. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Watson, J.D., N.H. Hopkins, J.W. Roberts, J.A. Steitz, and A.M. Weiner. 1987. *Molecular Biology of the Gene.* Benjamin/Cummings, Menlo Park, CA.

The history of molecular biology was comprehensively surveyed by: Judson, H. F., 1979. *The Eighth Day of Creation.* Simon & Schuster, New York. The early history of *E. coli* mapping is the subject of a memoir: Lederberg, J. 1987. Genetic recombination in bacteria: a discovery account. *Ann. Rev. Genet.* **21**: 23-46.

Special thanks are due to Professor John A. Moore, University of California at Riverside, for the photographic material on the Fly Room at Columbia University. See Moore, J. A. 1986. Science as a way of knowing. III. Genetics. *Am. Zool.* **26**: 583-747.