

7

**GENOME SEQUENCING WORKSHOP**

**MARCH 3 & 4, 1986**

**SANTA FE, NEW MEXICO**

**SPONSOR**

**DOE**

**OFFICE OF HEALTH AND  
ENVIRONMENTAL RESEARCH**

**HOST**

**LIFE SCIENCES DIVISION  
LOS ALAMOS NATIONAL LABORATORY**

GENOME SEQUENCING WORKSHOP  
MARCH 3-4, 1986  
PARTICIPANTS

Norman Anderson	Proteus Technology Rockville, MD
Ben Barnhart	OHER Washington, D.C.
George Bell	Los Alamos National Laboratory Los Alamos, NM
Mark Bitensky	Los Alamos National Laboratory Los Alamos, NM
Fred Blattner	University of Wisconsin Madison, WI
Albert Branscomb	Lawrence Livermore National Lab Livermore, CA
Sidney Brenner	Medical Research Council Cambridge, England
Christian Burks	Los Alamos National Laboratory Los Alamos, NM
Charles Cantor	Columbia University New York, NY
Anthony Carrano	Lawrence Livermore National Lab Livermore, CA
Thomas Caskey	Baylor College Houston, TX
George Church	University of California SF San Francisco, CA
David Comings	City of Hope Medical Center Duart, CA
Scott Cram	Los Alamos National Laboratory Los Alamos, NM
Joseph D'Anna	Los Alamos National Laboratory Los Alamos, NM
Larry Deaven	Los Alamos National Laboratory Los Alamos, NM

9

James Fickett	Los Alamos National Laboratory Los Alamos, NM
Richard Gelinas	Fred Hutchinson Cancer Res. Ctr. Seattle, WA
Walter Gilbert	Harvard University Cambridge, MA
Walter Goad	Los Alamos National Laboratory Los Alamos, NM
Gerald Guralnik	Los Alamos National Laboratory Los Alamos, NM
Joyce Hamlin	University of Virginia Charlottesville, VA
John Hearst	University of California Berkeley Berkeley, CA
Elaine Heron	Applied Biosystems Foster City, CA
Ed Hildebrand	Los Alamos National Laboratory Los Alamos, NM
Maurice Kashdan	Dupont, New England Nuclear Res. Boston, MA
Hans Lehrach	European Molecular Biology Lab Heidelberg, W. Germany
Sankar Mitra	Oak Ridge National Laboratory Oak Ridge, TN
Jane Moores	University of California SD La Jolla, CA
Robert Moyzis	Los Alamos National Laboratory Los Alamos, NM
Ted Puck	University of Colorado Denver, CO
Richard Roberts	Cold Spring Harbor Long Island, NY
Francis Ruddle	Yale University New Haven, CT
David Smith	OHER Washington, D.C.

Hamilton Smith	Johns Hopkins University Baltimore, MD
Gary Stein	University of Florida Gainesville, FL
Janet Stein	University of Florida Gainesville, FL
F. William Studier	Brookhaven National Laboratory Upton, NY
Eldon Sutton	University of Texas Austin, TX
Robert Wagner	Los Alamos National Laboratory Los Alamos, NM
David Ward	Yale University New Haven, CT
Ronald Walters	Los Alamos National Laboratory Los Alamos, NM
Sherman Weissman	Yale University New Haven, CT

---

## GENOME WORKSHOP AGENDA

MONDAY MARCH 3

		<u>Speaker</u>	<u>Room</u>
8:00 - 9:30a	Breakfast and Introduction	Mark W. Bitensky Frank Ruddle	Mesa C
9:30 - 11:00a	Workshop I: Technology		
	1. Sequencing	David Ward	Aspen
	2. Chromosome Isolation & Cloning	Larry Deaven Richard Gelinas	Juniper
	3. Ordering	Hans Lehrach	Mesa B
	4. Computation	George Bell Jim Fickett Christian Burks	Mesa A
11:00 - 11:15a	Break		
11:15 - 12:00p	Present Viewgraphs & Discussion of Workshop I	Richard Gelinas David Ward	Mesa A
12:00 - 1:30p	Lunch and Business	Frank Ruddle	Mesa C
1:30 - 2:30p	Break		
2:30 - 4:00p	Workshop II: Benefits & Liabilities		
	1. Clinical and Sociological	Tom Caskey David Comings	Aspen
	2. Basic	Sherman Weisman Robert Moyzis	Juniper
	3. Economic Impact	Mark W. Bitensky Frank Ruddle	Mesa B
4:00 - 5:00p	Present Viewgraphs & Discussion of Workshop II	David Comings Sidney Brenner	Mesa A

## WORKSHOP AGENDA CONTINUED MARCH 3

		<u>Speaker</u>	<u>Room</u>
5:00 - 5:30p	Break		
5:30 - 7:00p	Dinner		Mesa C
7:00 - 8:30p	Workshop III: Model of the Enterprise/ Strategies and Approaches/Costs		
	1. Map Now & Sequence Later	Charles Canter	Aspen
	2. Sequence Intensively with "Special" Foci or Randomly	Frank Ruddle Fred Blatner	Juniper
	3. Coordination/Integration	Walter Gilbert	Mesa A
	4. Cost Estimates	Christian Burks George Church	Mesa B
8:30 - 10:00p	Present Viewgraphs and Discussion of Workshop III	Walter Gilbert Walter Goad	Mesa A

GENOME WORKSHOP AGENDA

TUESDAY MARCH 4

		<u>Speaker</u>	<u>Room</u>
8:00 - 9:00a	Breakfast		Mesa C
9:00 - 10:30a	Workshop IV: Participants, Funding, Unfinished Business	Norman Anderson Frank Ruddle	
	1. Participants: Organizations and Individuals	Norman Anderson	Aspen
	2. Funding Sources	David Smith	Juniper
	3. Coordination & Intergration Part II	Walter Goad	Mesa B
	4. Open		
10:30 - 12:00p	Summary Plenary Session Document Completion	Frank Ruddle	Mesa A
12:00 - 1:30p	Lunch		Mesa C

GENOME SEQUENCING WORKSHOP  
MARCH 3-4, 1986  
PARTICIPANTS

Norman Anderson	Proteus Technology Rockville, MD
Ben Barnhart	OHER Washington, D.C.
George Bell	Los Alamos National Laboratory Los Alamos, NM
Mark Bitensky	Los Alamos National Laboratory Los Alamos, NM
Fred Blattner	University of Wisconsin Madison, WI
Albert Branscomb	Lawrence Livermore National Lab Livermore, CA
Sidney Brenner	Medical Research Council Cambridge, England
Christian Burks	Los Alamos National Laboratory Los Alamos, NM
Charles Cantor	Columbia University New York, NY
Anthony Carrano	Lawrence Livermore National Lab Livermore, CA
Thomas Caskey	Baylor College Houston, TX
George Church	University of California SF San Francisco, CA
David Comings	City of Hope Medical Center Duart, CA
Scott Cram	Los Alamos National Laboratory Los Alamos, NM
Joseph D'Anna	Los Alamos National Laboratory Los Alamos, NM
Larry Deaven	Los Alamos National Laboratory Los Alamos, NM



James Fickett	Los Alamos National Laboratory Los Alamos, NM
Richard Gelinas	Fred Hutchinson Cancer Res. Ctr. Seattle, WA
Walter Gilbert	Harvard University Cambridge, MA
Walter Goad	Los Alamos National Laboratory Los Alamos, NM
Gerald Guralnik	Los Alamos National Laboratory Los Alamos, NM
Joyce Hamlin	University of Virginia Charlottesville, VA
John Hearst	University of California Berkeley Berkeley, CA
Elaine Heron	Applied Biosystems Foster City, CA
Ed Hildebrand	Los Alamos National Laboratory Los Alamos, NM
Maurice Kashdan	Dupont, New England Nuclear Res. Boston, MA
Hans Lehrach	European Molecular Biology Lab Heidelberg, W. Germany
Sankar Mitra	Oak Ridge National Laboratory Oak Ridge, TN
Jane Moores	University of California SD La Jolla, CA
Robert Moyzis	Los Alamos National Laboratory Los Alamos, NM
Ted Puck	University of Colorado Denver, CO
Richard Roberts	Cold Spring Harbor Long Island, NY
Francis Ruddle	Yale University New Haven, CT
David Smith	OHER Washington, D.C.

Hamilton Smith	Johns Hopkins University Baltimore, MD
Gary Stein	University of Florida Gainesville, FL
Janet Stein	University of Florida Gainesville, FL
F. William Studier	Brookhaven National Laboratory Upton, NY
Eldon Sutton	University of Texas Austin, TX
Robert Wagner	Los Alamos National Laboratory Los Alamos, NM
David Ward	Yale University New Haven, CT
Ronald Walters	Los Alamos National Laboratory Los Alamos, NM
Sherman Weissman	Yale University New Haven, CT

Norman G. Anderson  
Proteus Technologies, Inc.  
12301 Parklawn Drive  
Rockville, MD 20852 U.S.A.  
Tel: 301/231-5528

# Los Alamos

Los Alamos National Laboratory  
Los Alamos, New Mexico 87545

Dear Colleague:

A cordial welcome to participants in this OHER sponsored workshop on sequencing the human genome. In addition to the principal sponsorship of OHER and Los Alamos, New England Nuclear is contributing to the support of this meeting without commercial exposure beyond this simple acknowledgement. We are called together by the new head of the Office of Health and Environmental Research, Dr. Charles DeLisi. Dr. DeLisi wishes to carefully and prudently examine the potential importance of this bold initiative. This enterprise is clearly appropriate to the mission of OHER which has given strong support to the National Gene Library project and the sorting of human chromosomes. As part of his initial assessment of future OHER directions, Dr. DeLisi has had the hubris to ask whether and to what extent biologists might consider entering the arena of big science, an arena which until now has been peopled almost exclusively by those in high energy and physics cosmology. Is it possible that accelerated sequencing of the entire genome and the resulting insights into its organization might produce tangible health and research benefits which are even more compelling than those findings which have emerged or might emerge from on-going and planned mega scale physics projects?

It is thus important that we identify here what real benefits and liabilities might emerge from the contemplated sequencing activity, which would aim at capturing the entire human genome in a period of 10 or 12 years. Do we have the technologies necessary to do this, and do we have the computational power and algorithms needed to integrate and analyze this data? Will this information provide both clinical and basic benefits of such magnitude to warrant an accelerated effort?

We need to generate a succinct and focused document which summarizes our conclusions with regard to these fundamental issues. The conference has been designed in an oscillating mode around four workshops. These workshops will be dedicated to questions concerning: 1) new and emerging sequencing technologies; 2) the overall benefits and liabilities which would arise; 3) approaches to overall cost and coordination of the enterprise; 4) generic identification of participants and funding sources. Each of these workshops will be divided into subsections, addressing particular aspects of the workshop topic. In addition, each of the workshops will prepare a vu-graph for presentation at plenary sessions. Therefore, everyone will have the opportunity to evaluate each issue. All points of view will be heard in order that the final document represent impartially the combined wisdom of the participants. To save time, the planned agenda has a designated Chairman, Frank Ruddle, and workshop session leaders. The latter will be asked to develop a summary vu-graph (or two) for each of the workshop sessions for presentation before all assembled participants in plenary sessions. Participants will assemble in the workshops of their own choosing.

18

We are committed to the generation of a concise and impartial document which can be used as a resource (by Dr. DeLisi) for formulating recommendations to the Department of Energy and the Congress. Again, welcome! We face a substantial challenge with compelling time constraints. Let us enter this enterprise with excitement, and with the clear conviction that the deliberations and conclusions of this workshop will be of value to the biomedical community, the Department of Energy and the Nation.

Very truly yours,



M. W. Bitensky, M.D.

Senior Fellow

Los Alamos National Laboratory

P.S. There are in your folder in addition to the Agenda and information materials three sets of comments on our enterprise, from Walter Gilbert, Richard Sinsheimer, and a letter from Los Alamos biologists to Charles DeLisi. These are only included for your reference and are not intended to influence your judgements.

## MEMO ABOUT THE HUMAN GENOME INSTITUTE

Walter Gilbert  
2/9/86

1) I would see the scientific work split about 50-50 between a highly programmed, mission-oriented mapping and sequencing of the human genome, which I think needs to be fully supported by institute funds, and a highly innovative program in basic research on human genetics and biology (in the sense of human genes and development). The basic research side also should involve work in model systems, especially mouse, but also possibly nematodes, yeast, etc. where direct work on genes related to development and direct experimental genetics is possible.

Within the institute, the basic side might be largely staffed with scientists on rolling 5 year appointments, being supplied by the institute with a technician and core research support but encouraged to apply for outside funding for postdocs, further support, etc. This would permit the development of major research groups going beyond the institution's own funding resources and provide some outside peer review on the quality of the basic research. The research on the mission-oriented side needs review and targeting from time to time--some outside review system for this should be set up early (to report to a board of directors.)

2) Size--maybe 10 scientists and 40 technicians in the mission oriented part: covering physical bank and map, rflp map, sequences, technique development, and data-base maintenance and creation. On the basic side, genetics, molecular and developmental biology, engineering and robotics development, advanced statistics and program development--20 scientists, 20 technicians (one for each, 10 postdocs (on institute funds but free to get involved anywhere). This group would be able to expand its effort by getting grants--to maybe 5 people/scientist on average--another 80 students, postdocs, or technicians.

Support staff--administrators, secretaries, accountants, purchasing and stockroom, workshop, building maintenance, etc. is probably another 20 people.

Overall, 120 people on institute support and a running budget of 10-12 million a year--going up to 200 people with grant support and a budget of 15-20 million. (The budget of 12 million for 120 people is fully loaded with the depreciation of the building etc. This is what it would cost with leased building and equipment).

This is a total endowment of 240 million (assuming 5%/year above inflation). If from the Hughes foundation, for instance, this might come in over 4-5 years, as the institute built up. I would expect the mission oriented side to build up over three years, work on the map, both rflp and physical, and on sequencing

technique improvement should begin immediately, while the basic side should build up over 8 to 10 years, with the continuing choice of good people.

Other funding possibilities--a group of private donors for the whole endowment. Or a private donor for the building along with an endowment for its maintenance, the building would be about 25M (200 benches, 25,000 net sq. ft. of lab space plus instrument rooms, other support, probably 60-80,000 sq ft gross, with a maintenance cost of 2-2.5 M or an endowment of 40-50M) along with a state pledge of the rest of the running budget-- 10-20M/year.

### 3) Scale of effort and timetable.

The physical map involves ordering, storing, and analyzing some 100,000 cosmids. At current techniques, some 40 person-years, with improvements it might be done by ten people in two years. The rflp map, to be correlated with the genetic map will constantly improve, but the major portion might also be done, in collaboration with other groups by 20 people in 2-4 years.

Sequencing, since the current technology is about  $10^5$  bases/person year, the whole project looks like 30,000 person years and a cost of 3 billion dollars (one \$ per base). Technology will improve.

I would suggest the sequencing should focus on selected gene areas, in the early years the institute may want to be a sequencing resource--taking genes and probes from outside and returning sequences, cosmids, and probes to the outside. Some selection may have to be done as to which are the most interesting genes to be done first. The gradual increase of sequence around genes, placed on the physical map, will mean that the information available at the institute is of great scientific interest and use throughout all the first phase of the institutes existence. I expect that the most rewarding information scientifically will be in the first 1% of total sequence, if the work is focused, that most of the information, in the sense of interesting differences will be in the next 10%, and the last 90% --of intron and intergenic region--will be the least informative, but the increase in speed of sequencing should make each of these three phases take roughly equal times---or possibly make the last faster than the first.

Phase 1) the first  $3 \times 10^7$

Even using current technology ( $10^5$ /person-year) thirty people can do this phase in 10 years.

Phase 2) This needs an overall rate of 30 megabases (mb)/year from the working group. I think just changes in strategy, along with automated reading and storage, will attain this rate. A directed strategy, coupled with a simple form of the genomic sequencing, along with the improvement made by mass production of the input clones and labeled fragments should get to 50,000 bases /5 people/week (as a group) which is 0.5mb/50 people/week.

Phase 3) This needs 300mb/year or about 1.5mb/working day from the sequencing group. Mechanization of the DNA fragment

handling technology (robotics), sequencing, and storage will be needed.

Thus in 30 years (at most, with the steady improvement of techniques) the institute would have sequenced a reference human genome. The continuing viability of the institute as a center for human biology--which will take unpredictable forms--would be assured by its basic research side--and by the use of its techniques and focus to provide other sequences, such as mouse, for comparison.

4) Benefits: The total human sequence is the grail of human genetics--all possible information about the human structure is revealed (but not understood). It would be an incomparable tool for the investigation of every aspect of human function.

More modestly: the correlation between the rflp map and a physical map makes it possible to isolate and identify (at the institute and elsewhere) the genes for all human genetic diseases--some 3000. Any human genetic marker can be tracked down and analysed.

The human development genes and nervous system genes will be analysed--probably found by comparison with the mouse genes.

By making available filter sets bearing all the ordered cosmids, (one hundred filters, one for each 30 centimorgans, each with 600 spots) suitable for blotting, the speed of cloning new human genes by the general community, both academic and industrial, will be greatly speeded (that step some 100 fold). The centralization of all such information, created by open exchange, is of great value.

## APPENDIX VI

THE COST OF SEQUENCING THE COMPLETE HUMAN GENOME  
CHRISTIAN BURKS

I've broken the cost of the project down into four categories: (1) mapping; (2) sequencing; (3) computing; and (4) administering. The cost of office and lab space as well as required materials are built into the salary estimates, based on the value suggested by W. Gilbert of \$100K per fte per year. (Personally, this seems like a lower estimate — especially in terms of national laboratory dollars — if it is to include physical plant and experimental materials.)

## (1) MAPPING

The consensus opinion was that this would take on the order of 55 man-years, which at a rate of \$100K per man-year, comes to a total of \$5,500K.

## (2) SEQUENCING

The estimate here depends on whether one uses the current rate of sequencing (the consensus opinion was that sequence can currently be determined at  $10^{*5}$  bp per man-year), a (likely) estimate that the sequencing rate will increase ten-fold in the next three-five years, or the (optimistic) predictor that the sequencing rate will increase one hundred-fold in the next 10-12 years. In the following, I have used the estimate of sequencing rate reflecting a ten-fold increase —  $10^{*6}$  bp per man-year. The human genome is  $3*10^{*9}$  bp, so sequencing it would take  $3*10^{*3}$  man-years, which would cost \$300,000K.

## (3) COMPUTING

The calculation here presumes a centralized computing facility with a Cray-class computer at its heart. The Cray would be required for similarity comparisons crucial to sorting the sequences submitted to the central facility by the various in-house and/or distant laboratories, and would also be crucial for analysis (e.g., detection of protein coding regions) of the assembled sequences. A Cray costs about \$15,000K initially, and on the order of \$3,000K (including maintenance staff) per year to maintain — so total cost of the hardware over (say) a ten year period would be \$45,000K. The cost of additional "front-end" machines, network machine, the network itself, many local workstations, and related hardware is harder to pin down, but I think an allowance of \$2,000K per year is not unreasonable: this yields \$20,000K over 10 years, for a total hardware estimate of \$65,000K. One also needs to consider the man-years required to: collect, manage, and distribute the data; write software for managing the database and the network; and write and execute software for sorting and analysing the sequence data. Based on our experience with GenBank, this could easily keep a centralized staff of 30 busy — over 10 years, this would come to 300 man-years, or \$30,000K. The total for hardware and staff thus comes to \$95,000K.



63

Note that this is ten to twenty-fold greater an estimate than we discussed the last day of the workshop — but I think this reflects a more reasonable estimate of the effort involved. Another way of looking at this is that GenBank is costing, over the first 5 year span, about 30 cents per bp; the numbers above applied to the human genome come to about 3 cents per bp; reflecting a projected increase of efficiency similar to that envisioned for sequencing technology, as discussed in the previous paragraph. Yet another way of backing up this total cost, which comes to about \$10,000 per year, is the following examples: C-Division costs about \$60,000K per year to run with 6 Crays (including amortization) and a large network; the Defense Nuclear Agency pays C-Division about \$12,000K per year in return for - 1 Cray, 1 Cyber, and use of CFS and other network facilities; finally, each NSF-sponsored Cray center get \$10,000K-\$15,000K per year.

#### (4) ADMINISTERING

Again, this is difficult to estimate, but there will be a number of people at the centralized facility, and some at the distributed institutes, required to organize and oversee the project, generate and keep track of the funding, and synthesize the understanding resulting from the project. I would make a guess at a staff of 15 for this, which over 10 years comes to 150 man-years, or \$15,000K.

#### TOTAL

Adding categories (1)-(4) yields a total of \$415,000K (or 1/2 billion dollars). I think, given the optimistic estimates above, especially with respect to advances in sequencing technology, that the actual cost could actually be up to five-fold greater. In summary, I would place the cost in the range of .5 to 2.5 billion FY 86 dollars.

## GENOME WORKSHOP I TECHNOLOGY

## Running Notes

(By Dr. David E. Comings)

Dr. Deaven summarized the National Gene Library Project for making chromosome specific libraries. The initial questions were: a) Is it possible to get enough DNA per chromosome by sorting techniques and b) Is the DNA of high enough quality to be utilized. Four approaches were tried 1) the ethylene glycol method, which resulted in DNA with low molecular weight, 2) the propidium iodide method, which has limited usefulness because although it stabilized the chromosomes they couldn't subsequently be restained for sorting, 3) magnesium sulfate technique, and 4) the spermidine stabilization technique. The latter is the one found most useful.

The flow sorting histograms gave a purity of somewhat less than ninety percent for each chromosome. The small chromosomes (between 13 and 22) were easily sortable. Chromosomes 14 and 15 were well resolved in some individuals and not others. For the larger chromosomes hamster/human hybrids containing a single or a few chromosomes had to be used. Because.

of the difference in AT/GC content the human chromosomes separated on different axes and/or were readily isolated. There was some problem with translocations in some of the hybrid cell lines. The efficiency of this sorting was low (in the order of 1-10%). A lot of chromosomes were lost in the isolation procedure. Using the standard sorting machine it was possible to sort 300,000 chromosomes per day; with high speed machines  $1.3 \times 10^6$  chromosomes per day. For the larger chromosomes it was necessary to obtain  $2 \times 10^6$  chromosomes/mg and for the smaller chromosomes  $10 \times 10^6$ /mg.

It was initially felt that making libraries of the small insert size was most feasible. When these libraries are completed, the second phase will provide cosmid or  $\lambda$  libraries using larger inserts.

The molecular weight of DNA is critical for making the larger insert libraries and here it has been demonstrated that sizes of 150Kb or greater are easily obtained. No additional sorting capacity is required for making the cosmid or  $\lambda$  libraries.

It was pointed out that the efficiency of sorting increased using Chinese Hamster human hybrids because of the small number of Chinese hamster chromosomes. Going to the Indian muntjac, which has only a few chromosomes would be even more efficient, but some individuals have tried that and the cell lines have proven to be rather unstable. The question was raised as to whether it wouldn't be possible to use a small-insert

chromosome specific library to probe random cosmid libraries and to segregate the cosmids, into chromosome specific sets. The repetitive elements in the libraries may make this difficult although they could conceivably be reduced by prehybridization.

Dr. Hans Lehrach addressed the question of ordering of the cosmids. He felt that any technique that required individual handling of cosmids would be very labor intensive. Thus he favored methods in which 100K cosmids were dot blotted on filters and hybridized to 50 or 100-9 base pair random probes. This size of probe was chosen because it would give hybridization to approximately 50% of the cosmids. Using 50 or 100 probes one generated a probe map for each cosmid. These were then entered into a computer using a binary code and this information made it possible to determine that cosmids were identical and which had moderate or significant amounts of overlapping, and those that showed some overlapping could be ordered utilizing this technique. He showed computer generated figures based on zero and 5% error rates for examining 100 to 500 cosmids using 50 to 100 probes. This could also help address the problem of library purities. There would be some problem with this procedure with very long DNA repeats but not with short repeats.

Dr. Walter Gilbert pointed out that there were other methods already in place, which could accomplish the same thing and this generated discussion

by Dr. Sidney Brenner of the technique, which he was using, called the restriction signature method for identifying and ordering the cosmid libraries of the nematode. This technique consists of taking the cosmid and cutting it with Hind 3, removing the insert and cutting a 40 Kb insert of approximately 10 Hind 3 fragments. These fragments were then labeled on both sides and then cut with a 4-base cutter. This produced a large number of small fragments, which were then separated on a standard DNA sequencing gel and autoradiographed. This technique labeled about 10% of the total number of fragments produced and the patterns were unique for each cosmid insert. It also allowed identification of overlaps between inserts in different cosmids. It was also pointed out that individuals who had cloned smaller pieces of nematode DNA and used the same technique to obtain a signature could send the results to him and then he could send back a cosmid containing a much larger genomic fragment from which that small piece of DNA came. These techniques would be quite valuable in the human genome sequencing procedure, especially when individuals could map segments of DNA of interest and, from the entire human cosmid library, receive a larger segment of genomic DNA covering that particular sequence. This would be a valuable asset for the program even in its early stages. The nematode has  $10^8$  base pairs and while it could be covered with about 6,000 evenly spaced cosmids with an average size of 40 Kb it would actually

require about 20,000 cosmids. By these approaches perhaps 4 megabases of DNA sequence could be ordered per month. About 19% of the cosmids examined are singles, that is to say have no similar sequence in other cosmids.

Dr. Richard Gelinas discussed  $\lambda$  vs. cosmid libraries.  $\lambda$  vector libraries contained inserts of 1 or 2 times  $10^4$  base pairs, while cosmid libraries can utilize inserts of 2 or 3 times this length. To cover the human genome would require several million  $\lambda$  clones but about 200k or 300K cosmids. He emphasized the psychological value of breaking the task into more manageable segments up by being able to sort out individual chromosomes. Cosmids have proven difficult to work with in some labs and can have poor life span in the lab. There may be a need to make new vectors to insure stability. He raised the question of utilizing herpes simplex virus as a vector, which would have the capacity to hold 150 Kb inserts. He pointed out that Dr. Lebo had made a cosmid library of the human Y chromosome based on flow sorting. Since  $\lambda$  vectors are proven the initial stages of the project should use a  $\lambda$  vector. It was also pointed out that initially sequencing cDNA libraries might be quite valuable.

Dr. Charles Cantor described the pulsed field gel electrophoresis procedure, which was capable of separating DNA fragments in the size range of 100 thousand to a million base pairs. The restriction endo-nuclease DMP

1, which is dependent upon methylation sites through the appropriate manipulation of DNA can be utilized to act as an 8-base cutter given very large pieces of DNA. Thus the restriction fragment map similar to the Smith-Bernstiel approach can be used but with much larger pieces of DNA instead. He felt it would be useful to start out with such a large size restriction map in the ordering of DNA pieces.

Dr. David Ward discussed the possibility of increasing the present rate of sequencing. In a dedicated lab with a highly motivated individual approximately  $10^5$  base pairs can be sequenced per year. At this rate it would require 3,000 persons working for 10 years to complete the sequencing of the human genome. He suggested that it was going to be necessary to increase this rate by approximately 2 orders of magnitude to bring the project down to 30 persons for 10 years. Some of the following approaches were suggested. 1) The technique described by Lee Hood utilizing 4 independent fluorescent primers, which are then all run in a single lane in a continuous elution system with scanning of the aliquot by fluorescence and the use of a computer to direct the results. Dr. Elaine Heron of Applied Biosystems spoke to this procedure and pointed out that approximately 250 or very optimistically 350 base pairs per lane could be sequenced with their new instrument and that the machine read 8 lanes simultaneously in a twelve hour period. This represents approximately

4,000 base pairs per day or an approximate 10 fold increase over the  $5 \times 10^5$  pairs per person per year. 2) Dr. Ward discussed fluorescent techniques to read standard gels. This would have the benefit of multiple lanes per gel but would not be a significant improvement over the above procedure. 3) Chemo luminescence reactions of specific molecules attached to the primers could potentially allow a single treatment of the gel to make chemo luminescence reading of the sequence ladders possible. 4) The George Church multiplex sequence analysis technique. Dr. George Church described his technique whereby 50 different vectors were utilized with 2 probes each and each vector containing a 1000 base pair insert. This was made into 50 libraries and 130 pools were combined for 6,500 different 1,000 base pair inserts. These were then treated with the 4 sequence reactions and after electrophoresis transferred to 5 membranes. These membranes were then labeled with 5 of the 100 probes, the results read, the membranes stripped, and labeled with 5 more probes. Each probe requires approximately 1 day exposure and the entire process would read  $6 \times 10^6$  base pairs. It would require automated equipment in order to have this technique begin to approach the 2 orders of magnitude increase in sequencing speed required. Dr. Ward discussed the potential of using biotin labeled probes, which are then probed with avidin linked to peroxidase or a chemo-luminescent compound. He described an automated



machine for fluorescence in situ hybridization (developed by David Righetti) and soon to be sold by Fisher Allied Company. Some modifications of this could allow to it to be used for southern blots hybridized with fluorescent probes.

The reliability of sequencing will be discussed and it was felt that careful workers had an error rate of approximately 1 in 1000.

GENOME WORKSHOP I CONTINUED

COMPUTATION: Problems with computation and data base management were discussed by Dr. George Bell and Jim Fickett. When such huge data bases are being accessed there arises a requirement for supercomputer use, such as, a Cray and or utilization of parallel processors. It was suggested that parallel processing might allow the data base for each chromosome to be handled on a different microprocessor and this would speed up the accessing of such large amounts of information considerably. Presently information is fed into GenBank and it was pointed out that already, 2 to 3 million base pairs per year are going into that bank and approximately 1/6 of those are for the human genome. Dr. Brenner pointed out that there is much sequence data that is never published and methods were discussed as how to stimulate individuals to publish this data. Journals are restricting the publication of sequences only to include those parts required to make the point of the paper. In the future much data will go directly to the data banks rather than be officially published.

Afternoon Workshop II: Benefits and Liabilities

Basic science advantages were discussed by Dr. Sherman Weissman. He

identified the four levels of information that the human sequence project would identify. 1) A large scale restriction map in the human genome. 2) The availability of cosmids. 3) Total sequence and 4) cDNA sequences. He discussed predominately the advantages of the initial phase just having a large scale restriction map of the genome available. This would be quite valuable for investigators studying various human genetic abnormalities, such as, translocations, deletions or genes localized to specific parts of the chromosomes by linkage mapping. The availability of such restriction map and access to cosmid libraries would significantly enhance the rate at which individuals could then zero down upon the gene in question. 6,000 DNA fragments (on average 500Kb) could be generated with appropriate restriction enzymes and separated by the pulsed gel electrophoresis technique.

Dr. Robert Moyzis talked about the advantage the gene sequence would provide toward our understanding of coordinate gene expression, both in regards to individual structural genes and gene clusters. He also discribed some studies of his on chromosome specific repetitive sequences detected by in situ hybridization, i.e., to chromosome 16.

Dr. Comings discussed the clinical and sociological aspects of the project. He emphasized that in order to gain wide spread public acceptance and especially acceptance in Congress and higher levels of government it

would be essential to demonstrate its broad appeal, not only for the unusual genetic diseases, but for the large number of common disorders that have a very strong genetic component to them. These include the following:

1) Cancer, which is basically a genetic disease of somatic cells.

Relevant to cancer would be the detection of additional oncogenes, growth factors and cell differentiation. 2) Birth defects, studies of trisomy

21, mongolism or Down's syndrome, trisomy 18 and 13 in the broad area of mental retardation. (Twenty-five percent of males with inherited mental

retardation have fragile X associated mental retardation.) 3) Heart disease, the hybrid lipidemias and familial hypertension are significant

contributors to heart disease. 4) Diabetes - Type II diabetes is a strongly genetic and Type I diabetes has genetic elements attributing to

its expression. 5) Alzheimer's disease approximately 50% of all

Alzheimer's disease is probably a genetic disorder and affects 6 or more percent of the population. 6) Mental disorders - 1% of the population has

schizophrenia and 2 to 3% major depressive disorders, 2 to 3% severe alcoholism, 2% panic attacks, 4% attention deficit disorder of childhood;

all of these have very strong genetic components to them and mapping of the human genome would very likely provide very strong clues as to their

etiology. 7) There are many less common genetic disorders, which are

important to many individuals affected by them including: Huntington's

disease, Cystic Fibrosis, Duchenne's Muscular Dystrophy, Pompe's, Tay Sachs and many others. There are also many common diseases, which have strong genetic components, such as the arthritides. A second aspect that is important is to demonstrate that improved health would come more quickly with this knowledge and the third aspect is that it would in fact save money in two ways. 1) It would save money in that each individual investigator would not have to do extensive sequencing and that centralized sequencing facility would probably do it about 10 times faster and 2) it would save health dollars through improved understanding of various genetic diseases both those that are common and the less common.

Dr. Comings also pointed out that when heterogeneous nuclear RNA is labeled and hybridized back to single copy DNA that the brain shows 25% hybridization, liver 15%, kidney 9%, lymphocyte 6%. This indicates that there are 2 to 3 times more genes expressed in the brain than in any other organ. Very few if any of these brain specific genes have been identified and the sequencing of the human genome would go along way toward a much more thorough understanding of the functioning of the human nervous system. Genes of related function could also be identified through their sequence similarity. Such important gene families presumably have a role in cell recognition, memory, gene regulation, proteins, neuropeptides and oncogenes to name a few.

It was also suggested that consideration be given to the sequencing of the cDNA libraries especially from brain and liver. The advantage of this would be that it would identify the expressed portion of the genome, which constitutes only about 1% of the total sequences and that probably 75% of the critical genetic information would be found in this small subset of the total.

There was somehow some sentiment against this approach stating that many investigators are involved in sequencing of cDNA sequences and the project should not step on the toes of those investigators and that the initial part of the human genome sequencing approach, i.e. ordering cosmids in a large scale restriction map would be very valuable and that the sequencing of the whole genome project would not conflict with any projects currently underway, i.e., it would be unique.

Dr. Eldon Sutton reemphasized the need to have a broad based appeal to generate funding for this project and reemphasized its importance in more common diseases with a strong genetic basis, such as, cancer, heart disease, and diabetes.

Dr. Puck emphasized that we should be careful not to convey the idea that we are involved in any way with cloning of the human genes. This is a common misconception among the public that we should very carefully indicate that this is not the case.

Dr. Mark Bitensky discussed the economic impact of such a project and pointed out that the funds would not compete for existing NIH funds and that there should be encouragement of the private sector to invest in this project as well. Dr. Bitensky pointed out that the health benefits associated with genomic sequencing should substantially decrease the profound losses in productivity which arise from acute/catastrophic illnesses and chronic/debilitating illnesses. He also pointed out that the American public will spend about 400 billion dollars on health care costs in 1986. The likelihood is profound that sequencing data will bring about a savings in these expenditures that would be likely to more than equal the cost of the project. Dr. Bitensky also suggested important private sector and economic benefits from the proposed sequencing activity in the form of new biomedical products and new employment opportunities.

Summary Remarks on Workshop II

Dr. Weissman pointed out again that there was significant benefits of the mapping alone along the way of getting a complete sequence map. Dr. Comings reemphasized the need to have the broadest possible base of support by the ability to investigate common diseases with important genetic components. Dr. Bitensky reemphasized the aspect of improved quality of life and productivity. Dr. Ruddle stated that a concerted effort would bring about improved treatment of genetic diseases a generation earlier than would otherwise occur.

Workshop III Monday

Model of the enterprise/strategies and approaches/costs

Dr. Cantor emphasized the importance of making a macro map by taking chromosome specific DNA, cutting it into very large pieces with methylation specific restriction endo nucleases and ordering these pieces by hybridization techniques. The cosmids from specific chromosomes would also be mapped by the signal sequence technique of Dr. Brenner and when questions arose about ordering the hybridization of these sequences to the



ordered large pieces based on pulsed separation could be utilized to fill any gaps that might occur if a critical cosmid sequence was missing. The cosmids would then be sequenced.

Dr. Blattner proposed the possibility of a highly labor intensive effort to sequence random clones from the genome of hydatiform mole and then enter all the sequences into the computer. This would result in ambiguities around repetitive DNA and where pieces could not be ordered properly because of these, branches mapping of the cosmids would have to be done to settle the order. This is a sequence first map later strategy as opposed to Dr. Cantor's approach, which is to map first and sequence later.

Dr. Walter Gilbert reemphasized the incalculable benefit of a total sequence map of the human genome and emphasized again that the cost would be proportion... to how rapidly the project was completed.

There then followed considerable discussion concerning the map now sequence later vs. the sequence now map later philosophy and the general consensus was that a map should be generated first, followed by sequencing. The best solution would be to have the Los Alamos and Lawrence Livermore Laboratories make overlapping cosmid and  $\lambda$  clones of each individual chromosome and make junction libraries and then make these sets of libraries available throughout the scientific community. Individuals interested in specific chromosome would then begin mapping on their own.

The next step would be to pick an individual small chromosome, such as chromosome 21, and produce a map of that through a combination of the pulsed gel technique of Dr. Cantor and the signal sequence of Dr. Brenner. Soon after this approach was initiated sequencing of cosmid clones could begin. It is important to get a commitment to finish the entire genome and the best approach would be to proceed by sequencing a chromosome at a time first with the ordered map then completing the sequences. It is anticipated that other institutions would probably jump in and do some of the work on their own when the cosmid and junction libraries became available. These supporting efforts would become contributory when the DOE laboratories got to those particular chromosomes.

The cost estimates were discussed again. Estimates for the mapping phase ranged from 55 to 200 man years. At an estimate of 100K per man year this ranged from a 5 to 20 million dollar effort. With explosive improvement in techniques it was projected that the entire sequencing might be completed with a 500 million dollar investment or approximately 16 cents per base pair. Many thought that even further declines in sequencing costs were quite feasible.

The workshop ended with a discussion of cost estimates and funding. The Hughes Medical Research Institute has also expressed interest in this project and infusion of money from Hughes would certainly help in the

overall work. It was hoped that funding could also be obtained from other sources as well. (NIH, other foundations, NSF.)

Session of 3-4-86 Summary of previous Workshops

Dr. Cantor reemphasized the concept of a rapid push of the physical mapping of the genome and gradually phasing into sequencing. Dr. Fred Blattner pointed out that there are multiple units of the task and that there is the genome consisting of one unit, the chromosomes consisting of 24 units, cantor fragments of 3K, cosmids of 100K and individual cloning pieces of 2 million. By giving his estimate of  $10^5$  residues per day per technician, 500 technicians working 300 days per week or 5 years could map  $1.5 \times 10$  bases, offering some redundancy to the genome. What should be sequenced? The genome of an individual vs. a hydatiform mole. The latter has the advantage of being homozygous. The sequence options ranged from cosmids to cDNA, every single chromosome or hop junctions. General consensus was that a reasonable approach was to take the project a chromosome at a time.

Dr. Gilbert discussed three levels of the map namely, ordered cosmids, RFLP maps and hoped that this could be done by 1990. The sequencing part would have a goal of being finished by the year 2000. Dr. Christian Burks

reiterated the cost again placing it at 100K dollars per man per year assuming a 10 fold increase in the present rate of sequencing gave a figure of 500K dollars as a very conservative lower estimate. There was some very significant disagreement with this low estimate and if the sequence rate could only be increased 5 fold the cost would be 2.5 billion dollars instead of 500K.

Dr. Ruddle restated that sequencing was the ultimate goal; that the mapping and then the sequencing should be done on a chromosome per chromosome basis and doing the physical cosmid map and the RFLP map would allow melding of the physical and Mendelian maps and that the cost would be somewhere between 1/2 and 1 billion dollars.

#### Workshop IV Participants, Funding, Unfinished Business

Dr. Norman Anderson started off by recommending 2 books, one by James Webb entitled "Space Age Management: Large Scale Approach" McGraw Hill, NY, NY 1969 and another by Weinberg A. entitled "Reflections on Big Science" MIT Press, Cambridge, MA 1967. The book by Webb described how he set up NASA and the book by Weinberg went into the problem of choices between large programs.

Dr. Anderson pointed out that it was critical that we not ask Congress

whether to do it, but simply to say that the decision to do it has already been made, the project is underway, and that funding is needed to finish the project.

He pointed out that ARPA Net already has a contract to investigate parallel computing. This is already related to the GenBank and we should point out the similarity between the computation requirements for this project and that of Star Wars projects. He also pointed out the question of who has the "license" to do this project. He stated that the way to get the license is simply to take it and DOE should assert itself in this area and take the lead.

Dr. Ruddle addressed the administrative aspects of the task. He recommended a Science Advisory Committee and then an administrative unit and on one side a computational unit for data basing, programming analysis, and storing of the sequence material and on the other side a wet lab to work on methods, information, chromosome sorting, and other aspects of technique. He then visualized that there would be multiple regional centers from around the country involved in the process.

The aspect of a centralized institute vs. multiple regional centers was discussed in great detail. The question of morale was brought up and various people felt that because of the importance of this project morale might not be a problem, but that decentralization might in general be more

welcome.

Dr. Smith pointed out that he didn't want to turn off present workers in the field by seeming to preempt their tasks and that a cooperative interaction with scientists around the world was critical. He cited the National Acid Rain Precipitation Program, which is predominantly coordinated by the National Research Program in conjunction with regional laboratories.

Dr. Hamilton Smith suggested that an approach be devised in 2 phases. Phase 1 mapping, which is relatively small in scope, low cost, and potentially doable now and phase 2, sequencing, which is high cost, and requires multiple centers and decentralization.

It was felt it might be possible to sell part of the program as a joint program with the Soviets or possibly the Japanese. Dr. Anderson pointed out that Fuji has launched a program using thin films for sequencing, KAKO a program for sequencing robotics and Hitachi for the mechanics of the process.

Frank Ruddle concluded that there was a need for a map and sequencing and a strong need for a continuous group of advisors so that the concept would not die.

In the final session Dr. Brenner stated that it was important to focus on how to get started and not try to do the whole project at once.

Larry Deaven pointed out that the project of starting the chromosome specific cosmid or  $\lambda$  libraries had already been begun and Dr. Carrano from Lawrence Livermore Laboratory stated that they had also initiated this phase of the work. The vector they were using was Homer IV.

Dr. Ruddle emphasized step wise increases in the funding and cooperation with Hughes Institute. There was some discussion of which agency should be involved and it was generally felt that DOE should continue with its present strong involvement.

The session was summarized by Dr. Ruddle, who pointed out that in a 1981 Gene Mapping Conference, Dr. Southern had first discussed the concept of an overlapping cosmid map of the human genome and that this subject has increasingly come up and has generated a lot of interest and excitement in the genetic community. An ordered cosmid library would be of tremendous value and would stimulate a great deal of research in human genetics. There was a very strong consensus that the mapping and the sequencing project should unquestionably be done.

Dr. Bitensky thanked the participants on behalf of Dr. Charles DeLisi for their lively and substantive contributions and requested an accelerated mailing of comments in order to expedite completion of the Workshop report.

19.

Reprinted from  
*American Biotechnology Laboratory*  
Sept/Oct. 1985

AMERICAN  
**Biotechnology  
Laboratory**

GUEST EDITORIAL

**A policy and program for biotechnology**

BY N.G. ANDERSON AND N.L. ANDERSON

**Present Address:**  
Proteus Technologies, Inc.  
12301 Parklawn Drive  
Rockville, MD 20852 U.S.A.



# Editor's Page

## GUEST EDITORIAL

### A policy and program for biotechnology

BY N.G. ANDERSON AND N.L. ANDERSON

If the biotechnology enterprise in the United States is to succeed in the long term, a unifying policy based on well-defined objectives appears necessary. And if biomedical research and biotechnology are to achieve and deserve major government and private funding comparable to that provided for aerospace, nuclear energy, and nuclear physics, the objectives must be fully comprehensible to the average person, fill a deep-felt need, and be sufficiently broad to encompass most of biology.

Man is the most complex entity in the physical universe. Although explorations of space and the atomic nucleus have achieved and deserve large-scale funding, huge facilities, strong mandates, and continued public attention, the average citizen is more interested in human mysteries: reproduction, development, birth, disease, aging, and death. This public interest and concern is reflected in the position of health care as our largest single industry.

Assuredly, the exploration of man and the cells of which he is composed will ultimately require the best minds, the most sophisticated technology, and a large and superlative organization.

Only two objectives appear to us to offer the possibility of long-term support on the scale required to maintain national leadership in biomedical research and in biotechnology: the complete sequencing of human DNA and the separation, cataloging, and characterization of all human gene products. The first objective might be called the "Plan for Man," and the second the "Parts List for Man." These objectives encompass, as necessary prerequisites and as defensible ancillary projects, much of current biomedical research and engineering and are (or can be made) readily comprehensible to nearly everyone. Both efforts are now technically feasible, both have already been organized on a small scale, and both offer means for the acquisition of new knowledge and for the

solving of numerous practical problems. Progress toward reaching these objectives can be measured objectively. Overpromising need not be a problem.

DNA databases are currently organized at the European Molecular Biology Laboratory, and jointly by the firm of Bolt, Beranek, and Newman in Boston and Los Alamos National Laboratory of the U.S. Department of Energy. In both instances, reliance is largely on published sequences resulting from investigator-initiated studies, as should be the case initially. However, the Los Alamos Laboratory is now developing a human gene bank based on chromosome isolation. Hence, the chromosomal origin of each cloned DNA segment will be known. Sequencing long DNA segments on a routine basis, assembly of large numbers of sequences in order, and systematic assignment of the sequences to chromosomes will probably be beyond the interests, funding, or motivations of academic laboratories. If sequencing is fully automated (and rumors suggest industrial attempts here and abroad to do this), then the rate of data acquisition may increase by several orders of magnitude. If the automated sequencing systems are large and expensive, sequencing may need to be centralized in a few facilities. The organization and management of these facilities may be patterned after that for contemporary large accelerators.

It may also be that the major contribution of the so-called fifth generation of computers and artificial intelligence will be to elucidate in detail how human DNA is organized and how gene expression is controlled.

The new knowledge to be gained from global sequencing is breathtaking indeed. Computerized analysis of genomic architecture is already a new and exciting field in itself and will probably be the first one in biology where theory will play the key role it does in physics. The details of human development and gene regulation ultimately lie in DNA sequences. In addition, the library of genetic variants will become completely accessible as the sequences for mutant genes associated with the more than 2000 described genetic diseases become known. The key question of whether new

*Dr. Norman G. Anderson is Head, and Dr. N. Leigh Anderson is Director of Research, Molecular Anatomy Program, Division of Biological and Medical Research, Argonne National Laboratory. This work was supported by the U.S. Department of Energy under contract no. W-31-109-ENG-38.*

mutations are produced in the human population by radiation or chemical mutagens will be answerable in detail. Hence, the sequencing effort is fundamental to the Environmental Protection Agency, the U.S. Department of Energy, the National Institute for Environmental Health Sciences, the National Cancer Institute, and the National Institute for General Medical Sciences, but is beyond each institution individually.

The ultimate intellectual challenge and goal of a DNA-sequencing project is to deduce man from the sequence (or show definitively that this cannot be done). Progress toward this goal would require that the mechanisms for the control and programming of gene expression be understood, that the amino acid sequence of protein gene products be systematically deduced from the corresponding DNA sequence (as can be done), that the three-dimensional structure of proteins be deducible from the sequence, that protein function be deducible from tertiary structure, that cell structure be deduced from the sum total of properties of cellular macromolecules, that human anatomical development be deduced from cellular behavior, and that genetic disease be describable purely from knowledge of sequence errors, etc. Although the challenge to deduce man from DNA sequencing will persist and be the subject of long and interesting arguments, progress toward reaching the more exotic goals listed above will be slow. However, most of them will be met, we believe, in the next or the following century. While DNA sequencing will ultimately provide much of the basic data required for a human protein database, decades are too long to wait. Hence, we currently have the absolute necessity for a parallel program to characterize human gene products (i.e., proteins).

The second major objective must be to characterize and catalog cellular proteins, the precision-made molecular machines underlying nearly all cell functions. In fact, for the near future we propose that the Human Protein Index Project<sup>1,2</sup> be *the* major effort, both because of the many short-term applications and because the feasibility studies for it are now almost complete.<sup>3</sup>

Estimates of the number of structural genes in man vary, but generally range from 30,000 to 50,000. Extracellular proteins are generally modified post-translationally, while a fraction of intracellular proteins are processed to yield two or more different forms. Therefore, any attempt to characterize and catalog human proteins systematically must deal with more than the number of structural genes. It is estimated that approximately 10% of structural genes are turned on in any

one cell type; hence, if one cell type is to be analyzed, 3,000 to 5,000 proteins, plus modified proteins, must be dealt with.

Recent improvements in high-resolution, two-dimensional electrophoresis<sup>4</sup> give theoretical resolutions as high as 40,000 proteins, suggesting that the Human Protein Index Project is feasible.

Several ancillary techniques have been evolved that allow further study of the very small amounts of proteins separated on individual gels. Using submilligram starting samples, we can now identify groups of proteins that are coregulated, do global amino acid determinations on large numbers of proteins simultaneously, determine relatedness of wild-type proteins and genetic variants by mapping of partial digests, identify proteins immunochemically, produce antibodies to each protein, and do sufficient sequencing to enable synthesis of DNA probes so that the genes can be cloned and the proteins in question produced in quantity. In addition, nondenaturing systems are being incorporated into the mapping systems to allow histochemical methods to be used for enzyme identification. These methods require very small amounts of starting material; for example, high-resolution maps can be made using the proteins from a single frozen section.<sup>5</sup> In addition, new very high-resolution centrifuges are under development for precision cell fractionation to allow every protein detected to be assigned an intracellular location. Proposals for the development of such a system are now under consideration by the U.S. Department of Energy and are based, in part, on gas centrifugation technology developed for uranium enrichment.

As part of the Human Protein Index Project, a library of both monoclonal and polyvalent antibodies will be required so that the distribution of each protein in cells and tissues may be confirmed and so that clinical tests for the proteins that vary in disease can be rapidly developed. Parallel programs to develop comparable, identified protein data bases for mice, rats, corn, and wheat (among others) will ultimately be required. Comparative studies will also be needed to attempt to resolve many interesting problems in both human and animal evolution. It is, in fact, difficult to imagine very much research in biochemistry, molecular biology, genetics, medicine, or in many areas of zoology that will not ultimately contribute to the two main objectives suggested here.

Large national efforts do not generally arise by consensus of the scientists and technologists concerned. Neither the space nor nuclear energy programs in their present forms would have been ap-

proved by the scientists ultimately involved, if that vote had been taken before the programs were originally established. Hence, the fundamental decision to establish a DNA sequencing or a Human Protein Index effort is an almost purely political one. C.P. Snow pointed out that, in the past, most decisions in science and technology were made in secret, unknown to the majority of those ultimately affected. (He also pointed out that in government, scientists should be "on tap but not on top".) Secrecy relative to very major decisions such as those suggested here was never desirable and is no longer possible. If, as George Keyworth (Director, Office of Science and Technology, The White House) has suggested, biotechnology is the most important high technology for the future of the United States, and if (as nearly all major countries have concluded) future economic progress depends on mastery in high technology, then it is essential that the ideas described here have the widest possible discussion.

The transition in physics from small-scale efforts to large-scale problem solving occurred under wartime pressure, while the parallel transition in space exploration was largely driven by Sputnik and defense requirements. A similar transition and redirection may now be required if the United States is to win the so-called biotechnology race. Whether this transition can be made for nonmilitary, humanitarian, and economic reasons provides an interesting test of our civilization.

The policy we propose is to organize biomedical and biotechnological research around two major, understandable, and technically feasible very long-range goals, which can logically encompass nearly all present program objectives and which can also support the engineering and development now required to master biology and biotechnology. Few existing genetic engineering and biotechnology companies will fail to see at once where they might fit in.

To be realistic, a national biotechnology program must be planned for at least a twenty-year period, with major milestones spelled out for the

first eight years. The average length of time between innovation (either administrative or scientific) and success or failure in achieving an end result or use appears to be at least five years.

An eight-year span, therefore, would give three years for start up and time for at least one innovation-application cycle.

The best guides to the organization and philosophy of sizable multidisciplinary efforts remain the books by Webb, who organized NASA,<sup>6</sup> and Weinberg,<sup>7</sup> who supervised the maturation of the Oak Ridge National Laboratory under the Atomic Energy Commission.

Success or failure will *not* be solely dependent on our own efforts; we have competition. A day of judgment lies ahead. There is the very real possibility that, within a decade, biomedical scientists and technologists in the U.S. may join their fellows in the electronics and automotive industries as spectators. Then neither policy nor programs will help. The U.S. biotechnology die is being cast now.

**References**

1. ANDERSON, N.G. and ANDERSON, N.L., "Molecular anatomy," *Behring Inst. Mitt.* 63, 169-210 (1979).
2. ANDERSON, N.G. and ANDERSON, N.L., "The human protein index," *Clin. Chem.* 28, 739-748 (1982).
3. ANDERSON, N.G., and ANDERSON, N.L., "The human protein index project and the molecular pathology data base," *Medical Laboratory* 11, 75-94 (1982).
4. ANDERSON, N.L., NANCE, S.L., PEARSON, T.W., and ANDERSON, N.G., "Specific antiserum staining of two-dimensional electrophoretic patterns of human plasma proteins immobilized on nitrocellulose," *Electrophoresis* 3, 135-142 (1982).
5. GIOMETTI, C.S. and ANDERSON, N.G., "Muscle protein analysis, III: Analysis of frozen tissue sections using two-dimensional electrophoresis," *Clin. Chem.* 27, 1918-1921 (1981).
6. WEBB, JAMES E., *Space Age Management: The Large-Scale Approach* (McGraw Hill, New York, 1969).
7. WEINBERG, A., *Reflections on Big-Science* (MIT Press, Cambridge, Mass., 1967).

Draft #1  
March 18, 1986

A

PROGRAM FOR LARGE-SCALE ANALYSIS OF NUCLEOTIDE SEQUENCES  
(P.L.A.N.S)

Contributions to the Development of a Position Paper on the relationships between current and future DNA sequencing technology and the mission and plans of the Office of Health and Environmental Research of the U. S. Department of Energy

Norman G. Anderson  
and  
N. Leigh Anderson

Proteus Technologies  
12301 Parklawn Drive  
Rockville, MD 20852

It is now technically feasible to sequence the entire human genome, to determine where the coding regions are and to index the genome to the set of all gene products (RNAs and proteins). As a result the sequences of all proteins will be known, it will be feasible to find the root cause of all of the estimated 3,000 human diseases, the structural basis of cancer as a somatic genetic disease, the genetic basis of aging, and to provide a wealth of new information on the effects of radiation and chemical mutagens on the human genome. It will also be possible to begin to write the program of human development in detail, once the complete sequence (i.e., the PLANS for MAN) are known.

The Program for Large-scale Analysis of Nucleotide Sequences (PLANS) has far-reaching implications. No major research program in biological research is unaffected by it. Without links to PLANS, and support of it, programs in nearly all institutes of the NIH including those in cancer, human development, genetics, and aging, and studies in radiation research and birth defects are not credible. Defense of indirect approaches to core human health problems will be difficult when direct, and in many cases final answers can be obtained. As the technology for PLANS develops, major revisions of programs in agriculture, animal breeding, and biotechnology generally will be required.

Many forcing functions are at work to drive human genomic sequencing. International competition including the emergence of a major Japanese large-scale sequencing effort, and the central role of the PLANS effort in identifying new genes from which diagnostic and therapeutically useful products may be derived all suggest a major new U.S. sequencing initiative. The drift of the economic fruits of high technology abroad is of central national concern. The manufacture of automobiles, consumer electronics, and micro-electronics has moved increasingly overseas. It is evident that the Program for Large-scale Analysis of Nucleotide Sequences will play a major role in attempting to reverse an economically dangerous trend by retaining leadership in this key area of biotechnology in the United States.

Major new science based programs, once feasibility is demonstrated and need identified, arise from political decisions, which are in turn driven by public expectations. An initial task, therefore, is public education and public discussion. There are no "end runs" by which large budgets can be quietly obtained, and new large programs initiated. The elements of the required education process are symposia (to justify the program in detail to the scientific community), workshops for science writers, popular articles, TV specials, interviews - in short, all those activities which we associate with early AEC and NASA success. The objectives are simple: To demonstrate to each citizen that the human genome can be completely sequenced, and that his own future health and part of his economic well being depend on the project. In addition he must be shown that an entire new universe is now open to exploration and discovery, that exciting milestones have already been identified, and that everyone will share in the excitement.

If the case made to the public is to be credible, the supporters of the PLANS project must accept that funding involves choice between competing programs, and must ultimately be willing to challenge directly NASA, the superconducting supercollider, some existing health programs, and many programs in the military. There is no "new" money.

Role of DOE

Once public expectations provide the necessary motivation, what next? All previous large science-based programs have grown from existing roots. There are many roots for PLANS, and the question is, which existing root can support large growth? If new legislation is to be passed, who gets the funds? Almost invariably they go to the agency already having the mission, the most credible program in being, the courage and leadership to announce aims and goals, and the requisite management skills.

We assume that the Office of Health and Environmental Research desires a lead role. The case it can make to the Congress must be based on already having announced the mission to explore large-scale sequencing, on already having in its laboratories major segments of the program (for example GENBANK), or having the management flexibility to move rapidly by integrating commercial contract activities, academic research, and programmatic research in National Laboratories, and by having unique talent and facilities.

It would appear that the very first order of business is to fold GENBANK, chromosome sorting, support for the Human Gene Library, and work on the closely allied Human Protein Index into DOE as line items, and then to announce the objective of sequencing the entire human genome, spelling out the details of the mission, including short term objectives.

PLANS and Doe Missions

DOE, Radiation Biology, and Genetic Toxicology:- DOE owns radiation biology and the genetic toxicology of energy-related pollutants. It is estimated that approximately two billion dollars has been spent on radiation effects to date, and a very large and valuable body of information has been accumulated. The major and overriding reasons for public non-acceptance of nuclear power and nuclear energy, and the major concern with nuclear weapons are obviously human health effects. While numerous and credible extrapolations from animal models to man have been made, in no instance has a human radiation-induced mutation been identified, the mutant protein characterized, and the mutant gene sequenced. Hence a large number of central questions, some of which are listed below, remain unanswered. The challenge is therefore to extend present knowledge, especially of human radiation and chemical mutagen effects, to the DNA level. To answer the central-mission questions of the Office of Health and Environmental Research of the U. S. Department of Energy, it is necessary to sequence the entire human genome. This goal should be

stated as the major objective of OHER.

Basic Radiation Biology and Genetic Toxicology Questions:- Study of the genetic effects of radiation began at the whole organism level and proceeded to the cytological level with the study of chromosome breakage, translocation, etc. Genetic effects which were below the level of the light microscope were termed "point" mutations, and the exact nature and dimensions of these events remains obscure. Point mutations range from translocations or deletions involving several genes down to single base substitutions, and now require explicit redefinition.

There is no table or graph giving the sizes of deletions in mammalian DNA as a function of dose rate, linear energy transfer, stage in the cell cycle, or position on the chromosome. One does not know in man the exact nature of the initial radiation damage prior to repair, and the characteristics and details of repair as a function of the spectrum of radiation injury. There is no map indicating DNA segments which are truly nonsense and can be modified as to sequence with impunity, sequences which can be modified to yield null variants, and sequences of base pairs whose exact preservation is essential to survival.

Different genes, different parts of genes, different cell types, different stages in cell division all can exhibit large differences in radioresistance. Some apparent resistance is due to repair, some to elimination of injured cells, and in other cases the actual sensitivity of DNA itself could vary due to differences in its physical state and degree of condensation. We conclude that a very large fraction of the core questions concerning energy-related human health effects can be answered by large-scale genomic sequencing.

Thousands of individuals now exist at Hiroshima and elsewhere who have been either exposed to fairly high doses of radiation or are descendants of those who have been exposed. Using tissue culture techniques, cells from these individuals and their parents may be cultured and preserved in a frozen state so that their DNA may be examined, and the nature and extent of genetic damage determined in detail.

The limiting factor in human radiobiology is the rate of DNA sequencing.

We propose subsequently the organization of a contract laboratory, which may be called the MbL, dedicated to the modest objective of sequencing one megabase of DNA per year using the best available technology. In parallel with ongoing sequencing, MbL would develop automated and/or robotic systems for larger scale genomic analysis together with cost and design studies aimed at defining a laboratory to analyze one gigabase of DNA each year (GbL). GbL could then be repeatedly reduplicated if necessary, or could serve as the test bed for the development of one or more even larger laboratories. The goals and objectives would initially lie entirely within the mission of the U.S. Department of Energy, as would all funding, management,

and review. Possible enlargement of the mission to encompass human genetic disease, cancer, ageing, birth defects, evolution, sequencing of DNA from bacteria and viruses, production of probes for clinical use, should be widely discussed, but implementation should be deferred until the DOE mission-oriented objectives are being met.

The Test Segment:- Both radiation biology and genetic toxicology could be revolutionized if one large segment of human or mouse DNA could be routinely, easily, and repeatedly sequenced. It would be best if this segment were obtained from the Y chromosome since segments obtained from other chromosomes would be heterozygous. It is proposed that the activities of MBL be limited almost entirely to the repeated sequencing of one DNA approximately 10-50 kilobases long of human origin, and a similar segment of mouse origin. If possible, these segments should be made with restriction enzymes which cut in the middle of two structural genes, in regions which appear to be evolutionarily invariant, thus avoiding the problems which might arise if the ends are in capriciously varying regions.

By defining initial objectives which are well within DOE's mission, and by starting the projected without competing with grants, opposition can be minimized.

We conclude that DOE has many parts of the PLANS project within its grasp, needs to quickly fold them into its support structure, complete preliminary organizational studies, and announce their intentions to assume a lead role in genomic sequencing.

#### Organization of PLANS

It is suggested that an overall management task group be organized by DOE including DOE staff and individuals from academia and industry, with a full time executive chairman. Reporting to this task group would be (1) the Sequence Objectives Committee (largely geneticists), and (2) the Sequence Strategy Committee (largely molecular biologists, computer, AI and instrumentation specialists including some from industry). The turnover rates in top management suggested by Webb for NASA should be adopted.

At the outset an inventory of DOE programs which are relevant should be prepared.

One of the first problems to be considered is the use of National Laboratories vs industrial contracts. It is strongly suggested that in a highly competitive international environment, and with many proprietary processes and systems already in existence, that a firm policy in favor of competing industrial contracts to operate MBL, and for sequencing technology development be established. It is absolutely essential to use the best technology, and to entrain the best groups and minds in this effort. The efforts of industry can be utilized through contracts, and the best university minds can be obtained through consultantships. Problems which can be solved by neither of the above, or are of no interest to them, or where the best minds are already found in National Laboratories should be in



National Laboratories. It is essential that the sequencing project not become a welfare program for ailing Biology Divisions in DOE. The lesson of redirection is that it does not work. It is also essential to have industrial backing for the entire venture.

IF THIS PROGRAM IS TO BE A KEY VEHICLE FOR MAINTAINING U.S. LEADERSHIP IN BIOTECHNOLOGY, THEN AS MANY ASPECTS OF THE WORK AS POSSIBLE SHOULD BE THE PROPERTY OF U.S. FIRMS.

The ability of DOE to carry through this program with the end result (i.e., the sequence) being in the public domain, but with the bulk of the technology remaining private will be a key element for continued public support. The bottom question is, can the U.S. collectively compete with Japan Inc.

#### Organization of the Megabase/Year Laboratory (MbL)

Among the first contracts let should be one or more prime contracts for a sequencing laboratory scaled to sequence one megabase of DNA per year, and in parallel to develop automated and/or robotic systems for speeding up analysis and for reducing cost. Separation of development from actual process is often fatal, hence the inclusion of both activities under one contract. The technical objective is to develop the procedures and systems to be used in the larger one gigabase/year laboratory (GbL) for whole chromosome sequencing. Subcontracting should be extensively employed by MbL to be sure that the best U.S. technology is used.

Mutations are rare events, and the credibility of a sequencing approach to mutational studies depends on showing that sequencing can be done reliably and with an error rate well below the background mutation rate.

Hence Task 1 for MbL is to determine the noise level of the entire DNA isolation and sequencing operation by isolation and sequencing the same segment repeatedly. If the error rate, with existing procedures, is higher than the expected background mutation rate then Tasks 2 to 4 must be deferred until the basic analytical problems have been solved, and acceptable reproducibility has been demonstrated.

Task 2:- The second task is to determine the differential mutability of different parts of the chosen test segment. This would be done by repeatedly sequencing the same DNA segment isolated from different individuals. Are there nonsense regions which can accumulate background mutations with impunity? Can the length of intergene regions vary in a capricious manner? Do some base substitutions occur more frequently than others? Can genetic distance be measured by simply comparing non-coding regions? Are there constant or restricted variability non-coding and non-control regions which have to do with chromosomal organization and structure?

Task 3:- Begin the systematic analysis of the effects of ionizing radiation as a function of linear energy transfer, dose, dose rate,

dose fractionation, cell type, and length of recovery. Comparative studies would also be initiated using chemical mutagens on human cells in culture or on mice.

Task 4:- In parallel with biological studies, develop both through subcontracts and in-house R&D robotic and/or automated systems of analysis, sample preparation and management, and demonstrate that each new system is superior to methods then in use doing the tasks outlined.

Task 6:- Initiate design of the those portions of GbL not already under development (waste disposal and other problems associated with operating at a larger scale).

MbL should be set up immediately by diversion of existing funds to commercial contractors to obtain operational and personnel flexibility. If there are any serious problems, a competing parallel contract should be let. If the overall problem is deemed very urgent, then three contracts should be considered. The actual operations may be distant from, near, or actually in a National Laboratory. Extensive use must be made of the best personnel, groups and concerns now available. The MbL laboratory will set the tone for the technical level of the entire operation, and must be "state of the art" throughout.

Note that MbL may evolve into an efficient unit which can be repeatedly reduplicated in universities and institutes to work on selected DNA sequence problems. The MbL unit would include sample management and sample preparation systems and programs, sequencing machines, and computing systems, and MbL units could be scaled up by factors of up to 10 if necessary.

#### Ancillary DOE Programs

The objective of the PLANS operation is information (i.e., sequence data). Sophisticated analysis of large sequences will be an entirely new field in biology, and will, we believe, lead to the establishment of a new and central discipline of Genomic Theoretical Biology. Theory has played almost no role in biology previously. We believe this will change radically as large scale computerized genomic analysis develops. Provision must be made from the very outset for development of genomic theoretical biology through grants, graduate student support, access to large computers, and arrangements for personnel exchange. Specifically Walter Goad and associates should be line item DOE programs at the expense of almost anything else. Chromosome sorting, the physiology of chromosome condensation, techniques for producing non-aggregated chromosome suspensions, high-resolution centrifugal methods for chromosome separation, production and maintenance of suitable cell lines - all these should be DOE-supported programs.

GENBANK is now overwhelmed with data. The published information available through it is widely used and highly valued by the scientific and industrial communities. It too should be made a DOE

line item as soon as possible. In addition, distribution of cloned fragments from sorted chromosomes should have DOE support. The National Gene Library Project is already partially DOE supported, and the support should be enlarged.

#### Gene Identification and Indexing

Genes are named by the proteins they make, or the effects mutations of them produce. In the end it is essential to face up to the problem of providing a protein index which is also an index of all structural genes. Advances in two-dimensional electrophoresis now allow thousands of proteins to be separated and characterized, and a large fraction of them partially sequenced. This sequence data can be reverse translated to provide identification of the genes producing the proteins. Hence the technology now exists for linking spots to genes in a systematic manner. In addition, antibodies can be routinely produced against proteins recovered from 2-D gels. These antibodies can in turn be used to isolate undenatured proteins for functional studies, i.e., to find out where known enzymes are in patterns. In addition, the antibodies can be used to determine the intracellular location of each protein, and the cell types in which each occurs.

Genomic sequence data becomes valuable as it is linked to masses of existing biological data. The sequences for structural genes should therefore be linked to the highest resolution protein analytical system presently available (2-D electrophoresis) which system is now in turn linked to a data base including protein names, activities, sequences, disease-associations, and literature citations (The Human Protein Index).

Note that 2-D electrophoresis offers the possibility of detecting point mutations at an extraordinarily large number of loci, and hence over large regions of coding DNA. One 2-D pattern can serve to "report" single base substitutions in more than one megabase of DNA! Protein variants seen on these gels can be partially sequenced, and the mutant genes cloned and sequenced. Hence by combining 2-D electrophoresis, protein sequencing, and gene cloning and sequencing the entire system can be made to yield important experimental results very quickly.

The Human Protein Index project therefore will also provide the index for GENBANK.

#### GbL - Phase I.

The objective of GbL is whole chromosome sequencing. It is evident that GbL could continue on doing purely DOE programmatic studies, examining the radiobiology of the entire human and mouse chromosomes and ultimately genomes.

Phase I planning and initial program expansion should proceed in parallel as was done in the early Manhattan Project.

Task 1:- Organization of GbL DNA sequencing operations. Based on the experience of MBL, sequencing at approximately the megabase per year level without engineering development (but in close collaboration with MBL) should be started. The samples to be analyzed would be those chosen collaboratively as of most medical importance. Some time and effort will have to be expended initially on site choice, on selection of support personnel, and on arranging the base for what will become a relatively large operation.

Task 2:- Organization of collaborative studies. Many competent investigators have identified restriction fragments which are near to or include clinically important genes, have probes for isolating these genes, or actually have cloned the genes and need to sequence them. Using accelerator management experience, and with the inclusion on the GbL staff of specialists or curators in different areas (diseases, chromosomes, restriction fragments etc) it should be possible to organize these studies efficiently. Only a few percent of the human genome codes for proteins, and only a few percent of the coding region is concerned with important genetic diseases. Every effort should be made at the outset to solve as many human genetic disease problems as possible, or to move them into propose clinical settings with the proper equipment. Collaborative studies should be planned as an ongoing activity but not as the central effort.

Task 4: Planning the Core Activities of GbL. The product of GbL is information. Massive sequencing is required to obtain that information. While sequencing of a single human genome may be the initial stated objective, the long term objective is the sequencing of many human genomes, followed by the complete sequencing of many other genomes as well in decades to come. Massive data storage will be required, and computing systems not now envisioned will be required to obtain from sequence data the types of information we believe it to contain. Obsolescence, staff turnover, and acquisition and assimilation of new concepts, procedures, and ideas will, along with funding, be central problems for the planning staff. It is suggested that contracts be let with university groups to explore further exotic alternative approaches to large scale sequencing, including development of new restriction enzymes, new separation methods applicable to large fragments, new mapping and ordering methods etc - all with the express intent of incorporating the successful ones into GbL. Originators should be retained as consultants. As the ordering and sequencing philosophy and approaches become clear, and the ratio and relationships between university and industry sequencing and the GbL efforts is seen, the sequencing effort at GbL may be increased through collaborative efforts. As the "interesting" sequences are done, and interest in actual sequencing diminishes in the academic community, a gradual shift is foreseen to reliance on a central facility to drive the task to completion.

Task 5:- Support. MBL is viewed as an intramural program which can proceed from largely internal decisions within DOE, does not produce interagency problems, and justifies DOE's mandate in the sequencing area. GbL is big science, and can only succeed as such. GbL will require major decisions in full public view. There is zero chance

12

that GbL can succeed to a high level of funding in a short period in this era without very high visibility and a willingness to compete for funds with other big budgets in the public arena. Hence organization of a public education group outside the government is essential as well as one inside GbL. MBL and the GbL therefore should be the subjects of NOVA and as many other public television programs as possible, should be discussed widely, be presented to the Congress, and be described in the press. The objective is to provide detailed correct information and to build expectations for what we know can be done. Expectations can be built without creating opposition. Once the public is fully aware that the entire human genome can be sequenced, that all human genetic diseases can be fully understood (and some of them treated), that cancer is a genetic disease of somatic cells, that aging may be preprogrammed and partially preventable, that many previously undiscovered gene products have therapeutic uses, that that limiting factor in the promised genetic engineering revolution is finding new and useful proteins to make - then massive support for the PLANS program will emerge. Public discussion also converts many unthoughtful skeptics before they can do damage.

It is essential that MBL not compete with grant support at universities. Rather every effort should be made to point out instances where the results from PLANS will be useful to individuals with NIH and DOE grants. By the time that GbL is in operation, it is believed that the concept of large central facilities will have been accepted, and that a new constituency of data users will have evolved.

The public (and correct) image is of a program which has invented itself, exist now naturally in DOE, will continue to grow, and can benefit nearly everyone if fully supported.

Information contained in the sequences of DNA in the 46 chromosomes of human somatic cells constitute the PLANS for MAN. Phase II of GbL is therefore the expansion of this effort up to the full gigabase per year level.

### The Transition to Big Science

The unique features of so-called "big science" which are relevant to the PLANS project are:

1. Simple, well defined, widely understood goals in the national interest.
2. Cohesive support from academia, industry, and one or more government agencies (even where a new agency is ultimately organized), i.e., there is a constituency for the project.
3. Spokesmen and administrators completely committed to the project who place program success well above personal credit.

4. A well thought out plan which meets the program objectives and the interests of the major participants.
5. Active management (vs passive management of investigator-initiated research).
6. Innovator-translators who can subdivide major problems in biology including biochemistry and genetics into small tractable efforts and define them in such a manner as to interest and effectively utilize specialists in other disciplines including chemistry, mathematics, computer sciences including graphics, and many phases of engineering and construction. Lack of these key individuals accounts for many failures in industrial biotechnology.

GENOME WORKSHOP AGENDA

MONDAY MARCH 3

		<u>Speaker</u>	<u>Room</u>
8:00 - 9:30a	Breakfast and Introduction	Mark W. Bitensky Frank Ruddle	Mesa C
9:30 - 11:00a	Workshop I: Technology		
	1. Sequencing	David Ward	Aspen
	2. Chromosome Isolation & Cloning	Larry Deaven Richard Gelinias	Juniper
	3. Ordering	Hans Lehrach	Mesa B
	4. Computation	George Bell Jim Fickett Christian Burks	Mesa A
11:00 - 11:15a	Break		
11:15 - 12:00p	Present Viewgraphs & Discussion of Workshop I	Richard Gelinias David Ward	Mesa A
12:00 - 1:30p	Lunch and Business	Frank Ruddle	Mesa C
1:30 - 2:30p	Break		
2:30 - 4:00p	Workshop II: Benefits & Liabilities		
	1. Clinical and Sociological	Tom Caskey David Comings	Aspen
	2. Basic	Sherman Weisman Robert Moyzis	Juniper
	3. Economic Impact	Mark W. Bitensky Frank Ruddle	Mesa B
4:00 - 5:00p	Present Viewgraphs & Discussion of Workshop II	David Comings Sidney Brenner	Mesa A

WORKSHOP AGENDA CONTINUED MARCH 3

		<u>Speaker</u>	<u>Room</u>
5:00 - 5:30p	Break		
5:30 - 7:00p	Dinner		Mesa C
7:00 - 8:30p	Workshop III: Model of the Enterprise/ Strategies and Approaches/Costs .		
	1. Map Now & Sequence Later	Charles Canter	Aspen
	2. Sequence Intensively with "Special" Foci or Randomly	Frank Ruddle Fred Blatner	Juniper
	3. Coordination/Integration	Walter Gilbert	Mesa A
	4. Cost Estimates	Christian Burks George Church	Mesa B
8:30 - 10:00p	Present Viewgraphs and Discussion of Workshop III	Walter Gilbert Walter Goad	Mesa A



## GENOME WORKSHOP AGENDA

TUESDAY MARCH 4

		<u>Speaker</u>	<u>Room</u>
8:00 - 9:00a	Breakfast		Mesa C
9:00 - 10:30a	Workshop IV: Participants, Funding, Unfinished Business	Norman Anderson Frank Ruddle	
	1. Participants: Organizations and Individuals	Norman Anderson	Aspen
	2. Funding Sources	David Smith	Juniper
	3. Coordination & Intergration Part II	Walter Goad	Mesa B
	4. Open		
10:30 - 12:00p	Summary Plenary Session Document Completion	Frank Ruddle	Mesa A
12:00 - 1:30p	Lunch		Mesa C

# Los Alamos

Los Alamos National Laboratory  
Los Alamos, New Mexico 87545

64

DATE December 23, 1985  
IN REPLY REFER TO M881  
MAIL STOP LS-DO-85-1.14-132  
TELEPHONE 505-667-2690

Dr. Charles DeLisi  
Office of Health and Environmental Research  
ER-70, F-208, GTN  
Washington, D.C. 20545

Dear Charles:

## Prologue:

We have begun to seriously examine how we might assist in the process of evaluating and supporting your proposal for an initiative that aims at sequencing the entire human genome. There can be little question that this is a challenging and vital enterprise, aspects of which resemble in scope and impact initiatives developed under the aegis of high energy physics. Moreover, the important results for human health and the consequences that must accrue from such an effort are far more immediate and significant than many "Big Science" activities of far greater cost. We do not question the significance of such initiatives, but rather emphasize that sequencing the human genome is an undertaking of protean significance and compelling importance.

## Subjects Addressed In This Letter:

Here we examine the extraordinary and tangible benefits arising from this initiative; attempts to accurately estimate cost and time for completion of the work; the need for advanced instrumentation and automation of the sequencing process; a variety of approaches that could be taken; and ways in which to coordinate and orchestrate the program from a central location.

It seems more than appropriate that such integration would be closely associated with GenBank functions. We note that theoretical biologists in GENBANK routinely undertake analyses of sequences and do research into new algorithms for sequence analysis. Furthermore, a preferred experimental approach to this work could be expedited substantially by the National Laboratory Gene Library Project (at Los Alamos and Lawrence Livermore Laboratories), which is producing chromosome specific recombinant DNA libraries for the entire human genome.

## Significance and Benefits

It is impossible to anticipate at this point in the history of molecular genetics, all significant consequences of this initiative. Clearly we will have defined the organization of the whole human genome. Also we will have developed a powerful informational infrastructure in order to understand more perfectly the mechanisms utilized in the coordinate expression of gene families i.e., proteins, which by themselves are without significance, but when accompanied by a matched set of gene products serve as components of a

protein cascade, which subtends a particular function in a specialized cell. For example, proteins involved in muscle contraction; proteins involved in the synthesis and degradation of glycogen and proteins involved in insulin or steroid synthesis, and a multitude of other "matched sets".

Yet another benefit would be having access to (it would probably take a long time to work it all out) the complete interactive network of all the genes or subgroups of genes in a 'high level' organism. Though there are currently several complete genomes known, they are for extremely parasitic organisms (viruses) that rely extensively on the host gene network. Although *E. coli* will probably be sequenced within the next decade or so, it would obviously not suffice for understanding animal genes and their regulation. In summary, the theory of the operon told us (many years ago) that genes are interrelated, often in quite extensive networks. However, there is no single example of a relatively free living organism where we can say that we have the complete interdependent network at hand to analyze or understand.

Having the entire sequence of the human genome will have profound implications for the understanding of evolutionary processes; the appearance and disappearance of genes, which are either soon to be deployed or are no longer needed. Also evolutionary questions relating to species proximity and evolutionary pathways will be more readily resolved. It should also be emphasized that this sequencing initiative will have a very important impact on our understanding of development, embryogenesis and specialization of undifferentiated cells.

The sequencing will have profound impact on our understanding of the role of oncogenes in cancer and the abnormalities of gene expression that are associated with cancer.

Additional benefits in cost and time would derive from using the complete sequence as a reference for which one would need minimal information to readily access some desired portion of the genome in a relatively localized fashion. An example: suppose someone notices an unusual protein which appears on a gel in a clinical or basic study — determination of the amino acid sequence of a relatively short portion of the protein would allow one to scan the translated human genome data for regions of similarity or identity, thus narrowing in on an already known or previously unknown gene(s). This would eliminate the need for all the molecular biochemistry involved in making probes and attempting to isolate genes. If an investigator wants to acquire and amplify a copy of the gene, it would be easy to go from the sequence data entry (which would in all likelihood cite a fragment in an established library such as the Gene Library) to the place where the fragment corresponding to that gene is stored and available.

Yet another benefit, perhaps more abstract, would arise from answers to questions concerning the richness of sequence information, and how unique individual sequences are. It would provide an important base for calculations that are often used in, for instance, designing probes, estimating expression levels (through codon usage), etc. It would tell us whether or not there are chromosome-specific sequences; and the same for other levels of "locale".

The proposed sequencing will have a profound impact on questions of prenatal diagnosis, genetic counseling, inborn (genetic) errors of metabolism, and

the identity of various "abnormal" or absent genes (e.g., cystic fibrosis, Pompe's disease) and an even more profound impact on the whole question of genetic engineering in terms of how and where in the genome to introduce nucleotide sequences for the correction of genetic abnormalities.

Moreover, sequencing the entire human genome will provide a powerful understanding of the architecture that governs distribution of repetitive sequences and the potential role of both repetitive and intervening sequences with regard to gene structure and gene expression. The structural implications of the entire sequence ought to help us to understand ways in which the genetic information is made available especially with regard to the unfolded forms deployed in interphase, or when the chromosomes are fully condensed, and when the genetic information is being replicated. It is more than likely that a number of surprises (totally unexpected benefits) will emerge from such sequence information. In this regard insights into genetic heterogeneity within and between families or larger population clusters will have profound implications for individualized medical care.

Estimates of Time and Cost

Although the benefits are dramatic and unambiguous the issues of cost and time must be carefully considered. There are a variety of approaches to estimating cost. One can take the average rate at which DNA is sequenced at one of the more advanced laboratories (1000 bases per man-day) and estimate the cost by multiplying the cost per 1000 bases by the total number of kilo bases in the human genome (i.e.,  $3.5 \times 10^6$  kb). A problem with such an estimate results from the fact that it is based on current technology and current approaches all of which may become rapidly outmoded by technological advances. Automation of the sequencing technology will make it more rapid, less labor intensive, and less costly. Also such estimates do not include the need for multiple sequencing passes to accommodate ambiguities and enhance accuracy and the costs for preparing specific DNA segments for sequencing and the costs of organizing these sequences in a linear or overlapping array.

In the 1000 bases/day approach one arrives at the following estimate.

1. 1,000 bases per day sequenced by a "state of the art" lab.
2. 
$$\frac{3.5 \times 10^9 \text{ total bases}}{1000 \text{ bases per diem}} = 3.5 \times 10^6 \text{ man days}$$
3. 
$$\frac{3.5 \times 10^6 \text{ man days}}{350 \text{ man days/annum}} = 10^4 \text{ man years}$$
4. 
$$\frac{10^4 \text{ man years}}{500 \text{ laboratories (2 sequencers/lab)}} = 10 \text{ years}$$
5.  $10^4 \text{ man years} \times 10^5 \text{ \$ per man year} = \$10^9 \text{ (1 billion dollars) over 10 years.}$
6. Given ambiguities with respect to error rate and potential technical problems ameliorated by probable technical advances the \$1 Billion/10 year estimate should more properly be cast as 1-3 Billion over 9-14 years (at the outset).

7. We should plan a DOE sponsored workshop (or workshops) to begin to examine the best approaches, players, centers and technologies.

8. Such funding might well involve National and International participation including current adversaries, such as the Soviets, in order to provide a DNA centered mechanism for international cooperation and reduction in tension.

9. Such funding might thus be derived nationally (OHER, Howard Hughes, NIH and other foundations) and internationally (UN/WHO participants). There may be useful opportunities here for the nation regarding possible initiatives with the Soviets.

10. The phasing in of the project must occur gradually with careful planning and in ways which are perceived by individual investigators as supporting and non-threatening. The 500 laboratories used in the calculation begins to capture the need for distributing the task over many centers.

#### Need for Advances in Instrumentation and Automation of Sequence Analysis

There is clearly a formidable need for improvement in instrumentation/automation methodologies that are now employed for DNA sequencing. This extends from gel column technologies to the preparation of DNA. It is likely that a variety of new approaches could emerge in the next 5-10 years. The last decade has witnessed an order of magnitude gain in the speed of sequencing! Of course it is a profound and central question as to whether and how the next order of magnitude gain in sequencing speed and cost will be accomplished, e.g., multiplexing. In the context of such sequencing technology, the National Laboratories are one potential locus in which to develop advanced instrumentation both because of personnel, experience, the association with the National Laboratory Gene Library Project and because of the computationally intense nature of the task.

#### Experimental Approaches

At least three approaches have been considered. In discussions with Frank Ruddle we have learned that at a recent Gordon Conference (this past summer) one approach suggested was sequencing an entire human chromosome. This was seen as an attractive "demonstration" possibility. Another possibility that Ruddle favors is to do extensive sequencing in regions of clinically significant RFLPs. Yet another approach that Bob Moyzis has suggested is to use the shotgun approach and to sequence all available clones in the recombinant libraries with the idea that this minimizes preparation time (the clones are already available). This would require an intense retrospective and computational effort to array all domains being simultaneously sequenced. There are various technical approaches and it is agreed that many of these issues (once we have developed an initial position) will benefit from discussion with colleagues at an OHER sponsored conference dedicated to this purpose.

#### Integration

GenBank appears to be a logical organizational unit within which to organize this initiative. The computational base at GenBank is a valuable receptacle for the information that will emerge as well as the perfect tool with which to arrange, and integrate this flow of information. GenBank could also

68

participate in efforts to identify overlaps of sequenced domains and to develop new algorithms with which to explore, study and analyze the total sequence.

Concluding Remarks

In summary, there is considerable enthusiasm for this concept. A note of caution has been voiced concerning uncertainties in time and cost estimates. They may be too sanguine or too conservative. (More data are needed). We find strong unanimity with regard to the need for advanced instrumentation and automation and enthusiasm with respect to the basic importance of clinical and applied spinoffs associated with a project of this magnitude. Finally, there was enthusiasm for your initial point that the Life Sciences have still to learn to address the very effective large scale enterprises that so often have graced and strengthened the arena of high energy physics.

Very truly yours,

Christian Burks

Walter Goad

Jim Fickett

Ed Hildebrand

Bob Moyzis

Larry Deaven

Scott Cram

George Bell

Mark Bitensky

# Los Alamos

Los Alamos National Laboratory  
Los Alamos, New Mexico 87545

## memorandum

TO Mark Bitensky, Senior Fellow, MS M881          DATE March 13, 1986  
FROM G. Bell *GB*    MAIL STOP/TELEPHONE B210/7-4401  
SYMBOL T-DO/86-28  
SUBJECT COMPUTATION ASPECTS OF THE HUMAN GENOME PROJECT

A summary of this subject is attached for use in the workshop report. The costs presented are in rough agreement with the more detailed numbers presented in Christian Burks' section.

GB:lam

Attach. a/s

Sequencing the Human Genome - Computation Requirements

A central facility is required to collect, correlate, store, and distribute the sequence database and associated information (source, function, relation to other sequences and to physical and genetic-map, etc.). It is essential that this be linked in a network with generators and users of the data. The collection, storing, and distribution of the data could be handled using an advanced minicomputer at the network core. However, by the time the database reached  $\sim 10^8$  bases, we believe that a dedicated state-of-the-art supercomputer would be required in order to handle problems of sequence alignment and similarity searches.

Computer memory requirements associated with the database are modest,  $\sim 10^{10}$  sequence bits or  $\sim 10^{11}$  bits overall. This is easily accommodated within current state-of-the-art. For example, the Los Alamos "Common File System" will soon store  $\sim 10^{13}$  bits. Fast disks storing  $10^9 - 10^{10}$  are common.

The rate of data acquisition requires direct input of the sequence and other data into the database and hence an efficient data transfer network coupling data generators to the database. In addition, the community of users must have ready access to the database and to analysis tools presumably over the same network. Facile communication over this network should make it a vital tool for molecular genetics research.

Sequence comparisons for purposes of data assembly do not pose major problems, at least if partially ordered cosmids or other fragments are used in generating sequence data. This conclusion may be reached as follows: A single search for nearly identical subsequences between a "gene" of  $10^3$  bases and GenBank ( $6 \times 10^6$  bases) would take a few seconds on a current supercomputer, e.g., CRAY X-MP/48. Thus aligning a unique sequence with a 100 base overlap to the human genome library ( $3.10^9$  bases) would take a few minutes and many ( $\sim 10^5$ ) such alignments could be done per year using current technology.

Similarity searches allowing for mismatches, insertions, and deletions are much harder and will pose formidable challenges to the designers of algorithms and of supercomputer hardware. Fortunately, current trends towards increasing parallelism in computer design will benefit database

Searches for other



will similarly challenge both computer scientists and molecular geneticists probably for many decades. We believe the network should make available algorithms and computer time for scientists doing computationally intensive work.

Costs of the computing facility and network and management of the database are estimated to average \$10M per year, being about half for equipment and half for people. The rate of expenditure would increase somewhat with the size of the database but an early investment should be made in establishing the communication network, database framework, and in algorithm research. Once the network is delivering a reasonable level of service, user demands may require substantial increases of communication or computing resources beyond that implied by the above estimate.

## FRED HUTCHINSON CANCER RESEARCH CENTER

Area Code 206  
467-5000

1124 Columbia Street  
Seattle, Washington 98104

17 March 1986

Dr. Mark Bitensky M-881  
Los Alamos National Laboratory  
Los Alamos NM 87545

Dear Mark,

Thank you for the chance to attend the meeting about the human genome sequencing project. I thoroughly endorse the idea of sequencing an entire reference human genome within a ten year period as did everyone I spoke with at the meeting. The largest problems I believe are political and social, not scientific. I am convinced that at least two methodologies exist (the ABI machine and the multiplex sequencing approach) to improve today's sequence data acquisition rate of  $0.5 - 1.0 \times 10^5$  bases person<sup>-1</sup> year<sup>-1</sup> by a factor of 20 to 100. This is the sole technical advance necessary to sequence the genome (or the parts which can be stably carried in current vectors, probably 99%) within the tentative time frame of ten years.

I also endorse the preliminary goal of working out a coarse physical map (as with Sfi I and Not I) of the genome. Again, no new technical advances are required; when such a map is integrated with the growing collection of restriction fragment length polymorphisms, medical genetic analysis would enter a new era, when phenotypes based on one (or a few) genetic determinants could be rapidly assigned to a physical locus, if a suitable family group is available for study. I think the possible goal of sequencing a complete complementary DNA library (message library) though useful, would not command the same high priority as the basic map. Arbitrary decisions about the choice of a cell from which to make the library and the notorious inability of the reverse transcriptase to copy certain messages would weaken the general usefulness of such a library. Were a map to be published, individual labs around the world could map cDNAs on their own. The coarse restriction map would be achievable within a few years, would start paying benefits immediately, and would continue to do so for the duration of the sequencing project. The mapping project should begin with total human DNA, not flow sorted chromosomes, since mapping would be carried out by Southern blots, the complexity of the human genome is so great that simplification by a factor of 20 or so is irrelevant to the mapping process.

My recommendation would be to sequence the genome with cosmid libraries on individual chromosomes. This approach would assume one could obtain enough euploid cells from one source to yield sufficient material for the high speed sorters to produce several micrograms of each of the 24 different chromosomes. I think there are several advantages to this approach. First, useful results would start very soon, since existing probes to important genetic loci, translocations, or oncogenes could be followed in. Second, dividing the genome into 24 segments will help the closure problem to a certain degree. Third, the motivation of the teams doing the actual work could be positively influenced with a focus ("we've finished 21p now on to 21q") which could lead to a competitive challenge between different groups. Fourth, it is politically and practically sensible to make use of the expertise of LANL and LLNL at high speed sorting of human chromosomes.

Next, libraries of cosmids would be ordered by any of the several approaches: the restriction 'signature' method of S. Brenner, the conventional mapping approach, or the differential hybridization to random oligonucleotide probe scheme of H. Lehrach. Additional shortcuts and contiguous groups of cosmids could be derived from existing sets of cloned probes (which now number in the thousands), as well as probes derived from the coarse physical mapping studies and the several hundred existing RFLP probes.

Finally, the sequences of individual cosmids would be taken with one of the available methods. Both the automatic sequencing device being developed by Applied Biosystems and the multiplex sequencing of G. Church are close to being 20 to 100 times more efficient than a person working in a lab today. A particularly exciting possibility would be a combination of the multiplex sequencing with chemiluminescent probes and exposure times of a few seconds (D. Ward) and thus the elimination of radioactive tracers and x-ray film. Ultimately such data could be taken directly on a video camera screen as an input device, the pattern of bands stored on magnetic media, and the sequence deduced by computer.

I urge the OHER to 'assert a license' to carry out this research and start active planning. The national labs, as opposed to the NIH would be natural choices because of their excellent computing resources, the high speed chromosome sorting projects, the excellent data communications (ARPA/ET), GENBANK, and the fact that the national labs do have a history of doing active rather than passive research. The NIH has still never organized its own computing and finally set up contracts with third parties (Intelligenetics) to make computer services available (at great and growing cost) to the peer-reviewed community. On the other

hand OHER will have to overcome 1) a lack-luster reputation in the general biology community (though unfair, I am afraid the accomplishments of the National labs are not yet fully appreciated; 2) jealousy from the NIH if it perceives that the funding would compete with their own- better to invite them to help in the early stages; and finally the difficulty of getting new money ( the incremental funding) out of Congress these days.

I would also favor the idea of a central location for the work to begin. I think this would be invaluable to help focus and motivate the scientists involved. As many conferees suggested, the projects could be divided up during the later years or to nurture specific technologies as warranted, but the main work would be expedited in the beginning if it were at one location. I think Wally Gilbert's plan would be quite practical.

I hope this summary will be of use. I enclose reprints on the collagen molecular genetics work from my lab.

Sincerely,

*Rich Gelinas*

Dr. Richard Gelinas  
Associate Member  
Program in Molecular Medicine

## College of Physicians &amp; Surgeons of Columbia University | New York, N.Y. 10032

DEPARTMENT OF HUMAN GENETICS AND DEVELOPMENT

701 West 168th Street

March 6, 1986

Dr. Mark W. Bitensky  
Los Alamos National Laboratory  
Life Sciences Division  
Los Alamos, New Mexico 87545

Dear Mark:

I enjoyed our workshop very much. This letter summarizes my views on some of the scientific and organization issues raised at Santa Fe, particularly focusing on items on which I feel sufficiently close enough to have strong views.

I favor a staged approach to sequencing the human genome. In the first few years I would focus attention on:

1. Macrorestriction mapping the entire genome with PFG electrophoresis, rare site libraries and somatic cell genetics and cytogenetics.
2. Developing optimal strategies for cosmid link up based on the premise that a full macrorestriction map will become available, and then actually implementing these strategies, chromosome by chromosome.
3. Aligning the physical maps on the RFLP linkage map of each chromosome.
4. Developing improved automated DNA sequencing methods.
5. A pilot, larger scale sequencing study, focusing on regions of current interest and aimed at gaining experience in automated data input to a data base, on error analysis, on improved data analysis, and on maximizing cloning efficiency prior to sequencing.
6. Selecting the samples, e.g. mole DNA, for further study.

In the next few years I would concentrate on:

1. Sequencing of the most interesting 10% of the human genome, including immediately important disease gene regions located through the combined linkage and physical maps.
2. Further improvement of automated sequencing.
3. Some pilot studies on physical map and sequence diversity among distantly related humans and within disease pedigrees.

In the last 5-7 years I would concentrate on:

1. Completing those regions of sequencing actually feasible.
2. More elaborate studies on diversity, including some mouse-human comparisons in interesting multigenic disease regions.

In terms of organization, I favor a strong central administrative structure with clout, but physical dispersal of most of the actual project to go to about 10 Centers located at or adjacent to existing institutions. I feel this is both politically and scientifically optimal. I would see these Centers as starting at around 10 people each and growing to 30-50 as each project matures. I am pleased that DOE might be the lead agency in administering the program and it seems highly desirable that the national laboratories play major roles, possibly with Los Alamos continuing to serve as the center for data storage analysis and distribution and for Livermore and Los Alamos involved in the development of automated instrumentation as well as cytogenetics. I believe the optimal sites for most of the mapping and early sequencing are at medical centers or universities where intense interest in the emerging data will keep the project focused on important goals and maintain morale. I would contract most of the less interesting sequencing to companies but I lean towards having intense communication between each contract and a particular Center.

The leadership of the endeavor must be visible, highly interactive and unusually flexible. I suspect that technology will continue to advance rapidly and that many major mid-stream course corrections will be required. A sub-directorate composed of the leaders of the Centers and participating national laboratory groups seems an appropriate governing board with one or two individuals additionally designated as full-time director (and possibly co-director) of the overall project. I believe the overall endeavor is sufficiently attractive that it should be possible to recruit outstanding individuals to head the program and its Centers.

Please let me know how I can be of further help since I support this overall program quite strongly.

Sincerely,



Charles R. Cantor  
Professor and Chairman



CITY OF HOPE NATIONAL MEDICAL CENTER

1500 EAST DUARTE ROAD • DUARTE, CALIFORNIA 91010 • <sup>575</sup>~~(714)~~ 359-8111

DEPARTMENT OF MEDICAL GENETICS  
David E. Comings, M.D., Director

March 11, 1986

Dr. Mark Bitensky  
Los Alamos National Laboratory  
Los Alamos, New Mexico 87545

Dear Mark:

Just a note that you requested of me. 1) I am strongly in favor of the proposal to map and sequence the human genome. 2) I believe that the best approach would be to finish the cosmid libraries for each individual chromosome along with a Not and lambda library of each chromosome. When that phase is complete, then progress with the mapping and aligning of the cosmid clones, then finally, the third phase of sequencing. The first part is apparently already being funded and begun. The second part of mapping the cosmid clones would probably cost much less than the final sequencing. Multiple center support might allow the mapping part to be done without needing to approach any government agency about spending \$500,000 to \$1,000,000.00 for the whole project. Between the Department of Energy, NIH, Hughes, and other institutes, it is possible that this part could be funded without any stirring up of any congressmen or other related creatures. By that time, there is a possibility that newer sequencing techniques would give us a better estimate of the third phase and that would be the time to approach congress for funding.

I would be happy to serve in any capacity, in the future, in any way you might need help.

Best regards,

David E. Comings, M.D., Director  
Department of Medical Genetics

DEC:km

Lawrence Livermore National Laboratory



March 13, 1986

Dr. Mark W. Bitensky  
Life Sciences Division, M881  
Los Alamos National Laboratory  
Los Alamos, NM 87545

Dear Mark:

In response to your request for input about the Santa Fe meeting on sequencing the human genome, we wanted to provide you with some of our thoughts and those of our colleagues at Livermore. Before we do that, however, we wish to thank you for the invitation and congratulate you on hosting and co-chairing an exciting and stimulating meeting. We believe that this meeting has clearly fulfilled the mandate of Dr. DeLisi. We concur strongly with the major points of consensus that emerged there. A summary of our ideas follows and a table giving a proposed time and effort schedule is also offered for your consideration.

We envisage this project as having perhaps four stages: (1) a 2.5 to 3.5 year "prefunding" period, (2) a five year period during which the main project is initiated at about one fifth its projected size and then grows at approximately 40% per year, (3) and (4), two subsequent five year periods with constant effort but changing emphasis. Since it seems certain that significant funding for this effort would not be available until October of 1988 at the earliest, we have about 2.5 to 3.5 years in a prefunding period. With only the preexisting funds or modest additional support, this time could be used very profitably to improve and test sequencing and ordering methods. In addition we could begin, based at least in part on the fruits of the "large insert" phase of the Gene Library project, to produce some partial, ordered libraries of a few chromosomes. This preliminary work should be very valuable in motivating and defining the main project itself.

We propose that the main project be initiated at the level of about 100 FTE, devoted to four major components: (A) ordering, (B) sequencing methods development, (C) computation related development efforts, and (D) administration, planning, and recruiting. During the first five year period, the manpower levels would be increased at about 40% per year to attain a level of 500 FTE in year six. This increase would be almost entirely devoted to bringing up two additional very large areas of effort, sequencing and community interfacing. Approximately this level could then be maintained for the expected 15 to 20 year duration of the project. At the completion of this project, the manpower effort could be re-directed to the mouse or other appropriate genomes.

We strongly concur with the view espoused by Gilbert and others that ordering efforts and the consequent integration of the physical and genetic maps should receive the dominant initial emphasis and be well under way before any significant sequencing is undertaken. This partly reflects our view that the ordered libraries would make the sequencing efforts much more efficient. They would also permit the sequencing work to be more useful, especially in the early years, by being efficiently targeted to regions of greatest interest. The use of cDNA libraries to select cosmids for sequencing should allow the most interesting 1% to 3% of the genome to be sequenced first.

Of course, the RFLP and genetic mapping efforts are currently under way and would be greatly enhanced by information gained from the physical ordering maps. In particular, known RFLPs and genes could be mapped with much greater precision using the ordered libraries which should also be a rich source of new RFLPs. Reciprocally, genetically ordered RFLP's may well play an important role in providing closure for the physical map. The synergistic interaction which will take place between the RFLP linkage and physical ordering efforts may be of particular importance and profundity. While the ordering map provides a critical characterization of a "reference" human genome, the RFLP linkage analysis provides the strongest available tool to investigate the orthogonal dimension of genetic variability and individuality.

Simultaneously, but as a separate effort, we think it critical to begin to develop and validate fast sequencing methods. We suggest that a useful trigger point for initiating massive sequencing efforts



might be the demonstration of sequencing to better than one error per thousand bases at a cost of under ten cents per base (a minimum 50 fold increase over present methods). Because no clearly superior method for either ordering or sequencing has been identified, we support the idea that several approaches should be pursued in parallel.

It also seems valuable to begin considering at the earliest possible point the expansion of the current Gen Bank resource at Los Alamos to accommodate the needs of this project. It is plain that very substantial changes, both qualitative and quantitative, will be required (data base restructuring, algorithm development, parallel/array processing, remote access and facilitated IO methods).

This need relates closely to the need to have active planning and administering efforts in place from the beginning. We concur that a partially distributed structure with unified funding and fairly centralized management, like that proposed by Ruddle, may be the optimum way to organize this effort. We believe, however, that these organizational issues deserve very careful study and feel that it is too soon to be very specific about them.

Interfacing with the scientific community is essential and, as indicated by Brenner, requires up to 50% of the effort devoted to the laboratory and computational work. Our own experience with the Gene Library also suggests this is reasonable. The table reflects these estimates of the costs for this essential part of the undertaking.

We also want to support the views expressed by Norman Anderson, Sidney Brenner and others emphasizing the importance of proceeding carefully in how this effort is brought to the attention of Congress, to the affected scientific research communities, and to the American public. For example, we feel that a program whose announced purpose was simply to "sequence the human genome" might unnecessarily and incorrectly arouse fears of territorial and financial usurpation in the biomedical research community. That is partly why a plan to "order first" is valuable apart from the technical issues involved. An ordered library would be a tremendously valuable, and largely non-competing, resource to virtually every molecular-level investigator in the field. As researchers adapt to the new opportunities the ordered libraries provide, a centralized program which then undertakes, in cooperation with ongoing efforts, to "complete" the sequencing of the human genome, would then be much more welcome.

Along this line we feel the "catch phrase" by which this effort comes to be known is worth designing with some care. Ruddle's suggestion of "The Human Ultimate Resolution Map", while technically correct and appealingly all-embracing, seems unsuitable. We propose instead: "The Human Genome Program". Of course, almost all non-biologists would have to learn how "genome" goes beyond "gene". But just as with the "Superconducting Supercollider", that is part of it's appeal.

Livermore remains very interested in seeing this project through and in being an active participant in its planning and operational phases. We would specifically desire to be involved in the proposed steering committee. The organization and expertise our respective laboratories have developed in the ongoing Gene Library Project could provide an administrative and technical foundation for our involvement in this larger human genome project.

Sincerely yours,

*Anthony V. Carrano*  
Anthony V. Carrano

*Eibert W. Branscomb*  
Eibert W. Branscomb

Proposed Plan				
Task	FTE			
	yr 1	yr 2→5 +≅40%/yr	yr 6→10	yr 11→15
Ordering	55	55	10	5
Sequencing methods	20	20→50	20	10
Computation/ data base	15	20	20	10
Sequencing	0	20→200 hot 1% (cDNA?)	250 next 5%	275 the rest
<u>Community interfacing</u>				
for ordering:	0	5→45	40	40
for sequencing:	0	0→80	110	120
Administration	10	10→50	50	50
<b>Total</b>	<b>100</b>	<b>130→500</b>	<b>500</b>	<b>500</b>

Laboratoire Européen de Biologie Moléculaire  
European Molecular Biology Laboratory  
Europäisches Laboratorium für Molekularbiologie

Dr. Mark Bitensky  
Los Alamos National Laboratory  
Los Alamos, New Mexico 87545  
USA

14th March, 1986

Summary of session on chromosome linkup:

Contributions to the discussion on chromosome linkup and physical mapping techniques were mainly made by myself (H.L.), Sidney Brenner, Fred Blattner, Tom Gingeras and Charles Cantor. There was agreement that linking up the human genome by cosmid libraries was achievable using current techniques, and an essential step in any large scale sequencing effort. To overcome the problems of heterocycosity, the use of "mole" DNA was suggested. Alternatively, sorted chromosomes from cell lines carrying translocations could be used. Depending on the linkup technique, the use of sorted chromosomes could offer advantages, though depending on the scale of the effort, the alternative possibility of a total linkup of random cosmid clones would be quite feasible. Using a technique developed in MRC Cambridge and used for the linkup of the Nematode genome could be used, requiring an estimated effort of the order of 40 to 80 man years, a value which could however be reduced automating some of the steps involved. This protocol requires the growth of individual colonies, cleavage by e.g. HindIII, radiolabelling of the ends by Kleno polymerase, and recleavage with a second enzyme (e.g. Sau3A) generating small cleavage products. These fragments are then sized on sequencing gels, and their length is used to search a database for overlapping cosmids.

As an alternative I have described a protocol to link up genomes under development in our laboratory, which uses hybridisation of cosmids to a large number of different oligonucleotide probes to identify overlapping cosmids. This protocol is currently being developed using the yeast genome as model system. If successful, this alternative protocol should reduce the effort required in a genome linkup by up to 10 fold.

Also discussed were the use of restriction enzymes cutting rarely in the genome in PFG gel electrophoresis (Charles Cantor), the construction and use of special libraries (junction libraries), selected to contain fragments bridging a rare cutting site (Charles Cantor, H.L.), as well as the construction and use of chromosome jumping libraries (Sherman Weissman, H.L.).

ELEANOR ROOSEVELT INSTITUTE FOR CANCER RESEARCH, INC.  
FLORENCE R. SABIN LABORATORIES FOR GENETIC AND DEVELOPMENTAL MEDICINE

4200 EAST NINTH AVENUE, DENVER, COLORADO 80262, TELEPHONE (303) 394-7152

March 17, 1986

Dr. Mark Betinsky  
Los Alamos National Laboratory  
Diamond Drive  
P.O. Box 1663  
Los Alamos, New Mexico 87545

Dear Mark,

I enclose a set of remarks on problems raised in the course of the meeting. I thought it was a highly illuminating and successful meeting. I think the project is extremely important. We will do anything in our power to assist its progress.

With regards,

Sincerely,



Theodore T. Puck, Ph.D.  
Director, Eleanor Roosevelt  
Institute for Cancer Research  
and Professor of Biochemistry,  
Biophysics and Genetics and  
Professor of Medicine  
University of Colorado  
Health Sciences Center

TTP/vst  
Enclosure

## MEMORANDUM ON DR. BETINSKY'S HUMAN GENOME CONFERENCE

Gene sequencing and gene mapping are two different sets of operations and both are necessary for the achievement of elucidation of the human genome. Eventually, the gene mapping will greatly assist the sequencing. But without mapping, sequencing itself will not nearly accomplish all of the contributions to human biology which are to be expected from complete analysis of the human genome.

The gene sequencing procedures available are well known and each of the two major technologies has certain advantages. There is no question but that a concerted effort of the kind discussed here will lead to additional economies and shortcuts in the sequencing operations.

Some discussion is necessary about gene mapping. A variety of procedures have been developed: a) in situ hybridization; b) gene walking; c) the use of a bank of hybrids containing multiple human chromosomes incorporated in a rodent cell so that one can by construction of a square array determine precisely on which chromosome a given gene is located; d) the use of hybrids containing single human chromosomes in a rodent cell; e) the use of single chromosomes isolated by flow cytometry, and applied by means of the dot blot method; f) the RFLP method; and g) combination of the use of repetitive sequences with deletion analysis. If techniques of chromosome microdissection improve sufficiently in their efficiency, those may also be considered.

Gene mapping must be considered at different levels: assignment of the specific chromosome (or chromosomes) to which a given gene

belongs; determination whether a gene family exists; assignment of genes to particular arms of its carrier chromosome; assignment to a region on that particular arm; fine structure mapping; and ultimately exact linkage to its flanking sequences.

In this connection it is also necessary to discriminate between coding sequences, non-coding sequences and repetitive sequences. The latter two kinds of DNA may well be concerned with regulation of gene expression.

The RFLP method produces results which in cases like Huntington's disease cannot yet be duplicated by any other method. There is a very significant body of human genetic diseases and other markers for which, at present, the RFLP method appears to offer the only means for mapping of the genes responsible for the disease or marker in question. However, more than 1400 human genetic loci have been mapped by the other methods mentioned without any organized concerted attack. There is no question that these methods offer a powerful resource. Moreover, by combining elements from RFLP and other methods it seems reasonable to expect even more powerful methods to emerge which may be especially useful in certain kinds of situations.

The point of these considerations is that one must not confine the analysis of the total human genome to one approach alone. Different methods will complement each other. In addition, the confirmations obtained by the use of multiple methodologies are not to be despised.

In the case of repetitive sequences, it becomes necessary to distinguish between a variety of different situations. Dr. Fisher in

our laboratory has demonstrated that repetitive sequence probes isolated from human chromosome 11 produces multiple bands when blotted to the DNA of this chromosome by standard techniques. However, when the Southern blotting procedure was repeated at a series of different temperatures providing different conditions of stringency, it became possible successively to reduce the number of bands which formed until ultimately only a single band remained. Obviously, though each of these bands shared certain sequences of the specific probe used, they were not identical. It becomes necessary to work out the fine structure which is revealed by such experiments. Such elucidation may yield information of great value in understanding how the human genome is regulated, in health and disease.

The information that will be achieved by the complete elucidation of the human genome will add greatly to our understanding of the nature of the human genetic diseases. It should provide new fundamental understanding, and new methods of diagnosis; and should suggest whole new areas of treatment and prevention when delineation of the metabolic basis of these conditions will be achieved. However, important as are these accomplishments, even greater contributions to human health may be expected. In any cell of the body, each of the metabolic reactions carried out requires the mediation of specific structural and enzymatic proteins. These are produced and regulated by specific regions of the cell genome. Therefore, even diseases not due to an inborn genetic defect will be illuminated by understanding of the structure of the genome resulting from this program.

Understanding of the principles of genetic regulation of the various human genes will not be forthcoming until a much greater fraction of the human genome is mapped and sequenced than is now available. In addition, the structures of the whole gamut of the human proteins will for the first time become available. It is difficult to conceive of any single advance in biomedical science that will have as great an impact as this on the development of new understanding of all disease and the devising of new measures to diagnose, treat and prevent many human afflictions.

I think it would be a mistake to set up a single laboratory to carry out this program. Such a monolithic structure could not help but suffer from a certain amount of isolation from the rest of biomedical science and would be cut off from the generation of new ideas and unconventional approaches that can result when a number of different laboratories are encouraged to assume responsibilities for different parts of the program and to communicate intensively with each other. The development of the atomic energy project was divided among several centers at Columbia University, Brookhaven, the University of Chicago, Oak Ridge, Hanford, Berkeley, and Los Alamos and the results achieved were remarkable. I think that a similar situation would be found to operate effectively in the present case.

Our laboratory has developed a number of the different chromosomal, regional and fine structure mapping procedures. We also carry on highly effective sequencing. We would be delighted to take an active part in some phases of this program elucidating the human genome.





# Lawrence Berkeley Laboratory

1 Cyclotron Road Berkeley, California 94720

(415) 486-4000 • FTS 451-4000

March 17, 1986

Mark W. Bitensky, M.D.  
Senior Fellow  
Mail Stop M881  
Los Alamos National Laboratory  
Los Alamos, New Mexico 87545

Dear Mark:

First, I would like to extend my thanks to you for organizing such a provocative and interesting meeting on the genome sequencing project. It is clear that there would be great value in engaging that process which would lead to the approval by Congress of this highly worthwhile project directed at the human genome. The following is a list of impressions and opinions which I now have as a consequence of the meeting.

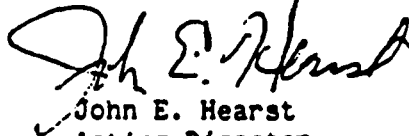
1) The first three years of activity on this project should be directed at ordering cosmid libraries of the various chromosomes, and setting up an information and material distribution procedure which would stimulate medical researchers around the world to focus those portions of the human genome with most immediate impact upon human health. Simultaneously, a substantial level of support should be provided in order to stimulate the development of instrumentation which would speed up the ultimate goal of a complete sequence of the human genome.

2) I favor a decentralized approach to the project. It is essential that approval of this activity be obtained from both the community of biological and medical scientists, and the political community. Such objectives will only be met by a decentralized and perhaps even an international effort.

3) The selection of centers where the sequencing activities will take place is also crucial to the success of the project. While Los Alamos National Laboratory might well be the center of the computer support of this effort, I am strongly of the opinion that the National Laboratories which lack an intimate connection with a university community should not be chosen as a site for the experimental work. Strong interaction with biochemists will be an essential ingredient for the success of this major project.

4) Finally, it is essential to realize that most of the work on such a massive project will be performed by highly trained technicians in a tightly controlled administrative environment. Such activity must be recognized to be distinct and different than basic research, and must not be supported at the expense of support for basic research in the biological sciences.

Sincerely yours,



John E. Hearst  
Acting Director  
Division of Chemical Biodynamics

cc Rosenblatt  
Delisi  
Varga



850 Lincoln Centre Drive, Foster City, California 94404 • (415) 570-6667 • Telex: 470052

March 7, 1986

Dr. Mark Bitensky  
Los Alamos National Laboratory  
Los Alamos, New Mexico 87545

Dear Dr. Bitensky:

Thank you for the opportunity of attending the recent workshop on sequencing the human genome. The discussions were especially helpful in confirming that the automated sequencing technology we are developing will make a contribution to molecular biology. Indeed our technology may be critical for the larger sequencing projects such as the one you are contemplating.

Although sequencing technology will no doubt be reviewed at the actual start of the project, I thought it might be helpful to summarize the capabilities and schedule for our instrument. Please feel free to use this information in your report to support the feasibility of this exciting project.

The instrument we expect to introduce in June will provide sequence data from eight clones in the time required for electrophoresis (four to seven hours). The operator runs the dideoxy reactions (about one hour) and loads the gel. From that point, no operator interaction is required. The data is stored on a floppy disk in a Gen-Bank compatible format. The initial raw data generated by the sequencer are available within one to two hours of the gel loading; complete data are available at the end of electrophoresis.

The cost of operation will be less than with conventional methods because no film or radioactive materials are used. The cost of the instrument has not been established, but will be less than \$65,000.

Although it is difficult to predict future improvements in this technology, we invest heavily in continuing product development. This has made possible dramatic improvements in every one of our existing product lines. Enclosed is information on our products under development and our most recent annual report; both include information on the automated sequencer.

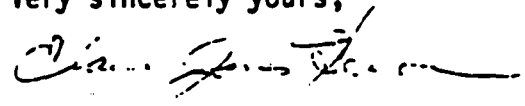
Also enclosed are two earlier annual reports which help to demonstrate the rapid success of the products which we created to fill the stated needs of scientists and researchers worldwide. Each instrument we introduced quickly became the standard because of our research and development and continuing enhancement.

March 7, 1986  
Page 2

Since nearly all of technologies required for the Human Genome Project have application to various segments of the worldwide scientific community which we serve, Applied Biosystems would be interested in developing other related automated instruments. The expertise we have developed and continue to expand includes all aspects of instrument systems engineering and software development. This resource could be a very economical alternative to assembling a group of engineers and a production facility at the center coordinating the sequencing.

Again, I appreciate being included in the workshop, the task you are contemplating will make an incalculable contribution to science. If there is any assistance which we can provide to you or to the realization of the Project, please do not hesitate to ask.

Very sincerely yours,



Elaine Jones Heron, Ph.D.  
DNA Group Product Manager

Enclosures

EJH/dlb

# CALIFORNIA INSTITUTE OF TECHNOLOGY

DIVISION OF BIOLOGY 150-20

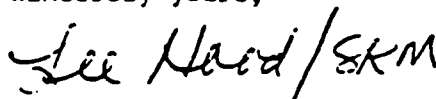
March 10, 1986

Dr. Mark W. Bitensky  
Division Leader  
Los Alamos National Laboratory  
Los Alamos, New Mexico 87545

Dear Dr. Bitensky:

I apologize for not getting back to you on your March 3 and 4 workshop on sequencing the human genome. It is on a subject that I am very interested in, and I would be delighted to hear what came of the meeting. Unfortunately I have just gone on sabbatical, and I had an enormous number of commitments to try and finish up so as to relieve myself of administrative duties. Again, don't take my absence at this workshop as a lack of interest. I hope the meeting went well.

Sincerely yours,



Lee Hood

LH:skm

## OAK RIDGE NATIONAL LABORATORY

OPERATED BY MARTIN MARIETTA ENERGY SYSTEMS INC

POST OFFICE BOX Y  
OAK RIDGE, TENNESSEE 37831

March 24, 1986

Dr. M. W. Bitensky  
Life Sciences Division  
Los Alamos National Laboratory  
P. O. Box 1663, MS M881  
Los Alamos, New Mexico 87545

Reference: LS-DO-86-1.12-50

Dear Mark:

I would like to thank you first for your superb organization of the Human Genome Sequencing Workshop. It was highly beneficial to me personally not only for developing an awareness of the debate on the feasibility and desirability of sequencing the entire human genome, but also for the opportunity to learn about current research on various aspects of sequencing strategies, since you had the leaders in these efforts at the Workshop.

I am planning to share the proceedings of the Workshop with my colleagues at ORNL through a seminar. I have also talked with Dr. Dick Griesemer, our Division Director, and Dr. Lee Russell, Head of our Mouse Genetics Research group about the general opinions of the Workshop participants.

Regarding my personal opinion, I agree completely with the consensus of the group that the sequencing of the entire human genome is not an unfeasible goal. At the same time, because the project will be so prohibitively expensive, many of us here are worried about the potential and long-range impact of such an endeavor on the current research projects being funded by DOE in national laboratories and elsewhere. Dave Smith emphatically expressed his conviction that such a project should be initiated only with the availability of new funds. However, even if the venture is blessed with the approval of OSTP and if the Congress provides it with some initial funding, can anyone be certain that in this Gramm-Rudman era, the Congress (or the OMB) in their quixotic wisdom will not propose some paring of already existing programs?

Many of my colleagues here share my views that there may be some merit in complete sequencing of one representative chromosome because some unusual and unexpected structure may be revealed. However, the immediate benefit of sequencing the complete human genome is not obvious. In view of OHER's

M. W. Bitensky  
Page 2  
March 24, 1986

missions on mutagenesis and other health-related research and the presumably large numbers of polymorphism present in the human population the whole sequencing process may have to be repeated over and over again for various individuals in order to make it useful. That is clearly impractical.

On the other hand, I also agree with the majority view expressed in the Workshop that the immediate objective should be the creation of genomic or chromosomal library and an ordering of the cloned segments. Once an ordered human genomic library is available, any selected segment of the genome could be targeted and can be sequenced in a large population sample.

Furthermore, in view of the rapidly emerging developments and breakthroughs in sequencing techniques, an immediate effort involving a large budget may be somewhat wasteful and premature because, a few years later, the same task could possibly be carried out more cheaply and expeditiously. Some participants in the Workshop made the argument that if the sequencing is left to individual laboratories instead of as a part of an organized and unified effort, the so called "uninteresting" regions of the genome may never be sequenced. However, it appears that there may be very few, if any, uninteresting segments in DNA.

I also want to raise another point. Dr. Ruddle pointed out the importance of sequencing the mouse genome. From my vantage, as a member of a Division with a large mouse colony of inbred and well characterized mutant strains, I strongly share Dr. Ruddle's view. In fact, an ordered library and extensive RFLP map of the mouse genome will be of immediate and major use in our Division and in the scientific community at large, and some of my colleagues will enthusiastically participate in such an effort. The mouse remains the most useful genetic model system for man and provides a means of performing genetic experiments either practically or ethically impossible to carry out in humans.

I strongly hold the view that even if OHER takes the lead in proposing this giant project, a perception of national consensus (and perhaps international collaboration as some speakers proposed) has to be conveyed. This may require a national debate in some kind of forum with participation of our leading scientists in biological disciplines. Because of the potential benefits from the availability of ordered libraries and extensive RFLP maps, I feel a majority will be favorably inclined and their blessing will improve the chances of successful funding.

In summing up, I recommend that the following course of action be considered.

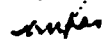
1. Development of a national consensus in a wider forum on the importance and desirability of genome sequencing including availability of ordered genomic libraries and RFLP maps.

M. W. Bitensky  
Page 3  
March 24, 1986

2. Emphasis, in the initial phase, on creation of these genomic libraries and RFLP maps in parallel efforts for man and a homozygous strain of mouse.

With warm regards.

Yours sincerely,



Sankar Mitra  
Senior Staff Member  
Biology Division

SM:ds

cc: R. A. Griesemer  
L. B. Russell  
File - SM



## Models of the Enterprise

To consider how the enterprise might be organized, we break it down as follows:

- 1) Central scientific management.
- 2) Laboratory operations; cloning, mapping, sequencing.
- 3) Data management, data distribution and computational support.
- 4) Technological support; development of new technology.
- 5) Biological materials distribution; clones, libraries.

Each of these activities could, in principle, either be centralized or distributed over several centers; laboratory operations, activity 2), could be distributed over a large number of centers - the 'cottage industry' approach. However the rest of the enterprise is organized, all of data should be collected and organized into a data bank that supports efficient retrieval and analysis, giving every other organizational element of the program instant access to its current state and to all of its data and computational resources. Thus adequate development and resources for the computational activity 3) makes it possible to at least consider decentralizing the other phases of the program, and at least one participant, Ruddle, suggested decentralizing the computational activity itself.

1) Since this is a program with definite overall goals, central and strong management seems a necessity, and a number of participants (Brenner, Gilbert, among others) spoke of this. The most straightforward model would charge a director with overall administrative and scientific responsibility; he or she would see to the direction of all elements of the program. There would be a strong and distinguished scientific advisory committee, in which there should be no problem in enlisting the most productive scientists in relevant fields.

2) Some participants thought it desirable to centralize and some to disperse the laboratory operations. Centralization would involve creating an institute on the scale of from several hundred to a thousand employees and would of course simplify management of this phase of the program. It seems likely that modest decentralization - into, lets say, ten or fewer centers - would increase the number of distinguished and productive molecular biologists that could be recruited into the program, and thus widen avenues into the program for new and innovative approaches. Several participants voiced concern that extreme decentralization would create great problems of quality control and of focussing the effort.

3) Sequence and map data, the state and availability of various libraries and probes, performance of various groups and strategies - all of the data needed for effective management and planning - should flow into a data bank and in turn be accessible to all parts of the project. I think this should be centralized in a strong data bank and computational analysis group with adequate computational and data communication resources to serve the entire enterprise. An important set of issues surrounds the extent to which this activity would serve to integrate the work of the scientific community as a whole in its use the data and its feedback to advance the project. In any case the computation group would both develop and make available from other sources software to support all needs of the project, including such

tasks as finding and ranking ambiguous relationships between sequence fragments, which will become especially important as the 'endgame' of whole genome sequencing is approached; identifying likely signals for sites and regulation of such genetic functions as expression, replication and recombination or integration at the DNA level; predicting likely structure and function for possible transcripts and translates of otherwise unanalyzed sequences; modeling of new techniques for sequencing and mapping.

4) The cost of the project is critically dependent on the extent to which sequencing technology can be improved. Thus, as a number of participants observed, it would be wise to invest some proportion of the project's resources in new technology. Thus it seems wise to commission at least one laboratory, with access to the full range of technological skills and resources available at institutions like the National Laboratories - robotics, digital engineering, physical analysis of separation processes, for examples.

5) Both for this enterprise itself, and to realize its payoffs for science, medicine and agriculture, adequate resources have to be devoted to distributing its products. This activity would provide the biological materials that, having been characterized, ultimately at the sequence level, would immensely speed the work of investigators and developers of applications all over the world. Again, this was spoken of by several participants.

*Walter Seal*

3/19/56

EXPRESS MAIL

Yale University

Department of Biology  
 Kline Biology Tower  
 P.O. Box 6666  
 New Haven, Connecticut 06511-8112

Campus address  
 Kline Biology Tower  
 219 Prospect Street  
 Telephone:  
 203 436-2571

March 17, 1986

Dr. Mark Bitensky  
 Los Alamos Scientific Laboratory  
 P. O. Box 1663  
 Los Alamos, New Mexico 87545

Dear Mark:

I greatly enjoyed the meeting on the human genome. I am sure that something will come out of it. I suspect by now you have seen Renato Dulbecco's editorial statement in Science which supports the general concepts arrived at in Santa Fe. Below you will find my rough notes regarding my opening statement at the conference. Feel free to modify this in any way you wish regarding your final report.

Some of the important milestones in gene mapping have been the following. The first human gene to be mapped was that of color blindness which has placed on the X chromosome by E. B. Wilson in 1911. Many human X chromosome assignments were made from that period forward and continue to be made on the basis of the special advantages in mapping to the X chromosome. Haldane performed one of the first human genetic linkage studies using the X chromosome, again dealing with the color blindness locus in 1936. In 1956, the correct human chromosome number was established by Tjio and Levan as  $2n$  equals 46. In 1968, the requirements for parasexual mapping in somatic cell hybrids were finally assembled and the period of parasexual mapping began. The important requirements were the use of chromosome segregation in mouse/human hybrids as reported earlier by Mary Weiss and co-workers, including the early work of Boris Ephrussi; the establishment of isozymes as genetic markers that could be employed in somatic cells in tissue culture by Ruddle and his co-workers; and the identification of all of the individual chromosomes using banding techniques as introduced by Caspersson and his co-workers. In 1973, there were sufficient assignments to initiate a gene mapping conference, the first of which was held at Yale University, under the organization of F. H. Ruddle. This meeting instituted the series of meetings which has persisted to this day, now being held every two years. It is interesting to note that the assignments to the map have doubled every two years. Parasexual techniques can also be used for establishing the regional position of genes making use of chromosome translocations and deletions. The first case in which the order of linked genes was established was that for phosphoglucose kinase (PGK), hypoxanthine phosphoribosyl transferase (HPRT), and glucose-6 phosphate dehydrogenase (G6PD), all located on the long arm of the X chromosome (Ricciuti and Ruddle). The practical resolution of such procedures is about  $1 \times 10^7$  base pairs. In 1980, physical mapping was advanced by the introduction of in situ hybridization techniques which permitted the identification of single copy genes on chromosomes. This method also has a resolution of approximately  $1 \times 10^7$  base pairs. Earlier in 1976, the identification and characterization of middle repetitive sequences was made. These middle repetitive sequences which were frequently species specific have been extremely useful in physical mapping studies and also in gene cloning. In 1978, the restriction fragment length polymorphism (RFLP) concept was introduced. This was essentially a Mendelian

mapping procedure. However, it is exceedingly practical because of the following features. (1) An efficient uniform DNA methodology is applied. (2) DNA restriction site variants are relatively common and serve as an excellent source of polymorphisms. (3) Reference families can be established and the data which is accumulated from these families becomes increasingly powerful as the data accumulates. This procedure has one great advantage in that it allows one to map any genetic condition expressed at the organismal level. There are at least 1,000 such conditions already known as recorded in Victor McKusick's book entitled "Mendelian Inheritance in Man". The resolution of the RFLP procedure is approximately  $1 \times 10^6$  base pairs which is equivalent to 1 centimorgan. Under ideal conditions, fractions of centimorgans can also be resolved. Also in 1978, parasexual methods of higher resolution ordering and distancing of genes was introduced. These include the following: DNA transfer which allows the transfer of pieces of DNA in the size range of 20-50 kb. This is somewhat too small to be useful for mapping purposes. Chromosome mediated gene transfer allows for the transfer pieces effectively in the range of  $10^7 - 10^8$  base pairs. Chromosome transfer also allows for the transfer of large unstable fragments, especially those with partially functional centromeres which may then undergo subsequent rearrangement by translocation into the recipient genome. This procedure provides even higher levels of resolution in the range of  $10^6 - 10^7$  base pairs. Chromosome sorting is the direct physical separation of chromosomes. It allows for the assignment of genes to particular chromosomes, as well as providing DNA for the isolation of genes and chromosome specific segments. Chromosome dissection, the actual manual subdivision of chromosomes using microdissection procedures, has also been applied to the purification of subchromosomal specific DNA. The resolution here is around the  $10^7$ -base pair range. Many of the above parasexual procedures have been reported, but not stringently evaluated. The chromosome transfer systems tend to be tedious and suffer the disadvantage that they may also be associated with fairly extensive random rearrangements of gene sequences which may disturb the real ordering of genes and DNA sequences. In 1984, individuals began to express interest in applying molecular techniques to the development of a physical DNA map. This approach makes use of the following techniques: One can make use of hopping techniques where sequences are isolated at distances of  $10^7$  base pairs. Pulsed field electrophoresis procedures have been introduced which allow one to identify large fragments in the range of  $5 \times 10^5 - 5 \times 10^6$  base pairs using analytical approaches. A number of techniques have been introduced to establish overlapping cosmid or phage insert maps. It is estimated that between 100,000 to 200,000 cosmids would be required to cover the whole human genome. Ultimately, overlapping cosmids would be subjected to nucleotide sequencing providing the ultimate resolution map. All of these procedures are relatively new in the context of mammalian gene mapping and much must be done to evaluate and develop them. Also, ancillary procedures must be developed, such as data bases dealing with map positions, sequences, RFLP, and probe availability. New computer techniques must be developed, both at the software and hardware levels. New instruments may also be required in order to speed up sequencing so that it is on the order of 50 to 100 times faster than presently available methods. In addition to the advancement and improvement of methods, new administrative innovations must also be introduced. These would include new administrative structures to facilitate large and interactive undertakings and the identification of new sources of financial support, perhaps including DOE, NSF, NIH, Hughes Medical Institute, and other foundations. One might ask what are the benefits? The establishment of an overlapping cosmid map would immediately allow the integration of physical and Mendelian maps.

Moreover, it would have the practical advantage that once a sequence had been isolated, the surrounding sequences could also be obtained. Or the sequences lying between two known sequences could be readily isolated or be made available. The primary use of a gene map is related to problem solving. Gene maps are especially powerful for excluding hypotheses. For example, one can ask a question such as: is gene X affected or inactivated by translocation? The gene map is also particularly important in generating hypotheses. For example, if different genes in a chromosomal region appear to have concerted functions, one might ask: are they regulated by a common transregulatory protein? Or, do they have a common evolutionary history? Such questions could be readily resolved by testing for sequence similarities between the genes in question. I believe the current 1,000 gene map has called our attention to these and even more exciting future possibilities. Some of the practical rewards of an ultimate resolution gene map are (1) isolation of the genes of medical interest as indicated to above; (2) understanding chromosome compaction in the context of mitosis and the mechanics of mitosis; (3) chromosome disposition and function in interphase cells in relationship to the nuclear membrane and nuclear pores; (4) functions such as recombination and DNA repair, and especially recombination in the context of meiosis; and (5) developmental control of gene regulation and expression. Others (particularly see a recent editorial in Science by Renato Dulbecco) suggest that an ultimate resolution map will provide extremely important information on modifications of the genome which are related to neoplastic cell inception and progression. It is clear that many benefits both general and practical would come from the complete knowledge of the human genome. A concerted effort would make the realization of this project materialize at least one to two decades before this problem would be worked out in the usual random methodology of scientific research. A directed effort may make these benefits available to our population at least one or two generations earlier than otherwise possible. Another point to mention is the structure of such a organization. My own feeling is that central authority would best serve the scientific public with dispersed regional centers, perhaps five to ten in number throughout the country. These could deal with regional laboratories and service collection points for data which is generated in a more random way. It is important to distinguish between a distributed system, such as the one that I advocate, and a diffuse system which would be essentially random in nature.

These are some of my thoughts. I have not taken a lot of time to polish them and bring them to a final stage of exposition, but I feel that it is more important for you to get these notes now in the two-week period that you indicated than to wait a longer period of time. Thank you again for a good meeting.

All the very best.

Yours sincerely,



Frank H. Ruddle  
Professor of Biology and Human Genetics  
Ross Granville Harrison Professor of Biology



# Lawrence Berkeley Laboratory

1 Cyclotron Road Berkeley, California 94720

(415) 486-4000 • FTS 451-4000

March 17, 1986

Mark W. Bitensky, M.D.  
Senior Fellow  
Mail Stop M881  
Los Alamos National Laboratory  
Los Alamos, New Mexico 87545

Dear Mark:

First, I would like to extend my thanks to you for organizing such a provocative and interesting meeting on the genome sequencing project. It is clear that there would be great value in engaging that process which would lead to the approval by Congress of this highly worthwhile project directed at the human genome. The following is a list of impressions and opinions which I now have as a consequence of the meeting.

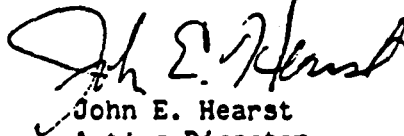
1) The first three years of activity on this project should be directed at ordering cosmid libraries of the various chromosomes, and setting up an information and material distribution procedure which would stimulate medical researchers around the world to focus those portions of the human genome with most immediate impact upon human health. Simultaneously, a substantial level of support should be provided in order to stimulate the development of instrumentation which would speed up the ultimate goal of a complete sequence of the human genome.

2) I favor a decentralized approach to the project. It is essential that approval of this activity be obtained from both the community of biological and medical scientists, and the political community. Such objectives will only be met by a decentralized and perhaps even an international effort.

3) The selection of centers where the sequencing activities will take place is also crucial to the success of the project. While Los Alamos National Laboratory might well be the center of the computer support of this effort, I am strongly of the opinion that the National Laboratories which lack an intimate connection with a university community should not be chosen as a site for the experimental work. Strong interaction with biochemists will be an essential ingredient for the success of this major project.

4) Finally, it is essential to realize that most of the work on such a massive project will be performed by highly trained technicians in a tightly controlled administrative environment. Such activity must be recognized to be distinct and different than basic research, and must not be supported at the expense of support for basic research in the biological sciences.

Sincerely yours,



John E. Hearst  
Acting Director  
Division of Chemical Biodynamics

cc Rosenblatt  
Delisi  
Varga

THE JOHNS HOPKINS UNIVERSITY  
SCHOOL OF MEDICINE

DEPARTMENT OF  
MOLECULAR BIOLOGY AND GENETICS



725 N. WOLFE STREET  
BALTIMORE, MARYLAND 21205

March 6, 1986

Dr. Mark W. Bitensky  
Division Leader  
Life Sciences Division  
Los Alamos National Laboratory  
Los Alamos, New Mexico 87545

Dear Dr. Bitensky:

I would like to take this opportunity, first, to thank you for inviting me to participate in the Genome Sequencing Workshop, and second, to summarize some of my conclusions from the meeting.

Perhaps the most impressive feature of the meeting was the unanimous consensus that sequencing the entire human genome is doable, even with current technology, and that enormous benefits would be forthcoming for medicine and basic science. Having the knowledge "up front" could accelerate every branch of biological and medical research.

How to implement such a heroic and costly undertaking is less clear. I currently divide the undertaking into two phases. Phase I would focus primarily on developing an ordered library of cosmid and/or phage clones for each human chromosome. Such a set of human chromosome libraries could be used to localize known genes, sequences, and RFLP's. Concurrently, a map of Cantor fragments could be assembled and related to the ordered cosmid library. These two concurrent physical mapping efforts would mutually facilitate and confirm each other. A third concurrent effort would be to choose an appropriate small chromosome for sequencing. This would allow development of systematic, automated sequencing strategies which could eventually be applied to the much greater task of the whole genome. The sequencing effort during Phase I would be relatively small. In addition, a small group of physicists, engineers, and electronics staff would interact with the biologists in order to develop new sequencing concepts, and to automate conventional methods. Finally, Phase I would establish a central computing and data bank facility with national networking.



Dr. Mark W. Bitensky

2

March 6, 1986

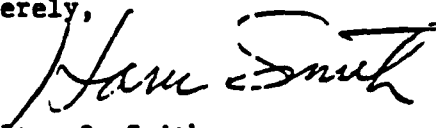
Phase I would be implemented in a national laboratory center, for example, Los Alamos. New or existing buildings could be used.

Phase II would be a large scale, dedicated, mission-oriented sequencing assault on the whole human genome. It would begin as soon as sequencing technologies and strategies are sufficiently developed. Again this would be done in a central facility (presumably an expanded Phase I facility). In my opinion, farming such work out to numerous university laboratories on a contract basis would be excessively inefficient and costly.

The central facility (Human Genome Resource Center, or whatever you want to call it) would be a repository for all the collected knowledge of the human genome. Consequently, it would need to provide an extensive service function to the rest of the nation. All scientists could readily obtain library clones and catalogued information through the center.

Good luck in preparing your report for Dr. DeLisi, and thanks again for an enjoyable and exciting two days.

Sincerely,



Hamilton O. Smith  
Professor of Molecular  
and Genetics

HOS:mk

DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY  
University of Florida • College of Medicine  
Box J-245, J. Hillis Miller Health Center  
Gainesville, Florida 32610



Telephone: 904 352 3500  
Room

Biochemistry  
and  
Molecular Biology  
Faculty

March 17, 1986

Daniel L. Purich, Ph.D.  
Chairman

Dr. Mark W. Bitensky  
Los Alamos National Laboratory  
MS 881  
Los Alamos, NM 87545

Dear Dr. Bitensky:

We want to express our appreciation for the opportunity to have attended and participated in the genome sequencing workshop. Sequencing the genome is an important endeavor that should be actively pursued. While execution of the project requires a large infusion of funds (estimates range from 0.5 to 1.5 billion over a 10-15 year period), the value and applications of the information to be gained make the project timely and cost effective.

With respect to approaches, we would strongly recommend the sequencing of libraries constructed from individual chromosomes using the largest clonable fragments. From our experience with the mapping of human histone genes (which constitute a multi-gene family), we know that similar or identical segments of DNA map to more than one chromosome. Other multi-gene families may be similarly organized. If the mapping and sequencing were to be carried out on a random rather than on a chromosome by chromosome basis, the problem of closure of the map will most likely be extremely difficult, if not impossible.

With respect to administration and execution of the project, a good case can be made for a primary, centralized facility with centers outside of the primary unit to carry out specialized aspects of the program. The chromosome sorting and cloning project is underway at Los Alamos and Lawrence-Livermore and has proven to be effective. A strong argument must, therefore, be made to incorporate this component as a major unit of the genome sequencing project.

If we can be of any additional assistance in the evaluation, planning or execution of the program, please do not hesitate to contact us.

Yours sincerely,

Yours sincerely,

Gary Stein

Janet Stein

# Yale University

Department of Human Genetics  
Yale University School of Medicine  
1-310 SHM  
P.O. Box 3333  
New Haven, Connecticut 06510-8005

Campus address:  
1-310 Sterling Hall of Medicine  
333 Cedar Street

March 7, 1986

Dr. Mark Bitensky  
Los Alamos National Laboratory  
Los Alamos, New Mexico 87545

Dear Mark:

Enclosed is the brief report you requested on the technical feasibility of sequencing the human genome. It deals exclusively with the technological aspects of the project; it does not address either implementation strategy or financial considerations which will be covered by the reports of other participants.

I really enjoyed the meeting and I sincerely hope the project can be implemented as soon as possible.

Best personal regards.

Yours sincerely,



David C. Ward

DCW:dsg  
Enclosure

Report of David C. Ward

Summary Statement of DNA Sequencing Technology.

It was concluded that sequencing the human genome using existing techniques would be an untenable project. However, there are novel technical improvements, foreseeable in the next 6-12 months, that should increase the speed of sequence data acquisition 10-100-fold. With the successful development of this technology, the human genome sequence project becomes a viable reality.

At present, a fully dedicated and totally competent scientist or technician can produce about  $2 \times 10^5$  -  $5 \times 10^5$  bp of new sequence data per year, or between  $5 \times 10^4$  and  $1 \times 10^5$  base pairs of definitive sequence information; this differential reflects an average 5-fold redundancy in data acquisition. At  $10^5$  bp/year, it would take ~ 30,000 man years of work to complete the full  $3 \times 10^9$  bp in the human genome. It is therefore mandatory that the through-put time for data acquisition be reduced markedly. With a 10-fold improvement in sequencing speed (i.e.,  $1 \times 10^6$  bp/person/year) it would take 300 fully dedicated personnel 10 years, while with a 100-fold rate increase ( $1 \times 10^7$  bp/person/year) it would take only 30 people 10 years to complete the project.

Two different technologies are presently under development. First, Applied Biosystems Inc. (Foster City, CA) is in the late stages of developing an instrument which uses a laser fluorometer to determine sequence data generated by a proprietary modification of the Sanger dideoxynucleotide primer extension method. The instrument can process 1 sequencing gel with 8 data lanes in a total thru-put time of 6-8 hours. With readable sequence at 250 residues per lane, each gel can generate 2000 nucleotides of sequence per run. Assuming two analytic runs per day, the instrument should generate 4 kb of sequence per day [i.e., 20 kb/week or 1 megabase (MB) per year]. Since

representatives of ABI state that one technician can easily handle 2 instruments, ~ 2MB of raw data (or 200-400 kb of complete sequence) could be obtained per person per year. This is a 2-4-fold improvement over present techniques. The first generation of the ABI sequence instrument is to be field tested during the summer of 1986 with market introduction expected by the late fall of 1986. Projected cost per instrument is presently \$50-60,000.

The first generation ABI instrument can be modified to process 40 sample lanes within the same 6-8 hr thru-put time without major retooling or technical development. This would expand the confirmed sequence acquisition rate to ~ 1MB/year/person, or 10-20 times existing rates. Upgrading of the ABI instrument capacity would be strongly stimulated by a commitment to the total genome sequence project.

A second sequencing strategy, termed multiplex sequencing, is under development by Dr. George Church (University of California, San Francisco). He has constructed 50 unique cloning vectors, each containing two unique 20 bp probe hybridization sites flanking a SmaI cloning site. Subsets of the genome can be cloned into each of the 50 vectors. One plasmid from each of the 50 clone sets are pooled and treated as one, using the Maxim and Gilbert base-specific chemical reactions. Several hundred such reaction pools are electrophoresed and then transferred from sequencing gels to nylon membranes. Up to 100 sequences should be obtainable from each pool by hybridizing the membranes with a succession of end-specific, strand specific probes. Once the membrane transfers have been completed, 10 membranes at a time can be processed manually by one individual, with each hybridization cycle taking ~ 30 hours (2-3 hr for hybridization and 24 hours for autoradiographic detection). Since each cycle of 10 filters should yield 80-100 kb of raw sequence data, it is conceivable that up to 15-20 MB of raw sequence or 3-4 MB of refined sequence could be obtained each year by one individual. Although this technique could

result in upto a 100-fold increase in the rate at which sequence data could be generated, it must be stressed that methodology has yet to be rigerously tested. The feasibility studies should, however, be completed within the next 12 months. Further reduction of the current detection time (24 hrs) may also be achieved by using peroxidase chemiluminescence to detect biotinylated oligomer probes. Preliminary data from the laboratory of David C. Ward (Yale University) indicate that as little as 1 pg of target sequence can be visualized by a 10-30 second film exposure using this chemiluminescence technique. Finally, relatively simple modifications of a prototype instrument for automated in situ hybridization (recently designed by David Brigati, Hershy Medical School, and under construction by the Fisher-Allied Corporation) it should be possible to fully automate the multiplex sequencing technique.

In conclusion, it appears that major improvements in sequencing technology will occur in the near future. The goal of sequencing the entire human genome within a 10-year period thus should be both feasible and highly desirable.

# Yale University

Department of Human Genetics  
Yale University School of Medicine  
1-310 SHM  
P.O. Box 3333  
New Haven, Connecticut 06510-8003

Campus address:  
1-310 Sterling Hall of Medicine  
333 Cedar Street

March 12, 1986

Mark W. Bitensky, M.D.  
Life Science Division  
Los Alamos National Laboratory  
P.O. Box 1663  
MS, M881  
Los Alamos, NM 87545

Dear Mark,

This is in follow-up to the SantaFe meeting about the project to sequence the human genome. This letter summarizes some of my personal views and is not in general an attempt to identify any consensus opinion.

First, I think there was a clear and enthusiastic consensus that the total description of a human genome would be a highly desirable thing to have, and that the cost to the nation to obtain it would be well justified, provided it did not come at the expense of diversion of funds away from other support for bio-medical research. In particular, there would be room for very serious debate if any of the funding were to come from NIH, and the impact on other research would have to be considered if bio-medical research support foundations or institutes such as Hughes were to provide any significant part of the money.

The three types of material describing the human genome are in order of increasing time and labor needed to obtain them:

- (1) A coarse level restriction map of the genome
- (2) A set of ordered clones covering the genome
- (3) An essentially complete nucleotide sequence of the genome.

My perspective on the technology is that (1) may emerge sooner than many people think, (2) may be much less costly as developing techniques solidify, and (3) might take longer than many of the participants estimated, unless (2) is in place and some structured approach to (3) is devised.

Goal (1) would be extremely useful as an end in itself, providing: (a) a rapid physical basis for accurate mapping of restriction fragments length polymorphisms (RFLP), (b) access to centromeric and telomeric sequences, (c) a rapid diagnostic and cloning approach for a number of oncogenes and genes involved in hereditary disorders, including some polygenic disorders, and (d) interesting insights into the organization and variation of midrange genome

structure in man, as well as insights into chromosomal evolution. Goal (1) would provide a firmer basis for execution of goals (2) and (3).

In addition to (1), (2), and (3) certain related or alternative goals should be considered as part of this (or another) national effort. These are

- (4) mapping and sequencing the mouse genome.
- (5) evolving a set of techniques (for which I have in mind specific proposals) to produce very extensive cDNA libraries in which all cDNAs would be represented at approximately equal abundance, and which could be used for sequencing the cDNAs and mapping them with respect to time and place of expression during the lifetime of the mouse.
- (6) Construction of strains of mice heterozygous for deletions which in toto cover large regions of the mouse genome. These would provide a general tool for proceeding from the nucleotide sequence to in vivo biology.

The completion of the first three of these goals would be almost revolutionary for higher eucaryote biology. Experimental methods already exist and would surely further evolve for using this data in a study of general and neural development, definition of human genetic disease, and of contributing genetic elements in diseases of complex etiology and environmentally conditioned disease, human evolution, and the design of proteins, peptides, and protein binding molecules for pharmaceuticals, disease prevention, etc.

I also share the common enthusiasm that this project be declared underway, funded at least initially by the DOE. I suggest that the initial phase of the work plan include goals (1) and (2) and pilot studies to evaluate the true overall efficiencies for the best DNA sequencing protocols. In addition further discussion of the technical strategies of (2) and (3) (random vs. chromosome at a time vs. other parsing methods) is sorely needed. I remain cautious about random approaches for (2) as well as (3). The end game could yet involve generating sufficient data to obviate the opening and mid game. The technology for accomplishing these goals is rapidly advancing.

I also share the almost general consensus that a national laboratory or institute, or at least a federally supported institute should be an important component of the effort. The unit should play an important role in development of instrumentation, software, and to not quite the same degree molecular technologies. It should be responsible for coordinating and distributing data and materials (including RFLP probes and other information and material not necessarily generated as part of the wet lab research effort.) How much of the cloning and sequencing effort should be housed in this central unit is presently unclear. Cloning could probably go forward by various approaches in academic laboratories until a consensus protocol is reached. I'd suggest that the initial efforts in sequencing be focused on automation of wet lab procedures, data reading, and data correlation procedures, and on an evaluation of the technical approaches i.e. progressive deletion of large fragment vs. shotgun sequencing, Maxam and Gilbert vs. dideoxy multiplexing, radioactive vs. luminescent detection, etc. before a serious effort on the genome is undertaken. My opinion is that a trial run on yeast would yield much more biology per buck in our lifetime than trying out the protocol on a human chromosome.



Mark W. Bitensky, M.D.  
Page 3  
March 12, 1986

111

The structure of the effort should be looked at from a "zero base" point of view as to what would be most desirable and efficient rather than what fits into currently existing structures, as it is likely that a successful effort would have indefinite extension for continuing new opportunities in large scale biology. For example, does something in the Rand Corporation structure offer advantages. Placing the central unit in a highly desirable area such as certain parts of California, and near a major academic center like Stanford could make recruiting top flight people easier and keep a close tie with developing trends in academia. Placing the organization outside of the government civil service or military systems has the obvious advantages of being able to offer competitive salaries and avoiding the sluggish apparatus of government.

As the discussion evolved it seemed possible that such an effort could consume a major part of national resources available for this type of work. In view of this, I believe a significant proportion of the funding of the project should be distributed to "peripheral laboratories." Regional centers have some advantages but several disadvantages with regard to flexibility, maximum quality of the workers in the trenches, etc. I fear they could also become self-perpetuating administrative rather than innovative structures, especially if they are large. Important contributions could also be made from institutional laboratories, scientific consortia, university based centers, and even bio-technology companies. One relatively unstructured way of administering extra mural money is to formulate goals and request investigator initiated proposals of various sizes subject to constraints about costs per unit data in cases where data generation rather than technique development is the goal. Criteria for evaluating these proposals would include their usefulness for the overall goal, freedom of access to materials, information and techniques developed under this sponsorship, etc. that do not as explicitly enter into evaluations for NIH funding. Funding could also be subject to annual or biannual zero base planning, and specified maximum periods. Review could be a function of an advisory committees which should contain a substantial majority of extra-mural academics. It is important that the process and structure be such as to obtain uniform support or at least lack of vocal dissent from the responsible bio-medical community. A part of this will involve convincing people that the goals are feasible and non-utopian, and that some of the benefits will be derived early in the process, rather than a decade or more in the future.

Sincerely, *Best regards,*  
*Sherman M. Weissman*

Sherman M. Weissman, M.D.  
Professor of Human Genetics,  
Medicine, and Molecular  
Biophysics and Biochemistry

SMW:amm



E.I. DU PONT DE NEMOURS & CO. (INC.)  
BIOMEDICAL PRODUCTS DEPARTMENT

March 18, 1986

Dr. Mark W. Bitensky  
Senior Fellow  
Los Alamos National Laboratory  
M881  
Los Alamos, New Mexico 87545

Re: LS-DO-86-1.12-50  
Genome Sequencing Workshop

Dear Dr. Bitensky:

I am in complete agreement with the general consensus of the workshop attendees- sequencing the entire human genome in one concerted effort will be highly beneficial. I believe that the project is highly desirable and could be justified for the following reasons:

- The project will increase our understanding of human genome organization and expression.
- Applications and extensions of this information will be extremely beneficial to society for diagnosis, drug design and therapeutics. These advantages also extend to agricultural applications, which are economically more significant. (GNP<sub>health care</sub> << GNP<sub>food</sub>)
- Most of the sequencing can be done by semi-skilled workers; project will provide jobs and/or training to this politically important sector of the workforce (Could one hire unemployed auto or steel workers?)
- Technology to do the project already exists. Why not reap the benefits when society/government has already expended research funds to develop these procedures?
- Similar expenditures to NASA have been very beneficial to U.S. taxpayers; (e.g. electronics industry). A large financial committment will help the U.S. to maintain its lead in biotechnology.

We are interested in continuing our involvement in this project. DuPont can provide management expertise, lobbying support, or any other assistance that may be necessary in getting this proposal approved. Please do not hesitate to call us if you need any help.

Sincerely,

Maurice A. Kashdan, Ph.D.  
Area Supervisor, Laboratory Operations  
DuPont NEN Research Products

NEN PRODUCTS

549 Albany Street, Boston, Massachusetts 02118 Telephone 617-482-9595 Telex 94-0996



THE UNIVERSITY OF TEXAS AT AUSTIN  
AUSTIN, TEXAS 78712-7818

The Genetics Institute  
Patterson Laboratories  
(512) 471-6268

13 March 1986

Dr. Mark Bitensky  
Los Alamos National Laboratory  
Los Alamos, New Mexico 87545

Dear Mark:

Enclosed are some comments and summaries on certain parts of the DNA sequencing workshop held last week. I never established that I was an "official" workshop leader, but that never stops me from making comments. Please use them however you see fit.

I thought that the workshop was very successful. The issues seem generally to be well defined. Those that are not should not alter the feasibility of sequencing the human genome. I do hope that DOE continues its interest in this project. I will be on a site visit with some of the DOE staff in early April and will do what I can to suppress any signs of waivering on their part.

You and your staff did an exceptional job of organizing and making arrangements for the workshop. We all benefitted.

Sincerely yours,

H. Eldon Sutton

NOTES ON SANTA FE WORKSHOP

Justification. Genetic disease is a major problem of human health. Of every one thousand recognized conceptions, approximately 150 result in miscarriage. Half of these are associated with visible chromosomal abnormalities that cause genic imbalance. An additional unknown number have genetic defects that cannot be seen cytologically. Many genetically abnormal embryos undoubtedly are lost very early and are not recognized as conceptions.

Some one percent of liveborn children have simply inherited defects that cause some degree of health impairment. The most common among populations of European descent are congenital adrenal hyperplasia (CAH) and cystic fibrosis (CF). Sickle cell anemia in blacks and thalassemia among several populations are other major inherited defects. Tay-Sachs disease is a lethal disorder that occurs with a frequency of some 1:3600 births among Ashkenazic Jews. Hundreds of other less common inherited defects are also known.

The total impact of inherited disease must also include morbidity and mortality among adults. Often the inheritance may be complex, and heredity may not be recognized commonly as a major contributing factor. Diseases such as diabetes, heart disease, and some forms of cancer may depend on inherited propensities. Mental disorders such as schizophrenia, which affects one percent of the population, affective disorder, and dyslexia seem to have underlying genetic bases, as yet poorly understood. Such disorders that cut short a person's productive years are a major health burden that adds substantially to the cost of health care.

An additional matter of both clinical and basic interest is the extent of genetic variability in normal persons and possible associations of these variations with risk of disease. There is presumably a standard array of genes, but there is no "standard" DNA sequence.

Any program supported by public funds must be justified ultimately by the public benefits. These include a decrease in human suffering and a reduction in cost of health care. It is not easy to express the dollar value of human suffering and premature loss of life. One can anticipate major advances in alleviating both these situations should the complete human DNA sequence be determined. For example, many inherited defects can be compensated for if the specific mutant genes involved are known. Our very limited knowledge of gene mutations indicates the great variety of DNA alterations that may impair or inactivate any one locus. Knowledge of the specific changes should help design effective therapeutic strategies. There are a number of

low activity enzymes, especially those that depend on vitamin B6 or biotin as coenzymes, that can be made to function within the normal range by high dietary levels of these vitamins. This is established at present by trial and error. There must be hundreds of additional loci for which similar therapeutic approaches are possible, but this will only be accomplished by learning the molecular details of the gene and its variants. The analysis of such details at the level of DNA confers the ability to predict risk much earlier, potentially at the 6 to 8th week of pregnancy, rather than waiting for a disorder to develop, a disorder that may not always be reversible, especially in the case of defects of the nervous system.

Priorities for study. Although the ultimate goal is to establish the DNA sequence of the entire human genome, there are various strategies that might be followed in pursuit of this goal. (1) One strategy is to start sequencing at random points, gradually filling in until there are no gaps. This may be the most efficient in terms of resources. (2) A second approach is to concentrate on one chromosome at a time. This has the advantage of providing information on chromosome structure relatively soon, information that may tell us some very interesting things about chromosome structure and function that we do not know or even suspect. This may be the most interesting approach from the point of view of basic science. (3) A third approach is to start with genes that are known to be transcribed and therefore functional, filling in between them over time. Such regions can be identified by cDNA probes. This would probably be less efficient than (1) or (2) but should yield the earliest benefits to health. A combination of (2) and (3) might be the best strategy, concentrating on each chromosome in turn starting with the known functional regions.

# Yale University

Department of Human Genetics  
Yale University School of Medicine  
1-310 SHM  
P.O. Box 3333  
New Haven, Connecticut 06510-8005

116  
Campus address:  
1-310 Sterling Hall of Medicine  
333 Cedar Street

June 11, 1986

Dr. Alan S. Rabson  
Director  
Division of Cancer Biology  
Building 31, Room 3A-03  
National Cancer Institute  
Bethesda, MD 20205

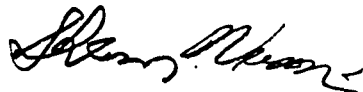
Dear Alan:

Just a brief note in follow-up of our phone call about the genomic sequencing project. There are a few points that I might make in writing. One is that I can see that this project could be divided into three stages. First is the development of a restriction map of the total genome. This would mean that one would get short segments of cloned DNA distributed over genome in an ordered fashion such that each segment would be separated on an average by several hundred thousand base pairs from the next and each segment would have on an average 4,000 base pairs within it. At that point, something between .5 and 2% of the genome would be cloned. The next stage of the project would be to obtain a set of overlapping cosmid clones covering the entire genome. This would certainly be expedited by having the ordered restriction map. Parts of the work could go ahead in parallel with, rather than subsequent to, the development of the restriction map. The third stage of the project would be to proceed with the sequence of the genome. Here there are a number of alternative technologies proposed including George Church's multiplex sequencing, the possibility of sequencing machines using fluorescent nucleotides and possibly other physical methods that are rumored. I'm not working with these procedures directly but from what I know from what had been published so far, my impression is that these procedures have yet to be fully established in the field and evaluated in terms of overall economics as compared with the current methods. At any rate, a sequencing project would be very large, long term, and expensive and a very major part of the benefit would be derived only in the latter stage of the project. In contrast, the restriction map of the genome is a relatively small scale project, requiring scarcely more than 2-4 of the conventional laboratories, and the cloning is no more than an order of magnitude larger, using currently the best available procedures. Having the restriction map together with partial digestion mapping procedures such as we are working on could among a

Dr. Alan Rabson  
Page 2  
June 11, 1986

great many other things provide very rapid scanning and identification of chromosomal deletion and translocation sites and also expedite very much the actual cloning of the DNA at these sites (something of potential relevance for the Cancer Institute). The restriction map also would be of great value for the study of genetic disorders in man and having the restriction map could accelerate cloning effort and according to my perception, skim off some of the cream that would eventually emerge out of a total sequencing project years before the sequencing data would provide these results.

With best regards,



Sherman M. Weissman, M.D.  
Professor of Human Genetics, Medicine  
and Molecular Biophysics and Biochemistry

SMW:rt

## College of Physicians &amp; Surgeons of Columbia University | New York, N.Y. 10032

DEPARTMENT OF GENETICS AND DEVELOPMENT

701 West 168th Street

May 27, 1986

Dr. Alan Rabson  
NCI  
Building 31  
Room 3A03  
Bethesda, Maryland 20892

Dear Alan:

I really enjoyed meeting you and discussing the possible merits of making a physical map of the human genome. I am glad you were enthusiastic about our approach in which a physical map would be constructed prior to any large scale sequencing. Here, as you requested, is a very brief outline of what is involved.

New Methodology Four types of techniques can be used in concert to provide rapid physical mapping of the human DNA. These are:

- a. Isolation of unbroken genomic DNA. This is done by suspending live cells in liquid agarose and after solidification, diffusing reagents into the agarose to remove all cell constituents except the DNA.
- b. Separation of large DNA molecules by pulsed field gel electrophoresis. This works well for DNAs from 10 kB to 10,000 kB and may be extendable to larger molecules.
- c. Specific fragmentation of genomic DNA into large pieces. This cannot be done by normal solution techniques because of the fragility of large DNA. Instead, restriction nucleases are diffused into DNA samples inside agarose. Enzymes are chosen that recognize very infrequent sequences. Not I, which cleaves at GCGCCGC, Sfi I which cleaves at GGCCNNNNGGCC, and Mlu I, which cleaves at ACGCGT, are particularly effective with human DNA. They yield fragments averaging in the range of 250 kB to 1000 kB. Methylase-nuclease combinations can yield even larger fragments.



d. Determining the order of large DNA fragments in the genome. Where a genetic map already exists, or large number of probes are available, direct Southern blotting of separated large fragments can reveal their order just as in conventional methods for mapping smaller DNA regions. However, a more powerful and efficient approach is to screen or select probes that contain just the ends of large fragments. There are two types of such probes: linking probes span rare cutting sites and thus contain just the ends of two contiguous large fragments; jumping probes contain just the ends of a single large fragment. Together, the two types of probes contain all the information needed to place the large fragments in order.

Overall Strategy Here is an outline of a simple scheme for physically mapping the entire human genome. One can work with single chromosomes, one at a time, by using hybrid cell lines containing only a single human chromosome and libraries made from flow sorted DNA from the same chromosome. A typical human chromosome is 150,000 kB in size. This will be cut into 300 fragments by the methods described above, and only 600 clones are required to place these in order. Available mapped markers and cytogenetic markers will serve as bench marks. If cell lines and flow sorted material representing only parts of a single chromosome are available the task will be even simpler, since it is easier to link up small patches of fragments into contiguous units and then order these, rather than work totally blind.

Organization The task requires merging the best available pulsed field gel electrophoresis, the most efficient construction of libraries of specialized clones, chromosome sorting, and state of the art cytogenetics and the construction and maintenance of specialized cell lines. This will best be achieved by orchestrating a tightly knit cooperative effort among a consortium of five to ten laboratories, each contributing specialized technology, and each supervising the part of the parallel effort on particular chromosomes once all of the common groundwork has been completed.

The overall task of constructing a complete human physical map should take about five years and would cost, I estimate, roughly \$10 million per year. It would be efficient if a parallel effort on mouse and Drosophila mapping were coordinated with the human map project, but that clearly should be funded separately.

Justification The first human physical map will consist of a

set of cloned DNA markers spaced at accurately known positions throughout the entire genome. The average resolution of the map will be 500 kB. This is ten times the resolution of the human genetic linkage map being developed by Ray White and others. It is also ten times the resolution of existing cytogenetic methods. This extra resolution should allow the visualization of many oncogene-mediated DNA rearrangements currently invisible cytogenetically. It will also dramatically speed the search for genes associated with inherited cancers. With even extremely tenuous evidence for genetic linkage for a particular tumor (or any other disease, for that matter), one will be able to use the map to select the appropriate DNA probes to provide a clear test of possible inheritance. If this is confirmed, then the map will serve to accelerate the search for the gene involved. The physical map will also serve to calibrate the genetic map. This will reveal whether that map is badly distorted by hot spots for DNA rearrangements, which in turn, allows the discovery of additional loci important in oncogenesis.

The first human physical map will also set the stage for the construction of a higher resolution map. All the techniques needed to proceed efficiently from a 500 kB resolution map to a 50 kB map are already developed. The higher resolution map is an ordered set of cosmid clones spanning the entire genome. This map in turn will set the stage for the sequencing of any or all regions of the genome.

I enclose preprints that describe some of the above issues in more detail. I would be happy to discuss all of these plans further with you and others at NCI. Please share this letter with Ruth and send her my regards.

Sincerely,



Charles R. Cantor  
Professor and Chairman

From: GENETIC ENGINEERING, Vol. 8  
Edited by Jane K. Setlow and Alexander  
(Plenum Publishing Corporation, 1986)

**ANALYSIS OF GENOME ORGANIZATION AND REARRANGEMENTS BY PULSED  
FIELD GRADIENT GEL ELECTROPHORESIS**

Cassandra L. Smith, Peter E. Warburton, Andras Gaal\*,  
and Charles R. Cantor

Department of Human Genetics and Development  
Columbia University, New York, NY

\*LKB-Produkter AB, Bromma, Sweden

**INTRODUCTION**

Conventional gel electrophoretic techniques for DNA analysis are effectively limited to molecules less than 20,000 base pairs (20 kb) in size. Above this size all DNA molecules have such similar mobilities in ordinary agarose gels that no separations can be achieved (1,2). Most recombinant DNA procedures inevitably involve one or more electrophoretic size fractionations. Many strategies for the molecular analysis of genome organization and rearrangements, and some strategies for purification of genes of particular interest, are rate limited by the size of the DNA pieces used in the procedures. Thus the size limitations in conventional electrophoresis place severe restrictions on scope and speed of many important biological experiments.

An alternative method for genomic analysis is based on cytogenetics through direct light microscopic observations of chromosome rearrangements and mapping of specific in situ hybridization of cloned DNA probes to chromosomes. This technique is extremely powerful in organisms like Drosophila where high resolution mapping by in situ hybridization to polytene chromosomes is a routine procedure. However the cytogenetic approach is typically much more limited. Many unicellular organisms fail to show condensed metaphase chromosomes, rendering the cytogenetic approach impossible. Mammalian cells in metaphase show well-condensed chromosomes that can be routinely identified. In situ hybridization, especially of single copy genes, is effective for determining which chromosome a gene is on, and provides some information about sub-chromosomal location, but high resolution mapping has not yet proven reliable.



Between ordinary electrophoresis and cytogenetics lies a DNA size range of almost a thousand-fold, from 10 kb to 10,000 kb that is almost totally unexplored. Such a size range is analogous to the difference between what is observable by the naked eye and the oil immersion lens of the light microscope. In between there is almost certain to be a wealth of important biological information inaccessible to observations at both extremes.

The new technique of pulsed field gradient gel (PFG) electrophoresis (3-5) allows high resolution separations of DNA molecules ranging in size from 10 kb to more than 2500 kb. Thus this technique neatly fills the unexplored DNA size range. The purpose of this paper is to summarize the state of the art of PFG, and to describe some of the potential applications.

#### PRINCIPLES

In aqueous solution DNA behaves like a free draining coil, with a frictional coefficient directly proportional to length. The charge on DNA is also directly proportional to the length. Hence the electrophoretic mobility in solution, which is dependent on the ratio of charge to friction, is molecular weight independent. Therefore DNA molecules in free solution cannot be fractionated by electrophoresis.

Electrophoresis of small DNA molecules in gels fractionates molecules of different size very effectively. The separation occurs because of gel filtration (6). All of the separation power is determined by the sieving properties of the gel matrix; the electrical field just provides for rapid DNA motion.

DNA molecules larger than 20 kb are not fractionated during ordinary gel electrophoresis because they cannot be sieved. These molecules are larger than the pores in a typical agarose gel. However, large DNA molecules can still enter the pores and move in the gel because, under the influence of an electric field, they can distort their shape to match that of the gel pores (Figure 1). The motion of the distorted DNA molecules is similar to what the polymer physicists call reptation. In concentrated solutions, polymer molecules move most easily along the polymer axis like a snake slithering along its path. Similarly the trailing sections of a DNA molecule will follow the leading sections on a complex path through the gel network.

Once the reptation process sets in, and sieving is eliminated, the ability of ordinary gel electrophoresis to fractionate DNA molecules by size is lost. The charge on DNA is still proportional to molecular weight as is the friction. The

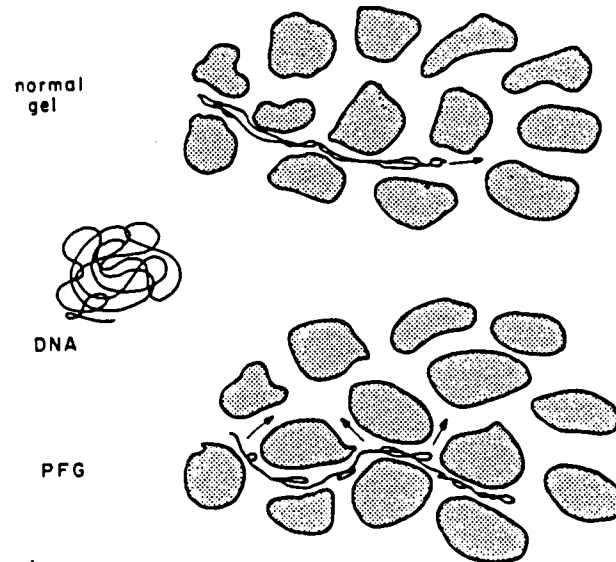


Figure 1. Schematic drawing of a DNA molecule free in solution, during electrophoresis in ordinary agarose gel electrophoresis, and during pulsed field gradient gel electrophoresis.

friction will be determined by the size and shape of the pores in the gel matrix. Since all molecules will sample similar pores, they will experience the same average friction per unit length. The total frictional drag will depend only on the length of pore occupied, which will be directly proportional to the molecular weight of the molecule. Thus all molecules will have the same ratio of charge to friction and thus all will have the same electrophoretic mobility.

In PFG electrophoresis, DNA molecules in agarose gels are forced to continually change the direction in which they are moving. The rate at which a distorted DNA coil can change its configuration is known to be extremely dependent on molecular weight. In free solution this forms the basis of the viscoelastic relaxation technique (7). In a gel matrix no detailed physical modeling of the reorientation of DNA coils has yet been reported. However it seems clear, intuitively, that a short DNA coil that occupies only a short path through the gel pores ought to be able to change direction fairly quickly, while a much longer molecule might face many false starts before it found a new distorted configuration that would allow net movement through the gel.

In a typical PFG electrophoresis experiment, molecules are subjected to alternate electrical fields in nominally perpendicular directions (3-5). Thus the net motion of the molecules will be along the line bisecting the two field directions. Usually the two alternate fields are applied for equal time intervals, called the pulse time (Figure 2). The switch between the two fields occurs in less than 0.1 second which is faster than the reorientation time of most DNA molecules. However the pulse time may be comparable to the reorientation time of a DNA molecule.

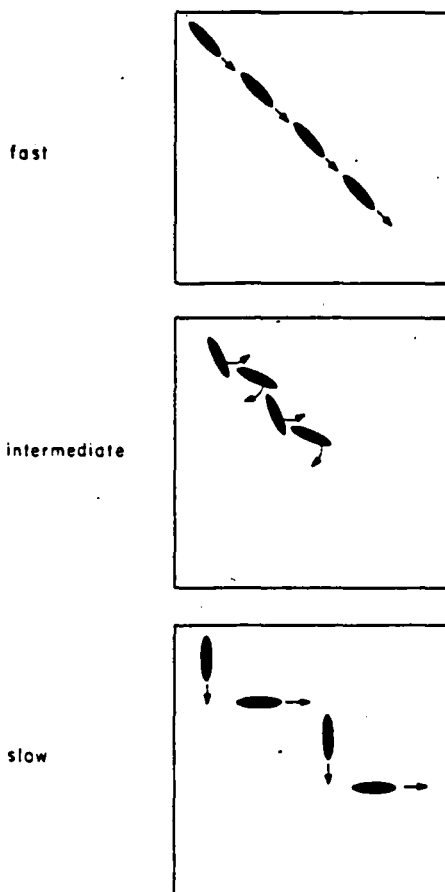


Figure 2. Schematic illustration of DNA migration in response to alternate pulsed perpendicular electric fields. In the top panel the pulse time is much slower than the DNA reorientation time; in the center panel the two times are comparable; in the bottom panel the pulse time is much faster than the reorientation time.

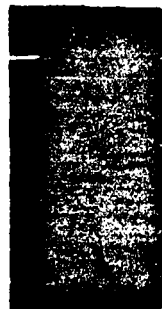
The relative mobility of different size DNA molecules depends critically on the electrical field strength and the pulse time. For example, if the pulse time is much longer than the reorientation time of the molecules in the gel, there is little or no size resolution. What happens is that the molecules move by ordinary electrophoresis first along only one field and then along the other. Since the relative time required for reorientation is negligible compared to the pulse time, the only process occurring is reptation, which is size independent.

The other extreme case occurs when the pulse time is much shorter than the reorientation time. When this occurs, the molecule simply experiences the rapidly fluctuating perpendicular fields as a net constant field midway between them. Ordinary electrophoresis occurs in response to this field and since reptation is occurring, there is no size resolution. At intermediate pulse times, effective size resolution occurs because, during each pulse, the molecules spend a fraction of the pulse time reorienting and the remainder of the pulse time migrating. This fraction, and thus the net motion, will depend directly on the reorientation time. Thus very high resolution size separations can be obtained, for those molecules in which the reorientation time is the same order of magnitude as the pulse time.

In practice, with overall electrical field strengths of 10 V/cm, pulse times of 1 to 10 seconds are optimal for molecules around 100 kb, while times of 50 to 100 seconds are optimal for molecules around 1000 kb. What typically happens for a given pulse time is a window of high resolution size separation extending over a five-fold size range. All smaller molecules move with similar, rapid mobilities while all larger molecules move with similar slow mobilities. The gel pattern which emerges shows a zone of effective separation spaced between two zones with a compressed DNA pattern. Experience thus far suggests that a critical feature in determining the overall resolution is the total number of applied pulses.

#### ELECTRIC FIELD SHAPES

There are currently two experimental configurations in common use for PFG. These are illustrated in Figure 3. In the single inhomogeneous field configuration (si) the vertical field electrodes consist of a single point near or at one corner and a linear array of points that for all practical purposes behaves like a line. This leads to an electrical field strength that continuously decreases as one moves away from the point electrode. The second field is homogeneous and acts to move molecules away from the point electrode. The net result is that molecules are always moving into regions of progressively weaker fields and increasing angles between the two fields (see below).



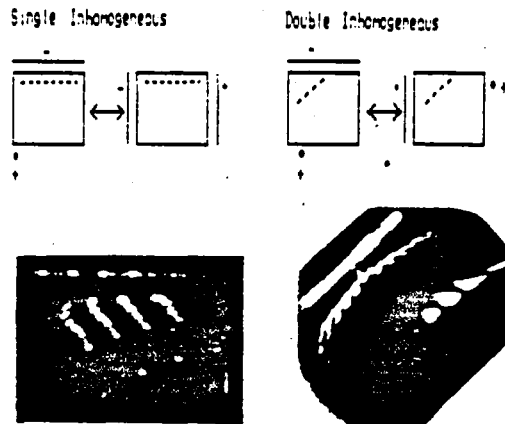


Figure 3. Illustration of the electrode configurations used in the single inhomogeneous and double inhomogeneous modes of pulsed field gradient gel electrophoresis. Also shown is a typical electrophoresis pattern seen for yeast DNA using the two modes.

The si configuration was used for most early PFG experiments because it allows a relatively large number of samples to be loaded for a particular size apparatus (4,8). It has moderately high resolution throughout a fair fraction of the entire gel area. Regions near the single point electrode should be avoided since the large field gradients present here lead to severe shear breakage of large DNA molecules. The major disadvantages of the si configuration are that the bands run at an angle, making quantitative size and mobility measurements difficult, and the resolution in different gel lanes varies quite a bit, so that the overall gel patterns are quite complex and require elaborate description. We have experimented with more complex analogs of the si configuration in which fields are periodically reflected so that molecules first run to the left and then to the right. In this way it is possible to produce gels in which large numbers of samples run vertically with very similar mobilities. However these more complex configurations, in general, show poorer resolution than the optimal simpler configurations.

The double inhomogeneous electrical configuration (di) was used for the very first PFG measurements (3). In this configuration both sets of perpendicular fields have one point electrode and one line. Thus the pulsing is symmetrical when viewed along the line bisecting the two fields, as shown in Figure 3. In the di configuration the DNA bands run along the diagonal and best results are obtained when the samples are loaded along a line perpendicular to this diagonal (5). The resolution of different lanes is somewhat variable and only the center few lanes run perfectly straight. However the symmetry of the overall gel



pattern is much more aesthetic and conducive to simple description and quantitative analysis.

If pulsed electrophoresis is carried out with perpendicular uniform (homogeneous) electrical fields, very poor size resolution is observed (4). DNA molecules of a particular size migrate as a relatively broad band and different size molecules have similar, though not identical, mobilities. If one or both electrical fields is inhomogeneous, like that generated by the electrode configurations shown in Figure 3, much sharper bands and much more molecular weight discrimination are usually observed. There are at least three reasons for this. When electrical field gradients are present in typical si and di configurations, the actual angle between the two alternating fields is greater than 90 degrees in virtually all regions of the working area of the gel. For configurations that afford particularly good resolution we calculate that the angles between the two fields range from 110 to 150 degrees. Although the optimum angle is still unknown, it is apparent that fields 90 degrees apart or less give very poor resolution (Figure 4) while fields 180 degrees apart would be expected to produce no net electrophoretic motion.

A second effect, with field gradients present, is that the component of the electrical field along the diagonal of the gel (parallel to the net direction of motion) keeps getting weaker. This occurs primarily because the angle between the two fields keeps increasing, but also because the fields themselves grow weaker. The net result is that the leading edge of each DNA band moves more slowly along the diagonal than the trailing edge. Thus each band is self-sharpening.

A third effect of field gradients can be understood by considering the effect of overall field strength on PFG. What matters during each pulse is the relative time spent reorienting and the actual distance covered during the remainder of the pulse. A weaker electrical field will decrease the distance covered, and it also retards the reorientation process. Both phenomena are additive and imply that at weaker fields, longer pulse times will be needed for optimum separations in a particular size class. When a field gradient is present, at constant pulse time, as molecules move through the gel they encounter progressively weaker fields. Thus the effective pulse time they feel becomes shorter and shorter. The field gradient acts like a constantly decreasing program of pulse times. The smallest molecules that move the fastest reach the weakest fields and largest angles where they will have the optimal size fractionation. Once the appropriate match between the field shape and the pulse time can be found it is possible to obtain high resolution size fractionations from 10 kb to 1500 kb with a single pulse time.

The electrical field gradient is a tensor quantity. The field shapes used in typical di and si configurations lead to



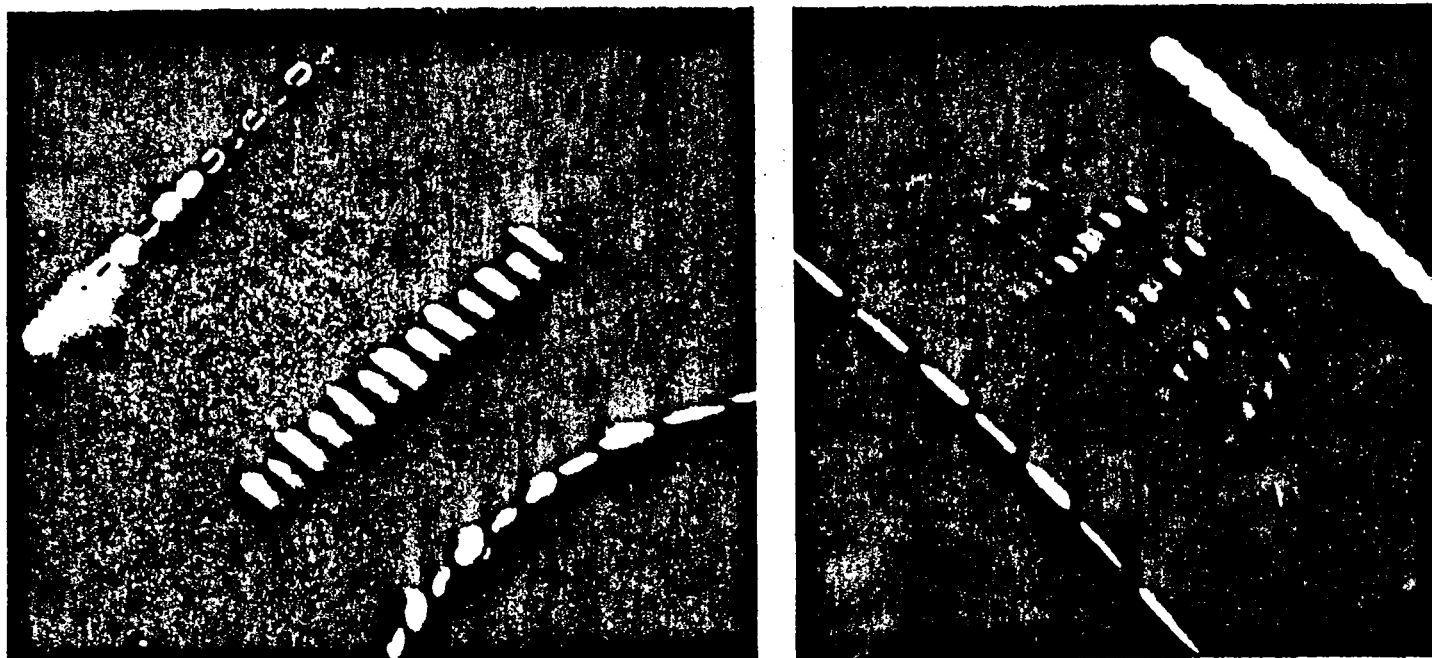


Figure 4. The effect of angle between the alternate electrical fields on resolution in pulsed field gradient gel electrophoresis. In both cases shown the samples are identical yeast chromosomal DNA and the double inhomogeneous configuration was used. In the left panel the point electrode was positioned at the midpoint of the edge of the gel producing angles between the two electrical fields that are less than 90 degrees of all locations within the gel from the sample wells to the center of the gel. In the right panel the point electrode was situated essentially as shown for the di configuration in Figure 3, producing angles between the fields that are much larger than 90 degrees for all locations in the gel from the sample wells to the position of the fastest moving bands.



gradients parallel to the net direction of DNA motion and gradients perpendicular to it. The parallel components of the field gradients lead to sharpening of each DNA band as discussed above. The perpendicular components lead to spreading and thinning of the band along the direction perpendicular to net motion. Some bands, especially those that have migrated a long distance, become much wider than the original applied DNA samples. This band spreading is particularly pronounced for most di configurations. Thus the di configuration is usually limited to a smaller number of samples than the si configuration and it was avoided in most of our early studies. However, as more systematic studies of electrical field shapes were carried out, electrode positions could be found that minimized lateral spreading of the bands while preserving considerable band sharpening. This makes the di configuration superior for most applications.

In summary, from numerous experiments with different electrical field geometries, it is clear that the overall pattern of PFG separation is very sensitive to electrical field shape. The resolution is extremely sensitive to the position of the point electrode and also the length and position of the line electrode. As a result, the performance of a particular apparatus can be optimized by judicious tuning of the electrical field shapes.

#### EXPERIMENTAL APPARATUS

Although it is potentially possible to build PFG apparatus with a wide variety of different electrode types, we have found the use of diode-linked electrodes particularly convenient. Since a diode will conduct in only one direction, a diode-linked electrode is active when voltage is applied, but nonconducting otherwise. In this way the field generated by one electrode array cannot be distorted by the presence of another array.

Using diode electrodes we have constructed square horizontal submarine PFG apparatus in a range of sizes including 10, 20, 24, 28, 36, and 55 cm electrode arrays (Figure 5). In practice the largest convenient gel surface used inside these arrays has been 20 x 20 cm. Larger gels are expensive, difficult to handle, and hard to immobilize in a typical submarine gel apparatus. There seems to be no compelling reason to use larger gels at this time.

The general behavior of the different size PFG boxes is similar. However there is some tendency for larger boxes to perform better for larger DNAs. The major reason for this is probably shear breakage. In a small box it is very difficult to avoid regions of the gel with high field gradients. In a very large apparatus, one can place a gel in the center where the field gradients are small and fairly regular across the gel lanes. Thus shear breakage is minimized, and, in addition, the behavior of different gel lanes is quite similar.

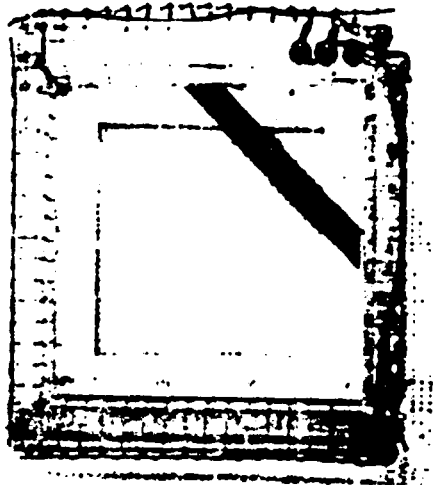


Figure 5. Photograph of a typical 20 x 20 cm pulsed field gradient gel electrophoresis apparatus.

Running times for the various boxes at overall fields of 10 V/cm are typically 1 hour per cm of box size, with 1% agarose gels. However improved resolution can often be obtained with longer running times. Increasing agarose concentrations require still longer running times but frequently produce sharper DNA bands. Lower agarose concentrations decrease resolution and have not been very successful. The typical buffer used is TBE (100 mM Tris-borate, 2 mM EDTA) but half this concentration is equally effective. There are some indications that different anions can have a marked effect on electrophoresis patterns and sample loading, similar to what is observed with ordinary DNA gel electrophoresis. However no systematic studies exist yet. Temperature is a very important variable and preliminary indications suggest that temperatures around 15°C are optimal for most applications.

#### DNA SAMPLES

DNAs from 10 kb up to a few hundred kb can be loaded into gel wells as aqueous solutions just as is done in ordinary agarose gel electrophoresis. With such samples the usual practice is to apply the pulsed field in only one direction for 20 to 45 min to run the samples into the running gel before applying the pulsed perpendicular fields. Larger DNA molecules cannot be handled in solution without severe degradation due to shear breakage. The best way to circumvent this is to prepare the DNA

directly in a solid agarose plug, called a gel insert (4). The insert is put directly into the gel slot to load the sample for electrophoresis. With the di mode pulsed perpendicular fields can be applied immediately. With the si mode a 20 to 45 min run in period with only a single pulsed field is still recommended.

To prepare high molecular weight DNA, intact live cells are suspended in liquid low-gelling agarose which is poured into a mold and quickly cooled (Figure 6). The concentration of cells is chosen to provide typically 0.5 to 20  $\mu$ g DNA in a 10x5x2 mm gel insert. Ultimately anywhere from 1/6 to a whole insert will be used for each electrophoresis sample. Earlier protocols employed higher cell concentrations but subsequent work has shown that this leads to significant loss in electrophoretic resolution. A detailed protocol for preparing DNA samples is provided in the Appendix.

The gel insert is treated with whatever combination of detergents and enzymes is necessary to remove cell walls, cell membranes, RNA and proteins and leave, eventually, naked DNA. The first procedures were worked out for yeast. These used high concentrations of EDTA to inactivate cellular nucleases and extensive treatment with proteinase K in the presence of

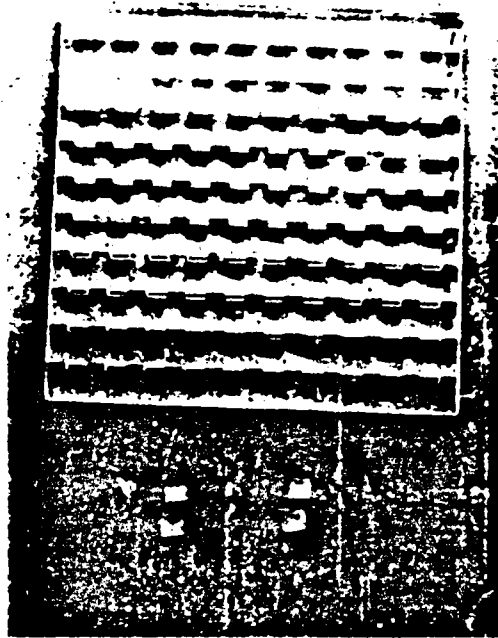


Figure 6. Photograph of a mold used for preparing gel inserts and several inserts.

detergent to remove all proteins. Modifications of these procedures have now proven successful in E. coli and other bacteria, trypanosomes, Leishmania, Plasmodium and other unicellular protozoa, Drosophila and mammalian cell lines. There is no reason to think that any source of DNA will be refractory to this kind of a procedure.

The gel insert method for preparing DNA has numerous practical advantages. No centrifugations or precipitations are required. Hundreds of replica samples can be prepared simultaneously, each containing the identical amount of DNA. In agarose inserts, at 0.5 M EDTA concentrations, high molecular DNA shows amazing stability. No detectable double strand breakage is observed for molecules as big as 1000 kb over several years at 4°C or several weeks at room temperature to 37°C. Thus it is routinely possible to mail DNA samples in gel inserts with no degradation. One clear advantage of the gel insert procedure is that it could be carried out in the field with relative ease. The most complex piece of portable apparatus required would be a 50°C temperature bath for the proteinase K digestion.

DNA prepared in gel inserts is free from detectable double-strand breaks up to the current limits of PFG electrophoresis, around 4000 kb. There are some variations in the behavior of the inserts depending on the size of the cells used to prepare them. Although there have been few systematic studies, intuition suggests that it is best to use cells that are not rapidly dividing. It is known that topological complexity of DNA can have a profound effect on PFG electrophoretic behavior. Thus replication forks could lead to complications in the electrophoresis. There are other effects of DNA topology that are still unexplained. For example, DNA pieces containing tandemly repeated ribosomal DNA genes show retarded mobility in PFG electrophoresis (4,8, and unpublished results).

#### LENGTH STANDARDS

Yeast is the organism that has been most extensively characterized by PFG electrophoresis, and various yeast strains are convenient standards for monitoring the behavior of PFG apparatus. However the sizes of all of the yeast chromosomes in any strain are not yet known directly. Thus for quantitative studies it is preferable to use viral DNAs as length standards. While a few viruses like bacteriophages T7, P1, T2 or T4 and G provide convenient markers, the best set of length standards appears to be tandem linear concatemers of bacteriophage lambda. Various lambda strains are available with monomeric lengths from 40 kb to 50 kb. We have found it convenient to use lambda vir, which is 42.5 kb.

The ends of linear lambda DNA have 12 complementary single stranded bases. At low concentration the monomeric circle is the

thermodynamically stable form at room temperature or below. At concentrations approaching 0.5 mg/ml, an infinite head to tail linear concatemer is preferred (9). In practice, what will limit the growth of this concatemer is molecules with damaged ends which act as chain terminators. From such samples we have resolved linear concatemers as large as 31-mers and larger material is evident. The use of lambda ladders provides accurate length standards up to about 1500 kb. It remains to be seen what kinds of samples will provide satisfactory length standards for even larger sizes.

Lambda ladders also provide a test of the electrophoretic resolution at each molecular weight. In general, using the di configuration we find that the net mobility of DNA is roughly a linear function of molecular weight from lambda monomer up to some particular lambda n-mer where the value of n depends on the pulse time. Between that value and some larger critical length resolution is optimal and appears to be 5 kb or better. Above the critical length resolution drops off abruptly (Figure 7). By choosing an optimal pulse time it is possible to resolve large DNA molecules differing by less than 1% in length across a wide size range.

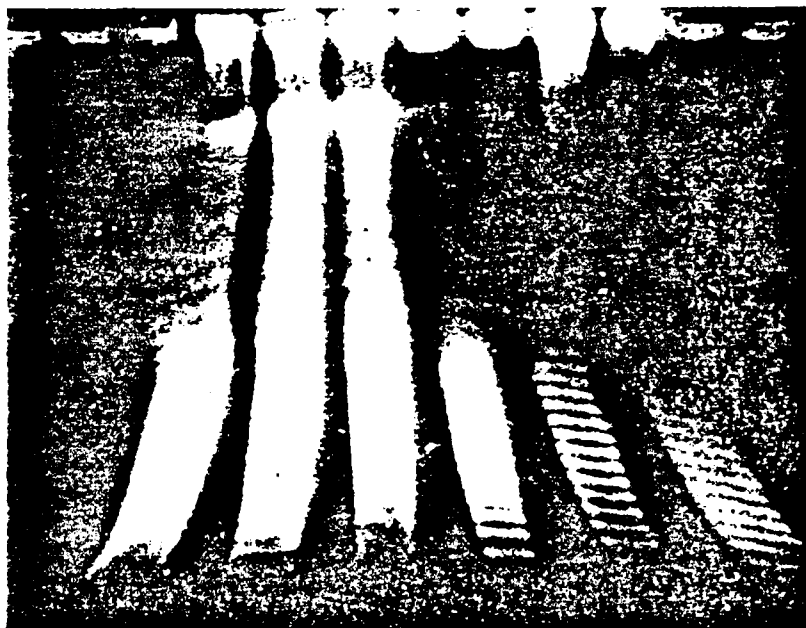


Figure 7. Pulsed field gradient gel electrophoresis of a lambda ladder formed by tandemly annealed lambda vir DNA. Each successive band is 42.5 kb larger.

SAMPLE ANALYSIS AND RECOVERY

In general most other procedures for handling DNA in PFG electrophoresis parallel those in ordinary agarose electrophoresis. After running, the gel patterns are stable, essentially indefinitely. DNAs can be visualized by staining with ethidium bromide and photographed under ultraviolet light. However great care should be taken to minimize the exposure because photobreakage of large DNA molecules is quite an efficient process. Molecules can be transferred to nitrocellulose or other membranes by adaptations of standard blotting procedures. In our hands, thus far, UV breakage of DNA prior to Southern transfer has proven most satisfactory. DNA molecules can also be electroeluted by placing an agarose slice inside a dialysis sack in the PFG apparatus and applying pulsed alternate perpendicular fields. The resulting material is suitable for subsequent digestion with restriction nucleases and cloning. In this way one can construct a DNA library from any large piece of DNA fractionated by PFG electrophoresis.



CYTOGENETICS BY ELECTROPHORESIS

Many unicellular organisms have chromosomal DNA molecules that fall into the size range currently accessible to PFG. This includes various yeasts (4,5,10,11) and numerous parasitic protozoa (8,9,12-16). For such samples PFG provides what is essentially cytogenetics by electrophoresis. Genes can be mapped to chromosomes by Southern blotting which is far easier than in situ hybridization, even if it were possible with these organisms, which it isn't. DNA insertions, deletions and reciprocal translocations can all be visualized directly by changes in the sizes of chromosomal DNAs and confirmed by watching the shift of known chromosome-specific DNA probes. A few chromosomes have been observed to show anomalous behavior which is presumably related to some kind of topological complexity. An example, discussed earlier, is chromosomes with large tandem arrays of ribosomal genes which frequently have difficulty entering the gel at all.

For many simple organisms PFG can provide an instant snapshot of the genome. It is especially useful for examining the number and chromosomal distribution of moderately repeated elements like yeast TY1 elements since one can study the entire family at once. The general picture that has emerged from studies thus far is that of genomes in considerable turmoil. Chromosome length polymorphisms and rearrangements abound, even among closely related strains. In some parasitic organisms these DNA rearrangements are exploited to generate antigenic diversity that is valuable to the survival of the parasite against the



hostile environment of the immune defense system of the host (8). For non-parasitic organisms, the role played by such karyotypic variations remains to be seen.

GENERATION OF LARGE DNA FRAGMENTS

Although the sizes of DNA molecules handled by current PFG technology are impressive, the chromosomal DNAs of most organisms are still too large to be handled as intact molecules. The solution of this problem is to cut intact chromosomal DNAs into well defined fragments that fall in the 10 to 2000 kb size range now accessible by PFG. This has proven possible through the use of restriction nucleases with relatively rare cutting sites. Since handling of large DNAs in solution would result in unacceptable shear breakage, the procedures that have proven effective are to diffuse enzymes into agarose inserts of DNA and allow digestion to occur in situ. The restriction enzymes are then removed by proteinase K treatment in the presence of detergent and high concentrations of EDTA. Failure to do this leads to serious loss in resolution. This is presumably caused by restriction enzymes remaining bound to DNA and altering its PFG electrophoretic properties. A detailed protocol for restriction nuclease digestion of DNA in agarose inserts is provided in the Appendix.

There are two restriction nucleases commercially available that have eight base pair recognition sites. These enzymes are NotI and SfiI (Table 1). Both work quite successfully when handled as described above, and both have yielded clean patterns of DNA digestion on samples that include E. coli, yeast, mouse and human cells. There is one restriction nuclease with a seven base pair recognition sequence, RsrII (17), but unfortunately this sequence seems to be present quite frequently in many organisms as determined by scanning nucleic acid sequence data banks.

There are a number of restriction enzymes that have six base recognition sites but these sites are relatively rare in particular genomes (Table 1). For example in the DNA of higher eukaryotes the sequence CpG is almost five-fold rarer than expected statistically from the base composition (18). Thus enzymes that have one or two CpG's in their recognition site tend to give large DNA fragments. When the target DNA is very AT-rich or GC-rich one can clearly pick enzymes to bias the pattern of cutting towards larger fragments. It has been our experience that most enzymes tried will eventually work in agarose protocols. However we have noticed that for some enzymes there is a marked, severe, batch to batch variation in the effectiveness of agarose samples in supporting total nuclease digestion. The reason for this is unknown but it is clearly a very significant experimental variable. At present the lack of additional

Table 1  
Restriction nucleases useful for generating large DNA fragments

Enzyme	Sequence	Source of DNA				
		Phages	Bacteria	Viruses	Mammals	Others
<u>NotI</u>	GCGGCCGC	<u>0</u>	<u>1</u>	<u>1</u>	<u>1</u>	<u>1</u>
<u>SfiI</u>	GGCCN5GGCC	<u>1</u>	<u>2</u>	3	<u>1</u>	<u>1</u>
<u>PvuI</u>	CGATCG	3	13	4	<u>1</u>	<u>2</u>
<u>SalI</u>	GTCGAC	<u>2</u>	7	6	<u>2</u>	4
<u>MluI</u>	ACGCGT	4	14	4	<u>1</u>	<u>2</u>
<u>SacI</u>	GAGCTC	4	5	<u>1</u>	4	6
<u>NruI</u>	TCGCGA	5	12	7	<u>1</u>	3
<u>ApaI</u>	GGGCC	7	4	11	9	4
<u>ScaI</u>	AGTACT	3	7	11	6	11
Files Searched		148	565	904	1846	1491

Genebank was scanned for the sequences indicated. Each entry shows the % of the files for each type of DNA that contains one or more enzyme cutting site.

Enzymes that cut especially rarely are indicated by underlined numbers.

restriction nucleases with relatively rare recognition sequences is the single most severe limitation on PFG analysis of large DNAs. This is not surprising since, before the advent of PFG, there was no convenient way to see if such enzymes even existed. Now one can anticipate that a thorough search will be made, and, more than likely, many highly specific useful nucleases will emerge.

It is possible to increase the specificity of nuclease action by combining restriction nuclease digestion and methylation, and thus to increase the length of the average size fragments produced. There are two different ways to do this. Some restriction nucleases are blocked by prior methylation. Thus by treating a DNA substrate first with the appropriate

methylase, one can increase the effective size of the DNA fragments a subsequent nuclease produces (19, and see many examples cited in the 1985 New England Biolabs Catalog). Alternatively, some restriction nucleases like DpnI require prior methylation. In this case, by using one or more methylases whose specificity overlaps that of the DpnI cutting site one can create a situation where two separate recognition events by methylases are required to generate a single DpnI cleavage (20). In this way one can generate recognition sites that have effectively, 8, 9, 10 or even more base pairs. To date such schemes have proven very effective when applied to small DNAs in free solution. However they have been less effective in generating stoichiometric cutting inside agarose, presumably because of the considerable fastidiousness of the DpnI enzyme.

There are other potential highly specific cutting schemes that remain to be tried to provide samples for PFG analysis. These include the use in nucleases known to be involved in site-specific DNA insertion-excision or recombination reactions (21,22). It also seems likely that methods will be found to exploit unusual DNA structures to generate specific cuts. Finally, the potential exists to use DNA itself as a reagent to direct highly specific DNA cleavages. Thus far such schemes have been limited to single-stranded cleavage but it seems only a matter of time until more general highly specific double-strand DNA cleavage techniques emerge (23-25).

MACRORESTRICTION MAPPING

The eight base pair-specific restriction nucleases NotI and SfiI have been used to cleave E. coli DNA, mouse and human DNA into large fragments. These range in size from less than 25 to more than 1000 kb. Individual fragments can be identified by Southern blotting and hybridization with cloned single copy DNA probes. It is then relatively trivial to tell, for example, whether two genes, believed to be closely linked, appear in the same size DNA fragment. This is suggestive but not conclusive evidence that they are physically linked.

A major goal for molecular genetics is to develop physical maps of complete genomes of organisms of particular interest. A first stage towards such a goal would be to piece together the order of the large fragments generated by restriction nucleases with rare cutting sites. Even the smallest mammalian chromosome is likely to consist of a few hundred such DNA pieces. Thus one can examine only one region at a time. As a model for such studies, and because of its own intrinsic interest, we have concentrated our initial efforts on trying to construct a physical map of the E. coli genome. This is a single 5000 kb circle. Thus it is the size of a typical band seen cytogenetically for a mammalian chromosome. The circle is an added complication;

however the much smaller complexity of E. coli, compared with even a single human chromosome, more than compensates for this.

Statistically, NotI would be expected to cleave the E. coli chromosome into about 80 pieces. However, in practice, NotI generates only 20 fragments from a typical E. coli strain (Figure 8). One is a nominally 1200 kb DNA piece (but this contains several rDNA genes and may have an anomalous mobility) while all the others fall into the size range of 25 to 400 kb. In some strains every single piece is resolved. Simply blotting such gels with DNA probes with known positions on the E. coli genetic map is allowing us to piece together the physical map of the genome.

Additional information about the physical map of E. coli DNA is provided from known DNA insertions and rearrangements. For example, lambda lysogens have been characterized with prophage DNA inserted at different loci around the genome. Lambda itself has no NotI sites. Thus depending on whether a given lysogen is a single or multiple tandem insertion event, the length of the NotI DNA fragment containing the insertion site will increase by approximately 50 kb or some multiple of this. Given the resolution of PFG such insertions are readily detected on ethidium

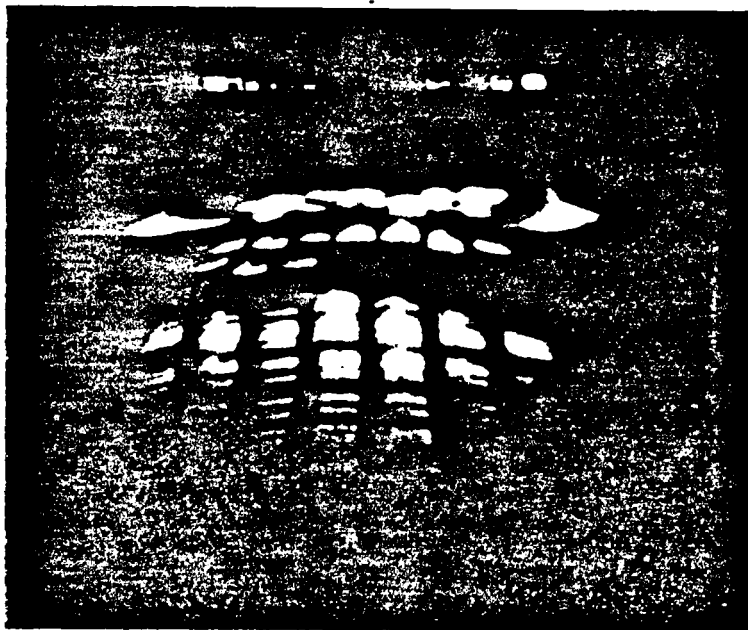


Figure 8. Pulsed field gradient gel electrophoresis of fragments of the E. coli genome generated by a complete digest with the restriction nuclease NotI.

stained gels and easily confirmed by hybridizing Southern blots with lambda DNA. In a similar way, F+ insertions have been characterized by monitoring the 90 kb size increase in particular bands in a variety of Hfr strains. Finally, a number of E. coli strains are available with known inversions among ribosomal cistrons, or known deletions. These are providing confirmation of band assignments made by blotting.

The approaches described above seem destined to provide a complete physical map of E. coli shortly. They are clearly applicable to any other prokaryote with a comparable sized genome so long as an ample supply of genetic mapping data exists. Clearly such an approach should also be successful for individual yeast chromosomes. However this approach could not be used for larger genomes or for organisms where no genetic maps were available. As more rare cutting schemes become available it should be possible to use these to construct macrorestriction maps by overlapping cutting patterns in exactly the same way as currently used on a much smaller scale for plasmids and phages. However this is a tedious approach and it would be highly desirable to have a more general approach.

The relatively large size of DNA fragments separable by PFG has encouraged us to devise a new simple mapping scheme that depends on large fragments for its efficiency. We call this approach a junction fragment analysis. The basis of this scheme is to generate large DNA fragments by an enzyme with a rare cutting site, separate these fragments by PFG electrophoresis and then analyze the PFG separation by hybridization with cloned DNAs containing one single rare cutting site each. Each clone will hybridize to two large DNA fragments and these must be adjacent in the physical map (Figure 9). In cases where more than one large fragment co-migrates, secondary cutting with another rare cutter should provide unambiguous identification, in most cases. In the remaining cases, rearrangements, or screening against a hybrid cell panel containing chromosome fragments, will be required.

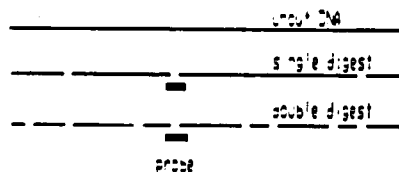


Figure 9. Schematic illustration of how a junction probe can be used to determine that two large restriction fragments are adjacent in the genome.

The major obstacle in the junction fragment approach is to obtain a complete set of clones representing all rare cutting sites in the genome. For prokaryotes this should be relatively straightforward since the number of clones required is small. Directional cloning procedures have been developed for proceeding from a complete genomic library to a sub-library containing only junction fragments. One first clones DNA fragments containing one rare site and one common restriction site. Next one uses these clones to screen a genomic library made from a common restriction site to pick out those clones that contain junction fragments. This is effective for enzymes like *NotI* that generate clonable ends. For enzymes like *SfiI* which do not produce sticky ends, one can, instead, differentially label *SfiI*-containing clones from one library and use these to screen a second library for the presence of junction fragments.

The junction fragments (and, in the case of enzymes like *NotI*, the directionally cloned fragments used to find them) serve two functions. First they are used to identify neighboring large fragments. Next they can be used in the macro-analog of the Smith-Birnstiel procedure (26,27) rapidly to map more common cutting sites within each large fragment (Figure 10). A total digest is made with the rare cutter, a partial digest, ideally with single hit kinetics, is made with the common cutter. The mixture of DNA fragments is fractionated by PFG electrophoresis and then a Southern blot is probed with one side of the junction fragment.

By this indirect end labeling, one will detect a ladder of DNA fragments progressing from the common cutting site closest to the rare site to the intact rare site fragment. The size of each band gives the location of each cutting site. Thus a single gel lane will provide a restriction map of the entire large fragment.

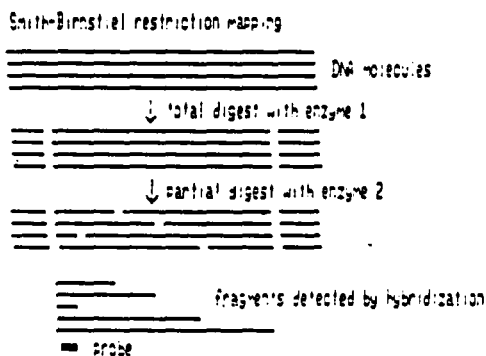


Figure 10. Scheme for rapidly mapping additional restriction enzyme cutting sites of a large DNA fragment. This is the Smith-Birnstiel method scaled to larger sizes.

This procedure is relatively simple when two rare cutters are used. For example, SacI will cut a typical E. coli NotI band into three sub-fragments. When more common cutters are used the patterns will become more complex but they should still be analyzable. For example, a 400 kb NotI fragment should contain 20 to 40 sites for an enzyme like BamHI with relatively uncommon sites. Given the current resolution of PFG it is likely that almost all of these sites will be spaced far enough apart to yield resolvable bands. The major potential problem is detecting the fragments. For a single copy probe, one will have to have sufficient sensitivity to detect 0.01 copy per genome to observe the pattern of fragments resulting from partial digestion. This is no difficulty with small genomes and it should still be possible with mammalian cells where such detection sensitivities have been reported in genomic sequencing and nucleosome footprinting. Thus far, optimized Southern blot hybridizations of fragments from PFG gels show signal to noise comparable with that seen in ordinary gels so one can be optimistic that macro-Smith-Birnstiel analysis should be practical.

All of the junction mapping procedures described above have been tested on E. coli and all appear to work quite satisfactorily. In principle there does not appear to be any reason why these procedures will not generalize rapidly to allow application to all genomes. For higher organisms one will have to start with single chromosome libraries and use hybrid cell lines with single chromosomes from one species on a background of non-crosshybridizing chromosomes from another species. Such materials are already available for many human chromosomes. Libraries and cell lines representing only part of a single chromosome will be even more desirable and procedures for constructing these are available. The major complication for the genomic analysis of higher organisms will undoubtedly be repeated DNA sequences. It is possible to screen libraries to remove clones containing highly repeated sequences. However the result is usually a library far too incomplete for the types of mapping approaches described above. However it should be possible to adapt the junction library approach to small enough DNA fragments to minimize the inclusion of highly repeated sequences. There is always the risk that a supposedly rare cutting site will be found in or associated with a repeated sequence. In this unfortunate case the site will be useless.

APPLICATIONS OF PFG TECHNIQUES

PFG is too new a technique and too different a technique for all of its major applications to have been tested yet, or even thought through. However some general classes of applications have been tested by preliminary feasibility studies and some

other applications seem to have no real obstacles although they have not yet been tested.

The methods by which DNA samples are prepared for PFG analysis are extremely gentle and effortless. They should be applicable to all DNA and RNA samples regardless of whether these are destined for the eventual analysis of large DNA fragments. Gel inserts work just fine in ordinary gel electrophoresis and the improved intactness of the sample can frequently produce pleasant surprises. The stability of DNA in agarose and the convenience of modular replica DNA samples should encourage all kinds of new automated procedures for DNA analysis.

PFG should allow the construction of a physical map of the DNA of any simple organism. This will provide many of the features of genetic analysis even for those organisms where no established genetic techniques exist. Even before a physical map is available, PFG will allow routine screening for DNA insertions, deletions, rearrangements, and gene amplification. A wide range of questions in evolution and regulatory organization can be answered just by knowing whether two or more particular genes are near by in the genome and PFG makes such information readily accessible.

Genes performing related functions quite commonly occur in physical clusters. Many genetic engineering tasks would be rendered much easier if such clusters could be handled intact. PFG greatly simplifies the analysis of such clusters. It can also provide purified DNA fractions in the 20 to 250 kb size range. These samples can be handled in ordinary solution. Such DNAs are an important intermediate in new advances in cloning technology such as the construction of jumping libraries for rapid chromosome walking. The availability of DNA molecules in this size range should encourage the construction of cloning vectors capable of handling large inserts. Such vectors would have major impact on a wide variety of biological and technological problems.

For organisms like parasites with very mobile genomes, it is often difficult to distinguish between closely related species or subspecies. It would also be desirable for epidemiological studies to be able to distinguish the geographic origin of particular parasite isolates. Preliminary studies suggest that PFG has excellent promise for providing all of this information. For example the pattern of chromosomal DNAs of a number of closely related Leishmania species is sufficiently distinct that it was possible to diagnose the organism infecting a patient by a single PFG electrophoretic analysis.

For higher organisms, PFG should pave the way for rapid physical mapping of the genome. Rapid progress in linkage mapping with restriction fragment length polymorphisms (28) will soon provide a coarse genetic map for all human chromosomes. Similar information is also becoming available in the mouse.



Here established genetic techniques will help lead from a coarse map to a fine map. In humans the task will be more difficult. PFG seems a natural way to fill in physical linkage between polymorphic markers spaced 5000 to 10,000 kb apart. It also provides a way to isolate large pieces of DNA neighboring polymorphic sites linked to particular disease genes. In this way PFG should greatly speed the search for more markers with better diagnostic value and ultimately provide DNA containing the disease gene itself. This possibility is currently being tested by starting from markers 5000 kb from Huntington's disease (29) and attempting to use them to find DNA closer to the gene responsible for the disease. The task of doing this is still quite difficult, but by combining PFG with improved cell culture and cytogenetic methods it now seems feasible.

It would seem that PFG offers considerable opportunities for improved clinical diagnosis. For example, DNA translocations near or at oncogenes are frequently associated with cancer (see 30,31 for typical examples). The difficulty in analyzing such translocations by ordinary Southern blotting is that they can occur over a relatively wide DNA region, say 50 kb or more, and still result in a malignant phenotype. To be sure of detecting and classifying all possible translocations one must be able to look at least 50 kb from the oncogene. It would be difficult to do this in one step by ordinary electrophoresis, without collecting a large number of different DNA probes, but it will be a relatively easy matter to accomplish the analysis by PFG.

Cytogenetics is useful for clinical diagnosis by providing an overview of the entire genome. Recombinant DNA probes allow a detailed view, one gene at a time or even one nucleotide at a time. Both approaches have found many useful clinical correlates. Three orders of magnitude in size stand between these two technologies. All of this is now accessible by PFG. There is every reason to think that this large domain of structural information also contains useful clinical correlates. All one has to do is look.

Acknowledgments: This work was supported, in part, by a grant from the U.S. Public Health Service, GM 14825.

REFERENCES

- 1 Fangman, W.L. (1978) Nucl. Acids Res. 5, 653-665.
- 2 Serwer, P. (1980) Biochemistry 19, 3001-3004.
- 3 Schwartz, D.C., Saffran, W., Welsh, J., Hass, R., Goldenberg, M. and Cantor, C.R. (1982) Cold Spring Harbor Symp. Quant. Biol. 47, 189-195.
- 4 Schwartz, D. and Cantor, C.R. (1984) Cell 37, 67-75.
- 5 Carle, G.F. and Olson, M.V. (1984) Nucl. Acids Res. 12, 5647-5664.

6 Cantor, C.R. and Schimmel, P.R. (1980) *Biophysical Chemistry*, pp. 676-682, W.H. Freeman and Company, San Francisco, CA.

7 Klotz, L.C. and Zimm, B.H. (1972) *Macromolecules* 5, 471-481.

8 Van der Ploeg, L.H.T., Schwartz, D.C., Cantor, C.R. and Borst, P. (1984) *Cell* 37, 77-84.

9 Wang, J.C. and Davidson, N. (1966) *J. Mol. Biol.* 19, 469-482.

10 Carle, G.F. and Olson, M.V. (1985) *Proc. Nat. Acad. Sci. U.S.A.* 83, 3756-3760.

11 Helter, P., Mann, C., Snyder, M. and Davis, R.W. (1985) *Cell* 40, 381-392.

12 Borst, P., Bernards, A., Van der Ploeg, L.H.T., Michels, P.A.M., Liu, A.Y.C., De Lange, T., Sloof, P., Schwartz, D.C. and Cantor, C.R. (1983) in *Gene Expression* (Hamer, D.H. and Rosenberg, M.J., eds.), pp. 413-435, Alan R. Liss, Inc., New York, NY.

13 Van der Ploeg, L.H.T., Cornelissen, A.W.C.A., Michels, P.A.M. and Borst, P. (1984) *Cell* 39, 213-221.

14 Van der Ploeg, L.H.T., Cornelissen, A.W.C.A., Barry, J.D. and Borst, P. (1984) *EMBO J.* 3, 3109-3115.

15 Kemp, D.J., Corcoran, L.M., Coppel, R.L., Stahl, H.D., Bianco, A.E., Brown, G.V. and Anders, R.F. (1985) *Nature* 315, 347-350.

16 Van der Ploeg, L.H.T., Smits, M. Connudura, T., Vermeulen, A., Meuwissen, J.H.E.T. and Langsley, G. (1985) *Science* 229, 658-661.

17 O'Connor, C.D., Metcalf, E., Wrighton, C.J., Harris, T.J.R. and Saunders, J.R. (1984) *Nucl. Acids. Res.* 12, 6701-6708.

18 Lennon, G.G. and Fraser, N.W. (1983) *J. Mol. Evol.* 19, 286-288.

19 McClelland, M. and Nelson, M. (1985) *Nucl. Acids. Res.* 13, r201-r207.

20 McClelland, M., Kessler, L.G. and Bittner, M. (1984) *Proc. Nat. Acad. Sci. U.S.A.* 81, 983-987.

21 Kostriken, R., Strathern, J.N., Klar, A.J.S., Hicks, J.B. and Heffron, F. (1983) *Cell* 35, 167-174.

22 Gold, M. and Becker, A. (1983) *J. Biol. Chem.* 258, 14619-14625.

23 Dreyer, G.B. and Dervan, P.B. (1985) *Proc. Nat. Acad. Sci. U.S.A.* 82, 968-972.

24 Chu, B.C.F. and Orgel, L.E. (1985) *Proc. Nat. Acad. Sci. U.S.A.* 82, 963-967.

25 Helene, C. (1985) *C. R. Acad. Sci. Paris* (in press).

26 Smith, H.O. and Birnstiel, M.L. (1976) *Nucl. Acids. Res.* 3, 2387-2389.

27 Saint, R.B. and Egan, J.B. (1979) *Mol. Gen. Genet.* 171, 103-106.

28 Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980) *Amer. J. Hum. Genet.* 32, 314-331.

- 29 Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakagucki, A.Y., Young, A.B., Shoulson, I., Bonilla, E. and Martin, J.B. (1983) Nature 306, 234-238.
- 30 Schwab, M., Ramsay, G., Alitalo, K., Varmus, H.E., Bishop, J.M., Matinsson, T., Levan, G. and Levan, A. (1985) Nature 215, 345-347.
- 31 Tsujimoto, Y., Jaffe, E., Cossman, J., Gorham, J., Nowell, P.C. and Croce, C.M. (1985) Nature 315, 340-343.

APPENDIX: PREPARATION OF SAMPLES

Procedures for yeast samples have been published elsewhere (4). The following are new procedures that have been developed working with E. coli.

Cell Preparation

E. coli was grown in M9 minimal medium plus supplements to  $1.5 \times 10^8$  cells/ml with vigorous aeration at 37°C. The chromosomes were aligned by adding 180 µg/ml chloramphenicol and incubating an additional hour. Cells were chilled by swirling on ice and concentrated by centrifugation at 8000 rpm for 15 min at 4°C. The cell pellets were resuspended in 10 ml of PIV (10 mM Tris-Cl, pH 7.6, 1 M NaCl) and repelleted by centrifugation. The cell pellet was thoroughly resuspended in PIV at about  $1.8 \times 10^9$  cells/ml. Since the inserts were about 100 µl this cell concentration produces inserts that contain about 1 µg of DNA assuming each cell contains about  $1 \times 10^{-14}$  g of DNA.

Insert Preparation

Freshly grown cells were warmed to 30° - 40°C, mixed with an equal volume of 1% low gelling-temperature agarose, and distributed into an insert mold covered on one side with tape. The insert mold was cooled on ice 5 to 10 min to allow the agarose to solidify. Inserts were pushed out of the insert mold with a bent glass-rod that had been alcohol flame sterilized into EC lysis solution (6 mM Tris-Cl, pH 7.6, 1 M NaCl, 100 mM EDTA (pH 7.5), 0.5% Brij-58, 0.2% deoxycholate, 0.5% Sarkosyl, 20 µg/ml DNase-free RNase and 1 mg/ml lysozyme). The samples were incubated overnight at 37°C with gentle shaking. The solution was changed to ESP (0.5 M EDTA pH 9-9.5, 1% sodium lauroyl sarcosine and 2 mg/ml proteinase K) and the inserts were incubated an additional 2 days at 50°C with gentle shaking. Organisms without cell walls, such as mammalian and parasite cells, may be lysed by

directly putting cells in agarose inserts into ESP. Insert preparations were stored in ESP at 4°C.

#### Restriction enzyme digestion

Ten inserts were washed on a rotator at room temperature for 2 hr, twice with 10 ml of TE buffer (10 mM Tris-Cl, pH 7.5, 0.1 mM EDTA) containing 1 mM phenylmethyl sulfonyl fluoride (PMSF) and three times with 10 ml of TE buffer not containing PMSF. Restriction enzyme digestions were carried out in an Eppendorf tube in a final volume usually of 250  $\mu$ l in the recommended assay buffer containing fresh sulfhydryl reagent, and an excess of 100  $\mu$ g/ml bovine serum albumin. The enzyme to DNA ratio was usually 20:1 (units of enzyme:  $\mu$ g of DNA). Reactions were usually allowed to proceed overnight at 37°C with gentle shaking. The next day the assay buffer was aspirated off and replaced with 1 ml of ES solution (ESP minus proteinase K) and the inserts were incubated at 50°C with gentle shaking. After 2 hr the ES solution was replaced with 250  $\mu$ l of ESP and the samples were incubated an additional 2 hr before loading onto a PFG gel. About 0.25  $\mu$ g/well was loaded.

Cassandra L. Smith and Charles R. Cantor, Department of Genetics and Development, College of Physicians and Surgeons, Columbia University, New York, NY 10032

Macroscopic methods of examining the map of the human genome include cytogenetics, somatic cell genetics, and linkage analysis. Each of these methods is powerful, and all are quite complementary. In practice, however, each becomes progressively more tedious when structural data is required at finer resolution than 10,000 KB. It seems unlikely that any of these methods will be extended, in the near future, to provide routine analysis of the human genome at 1000 KB resolution.

Molecular methods of examining the map of the human genome involve cloning, restriction mapping and sequencing DNA fragments. In practice the analysis of a 10 to 50 KB region by restriction mapping is nearly trivial. The extension of such maps to several hundred KB is possible by chromosome walking techniques. However these become quite tedious as the size of the region increases and as segments of DNA become dominated by highly repeated sequences. While walks of 1000 KB are practical, it would be hard to approach such a task with great enthusiasm unless the particular DNA region were of compelling interest.

To bridge the molecular and macroscopic techniques, what is needed is a way to construct a physical map with 100 kB to 1000 kB resolution. In this article, we will describe four techniques that, together, allow the construction of such a coarse restriction map. Each technique was developed or tested on simple organisms like yeast or bacteria but each has now been shown to be feasible for comparable studies on human samples. Then we will provide an outline a strategy that appears to be quite an efficient way to apply these techniques to construct complete physical maps of each human chromosome. Finally we will offer a brief discussion of the utility of such maps.

PHYSICAL MAPPING METHODS

Isolation of unbroken genomic DNA. High molecular weight DNA samples are prepared by suspending live cells in liquid low gelling agarose (Scwartz and Cantor, 1984). After solidification, extensive detergent, protease, and salt treatments are used to remove all cellular constituents except the DNA. This is possible because the pores of the agarose are large enough to allow rapid diffusion of proteins and other small macromolecules while genomic DNA is retained, quantitatively. We have found it convenient to use roughly cubic agarose samples, about 100 ul in volume, which we call inserts. However it will probably be equally effective to use cells suspended in agarose microbeads (Cook, 1984).

For bacterial samples, a typical insert will be made from 10 to 100 million cells resulting in 0.1 to 1 ug of DNA (Smith and Cantor, 1986a). For mammalian samples, we typically use 1 million cells per insert, resulting in 10 ug of DNA (Smith et al, 1986a). Detailed protocols for preparing such samples are described elsewhere (Smith et al., 1986b). DNA samples made in inserts show a negligible level of double strand breaks. They are stable indefinitely at 4 C and can be kept at room temperature for more than a month with no detectable damage. Thus, in practice, agarose inserts are transported routinely by ordinary mail.

Separation of large DNA molecules. The technique of pulsed field gel (PFG) electrophoresis allows high resolution size fractionation of DNA (Schwartz et al., 1983). In this technique, molecules are periodically forced to change their direction by a change in the applied electrical field direction. A number of variations in experimental geometry have been described in detail elsewhere (Schartz and Cantor, 1984, Carle and Olson, 1984, Carle et al., 1986). The general type of apparatus that we find optimal for physical mapping studies is shown in Figure 1 (Smith and Cantor, 1986b). This contains a 20 cm square 1% agarose running gel placed at 45 degrees between two inhomogeneous electrical fields. Two convenient sizes for the apparatus are 28 cm square, which is the smallest that will accommodate a 20 cm gel at 45 degrees, and 55 cm square. In practice, with typical applied electrical fields of 10 V/cm this apparatus fractionates DNA molecules with 5% or better size resolution through an order

15  
of magnitude in size. The frequency at which the electrical fields are switched will determine the size range of the fractionation which can vary from 5 kB for 0.1 second pulses to 5000 kB for 1 hour pulses (Smith, Hui and Cantor, unpublished results).

For PFG electrophoresis, the agarose sample block is inserted directly into a slot cut in the running gel and then the alternate fields are applied. Typical running times are 40 hours in the 28 cm apparatus and 70 hours in the larger apparatus. In general, resolution improves as the DNA concentration is lowered. We have found that use of halves or thirds of inserts at the sample concentrations described above represent a convenient trade off between resolution and sensitivity of resulting Southern blots for bacterial and mammalian DNA samples. In practice ordinary Southern blotting and hybridization techniques suffice for the analysis of PFG electrophoretic bands. However it is necessary to fragment the DNA prior to the Southern transfer, and our experience is that nicking with UV light has been consistently dependable while acid nicking has been unreliable (Smith et al., 1986a). In our hands the sensitivity of hybridization detection of DNA blotted from PFG gels has never been as high as ordinary Southern blotting, but the reason for this is unknown.

Accurate size standards are important in assessing the performance of PFG electrophoresis and vital in using the techniques for macrorestriction mapping as described below. We



15/

have found it most convenient to use tandemly annealed oligomers of wild type and deletion strain lambda DNA as finely spaced size markers (Smith et al., 1986b). The apparent sizes of concatomers of two different monomer lengths are consistent which supports the use of such samples as true size standards. In practice concatomers up to the 34-mer are routinely prepared.

Specific fragmentation of genomic DNA into large pieces. PFG electrophoresis has shown excellent ability to separate small natural linear chromosomal DNAs ranging in size from 50 kB parasite microchromosomes to multi million bp yeast chromosomes (Van der Ploeg et al., 1984, Schwartz and Cantor, 1984, Carle and Olson, 1984, Kemp et al., 1985). However, intact human chromosomes range in size from 50 MB to 250 MB, too large for direct PFG separations. Bacterial chromosomes fall into the size range currently accessible by PFG but their circular topology prevents their entry into gels. Thus, in practice, for the analysis of both bacterial and mammalian samples it is necessary to cut genomic DNA into discrete large fragments. This cannot be done by normal solution techniques because of the susceptibility of large DNA to shear breakage. Instead, restriction nucleases are diffused into DNA samples inside agarose, and digestion is allowed to occur in situ.

Enzymes are chosen that recognize very infrequent sequences. Not I, which cleaves at GCGGCCGC, Sfi I which cleaves at GGCCNNNNGGCC, and Mlu I, which cleaves at ACGCGT, are particularly effective with human DNA. They yield fragments

15-

averaging in the range of 250 KB to 1000 KB. Methylase-nuclease combinations can yield even larger fragments (McClelland et al., 1984, McClelland et al., 1985, Smith, Econome and Cantor, unpublished results).

Protocols that provide total digestion of DNA in agarose were first developed with bacterial samples (Smith et al., 1986a). An example of the appearance of such a digest is given in Figure 2. Twenty to thirty bands are seen ranging in size from less than 50 KB to more than 400 KB. The progressive increase in staining of bands as a function of molecular weight ( except for obvious multiples ) and the consistency of the sum of band sizes with known genome sizes indicate that the digests are complete. This can be verified by Southern blotting and hybridization with known single copy probes resulting in a single band.

The critical variable in preparing specific large DNA fragments was found to be the particular agarose batch. Once a suitable batch has been identified, complete digests of DNA can be obtained with 10 to 20 units of most enzymes per ug of DNA. A example of a typical digest of human DNA, analyzed by PFG is shown in Figure 3. Depending on the enzyme, the average fragment size varies from a few hundred KB to more than 1000 KB. In most samples, only a continuous smear of DNA is seen. This is reasonable since, from the haploid genome size of 3 billion base pairs, the digests should consist of 3000 to 10,000 discrete fragments. However some enzyme digests show discrete fragments in the highest molecular weight range examined. Elution of such

15

material from the gels could provide relatively pure samples of 500 KB to 1 MB regions of the genome. It remains to be determined whether any of these regions are particularly interesting ones.

Methylation is potentially a very serious complication for the production of unique large fragments of mammalian DNA. Many of the potentially most useful restriction enzymes with rare sites contain the sequence CpG in their recognition site. Indeed it is the relative rarity of this sequence, 20% the expected value (Bird and Taggart, 1980), that makes the sites of these enzymes rare. It is estimated that over 50% of the CpG's in mammalian genomes are methylated (Gruenbaum et al., 1981, Kunnath et al., 1982), and many enzymes will not cut at the methylated sequence. If the pattern of methylation at each site is all or none, the result will be an apparent reduction in the number of cutting sites but the digest will still appear to be complete. However an intermediate methylation pattern will result in incomplete digests which can seriously compromise some approaches for assembling the order of large fragments. Ironically, however, controlled incomplete digests are just what is desired for other macrorestriction mapping strategies.

In practice, the problem of methylation is unlikely to be as serious as it appears at first glance. The recognition sequences for the most useful restriction enzymes with rare sites appear to occur preferentially if not exclusively, in HTF islands. These regions of the genome are predominantly single copy DNA except

for the ribosomal RNA genes (Bird et al., 1985). In HTF regions the CpG sequences are generally not methylated. Thus the appropriate choice of restriction nucleases may largely eliminate both the problems of methylation as well as the problems of repeated DNA in the specialized junction libraries described below.

Macrorestriction mapping. Blotting a PFG electrophorogram of digested human DNA and hybridization with a single copy probe identifies a large DNA fragment that contains the DNA neighborhood of that probe. Where a genetic map already exists, or large numbers of probes are available, direct Southern blotting of separated large fragments can reveal their order just as in conventional methods for mapping smaller DNA regions. This approach has been used on E. coli to provide a nearly complete physical map in less than a year of effort (Smith and Cantor, unpublished results). The same approach has provided a map of much of the human major histocompatibility complex (Weissman, et al., 1987).

Ambiguities can arise when two fragments have the same size. However this is less serious than in ordinary restriction mapping by Southern blotting because the total number of fragments is much less. If the density of available probes is great enough, an unambiguous map can be assembled by overlapping two different enzyme digests. But often this is not the case, and then it becomes difficult to prove whether two particular fragments are actually adjacent.

In general, a more powerful and efficient approach is to screen or select probes that contain just the ends of large fragments. We call these probes junction probes (Smith et al., 1986a). There are two types of such probes: linking probes span rare cutting sites and thus contain just the ends of two contiguous large fragments; jumping probes contain just the ends of a single large fragment. These are shown schematically in Figure 4. Linking probes can be prepared by selecting just those clones from a complete small insert library that contain a particular rare restriction enzyme site. Probing a genomic digest of DNA generated by the same enzyme will reveal two large fragments (Smith et al., 1986b). These must be adjacent and thus a physical map of the distances between rare cutting sites can be generated systematically by sampling all of the linking clones in the library.

Ambiguities will arise in the analysis of linking clones when two similar sized pairs of fragments exist. For example if two different 250 KB fragments occur, each adjacent to an 840 KB fragment, it will not be possible to tell which is next to which. This problem can be eliminated, and the general utility of the linking library approach can be enhanced by the use of library of jumping probes. These are made by circularizing large fragments around selectable markers and then removing almost all of the genomic DNA before recyclizing (Lehrach and Poustka, 1986, Collins and Weissman, 1984). Each resulting small insert jumping probe should identify a single large fragment in a

macrorestriction digest and two unique linking probes in a linking library. Similarly, each linking probe should identify two unique jumping probes. In principle it should be possible to walk between linking and jumping probes and establish their order without any direct analysis of large DNA fragments. If the jumping library is made from size fractionated DNA, the distance covered by each step will be known.

Together, the two types of junction probes and the PFG electrophoretic analyses, contain enough information to place the large fragments in order, and determine their size. They provide sufficient redundancy to allow efficient discrimination against experimental errors. In practice it will be desirable to analyze at least two different enzyme digests in parallel. This will provide valuable overlap information as further protection against accumulated errors. It will also help walk through any areas where junction probes are too imbedded in highly repeated DNA to be useful. In addition, as described above, the optimal choice of restriction nuclease may yield junction libraries relatively free of repeated DNA. At present the strategies of constructing junction libraries have been tested and have been shown to work in limited cases. It will be desirable to optimize further the efficiency of creating such libraries before large scale applications to human genome mapping are initiated.

#### PROSPECTS FOR A HUMAN PHYSICAL MAP

It is possible to outline a simple scheme for physically

157

mapping the entire human genome. This scheme makes use only of existing technology outlined above and the known or presumed availability of human-rodent hybrid cell lines and small insert flow sorted single human chromosome libraries. If one could work with single human chromosomes, appropriate digests could fragment these into 50 to a few hundred pieces. Indeed the entire genome of *Drosophila melanogaster* is about the size of a single typical human chromosome and a good fraction of this genome can be visualized as discrete DNA pieces by PFG electrophoresis after digestion with a nuclease with rare sites (Smith and Cantor, unpublished results). Assembling these pieces into order would be no more difficult than assembling a macrorestriction map of a bacterial genome since the number of pieces is comparable even though the human pieces are five fold larger in size.

One cannot currently work directly with flow sorted human chromosomes because procedures for preparing these samples lead to too much DNA breakage. However one can reduce the problem of working with the human genome in practice to working with single chromosomes one at a time by using hybrid cell lines containing only a single human chromosome and libraries made from flow sorted DNA from the same chromosome. A typical human chromosome is 150 MB in size. This will be cut into 300 fragments by a single restriction nuclease with rare sites. Analysis of these will require 600 junction clones consisting of 300 linking clones and 300 jumping clones. Assuming that two nucleases are used, the number of clones needed doubles. Summaries of the numbers of

15  
clones needed for the largest and smallest human chromosomes are given in Table 1. These numbers are large but not unreasonable, especially in view of the many additional uses for the junction clones described below.

Available mapped markers and cytogenetic markers will serve as bench marks for the physical map. Wherever cell lines and flow sorted material representing only parts of a single chromosome are available, the task will be even simpler. Such samples will allow the physical map of segments of a chromosome to be completed first. Then these can be linked up to make entire chromosomal maps. Provided that the distribution of the human material in the hybrid cell lines is known unambiguously, it will always be easier to use the principle of divide and conquer.

It is tempting to consider focusing initial mapping efforts on selected interesting regions of the genome. The recent successes in identifying linked markers to Huntington's disease, polycystic kidney disease, cystic fibrosis, and Duchenne muscular dystrophy provide just a taste of the interesting regions likely to emerge over the next few years. To approach the mapping of such a region one must first identify the large DNA fragment on which the linked marker resides and then clone the ends of that fragment. This will not be particularly difficult. Assume one can start with a hybrid cell line containing only a single human chromosome, 150 MB in size, on which the marker resides. A digest of that sample with a rare restriction nuclease will yield 300 human DNA fragments averaging 500 kB in size.



Current PFG electrophoretic resolution can provide a gel slice in which the desired human DNA fragment, containing the linked marker, is contaminated on average, by only four other large human DNA fragments. Suppose one elutes DNA from the gel slice fragments it with some other restriction nuclease and clones it into a vector requiring pieces ending in the rare cutting site. Four of the twelve discrete human DNA-containing clones will represent an end of the fragment of interest. Cell hybrids containing partial deletions or translocations of the human chromosome of interest will serve to identify which of the six clones are desired. Then physical mapping can proceed as described above. In practice, however, it will probably be more efficient to map regions of individual chromosomes without regard to particularly interesting linked markers. These markers can be used to test the correctness of the emerging physical map. Then they will provide the basis for staging the construction of a finer map of the large DNA fragments estimated to be in the actual region of the particular disease gene of interest.

As outlined above, the task of making a complete physical map of each human chromosome is arduous but now practical. Particularly for the smallest or most densely mapped chromosomes, the time seems appropriate to actually proceed with the task. However several advances in technology seem likely to occur over the next few years that should accelerate the task still further. The size range of PFG electrophoresis has recently been extended up to 10MB (Smith and Cantor, unpublished results) and may be

160

extendable still further. Potential methods for cutting human DNA into pieces this large have been described in detail. As these begin to work in practice the possibility emerges of cutting the DNA of a chromosome into 10 to 50 pieces before fractionating these and then subcutting each into 5 to 20 smaller fragments. This would greatly facilitate physical mapping.

In all the strategies outlined above, the use of hybrid cell lines is a major limitation because one never actually has large human DNA fragments free from rodent contaminants. Problems with rodent-human crosshybridization will inevitably arise. In addition, the need to analyze all separations by blotting rather than direct DNA visualization is quite labor intensive. These problems would all be solved if it is possible either to (1) clone large DNA fragments directly, (2) purify of large fragment away from others of the same size by virtue of some feature of its sequence, or (3) separate intact human chromosomes or chromosomal DNA without significant double strand breakage. In view of the rate of progress over the past few years and the current level of interest in these problems it is hard to believe that one or more of these technical breakthroughs will not be achieved long before a physical map is complete

#### APPLICATIONS OF A HUMAN PHYSICAL MAP

The human physical map, constructed as described above will consist of a set of cloned DNA markers spaced at accurately known positions throughout the entire genome. The average resolution of

the map will be 500 KB. This is ten times the resolution of the human genetic linkage map currently being developed. It is also ten times the resolution of existing cytogenetic methods. This extra resolution should allow the visualization of many DNA rearrangements currently invisible cytogenetically. It will also dramatically speed the search for genes associated with inherited diseases. With even extremely tenuous evidence for genetic linkage for a particular disease one will be able to use the map to select the appropriate DNA probes to provide a clear test of possible inheritance. If this is confirmed, then the map will serve to accelerate the search for the gene involved.

The power of the physical map is best seen by its potential for analyzing variation in DNA structure in the human population. This is illustrated in Table 2. Each linking probe allows one to examine two adjacent large fragments. Suppose that PFG electrophoresis is used to search for restriction fragment length polymorphism. If 25 probes could be used per lane one could examine up to a third of the genome on a single gel at 20 KB resolution. This is two orders of magnitude better resolution than current cytogenetics. Furthermore the bands in a normal individual would always appear in the same place so no subjective image analysis will be required and the entire process is potentially automatable.

The physical map will also serve to calibrate the genetic map. This will reveal the relationship between average recombination frequency and chromosome position. It may provide

fundamental insights into the mechanisms of human meiosis. The physical map should be at high enough resolution to reveal whether the genetic linkage map is badly distorted by hot spots for DNA rearrangements. This may in turn allow the discovery of additional loci important in understanding human disease as well as human evolution.

The initial low resolution human physical map will also set the stage for the construction of a higher resolution map. All the techniques needed to proceed efficiently from a 500 KB resolution map to a 50 KB map are already developed. While it is premature to discuss higher resolution strategies in detail, one approach seems particularly powerful and is already bearing fruit in the analysis of bacterial genomes. This is the Smith-Birnstiel strategy which is shown schematically in Figure 5 (Smith and Birnstiel, 1976). Genomic DNA is digested to completion with an enzyme with very rare sites and then digested partially with a more frequently cutting enzyme. The digest is fractionated by length and then analyzed by hybridization with one half of a linking or jumping probe. The lengths of the resulting DNA bands reveal the positions of the more common restriction sites.

The power of the method is that many sites are mapped unambiguously on a single gel lane. The Smith-Birnstiel method requires accurate DNA sizing but this requirement is well met by the high resolution of PFG electrophoresis and the availability of reliable length markers. The method also requires that controlled partial digests be carried out in agarose. The ability

to do this is demonstrated by the example for E. coli DNA illustrated in Figure 6. The major disadvantage of the method is that partial digests inevitably require greater detection sensitivity. Current sensitivity in detecting single copy mammalian genes on macro-restriction fragments may have to be improved before Smith-Birnstiel mapping becomes a routine approach for human DNA mapping.

The higher resolution map is equivalent in information to an ordered set of cosmid clones spanning the entire genome. It can serve in practice to place an existing cosmid library in order. This in turn will set the stage for determining the DNA sequence of any or all regions of the genome.

#### ACKNOWLEDGEMENTS

This work was supported by grants from the NIH, GM 14825, the NCI, CA 39782, the Hereditary Disease Foundation, and LKB Produkter- AB. The assistance of Jason Econome and Peter Warburton was most valuable.

#### FIGURE LEGENDS

Figure 1. Photograph of typical pulsed field gel electrophoresis apparatus. The 20 cm square 1% agarose gel is placed at 45 degrees in the center of a 55 cm submarine gel box. For most applications the use of a 28 cm square box is equivalent.

Figure 2. PFG electrophoretic analysis of the E. coli genome after digestion with restriction nucleases. The lanes, from left to right are: S. cerevisiae, lambda vir, digests of E. coli with Not I, Eco RI, Hind III, Sfi I, Mlu I, Xho I, then lambda vir, and S. cerevisiae. The electrophoresis was performed in a 55 cm apparatus at 500 V with 25 second pulse times for 70 hours.

Figure 3. PFG electrophoretic analysis of human DNA digested with various restriction nucleases. The lanes, from left to right are: S. cerevisiae, lambda vir, digests of human lymphoblastoid cells with Not I, Sfi I, Sal I, Pvi I, Xho I, Mlu I, Apa I, then lambda vir, and S. cerevisiae. The electrophoresis was performed as described in Fig. 2, except that 120 second pulses were used.

Figure 4. Schematic illustration of two types of junction clones particularly useful in the rapid physical mapping of large DNA fragments.

Figure 5. Schematic illustration of the Smith-Birnstiel method of restriction mapping as applied to large DNA fragments.

Figure 6.- An example of Smith-Birnstiel mapping applied to macrorestriction fragments of E. coli. The outer lanes are S. cerevisiae, and lambda vir. The center lanes show a total Not I diges of E. coli, that was subsequently treated with progressively larger amounts (from right to left) of Sfi I. The separation pattern of this gel is unusual because a program of two

different pulse times (15 seconds for 36 hours, then 120 seconds for 36 hours) was used to generate high resolution separations at both small and large molecular weights simultaneously. Otherwise, electrophoresis conditions were as described for Figure 2.

References

Bird, A.P. and Taggart, M.H. 1980. Variable patterns of total DNA and rDNA methylation in animals, *Nucleic Acids Res.* 8: 1485.

Bird, A. Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. 1985. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* 40: 91.

Carle, G.F. and Olson, M.V. 1984. Separation of chromosomal DNA molecules from yeast by orthogonal-field-alternation gel electrophoresis, *Nucleic Acids Res.* 12: 5647.

Carle, G.F., Frank, M. and Olson, M.V. 1986. Electrophoretic separations of large DNA molecules by periodic inversion of the electric field, *Science* 232: 65.

Collins, F.S. and Weissman, S.M. 1984. Directional cloning of DNA fragments at a large distance from an initial probe: a circularization method, *Proc. Natl. Acad. Sci. USA* 81: 6912.

Cook, P.R. 1984. A general method for preparing intact nuclear DNA, *EMBO Journal* 3: 1837.

Gruenbaum, Y., Stein, R., Cedar, H., and Razin, A. 1981. Methylation of CpG sequences in eukaryotic DNA. FEBS Lett. 124: 67.

Kemp, D.J., Corcoran, L.M., Coppel, R.L., Stahl, H.D., Bianco, A.E., Brown, G.V., and Anders, R.F. 1985. Size variation in chromosomes from independent cultered isolates of Plasmodium falciparum, Nature 315: 347.

Kunnath, L., and Locker, J. 1982. Characterization of DNA methylation in the rat. Biochim. Biophys. Acta 699: 264.

Lehrach, H. and Poustka, A.M. 1986. Jumping libraries and junction fragment libraries: molecular tools for mammalian genetics, Trends in Genetics, in press.

McClelland, M., Kessler L.G. and Bittner, M. 1984. Site specific cleavage of DNA at 8- and 10-base-pair sequences, Proc. Natl.Acad. Sci. USA 81: 983.

McClelland, M., Nelson, M., and Cantor, C.R. 1985. Purification of Mbo II methylase (GAAGmA) from Moraxella bovis: site specific cleavage of DNA at nine and ten base pair sequences, Nucleic Acids Res. 13: 7171.

Schwartz, D.C. and Cantor, C.R. 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel



Schwartz, D.C., Saffran, W., Welsh, J., Haas, R., Goldenberg, M., and Cantor, C.R. 1983. New techniques for purifying large DNAs and studying their properties and packaging, Cold Spring Harbor Symp. Quant. Biol. 47: 189.

Smith, C.L. and Cantor, C.R. 1986a. Purification, specific fragmentation and separation of large DNA molecules, in Methods in Enzymology (ed. by R. Wu and L. Grossman), Academic Press, Orlando, in press.

Smith, C.L. and Cantor, C.R. 1986b. Pulsed field gel electrophoresis of large DNA molecules, Nature 319: 701.

Smith, C.L., Lawrence, S.K., Gillespie, G.A., Cantor, C.R., Weissman, S.M., and Collins, F.S. 1986a. Strategies for mapping and cloning macro-regions of mammalian genomes, in Methods in Enzymology ( ed. by M. Gottesman ) Academic Press, Orlando, in press

Smith, C.L., Warburton, P.W., Gaal, A, and Cantor, C.R. 1986b. Analysis of genome organization and rearrangements by pulsed field gradient gel electrophoresis, in Genetic Engineering (ed. by Setlow, J.K. and Hollaender, A.), Plenum, NY 8: in press

Smith, H.O. and Birnstiel, M.L. 1976. A simple method for DNA restriction site mapping, Nucleic Acids Res. 3: 2387.

Van der Ploeg, L.H.T., Schwartz, D.C., Cantor, C.R. and Borst, P.  
1984, Antigenic variation in Trypanosoma brucei analyzed by  
electrophoretic separation of chromosome-sized DNA molecules,  
Cell 37: 77.

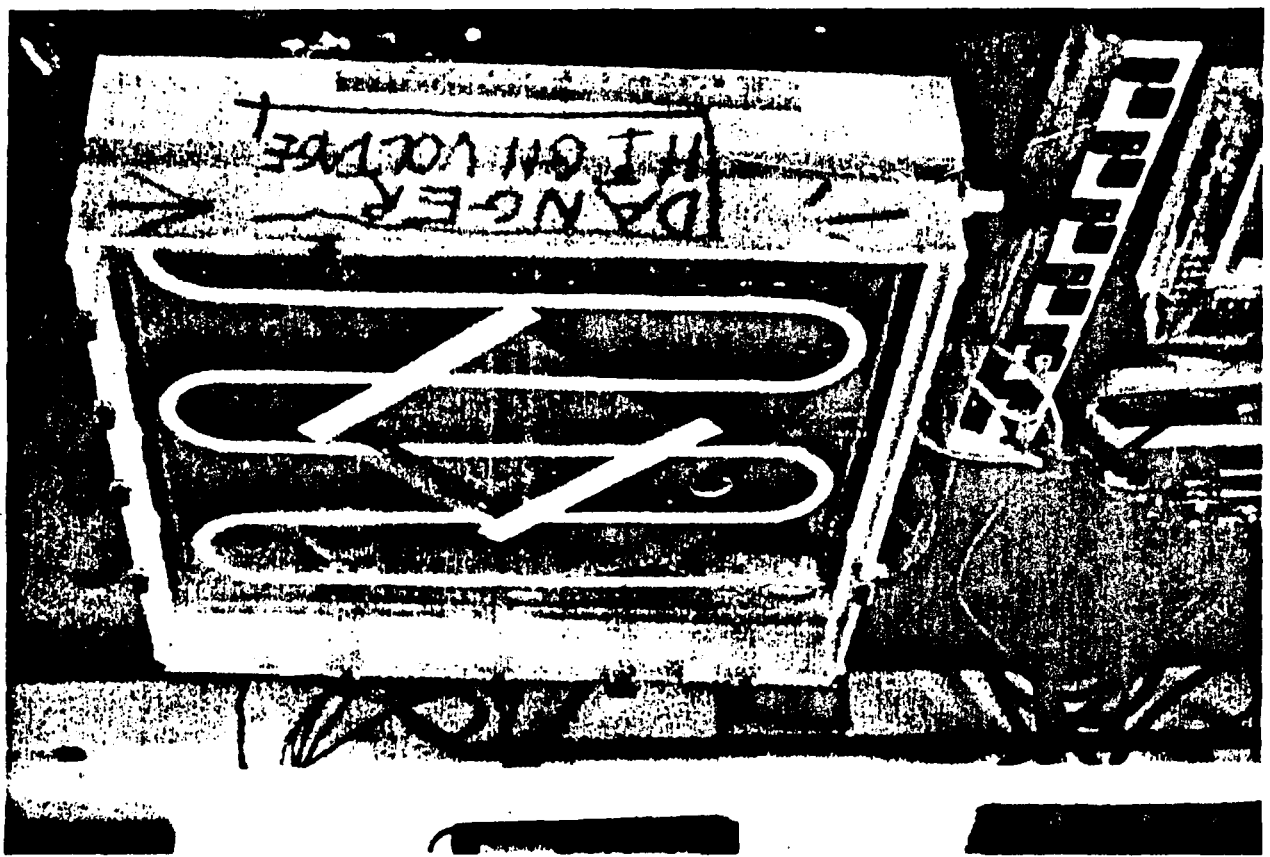
Weissman, S.M., Lawrance, S.K., Srivastava, R., Chorney, M.J.,  
Rigas, B., Vasavada, H., Gillespie, G.A., Smith, C.L., Cantor,  
C.R., and Collins, F.S. 1987. The Human MHC - Approaches to  
Characterization of Large Regions of DNA. This volume.

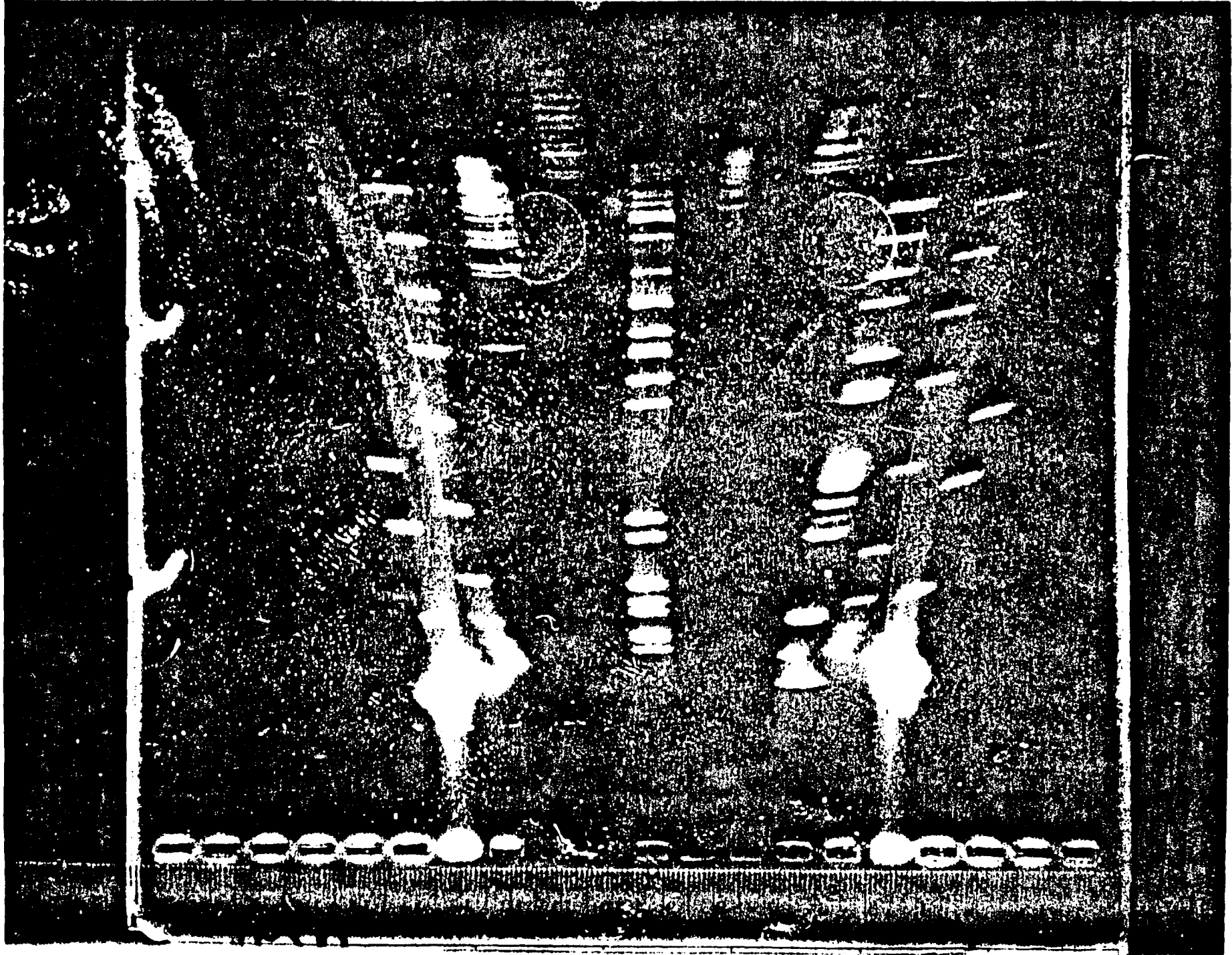
Table 1. Number of Clones Required to Construct a Chromosome Map with 500 KB resolution

Chromosome Size	One Enzyme	Two Enzymes
50 MB	100	200
150 MB	300	600
250 MB	500	1000

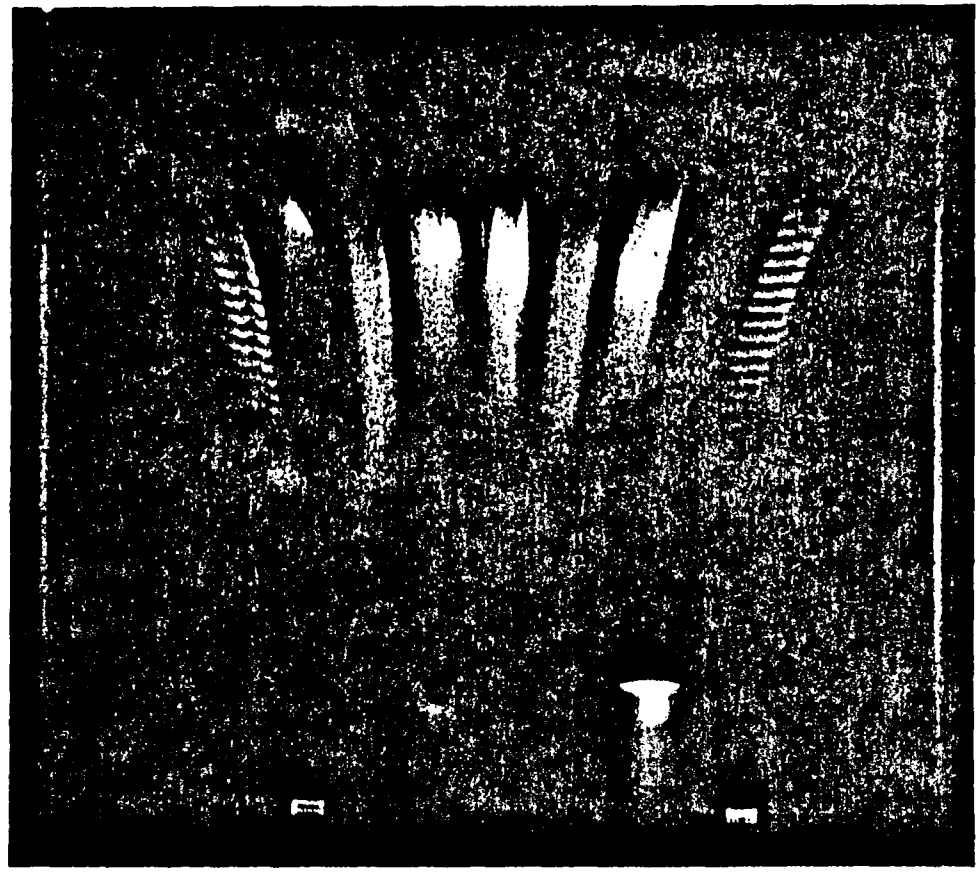
Table 2. Potential for PFG Analysis of Polymorphism or Genome Alteration Assuming 50 Bands Probed per Gel Lane

Average Fragment Size	Base Pairs Monitored per Lane	Resolution per Gel	Resolution
500 KB	25 MB	250 MB	5 KB
1000 KB	50 MB	500 MB	10 KB
2000 KB	100 MB	1000 MB	20 KB





172



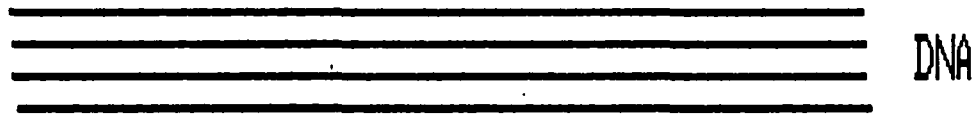
linking probes



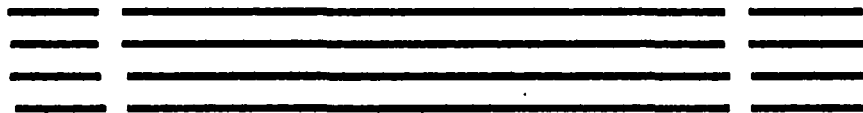
Large DNA  
fragments



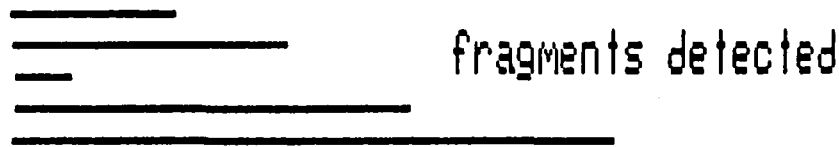
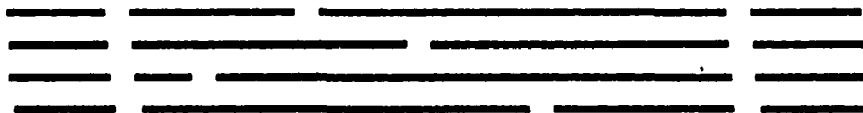
jumping probes



↓ total digest with enzyme 1



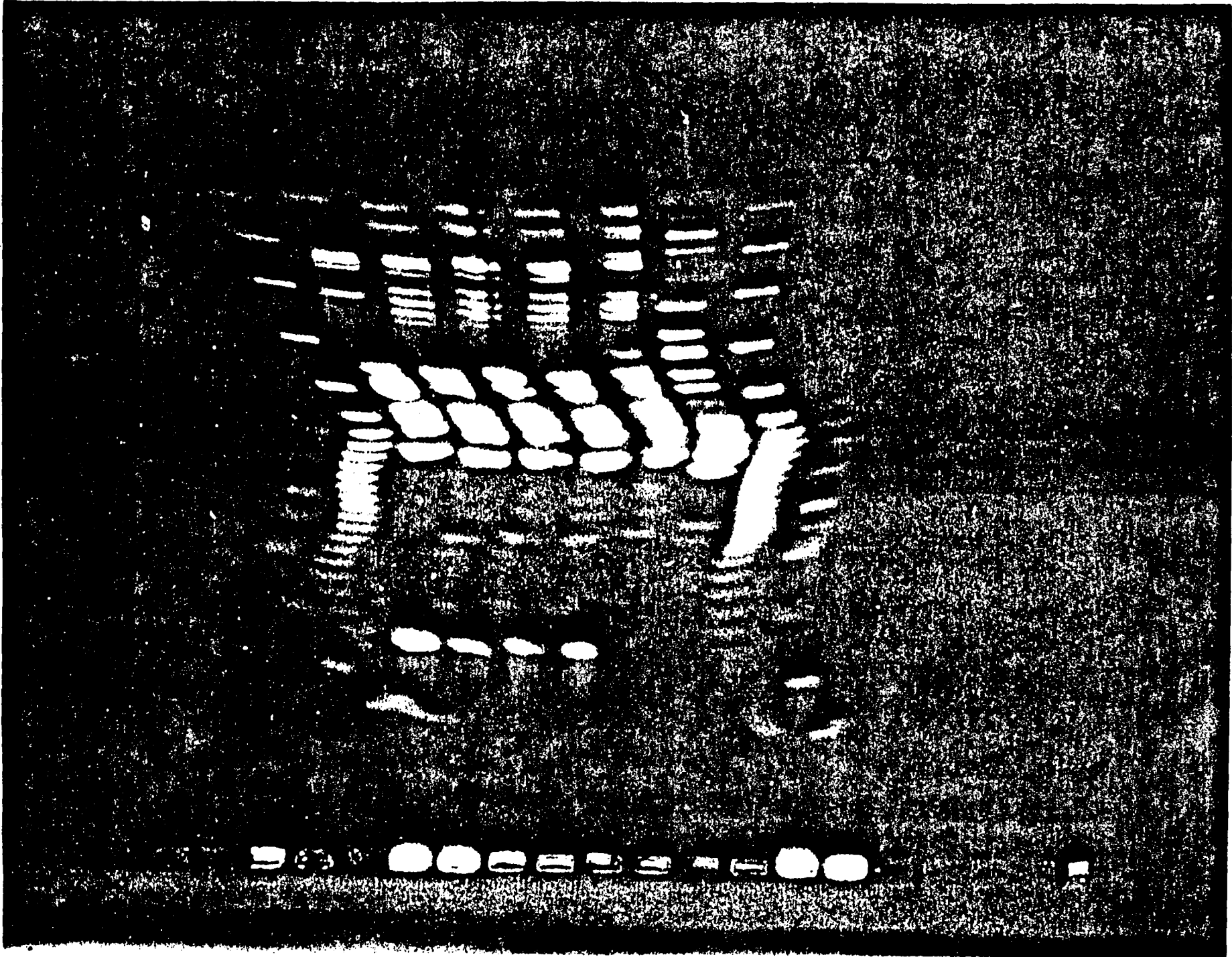
↓ partial digest with enzyme 2



■ probe

174





(M. Gellerman, ed.)

176

Mapping and Cloning Macro-Regions of Mammalian Genomes

Cassandra L. Smith<sup>§</sup>, Simon K. Lawrance<sup>\*</sup>, Gerald A. Gillespie<sup>\*</sup>,  
Charles R. Cantor,<sup>§</sup> Sherman M. Weissman<sup>\*</sup> and Francis S. Collins<sup>†</sup>,

Department of Human Genetics and Development  
Columbia University  
701 W. 68th Street  
New York, NY 10032

\* Department of Human Genetics  
Yale University School of Medicine  
333 Cedar Street  
New Haven, CT 06510

Departments of Internal Medicine and Human Genetics  
† University of Michigan Medical School  
4708 Med. Sci. II - Box 0618  
Ann Arbor, MI 48109

## Outline

### I. Introduction

### II. Pulsed Field Gel Electrophoresis and Blotting

1. Principles of PFG electrophoresis.
2. Practical considerations in PFG electrophoresis.
3. Preparation of mammalian DNA samples.
4. Restriction endonuclease digestions of mammalian DNA.
5. Length standards for PFG electrophoresis.
6. Blotting and detection of fragments from PFG electrophoresis.

### III. Chromosome Hopping: Generation of Jumping Libraries

1. Principles of the circularization method.
2. Use of markers and vectors.
3. Preparation of genomic DNA: Partial and complete digest jumping libraries.
4. Directionality.
5. Ligation concentration.
6. Avoiding noncircular ligations.

### IV. Cloning Fragments With Internal Rare Restriction Sites: Linking Libraries

1. Principles and construction of linking libraries.
2. Determining the organization and relationships of fragments detected with linking clones.

### V. Conclusion

### VI. Literature Cited

- A. List of Figures.
- B. List of Tables.

## 1. Introduction

A haploid set of human chromosomes contains about  $3 \times 10^9$  base pairs of DNA, corresponding to an average of about  $1.3 \times 10^8$  base pairs per chromosome. A number of techniques have aided the understanding of the organization of DNA sequences within chromosomes. Recombinational studies have suggested that the human genome corresponds to about 3,300 centiMorgans of genetic distance (1). The correspondence between base pairs and centiMorgans, however, is not linear. It may be influenced by sex, the position of the DNA relative to the centromere, and fine structure of the genetic material. Cytogenetic studies of chromosomes, such as in situ hybridization, can at best, resolve distances of the order of  $10^6$  base pairs (2). Mapping of gene location and distance by conventional pedigree studies has been limited by the availability of polymorphic markers. The application of studies of restriction fragment length polymorphisms (RFLPs) has been a major advance for these studies, but in particular families the number of closely linked polymorphisms may be limiting. Even where polymorphic markers are available, the resolution that can be obtained is limited by the number of individuals in informative pedigrees. Under most circumstances, it is difficult to obtain linkage information in a single family of a resolution much finer than one centiMorgan, a figure on the average equal to about  $10^6$  base pairs. Population studies are required to obtain finer resolution.

In contrast to the magnitude of the genome and the limitations of resolution of these approaches to structure, conventional cloning methods can only isolate DNA fragments of the order of 40 kilobases in a single step. In addition, the resolution of conventional gel electrophoretic methods is usually only adequate for fragments of about this size or smaller. Larger stretches of DNA have been

cloned by chromosome walking procedures that involve iterated steps of cloning and probe isolation, but this process can be time consuming, and is complicated by the prevalence of repetitive sequences in the human genome. Walking requires isolation of unique sequence probes from the end regions of a cosmid and using these probes to identify new, overlapping, cosmid clones. On the average, the new clone will contain only twenty kilobases of additional DNA and the maximum distance covered per probe will be somewhat less than forty kilobases. For higher eucaryotes, chromosome walking has been most successful when the region to be covered has contained multiple dispersed markers that can be probed for in parallel experiments.

Thus, as indicated in Figure 1, there is a size range between 100 and 2000kb which is too large to approach by standard molecular techniques and too small to resolve in cytogenetic and linkage analyses. There are, therefore, substantial applications for methods that can be applied in this size range to: (a) fractionate DNA fragments, (b) order such fragments, and (c) enable the rapid cloning of DNA located at distances away from an available marker. Success with such methods would expedite progress in identifying and cloning end points of DNA deletions and chromosome crossover points, and would provide valuable assistance in proceeding from loosely linked polymorphic DNA restriction fragments, such as those which have been shown to be associated with a number of inheritable diseases (e.g., Huntington's chorea, cystic fibrosis), to DNA that is more closely linked to the trait of interest.

Progress in the development of longer range mapping and cloning methods has been made with the development of pulsed field gel (PFG) electrophoresis (3-6) and "chromosome hopping" techniques (7,8). PFG electrophoresis is a method which

separates DNA fragments in a linear relation with molecular weight over a size range of several hundred thousand to over two million base pairs. This method has made it possible to fractionate and display chromosome sized DNAs of organisms such as yeast, trypanosomes, and plasmodium (4,9), as well as fragments, generated by infrequently cutting restriction enzymes, of E. coli, mouse and human chromosomes (8). Application of DNA hybridization procedures, in conjunction with PFG electrophoresis, has made it possible to detect fragments complementary to short unique copy probes, throughout this size range. The data generated in these studies has been used to construct physical maps of the E. coli genome (10) and the human major histocompatibility complex (S. K. Lawrance, C. L. Smith, S. M. Weissman, C. R. Cantor, unpublished).

Chromosome hopping techniques include several strategies for the isolation and mapping of DNA fragments separated in the genome by distances ranging from a few thousand to greater than a million base pairs. Chromosome hopping, as originally described by Collins and Weissman (7), is a technique that enables the isolation of DNA sequences within a defined range of distances in a specified direction from a given DNA probe. This is accomplished by circularizing large DNA fragments, generated either by random fragmentation or by complete digestion with infrequently cutting restriction enzymes, around a selectable marker and cloning the junction fragments so generated. These clones are referred to as jumping clones, or J-junctions. A similar approach has been described by Lehrach (11). Thus this technique may be employed to clone DNA throughout the resolution gap which can now be visualized in pulsed field gel electrophoresis and blotting experiments. For example, this procedure can be used to isolate, in a single step, several unique sequence DNA probes interspersed in a distance of 50-200kb

to each side of an initial probe. These jumping fragments can then be used, in batch or in parallel, to probe a cosmid library to isolate extensive DNA sequences covering a region of 400kb centered around the initial probe without resort to iterative cloning procedures.

A second type of library which is useful in the analysis of DNA fragments separated by distances within the resolution gap involves the selective isolation of genomic clones which contain internal sites for infrequently cutting restriction enzymes (8). These fragments are denoted linking clones or L-junctions. L-junction libraries can be employed to determine the orientations of the fragments detected by PFG electrophoresis as well as those isolated from jumping libraries. As shown in Figure 2, when jumping libraries are constructed from complete digests of genomic DNA with infrequently cutting restriction enzymes, the jumping and linking junction libraries are complementary and can be used in conjunction to walk rapidly along a chromosome.

The combination of PFG electrophoresis with chromosome hopping techniques, therefore, offers a considerable potential for generating restriction map data for significant portions of large genomes, including the human genome, and in addition, the cloning of gene family sized regions of DNA at pre-selected genomic locations.

## II. Pulsed Field Gel Electrophoresis and Blotting

Several detailed discussions of the principles of PFG electrophoresis have recently been presented (12). Here we will present a brief overview of the general methodology involved in this technique, and then we will focus more

specifically on aspects of electrophoresis and sample handling particularly relevant to mammalian DNA samples.

### 1. Principles of PFG electrophoresis

In conventional agarose electrophoresis of DNA molecules, the fractionation of different sized linear DNAs is based almost exclusively on the sieving properties of the gel matrix. DNA behaves like a free draining coil in electrophoresis. For such a coil, the friction is a linear function of size. The charge on DNA is also a linear function of size. Hence, the electrophoretic mobility, which depends on the ratio of charge to friction, should be independent of molecular weight. This is actually what is observed when DNA in gels is extrapolated to zero gel concentration. At finite gel concentration while all molecules continue to move at the same local velocity, a small DNA can travel through the gel in a relatively straight path because most of the pores are accessible. A larger molecule will have to take a longer path to find pores through which it can fit. Thus, the larger the molecule, the slower the net translation through the gel.

Sieving affords effective size separations of most types of molecules until the molecules become larger than the largest pores in the gel matrix. Then the molecules fail to migrate through the gel at all. However, DNA molecules behave like stiff coils. The shape of a DNA molecule is not constant. Under the influence of an electrical field, DNA molecules larger than the gel pores can distort their shape to enter the agarose gel. They can migrate through all the pores of the gel matrix by changing shape as the pores require. Once this process, called reptation, is initiated, all DNA molecules move through the gel



with the same net velocity because the gel is no longer acting as a sieve with selective pore sizes.

In pulsed field gel electrophoresis, the ability of a gel to retard larger molecules selectively is restored by requiring that the molecules periodically change their direction of motion (3,4). This is done by applying an electrical field in one direction for a fixed time period, the pulse time, then switching the field to a second direction, usually for the same pulse time, and then continuing to oscillate between the two directions. While no detailed mechanism for PFG electrophoresis has been proven, the following picture is consistent with the available facts.

When a field is applied, the DNA molecule must first distort its shape and orientation in the gel until it achieves a configuration that allows net translational motion. We will call this shape change DNA relaxation. The time required for relaxation is very sensitive to molecular weight; in fact over a wide size range it may scale roughly linearly in molecular weight. Thus, a larger DNA will require a larger fraction of the pulse time to relax, and have a smaller fraction of the pulse time for actual migration. This predicts that the resolution of PFG will actually improve as the molecular weight increases. Such increased resolution is observed in typical experiments until the DNA reaches a critical size where all larger molecules migrate with the same apparent velocity. What apparently happens is that these molecules have relaxation times longer than the pulse time. Thus, instead of continually relaxing in response to the changing fields, the molecules assume a time independent conformation in the gel in response to the sum of the two applied fields. Then, since no relaxation occurs, the molecules respond to the applied field as in ordinary

184  
electrophoresis. Since all have the same mobility in the reptation limit, no separation can be achieved. Thus, choice of the pulse time sets an effective upper limit to the separation range. While the choice of a pulse time depends on other experimental conditions, the following rule of thumb applies for the typical PFG procedures described below. One to two second pulses are optimal for separations of DNA less than 50kb; 10 second pulses are effective for DNAs in the size range of 50 to 200kb; 60 second pulses work well for DNAs from 200 to 800kb while 120 second pulses allow separations of DNAs as large as 1400kb.

## 2. Practical considerations in PFG electrophoresis

While a variety of apparatus designs for pulsed field gel electrophoresis have been effective, the device shown schematically in Figure 3 seems most generally useful for a wide variety of DNA sizes. This is a submarine horizontal gel box containing independent electrodes on all four sides. The gel is placed at 45 degrees relative to the square sides of the apparatus (5). Usually a 20cm square gel is employed while the actual box size can vary from 28cm to 55cm. The electrodes enter the gel vertically and are each connected to the power supply through a diode (4). Thus, when one set of electrodes is energized, the other is essentially invisible. In practice only 3 to 6 electrodes on a side are sufficient to provide fairly smooth electrical fields. The electrode geometry is absolutely critical to high resolution PFG separations. DNA reptation times are very sensitive to the angle between the applied electrical fields. While no systematic survey has yet been reported, preliminary studies indicate that larger angles are better, at least for molecules up to 1000kb. The particular geometry shown in Figure 3 provides angles between alternate fields that range from about 100 degrees to 150 degrees and appears to

have generally excellent performance for most DNA samples tested (Gaal, Warburton, Smith and Cantor, unpublished results).

In addition to the pulse time, and the angle, a number of other environmental variables are critical to high resolution PFG electrophoretic separations. Rather than go through these in detail, we present running parameters that appear optimal for mammalian DNA samples. We recommend an applied electrical field strength of 10 V/cm. If higher or lower fields are used, one will have to shorten or lengthen the pulse times proportionally. With fields of 10 V/cm the running time needed for excellent separations is roughly 1 hour per cm of box size. Temperature control is very critical for reasons that are completely unknown. The buffer in the submarine apparatus must be circulated to provide a uniform temperature. We have achieved excellent performance at 15 degrees but it is not clear if this is optimal for all samples. A 1% agarose gel concentration works reasonably well for most samples. Raising the gel concentration can improve resolution but increases the run time needed considerably. Lowering the gel concentration usually leads to markedly decreased resolution. No extensive studies on buffer variations have been described. We generally use 1 X TBE; however, others have reported that 0.5 x TBE is quite effective.

### 3. Preparation of mammalian DNA samples.

Intact mammalian chromosomal DNAs are generally 50 mb and larger except for organisms like chickens with microchromosomes. These molecules are much too large for current PFG technology. However, they can be broken into discrete fragments by treating unbroken chromosomal DNA with restriction nucleases that

have very rare recognition sites in mammalian genomes. The usual objective in PFG electrophoresis is DNA fragments in the 100kb to several thousand kb size range. Such large DNAs cannot be handled in aqueous solution because shear breakage caused by stirring or even thermal convection will be intolerable. To circumvent this problem we prepare DNA molecules directly in agarose as described below. All subsequent manipulations requiring the maintenance of high molecular weight material are carried out inside the gel.

Preparation of DNA in agarose is actually much easier than solution procedures. Freshly grown mammalian cells are spun at 1000 rpm in a conical centrifuge tube and resuspended in 5-10 ml of room temperature PBS. The cells are spun again and thoroughly resuspended in PBS at a final concentration of  $1-2 \times 10^7$  cells/ml. One to two ml of the cells suspension is mixed with an equal volume of 1% Low Gelling Temperature Agarose (SeaPlaque, FMC) cooled to 45-50°C and immediately distributed with a Pasteur pipette into a mold that makes 100µl blocks. The agarose is allowed to solidify by placing the mold on ice for 5-10 minutes. The agarose blocks are removed from the mold and incubated in ESP (0.5 M EDTA (pH 9-9.5) 1% Sodium lauroylsarcosine, and 1 mg/ml Proteinase K) for 2 days at 50°C with gentle shaking. The samples are stored at 4°C in ESP and can be shipped at room temperature with no detectable damage.

In most cases, we use a gel mold that makes 50-100 blocks that are 2 mm by 5 mm by 10 mm. We usually aim for 10µg of DNA per block. Thus, a million diploid mammalian cells are needed for each sample. A properly prepared DNA block sample will contain almost quantitative recovery of DNA. Inspection by ordinary gel electrophoresis or PFG should reveal little or no material entering the gel.

When first experimenting with new samples it is useful to make up inserts at several concentrations and to determine the best concentration empirically. We usually prepare agarose blocks at 2-4 times the DNA concentration that is used per run. Each block is cut into halves or quarters with a glass cover slip before loading onto a gel by simply pressing the piece of agarose into the well. The solid piece of agarose is sealed into place with liquid agarose. Thus, a single 10 $\mu$ g insert can be used for several PFG experiments such as runs of different pulse times.

#### 4. Restriction endonuclease digestions of mammalian DNA

The DNA samples in agarose in ESP can be loaded directly onto a gel if one is looking for microchromosomes. Further treatment of the samples is necessary if the DNA is to be cut by restriction endonucleases. Each insert is treated twice with 1 ml of 1 mM Phenyl methyl sulfonyl fluoroide (PMSF) in TE [10 mM Tris-Cl (pH 7.4) 0.1 mM EDTA] by slow rotation at room temperature for 2 hours. A fresh 0.1 M PMSF stock solution in isopropanol is made up each day. This is followed up by three 1 ml washes with TE buffer alone for 2 hours each. Restriction enzyme digestions are carried out in 1.5 ml microcentrifuge tubes usually in buffers recommended by enzyme manufacturers supplemented with 100  $\mu$ g/ml Bovine serine albumin in a final volume of 250  $\mu$ l. Inserts are usually added to the reaction buffer before enzyme is added. We routinely use 10 units of enzyme per  $\mu$ g of DNA. Reactions are usually allowed to proceed overnight at the appropriate temperature by gently shaking in a water bath. The next day the buffer is aspirated off carefully to avoid damage to the insert. One ml of ES (ES = ESP without Proteinase K) is added and the samples are incubated for 2 hours at 50 $^{\circ}$ C with gentle shaking. The buffer is changed to 250  $\mu$ l of ESP and

the incubation continued for another 2 hours before samples are loaded onto a gel.

Several features of the above protocol deserve further comment. Some restriction nucleases are very sensitive targets to protease activity used in the original DNA preparation. Therefore, this must be eliminated prior to adding restriction nucleases. Some restriction nucleases appear to be easily inactivated by EDTA. Thus, it is important to remove the EDTA used in storage of DNA-agarose plugs before attempting restriction enzyme digestions.

Small traces of restriction enzymes or other proteins left in the agarose plug may interfere seriously with PFG electrophoretic resolution. While the precise reason for this is unknown, a likely cause is that protein binding to the DNA will retard its mobility in PFG electrophoresis, just as is seen for ordinary electrophoresis. Thus, after any enzymatic treatment of the DNA, rigorous removal of any DNA binding proteins by proteolysis is recommended.

In general we find that the effectiveness of agarose in allowing restriction nuclease digestion is extremely batch dependent and the best procedure is to assay different batches until a reliable one is found. It is likely that special grades of agarose optimized for restriction nuclease cutting in situ will soon be available from FMC, Inc. In our hands, with proper agarose, roughly the same amounts of enzyme that suffice to produce a complete digestion of DNA in solution will yield the same result in agarose. Once a good batch has been identified, it is effective for a wide range of different enzymes. However, it may be necessary to alter the buffer and temperature conditions found optimal in solution if optimum performance is desired in the gel.

The choice of restriction enzyme will undoubtedly depend on the particular DNA sample and experiments. However, overall we have had consistently good success in generating large fragments of human and mouse DNA with the enzymes NotI, SfiI, MluI, SalI, and PvuI (see Table 1). These enzymes consistently yield average DNA fragment sizes for mammalian DNA that range from 100kb to greater than 1000kb (Figure 4). The relative size of fragments seen is quite consistent with relative frequencies observed for the appearance of these sites in the DNA sequence bank, GenBank. Thus, one can use this approach confidently to predict the behavior of enzymes. There are other restriction enzymes which have the potential to generate large fragments. In some cases we have tried these but have failed to detect bands upon hybridization with single copy mammalian probes (e.g. NruI). In other cases the commercial enzyme preparations have not been active enough or have been contaminated with nucleases (e.g., NarI, ScaI, NaeI, XmaIII, SacII). Still other potentially useful enzymes have not yet been tried, e.g., RsrII.

The average size of restriction fragments generated by digestion of DNA with a particular restriction enzyme cannot simply be calculated on the basis of the size of the recognition sequence. The base composition, nearest neighbor frequencies, and methylation pattern of the DNA sample as well as the methylation sensitivity of the restriction enzyme will influence the average fragment size obtained with a particular DNA sample. For instance, in human DNA the G+C content is 40% while the CpG frequency is only 0.8% (15). In principle, this allows one to use simple binomial statistics to attempt to calculate the expected restriction fragment sizes generated by enzymes with particular recognition sequences. Typical results are shown in Table 2. However, these calculations do not agree very well with the observed results for mammalian DNA (Table 1). There

are two reasons for the discrepancy. First although the sequence CpG is quite rare in mammalian DNA it occurs preferentially in HTF islands (16). While the exact preference is not known, the result is that restriction sites containing multiple CpGs within six or eight base pairs of continuous C+G are not nearly so rare as predicted by simple statistics. In contrast, CpGs embedded in A+T rich recognition sequences can be expected to be extraordinarily rare (Table 1).

The second problem with simple estimates of fragment sizes is DNA methylation. The major known methylation site in mammalian cells is the sequence CpG. It is estimated that over 50% of the CpGs are methylated (1). However, the CpGs in HTF islands are generally not methylated (16). Some restriction enzymes are inhibited by the presence of 5-methyl cytosine. Digestion of mammalian DNA samples with these enzymes will result in incomplete cutting. However, the extent of the effect of methylation will depend whether the particular restriction sites are clustered in HTF islands. There is yet one further complication. There are both tissue specific and cell specific differences in levels of methylation (17,18). Thus, it is difficult to predict a priori, the degree of incomplete digestion one may encounter with a particular sample. Table 2 summarizes what is known at present about the methylation sensitivity of particular restriction endonucleases. There is little available information about the occurrence of the sites of these enzymes in HTF islands. Thus, at present, not enough is known to predict the outcome of a particular restriction endonuclease digest on a particular tissue. One simple test one can perform is to determine whether a digest is complete by hybridizing such a digest with a single copy DNA probe. We have obtained single bands upon hybridization of single copy mammalian probes to PFG electrophoretic separations of digests with NotI, SfiI, MluI, SalI, PvuI, and XhoI. This indicates either that the sites of



these enzymes are not methylated, that a discrete subset of sites is stoichiometrically methylated, or that the enzymes are insensitive to methylation. While the three possibilities are indistinguishable at present the results indicate that the enzymes are clearly useful.

While the in vivo methylation of CpG described above leads to unfortunate complications, in vitro methylation can be used to improve the usefulness of certain restriction endonuclease to generate large fragments of DNA. For example, the inhibition of various restriction enzymes can be used to generate rare restriction enzyme cutting sites by increasing the apparent size of the site (19,20). Alternatively, the unique requirement of DpnI for methylated adenosines on both strands can be used to generate a large number of specific cutting sites by heterologous methylation of sequences that overlap the DpnI recognition sequence (21).

5. Length standards for PFG electrophoresis

Since mammalian DNA samples inevitably show a broad smear of DNA in most size ranges for PFG electrophoresis just as in ordinary electrophoresis, it is essential to have standards of defined DNA length both to assess the performance of the apparatus as well as to provide length markers. As primary standards, the only current available samples are tandemly annealed lambda DNA concatemers. These can be made by direct incubation of lambda DNA inside agarose plugs as described elsewhere (8). As secondary standards, yeast chromosomal DNAs like S. cerevisiae strains are recommended. These have discrete chromosomal DNAs that have been sized against lambda concatemers. Use of lambda DNA samples like those shown in Figure 4 indicate that the resolution of PFG electrophoresis can

be better than 5 kb for molecules as large as 1,400 kb and much better than this for smaller DNAs.

Several types of evidence suggest that lambda oligomers do provide accurate molecular weight markers applicable to other DNA species. When oligomer sets from two different sized lambda DNA monomers are compared, the resulting patterns shown the coincidences in mobility expected from simple arithmetic. When independent total NotI and SfiI digests of E. coli are fractionated by PFG electrophoresis and sized, relative to lambda standards, the resulting genome size calculated from the sum of all detected fragments is consistent to within a few percent.

#### 6. Blotting and Detection of Fragments from PFG electrophoresis

Large fragments of duplex DNA do not transfer efficiently from agarose gels to nitrocellulose or other media suitable for DNA immobilization and subsequent hybridization with cloned probes. We have experimented with a number of different protocols for DNA transfer. The object of these protocols is to break the large DNA into much smaller fragments which can be transferred more efficiently. In our experience, acid treatment has yielded unreliable results. It is possible that the size of the DNA fragments produced is extremely sensitive to the length of treatment. Introducing nicks into the DNA fragments by exposure of the DNA to short wavelength UV light in the presence of ethidium has been quite dependable. The DNA is stained by incubation in 1 µg/ml ethidium bromide for 10 minutes with gentle agitation on a platform shaker. We have then used 10 minute exposures to a very weak 245nm UV source during which time photographs of the gel are taken. It is, however, important to adjust the time of UV nicking to

fit the intensity of the particular light source available. The gels are protected from light during subsequent manipulations prior to and during Southern blotting (8). Denaturation is carried out for one hour in 0.5N NaOH, 0.5M NaCl and neutralization is carried out for one hour in 1.5M Tris-Cl pH 7.5 with gentle agitation. The gel is blotted to nitrocellulose (Schleicher and Schuell) or Zetapore (AMF-CUNO, Meriden, CT) membranes by ascending transfer overnight with 15 X SSC. The filter is then baked for two hours in vacuo at 80°C. Filters are stable for up to six months if stored in an air tight container.

Hybridization of DNA attached to filters is carried out for samples from PFG much in the same way as in ordinary agarose gel electrophoresis. An example is shown in Figure 5. In this case, an HLA-DR $\alpha$  probe was hybridized with two filters: one contained an EcoRI digest of human DNA electrophoresed and blotted in a conventional manner, and the other contained NotI, SalI, and MluI digests of human DNA electrophoresed in a pulsed field. The hybridization, washing and autoadiographic conditions were identical. The signal is markedly weaker in the PFG blot, however. This may be a result of suboptimal transfer of DNA from the PFG gels. Thus, it is important to take steps to insure sensitivity in autoadiographic detection. We have accomplished this by using probes nick translated to high specific activity, minimally  $10^7$  cpm per hybridization at  $10^8$  cpm/ $\mu$ g for a single copy mammalian probe. Probes labelled by mixed oligo priming may also be used (22). In general, we also find it advantageous to increase autoradiographic exposure times (e.g., one week in the case of the blot shown in Figure 5) and to wash the filters at lower stringencies (e.g., 3 X SSC at 50°C in Figure 5) than are commonly used in conventional Southern blotting experiments. To obtain accurate size measurements of hybridizing bands we have found it useful to probe the filters a second time with nick-translated lambda DNA. As shown in

Figure 5, by referring to the illuminated lambda ladder, the NotI, SalI, and MluI bands detected with the HLA-DR  $\alpha$  probe are readily assigned sizes of 920, 290 and 340 kb, respectively.

III. Chromosome Hopping: Generation of Jumping Libraries

1. Principles of the Circularization Method:

As shown schematically in Figure 6, if genomic DNA is broken into long linear fragments and each fragment is self-ligated to form a circle, the two ends of each linear fragment become covalently attached to one another. After digestion of the circularized fragments with a restriction endonuclease, a large number of restriction fragments are generated from sequences that were originally located internally in the large fragments. These restriction fragments are identical in sequence to restriction fragments that are obtained by digesting intact genomic DNA with the enzyme. However, digestion of the circular DNA produces an additional fragment, the jumping or J-fragment, that contains within it the sequence derived by ligation of the two ends of the original linear fragment. If the genomic DNA was initially broken into linear fragments of 100 kb, for example, then the J-fragments will each contain two segments of DNA that were originally separated in the genome by 100 kb of other sequences. Thus, a given J-fragment, isolated for example, by hybridization with a unique sequence probe, also contains a segment of DNA which was originally located 100 kb to one side or the other of the probe sequences position in the genome. Unique sequence J-fragments, derived in this manner, can then be used to establish the linkage relationships of restriction fragments detected in PFG blotting experiments. They can also be used to screen cosmid libraries so as to isolate larger segments

of DNA scattered over about 100kb in either direction from the original probe. Used iteratively, they can allow "hopping" along a chromosome to move from a linked marker to a disease gene, or across a translocation or deletion breakpoint.

## 2. Use of Markers and Vectors

A technical difficulty in screening a library prepared simply by complete digestion of large circular DNA molecules is that most of the clones in the library are derived from internal rather than J-fragments. Thus, the majority of clones isolated with a given probe will not serve the desired purpose. Additionally, the number of clones that must be screened for each J-fragment would be large and would increase with the size of the original fragments that were circularized. To overcome this problem, we have introduced a second step in which the linear DNAs are co-ligated with selective marker fragments to form circles in which the marker is incorporated between the two ends of the original linear fragment. The marker we have used to date is a synthetic tyrosine amber suppressor tRNA gene flanked by appropriate restriction cleavage sites (23). The suppressor tRNA gene is sufficiently small (219bp) that it is unlikely to undergo self circularization prior to ligating to other molecules. Following circularization, digestion and cloning, one can select biologically for clones that have incorporated the sup tRNA gene. This can be done by cloning either into phage vectors that require the suppression of nonsense mutants to grow, or with plasmids containing selectable antibiotic resistance markers which can only be expressed if nonsense mutants are suppressed. In this way, libraries can be prepared which contain only J-

fragments, reducing by one or more orders of magnitude the difficulty of screening for desired fragments.

Although we currently have less experience with their use, other selective markers could also be used. One possibility would be to enzymatically hemi-biotinylate a short DNA fragment with a dNTP derivative linked to biotin through a reversible bond such as S-S, ligate it into the junctions of the genomic circles, cut with a second enzyme, and then use the physical selection provided by biotin-avidin interactions to purify the J-fragments. Biologic selection by in vivo recombination with a plasmid containing a sup tRNA gene (24) could then still be used to isolate the desired junction fragments. Another hypothetical possibility would be to use the lac operator sequences as the marker and lac repressor as the means of physical selection. These physical methods have the advantage that one only needs to clone and package the useful fragments and would consequently represent a considerable savings in materials. A disadvantage of this approach is that generally one wishes to obtain as extensive a library of junction fragments and any additional preparative step prior to cloning would decrease the yield.

J-fragments can, in principle, be cloned in either plasmid, phage or cosmid vectors. When the J-fragments are prepared from size selected partial digests of DNA the yield is sufficiently low, however, that the high efficiency of cloning with packaged DNA and the lack of bias for insert size over a fairly large range makes use of the phage system advantageous. The size of plaques produced with some of the established phage vectors that require suppression for growth but have incorporated a suppressor tRNA gene in insert tends to be small. This can be a technical limitation but the use of lambda Ch3A  $\Delta$  lac (red+, gam+, provided

by Dr. Fred Blattner) has provided a satisfactory solution. We find the most satisfactory sup<sup>-</sup> host to be MC1061. The high efficiency transfection methods that have been developed in recent years may make plasmid cloning systems that depend on suppression competitive for some applications with phage vectors.

3. Preparation of Genomic DNA: Partial and Complete Digest Jumping Libraries

a. J-Fragments from Randomly Generated Genomic Fragments

The initial large linear fragments of genomic DNA for circularization can be prepared in several ways. A method which we have used successfully is to perform a partial digest of high molecular weight DNA using a restriction enzyme that has a recognition sequence which occurs frequently in mammalian DNA and that generates sticky ends which are suitable for the circularization step. For example, Sau3A1 or Mbo1, which cleave at the four base sequence 5'-GATC-3' can be used in conjunction with the sup tRNA gene flanked by BamHI linkers. To avoid shear damage of the partially digested high molecular weight DNA and to expedite the selection of the desired size ranges for hopping, PFG electrophoresis is the method of choice. DNA is prepared in agarose as described above. Standard inserts are tested with increasing amounts of enzyme or lengths of digestion to determine the optimal conditions for generating DNA of the desired size range. If any significant DNA enters the gel prior to restriction digestion, it is preferable to remove this by a short "pre-electrophoresis" of the blocks prior to enzyme digestion. This will remove sheared-end molecules which would be damaging to the protocol. These conditions are then scaled up for a preparative run. For preparative purposes, we have used

2mm X 10mm X 7cm inserts containing approximately  $5 \times 10^7$  cells (200 $\mu$ g DNA). Using the lambda ladders as a guide, the desired size range of DNA molecules is cut out of the gel, and electroeluted into 0.5X Tris-borate-EDTA (TBE) buffer. The electroelution is carried out in a dialysis bag in the PFGE box. The gel fragment is then removed from the bag and the DNA is dialyzed twice at 4 $^{\circ}$ C against 10mM Tris pH7.5, 1mM EDTA. Great care must be taken in subsequent manipulations of the DNA to avoid exposing it to shearing forces. For example, centrifuging, vortexing, or pipetting must be avoided. The size of the DNA can be re-examined by running an analytical PFGE gel. Its concentration is best assessed by evaporating a fixed volume of the solution to near dryness and comparing it with known standards on a conventional ethidium stained agarose gel.

b. J-fragments generated by infrequently cutting restriction enzymes

An alternative approach to chromosome hopping is to circularize genomic DNA which has been digested to completion with a restriction enzyme, such as NotI, that cuts mammalian DNA very infrequently. A library of J-fragments produced in this fashion contains fragments corresponding to the two ends of the linear genomic fragments produced by complete digestion (see Figure 2). In this instance, the DNA for circularization can be directly electroeluted from the inserts without size selection, resulting in a range of molecular sizes representing the range of genomic NotI fragments. For this approach, we have prepared a suppressor tRNA gene and an antibiotic resistance gene flanked by NotI linkers. J-fragments prepared in this way contain the two ends of a single NotI fragment of genomic DNA. Because of the limited number of NotI restriction sites in the human genome (perhaps 3-4,000), only a few thousand clones are required



for a complete NotI hopping library. A limitation of such a library is that it is of use only if one begins that hop with a probe which contains DNA flanking a NotI site. An advantage of this approach is that it can be used in conjunction with a library of linking fragments. This procedure is discussed below.

#### 4. Directionality

It is possible to employ chromosome hopping so as to clone in a predetermined direction from an initial probe. This can be readily seen if one considers the example of circles formed from partial Sau3A1 digests, which are then digested to completion with EcoRI to give EcoRI bounded jumping fragments. In the simplest case an EcoRI fragment from genomic DNA will have only one internal Sau3A1 site, and therefore, will be divided into a left half and right half by this enzyme. If the probe used for screening the jumping library was derived from the right half of the EcoRI genomic fragment, then all isolated J-fragments will contain this right half linked to a part of a different genomic EcoRI fragment. Prior to circularization, this other fragment would have been located at the distal end of a large linear fragment whose proximal end would have begun at the Sau3A1 cutting site within the original EcoRI fragment and would have extended from the Sau3A1 site to the rightward EcoRI site. The distal end could therefore only have been derived from DNA rightward in the genome. As shown in Figure 7, provided that the probe is chosen so as to lie close to an EcoRI site this obligate direction of cloning will still be true regardless of the number of internal Sau3A1 sites.

## 5. Ligation Concentration Parameters

To avoid ligation of two long DNA fragments together while permitting circularization to occur, the simplest expedient is to perform the ligation at low DNA concentrations. The concentration of DNA at which circularization will occur is calculable from physico-chemical estimates of the concentration of one end of a linear DNA molecule in the vicinity of the other end as a function of the length of the molecule. Theory predicts that this concentration should decrease as the square root of the length of the molecule (25). To favor circularization, for example, at a ratio of one hundred to one over intermolecular ligation, and molar concentration of DNA molecules in the solution should be one hundred fold less than the concentration of one end of any DNA molecule in the neighborhood of the other end. Theoretical calculations and experimental measurements appear to agree (7) and suggest that at a DNA concentration of  $3.3/(kb)^{1/2}$   $\mu\text{g/ml}$ , where kb is the length of the molecule in kilobases, 95% of the ligations will be circularizations.

The amount of DNA that must be used to obtain a complete hopping library increases with the length of the initial fragments to be circularized. This is true because a larger fraction of the total DNA is present in non-junction fragments as the length of the circles increases. This consideration together with those discussed above indicate that the total volume of DNA solution in which the circularization is conducted must increase as the  $3/2$  power of the original DNA fragments, to produce libraries of various size hops, all of which contain similar numbers of J-fragments embedding each probe. For example, if one begins with 100 kb fragments and wishes to obtain  $4 \times 10^6$  clones with greater than 95% true J-fragments, then it is necessary to ligate 5 micrograms of size

selected DNA in a volume of 25ml. These conditions are summarized in Table 3. For these sorts of efficiencies to be present, packaging extracts must be of high quality, yielding  $4 \times 10^8$  pfu/ $\mu$ g or better with wild type lambda DNA.

To increase the size of the hop, for example, to 200kb, while maintaining the number of junction clones produced, it is necessary to increase the amount of DNA to 10 micrograms, and the ligation volume to 70 ml. The DNA ligase concentration should be kept constant, at 1 unit/ml. To monitor the efficiency of ligation reactions and subsequent steps, we have found it useful to set aside small quantities of each reaction and to run these on 1.4% agarose gels. By probing Southern blots of such gels with the suppressor tRNA gene, the success of the reactions can be assessed. A successful ligation is indicated by ladder formation of the suppressor tRNA genes, and by the appearance of suppressor tRNA in the high molecular weight region of the gel.

Construction of hopping libraries, as outlined above, requires a 200:1 to 500:1 molar excess of marker DNA, such as the suppressor tRNA gene, in order to effect efficient recovery of J-fragments. To prepare large quantities of the marker, it is desirable to obtain a plasmid with multiple tandem inserts of the tRNA gene, so that the molar yield per mass of plasmid DNA is increased. Plasmids containing 8 copies of this gene were obtained by ligating a large excess of suppressor tRNA gene monomer into plasmid. The plasmid has been passed several times without deletion of copies of the suppressor tRNA gene.

6. Avoiding non-circular ligations

A major potential hazard in the preparation of hopping libraries is that they may include a troublesome percentage of fragments derived from ligation of the ends of two separate long fragments. This can occur if the concentration of long fragments in the ligation is too high, if the ligation does not go to completion in the first stage and end fragments are ligated together during the process of ligating genomic restriction fragments with the vector DNA, or if the original long fragments were damaged at one end so that they could not be circularized and therefore the other ends could only ligate to different DNA molecules. The first possibility can be minimized by use of low concentrations. The second possibility can be minimized by use of ample amounts of ligase and reaction times. To some extent both possibilities two and three can be reduced if the circularized DNA is treated with alkaline phosphatase prior to digestion and cloning. Phosphatase treatment will inactivate unligated ends and prevent their aberrant joining during the circularization step. The most important precaution, however, is to begin with very high molecular weight DNA such as can be prepared as described above by cell lysis in gels, and to avoid steps that might break DNA molecules prior to ligation. As noted above, brief PFG electrophoresis of blocks containing undigested DNA in the PFGE box can be employed to remove degraded material prior to partial digestion with Sau3A1 or complete digestion with enzymes such as NotI that cut sufficiently infrequently.

Even with these precautions, however, there is always a non-zero probability that a given jumping clone will represent the connection of two genomic fragments which were not originally closely related in the genome. An independent means of assessing the relationships should be employed, especially if hopping is to be

done iteratively. If an appropriate somatic cell hybrid or chromosome sorted DNA is available, one can determine whether the new fragment is on the same chromosome as the starting fragment. An alternative method is to prepare the library itself from a somatic cell hybrid which contains a single human chromosome on a background of another species. Hopping can be performed along the human chromosome, and any non-circular ligations will be recognizable by the presence of non-human sequences. A further advantage of this method is that it permits the general validity of the library to be assessed by using human-specific repeats before an extensive screening is carried out. For example, in a 100 kb hopsized hamster-human chromosome 4 hopping library, by screening for clones containing human repeat sequences and then re-screening with hamster repeats, we were able to rapidly establish that greater than 85% of the clones arose from circularizations (F. S. Collins, unpublished).

#### IV. Cloning Fragments with Internal Rare Restriction Sites: Linking Libraries

##### 1. Principles and Construction of Linking Libraries

Linking libraries (8) consist of all those clones from the genome containing a particular rare enzymatic cutting site of interest (See Figure 2). We refer to these as linking fragments, or L-fragments. A variety of techniques can be employed to construct linking libraries. One such strategy is outlined schematically in Figure 8. As indicated, the suppressor tRNA with NotI linkers described above may be used to isolate plasmid, lambda, or cosmid clones containing genomic fragments with internal NotI sites. Linking libraries specific for a chromosome of interest can be isolated by ligating selective

204  
markers into DNA prepared from a library of DNA prepared by chromosome sorting (26).

An alternative scheme is to circularize EcoRI or partial Sau3A1 digests of chromosome-specific DNA around a suppressor gene. Subsequent digestion with NotI will only linearize those circles with a NotI site, which can then be selected by ligation into an amber-mutated phage with a NotI cloning site, and plating on a sup<sup>-</sup> host.

2. Determining the organization and relationship of fragments detected with linking clones

When linking clones are used to probe a PFG electrophoretic separation of DNA fragments generated with the enzyme, each clone should reveal two DNA fragments, and these must be adjacent in the genome. In principle, with just a single library and digest, one should be able to order the rare cutting sites, although it will not generally be possible to distinguish between two fragments of the same size. Some ambiguities will be resolvable by performing a second digest with another rare cutting enzyme. Other ambiguities should be resolvable by examining a partial digest with the original rare cutter. However, in general it will be more efficient to use several different libraries, each for a particular enzyme and overlap the resulting patterns just as in ordinary restriction fragment analysis.

The jumping and linking libraries can be used in a complementary fashion (Figure 2). Jumping libraries, if constructed by digestion to completion with a rare cutter (see above), consist of clones containing the ends of each large

fragment generated this particular enzyme. Each jumping clone will detect only a single fragment on a PFG electrophoretic separation of fragments generated with the same enzyme. However, each linking clone, since it spans a restriction site, should cross-hybridize with two different jumping clones, and each jumping clone should likewise cross-hybridize with two different linking clones. Thus, by cross-screening the two libraries, it should be possible in principle, to walk from clone to clone and generate a complete ordered map without use of DNA blotting techniques. No information will be provided directly about the physical distances between the clones, but that information is readily available from PFG electrophoresis. Thus, it is clear that the two junction libraries outlined above are complementary and in combination with DNA blotting provide the necessary redundancy to allow error detection.

A major potential difficulty with the schemes outlined above is repeated DNA. This difficulty can be minimized in several ways but ultimately some regions of mammalian genomes will still contain too much highly repeated DNA to be mapped easily.

Once ordered large DNA fragments are available, it should be possible to generate finer restriction maps for each by an analog of the Smith-Birnstiel procedure (27,28). One performs a total digest of genomic DNA with the same enzyme used in generating jumping or linking libraries. Then one performs a partial digest with an enzyme that cuts somewhat more frequently so that each large fragment will be nicked on the average of once or less to produce smaller fragments that average roughly half the original size. The pattern of fragments generated is viewed selectively by indirect end labeling through hybridization

with one half of a junction clone. This will detect only those fragments extending from the rare cutting site in that clone to a site of partial digestion with the more frequently cutting enzyme as shown schematically in Figure 9. Thus, the pattern of fragment sizes reveals the order of the more frequent cutting sites. This technique has already been used successfully to examine the structure of large fragments of the E. coli genome. There is no fundamental reason why it should not be effective for mammalian DNA although it does require an order of magnitude more sensitive detection than ordinary Southern blots.

It may even be possible to map a larger region by using a "double barrelled" Smith-Birnsteiel approach, as diagrammed in Figure 10. In this protocol, a partial digestion is done with a rare cutter such as NotI, and parallel lanes are separately probed with the two halves of a linking clone. The pattern observed should allow mapping of several NotI sites to either side of the linking fragment probe.

None of the methods described above actually provide pure DNA internal to the large fragments. That DNA can be obtained in an ordered manner by preparing a jumping library where the jump starts from a rare enzyme cutting site and proceeds to a more frequent site. The same half junction or half jumping fragments needed for PFG analysis as described above can also be used to screen jumping libraries for the desired sets of clones adjacent to a particular sites of interest.

The linking library approach would only require 50-300 clones to cover each human chromosome. The identification of such clones and their use in creating a



physical map of chromosomes could certainly set the stage for subsequent total genomic sequencing.

### Conclusion

Mapping of complex mammalian genomes, especially those of mouse and man, has progressed rapidly over the past few years. New DNA probes, many of them polymorphic, are appearing in large numbers, and can now be readily mapped to a particular part of a chromosome using in situ hybridization or somatic cell techniques. An enlarging panel of such probes will soon allow the mapping of genes for most Mendelian disorders, through the method of linkage analysis. While such techniques are extremely powerful, they are not well suited for molecular physical mapping in the size range from 100 kb to 2,000 kb, as shown in Figure 1. Standard molecular cloning methods are also limited in their application to this size range. The ability to work in this macro-region is essential if one wishes to use map location to clone genes whose normal protein product is unknown; it is also necessary for mapping relatively large gene complexes or for identifying the specific breakpoints of chromosomal deletions or translocations, many of which have major biologic consequences.

By combining pulse field gel electrophoresis with the generation of jumping and linking libraries, it is now possible to outline strategies that will allow the complete physical mapping of large segments of mammalian genomes. Placing pre-existing polymorphic DNA probes on this map can then be readily accomplished by PFG blots, which will disclose the relationship between recombinational and physical distances. These mapped probes, together with the linking libraries,

can be used in combination with random-ended jumping libraries (Figure 6) to generate other clones in interesting regions of the genome in a truly directed manner. Thus, although the technical aspects of many of these methods are not trivial, the necessary tools are in place for a real conquest to be made of the detailed physical structure of mammalian genomes.

### Acknowledgements

This work was supported in part by grants GM34960 to F.S.C., GM14825 and CA39782 to C.R.C., and from the National Cancer Institute to S.M.W. Support from the Hereditary Disease Foundation is also gratefully acknowledged. Also, the authors wish to express their appreciation to Ann M. Mulvey for preparing this manuscript.

### Literature Cited

1. Shows, T. B., A. Y. Sakaguchi, and F. L. Maylor: *Adv. Hum. Genetic* 12:341 (1980).
2. Robins, D., S. Ripley, A. Henderson and R. Axel.: *Cell* 23:29 (1981).
3. Schwartz, D. C., W. Saffran, J. Welsh, R. Hass, M. Goldenberg and C. R. Cantor: *Cold Spring Harbro Symposia on Quantitative Biology* 47:189 (1984).
4. Schwartz, D. C. and C. R. Cantor: *Cell* 37:67 (1984).

5. Carle, G. F. and M. V. Olson: Nucl. Acids Res. 12:5647 (1984).
6. Carle, G. F., M. Frank and M. V. Olson: Science 232:65 (1986).
7. Collins, F. S. and S. M. Weissman: Proc. Natl. Acad. Sci. USA 81:6912 (1984).
8. Smith, C. L., P. W. Warburton, A. Gall and C. R. Cantor: In: Genetic Engineering Volume 8 (J. K. Setlow and A. Hollanender, eds.) Plenum Press, New York, In press.
9. Van der Ploeg, L. H. T., D. C. Schwartz, C. R. Cantor and P. Borst: Cell 37:77 (1984).
10. Smith, C. L. and C. R. Cantor: Manuscript in Preparation.
11. Lehrach, H. and A. -M. Poustka: In: Trends in Genetics (in press).
12. Smith, C. L. and C. R. Cantor In: Methods in Enzymology (Ray Wu, ed.) Academic Press, (In press, 1986).
13. Kessler, C., T. S. Neumaier and W. Wolf: Gene 33:1 (1985).
14. McClelland, M., and M. Nelson: Nucl. Acid Res. 13 Supple:r201 (1985).
15. Bird, A. P. and M. H. Taggart: Nucl. Acid Res. 8:1485 (1980).

16. Bird, A., M. Taggart, M. Frommer, O. J. Miller and D. Macleod: *Cell* 40:91 (1985).
17. Ehrlich, M., A. Gama-Sosa, L. -H Huang, R. M. Midgett, K. C. Kuo, R. A. McCune and C. Gehrke: *Nucl. Acid Res.* 10:2709 (1982).
18. Razin, A., and A. D. Riggs: *Science* 210:604 (1980).
19. Nelson, C., C. Christ and I. Schildkraut: *Nucl. Acid Res.* 12:5165 (1984).
20. McClelland, M., L. G. Kessler and M. Bittner: *Proc. Natl. Acad. Sci.* 81:983 (1984).
21. McClelland, M., M. Nelson and C. R. Cantor: *Nucl. Acid Res.* 13:7171 (1985).
22. Feinberg, A. P. and B. Vogelstein: *Analytical Biochem.* 132:6 (1983).
23. Dunn, R. J., R. Belagaje, E. L. Brown and H. G. Khorana: *J. Biol. Chem.* 256:6109 (1981).
24. Seed, B.: *Nucl. Acids Res.* 11:2427 (1983).
25. Jacobson, H. and W. H. Stockmayer: *J. Chem. Phys.* 18:1600 (1950).
26. Available from National Laboratory Gene Library Project, Lawrence Livermore and Los Alamos National Laboratories.

- 27. Smith, H. O., and M. L. Birnstiel: Nucl. Acids Res. 3:2387 (1976).
- 28. Saint, R. B. and J. B. Egan: Mol. Gen. Genet 171:103 (1979).
- 29. Lawrance, S. K., H. K. Das, J. Pan, and S. M. Weissman: Nucl. Acids Res. 13:7515 (1985).

List of Figures

Figure 1. Schematic representation of DNA regions analyzable by various standard techniques. Note that none of these methods are easily adaptable to the 100 - 2,000 kb size range.

Figure 2. Schematic representation of the relationship between NotI linking and jumping clones.

The jumping fragments are formed by ligating the two ends of a fragment from a complete NotI digest of genomic DNA. The linking fragments are segments of chromosomal DNA containing an internal NotI site.

Figure 3. Set up for double inhomogeneous pulsed field gel electrophoresis

Electrode arrangement used for pulsed field gel electrophoresis in the double inhomogeneous field configuration. The gel is placed at a 45 degree angle. Sample wells are indicated by the dashes.

Figure 4. Pulsed field gel electrophoresis of large DNA fragments.

Samples were run in a 55cm apparatus for 72 hours at 500 volts using a two minute pulse time. The lanes from left to right are: yeast chromosomal DNAs (S. Cerevisiae, strain DBY728), concatemers of lambda vir (42.5kb monomer), human DNA digested with NotI, SfiI, SaI, PvuI, XhoI, MluI, ApaI, lambda vir concatemer, and yeast.

Figure 5. Southern blot of DNA electrophoresed in a pulsed field.

Autoradiogram showing the hybridization of the 3.1 kb EcoRI fragment of HLA-DR $\alpha$  (29) with Southern blots of human DNA digested with (1) MluI, (2) SaI, (3) NotI and electrophoresed in a pulsed field; and digested with (5) EcoRI and electrophoresed in a conventional manner. The sizes of the hybridizing fragments were determined by reference to the lambda ladder (lane 4) illuminated by a subsequent hybridization of the filter with nick translated lambda DNA.

Figure 6. Scheme for chromosome hopping.

Principle of the cloning procedure. The heavy bar represents the starting probe, which in the final jumping clone is present along with the marker gene and another segment of DNA (open box) that was initially many kilobases away in the genome. In this particular example, a suppressor tRNA gene (*supF*) is used as the marker. Other marker DNA segments allowing biological or physical separation of the jumping pieces may also be of use. Horizontal arrows show the orientation of the jumping fragment pieces relative to their original genomic arrangement.

Figure 7. Directionality of jumping.

Since a jumping clone will only be represented in the library if an MboI site in that particular EcoRI fragment has been cut, choosing a probe which is immediately adjacent to an EcoRI site will make it very likely that the jump occurs in that direction. A jump to the other side can only occur (left side of figure) if the MboI cut occurs within the probe sequence.



Figure 8. Scheme for construction of a linking library.

Linking fragments are isolated by preparing a total genomic library or a library from sorted individual chromosomes in a plasmid vector with suppressible mutations in the tetracycline or ampicillin genes. DNA is isolated from an amplification of this library. The preparation is digested with NotI (which does not cleave the vector) and the suppressor tRNA with NotI linkers is inserted. Upon retransformation, only plasmids which contain the suppressor tRNA and, therefore, also a linking fragment, convert the recipient bacteria to tetracycline or ampicillin resistance.

Figure 9. Internal mapping of large DNA fragments.

Schematic of the Smith-Birnsteiel method for rapid restriction mapping as modified for large DNA fragments by detection through indirect labeling.

Figure 10. External mapping of large DNA fragments.

Scheme for restriction mapping by partial digestion without end labelled DNA.  $N_0$  represents a NotI site embedded in a linking fragment. To map NotI sites proximal to  $N_0$ , a partial NotI digest of genomic DNA is prepared, run on PFGE, and blotted. Probes  $\beta$  and  $\alpha$  are respectively, the left and right halves of the NotI linking fragment. The bands detected with these probes are designated by NotI sites at their two ends, e.g., ( $N^{-1} N^{-2}$ ) is the band derived by cleavage at position  $N^{-1}$  and  $N^{-2}$  during partial digestion.

Table I Restriction Enzymes That Generate Macrorestriction Fragments

Enzyme	Recognition Sequence <sup>a</sup>	Avg. Frag. Size (Kb) based on Sequence <sup>b</sup>	PFG <sup>c</sup>	Freq. in Genbank (%) <sup>d</sup>
Not I	5' G C G G C C G C 3' ↑      ↑	9750	1000	1.2
Sfi I	G G C C N N N N G G C C o          ↑      o o	390	250	1.4
Mlu I	A C G C G T o↑	170	1000	1.0
Sal I	G T C G A C ↑   ↑   ↑	35	500	2.0
Pvu I	C G A T C G o  ↑  ↑	170	200	0.9
Xho I	C T C G A G ↑   ↑   ↑	35	200	4.0
Apa I	G G G C C C ↑   ↑	16	100	9.0
Rsr II	C G G (A) C C G ↑ (T)		nd	1.6

a - † denotes inhibition of cutting by methylation; o denotes no effect of methylation on cutting; no mark indicates lack of knowledge; ↓ denotes site of cleavage. Methylation sensitivity is summarized from references 13 and 14.

b - Calculated by binomial statistics as shown in Table 2.

c - Weight average size seen by PFG electrophoretic resolutions of DNA from mouse and human cells.

d - Mammalian sequences in Genbank were screened for the occurrence of each of the restriction enzyme recognition sequences. Shown is percent files, out of 1842 mammalian DNA sequence files, containing at least one occurrence of the particular sequence.

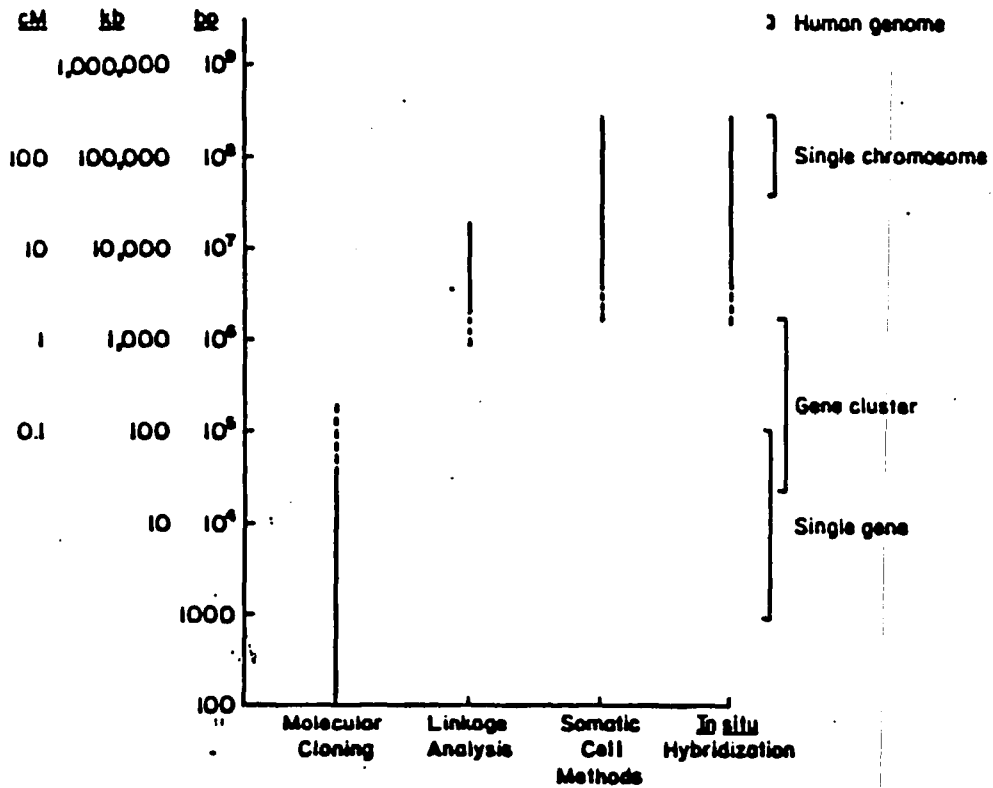
Table 2 Expected Sizes of Restriction Fragments From Mammalian Genomes

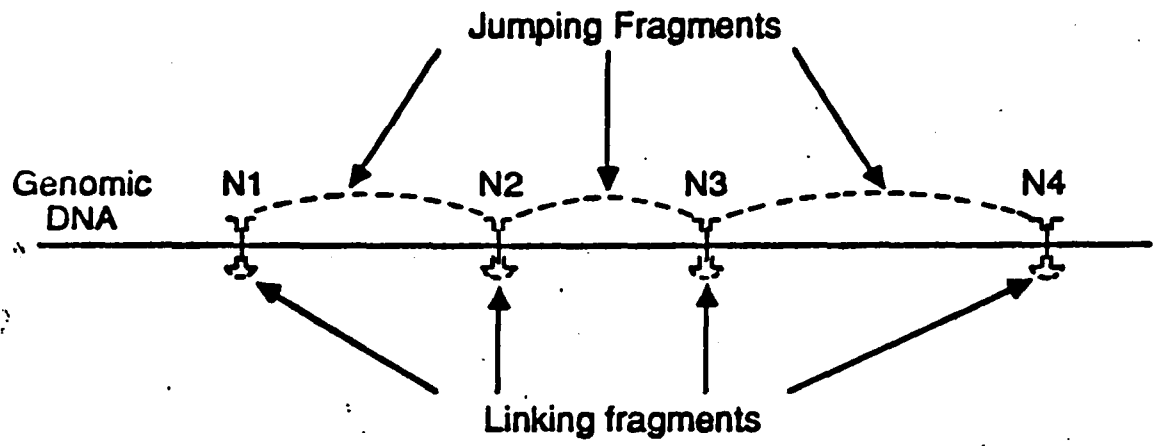
Site size (bp)	No. of C + G	No. of CpG	Avg. frag. size <sup>a</sup> (kB)
6	6	0	16
6	6	1	78
6	6	2	390
6	6	3	1950
6	4	0	7
6	4	1	35
6	4	2	170
8	8	0	390
8	8	1	1950
8	8	2	9750
8	4	0	77
8	4	1	380
8	4	2	1920

a - This was calculated by binomial statistics neglecting the clustering of CpG in HTF islands but including the observed occurrence of CpG at only 20% the frequency expected from the overall base composition.

TABLE 3: Parameters for 100 kb hopping library

Starting amount of high MW genomic DNA	100 ug
Amount of size-selected DNA	5 ug
Range of sizes included	80 - 130 kb
Amount of supF gene (BamHI ends)	2 ug
Molar excess of supF gene (220 bp)	200:1
Ligation volume	25 ml
Genomic DNA concentration	0.2 ug/ml
Amount of lambda vector	150 ug
Total plaques on sup <sup>+</sup> host	4 x 10 <sup>8</sup>
Insert containing plaques on sup <sup>+</sup> host	1 x 10 <sup>8</sup>
Total plaques on sup <sup>-</sup> host	4 x 10 <sup>6</sup>

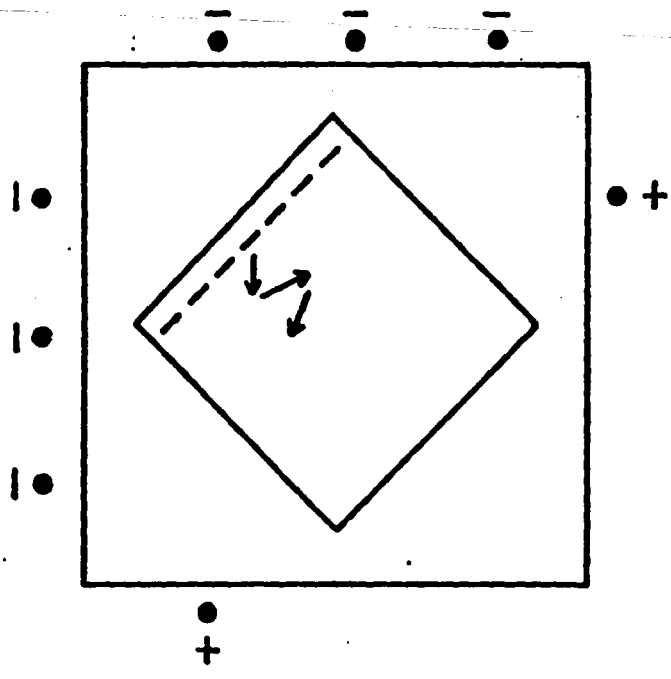




Linking clones include the ends of two adjacent Not 1 fragments  
Jumping clones join the two ends of a single Not 1 fragment

222

-G.





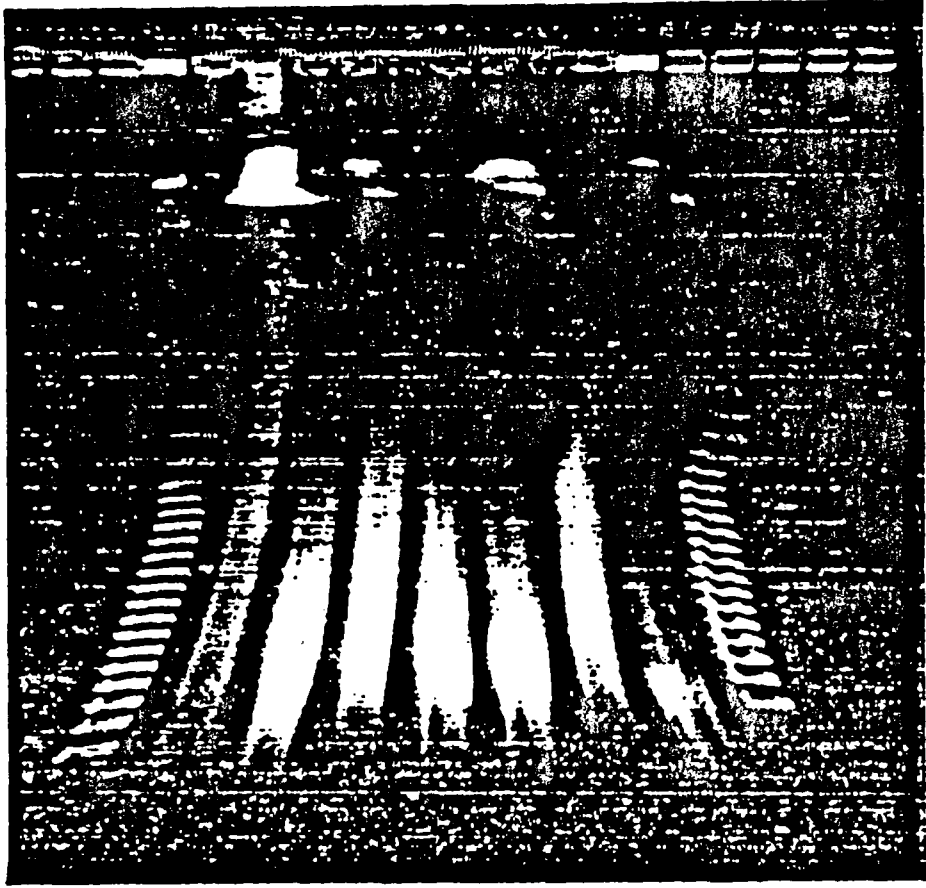
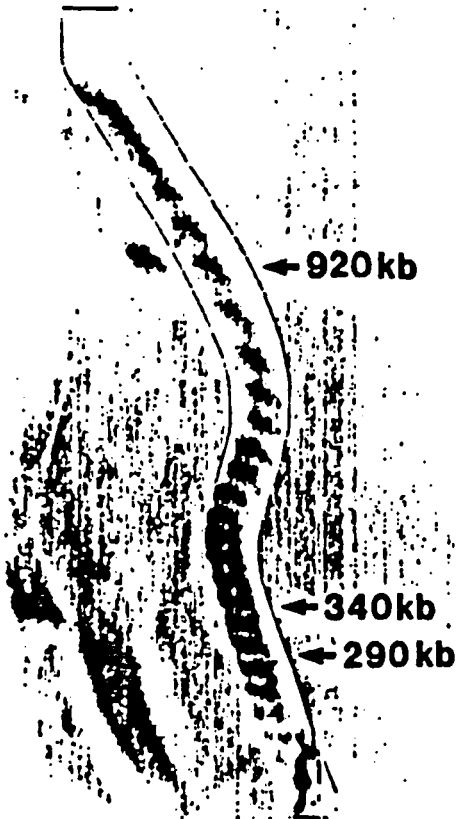


Fig 5

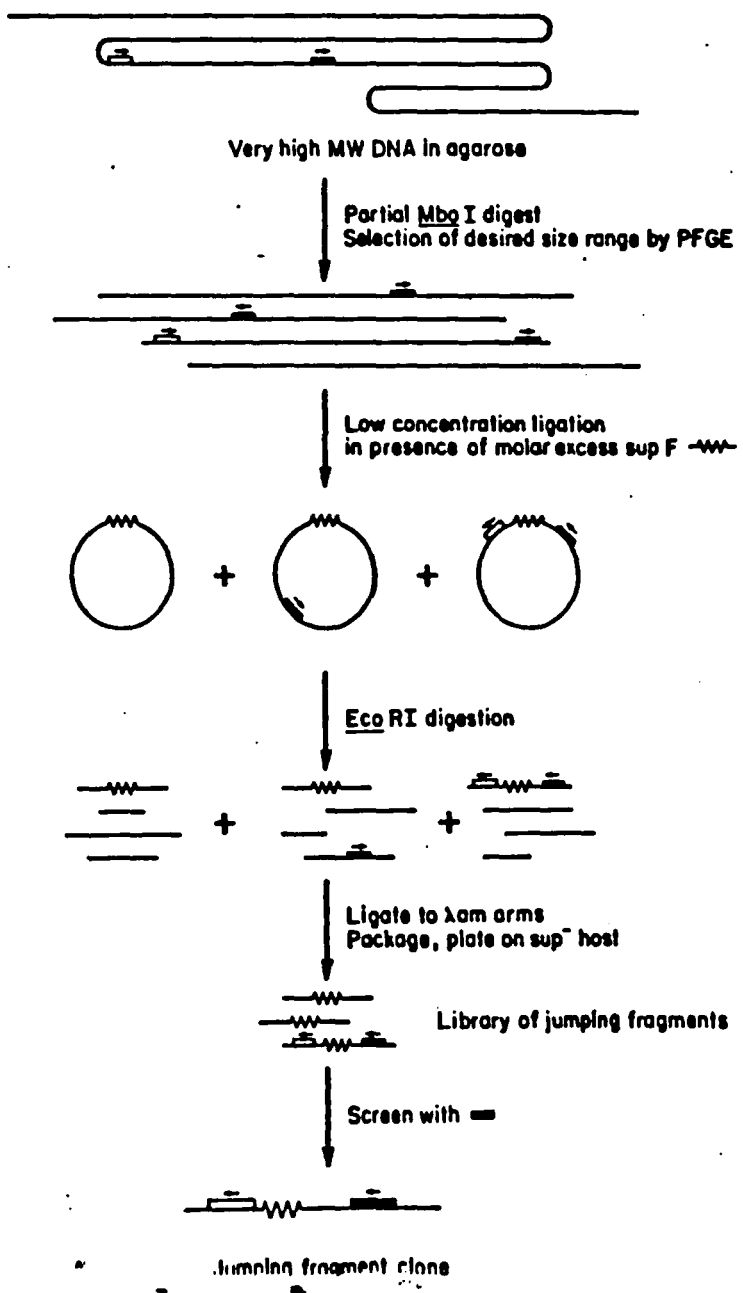
1 2 3 4

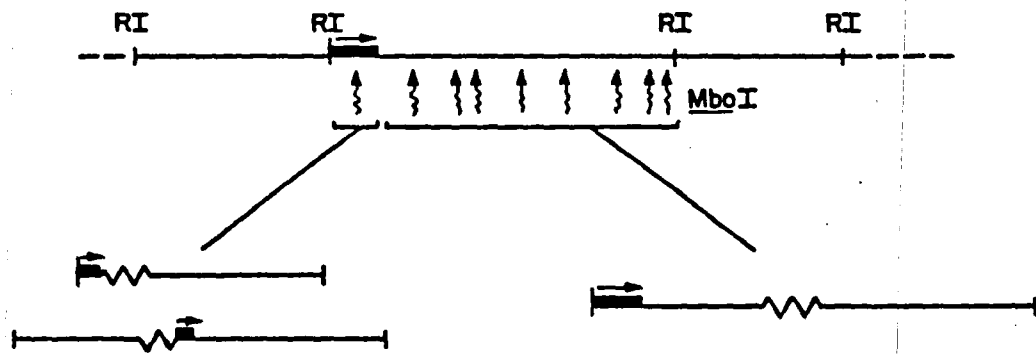
5



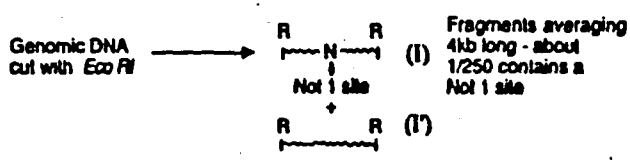
← 3.1 kb

Fig 5

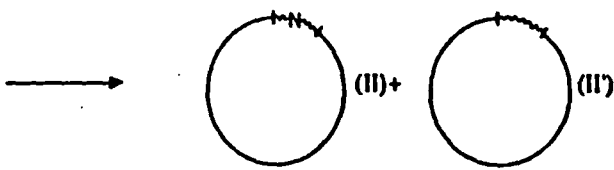




### ISOLATION OF LINKING FRAGMENTS



(I) + (I') are cloned in a plasmid that has only kanamycin resistance marker

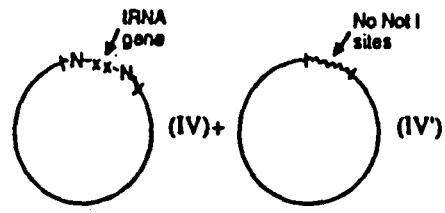


(II) is cut with Not I → (III)

(III) + N-xxx-N

↑  
Sup tRNA gene with Not I ends

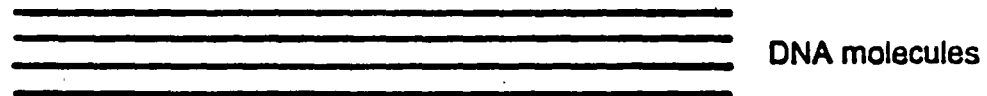
ligation at low concentration



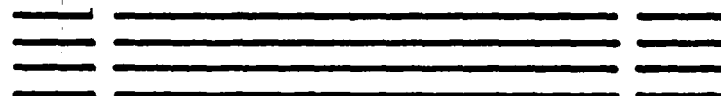
(IV) + (IV') are cloned into bacteria containing P3 plasmid with suppressible mutations in *cat* and *amp* genes on *cat* + *amp* selection (IV) grows, (IV') does not

→ library of linking fragments

# Smith - Birnstiel restriction mapping



↓ total digest with enzyme 1



↓ partial digest with enzyme 2

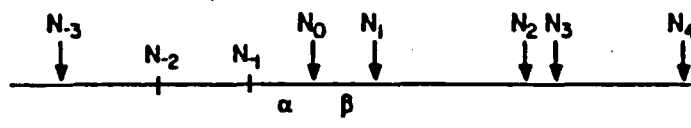


■ probe

228

### A GENERAL PARTIAL DIGESTION PROCEDURE FOR RESTRICTION SITE MAPPING

Genomic DNA



Relevant fragments on gel	Bands with probe $\alpha$	Bands with probe $\beta$	Bands extending in one direction from $N_0$	
			leftward	rightward
(N-2 N2)	—	—		
(N0 N4)		—		
(N-3 N0)	—			
(N-1 N2)	—	—	}	(N3-N4)
(N-2 N1)	—	—		
(N0 N3)		—	}	(N2-N3)
(N0 N2)		—		
(N-2 N0)	—		}	(N1-N2)
(N-1 N1)	—	—		
(N0 N1)		—		
(N-1 N0)	—		}	(N2-N1)
		—		



DEPARTMENT OF HEALTH & HUMAN SERVICES

Public Health Service

National Institutes of Health  
Bethesda, Maryland 20892  
Building : Shannon  
Room : 124  
(301) 496- 2433

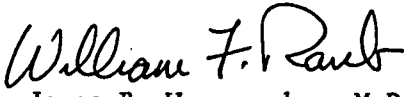
AUG 7 1987

David T. Kingsbury, Ph.D.  
Associate Director for Biological,  
Behavioral, and Social Sciences  
National Science Foundation  
Room 506  
1800 G St., N.W.  
Washington, D.C. 20550

Dear Dr. Kingsbury:

Enclosed is the Department of Energy and National Institutes of Health budget information concerning agency support for research related to the human genome, in accordance with the format developed at the May 8 meeting of the Domestic Policy Council, Working Group on Biotechnology, Subcommittee on the Human Genome. In addition, included are: (1) a list of organizations that may be contacted in order to determine the extent of support for similar research, (2) a draft letter for your signature requesting their cooperation in collecting this information, and (3) minutes taken from the Subcommittee meeting. Please let me know if I can be of further assistance.

Sincerely,

  
James B. Wyngaarden, M.D.  
Chairman,  
Subcommittee on the Human Genome

Enclosures

cc:

Dr. Kirschstein

Dr. Bentley

Dr. Danello

Dr. DeLisi

Dr. Fowle

Dr. Noonan

Dr. Smith

Dr. Wooley

✓ Ms. Levinson



## DEPARTMENT OF ENERGY SUPPORT FOR RESEARCH RELATED TO THE HUMAN GENOME

The Department of Energy Human Genome Initiative envisions a focused, coordinated program to create the technologies, tools and resources which will allow a complete characterization, at the molecular level, of the human genome. This coordinated focus on technology development is distinct from the current national effort in human genetics. The DOE Human Genome Initiative is not focused on mapping/sequencing around genes (even human genes) of particular disease interest, and such research in DOE is not included in that initiative. In the budget information provided below all projects which are part of this coordinated research effort are listed under I. A., Focused Research, Human, even though they may not deal directly with the human genome. The key criterion is that these projects will be coordinated and managed as part of the effort to develop the requisite technologies and resources for characterizing the human genome.

The DOE Human Genome Initiative officially originated as a focused program in the FY1988 Budget Submission to Congress which contains \$11.5 million for this activity. In anticipation of this program's beginning, we have initiated some projects in FY1987, including some reclassification of ongoing projects. In FY1986 the Genome specification did not exist, but DOE research has been developing along these lines for several years. The desirability of creating resources which can be used by the entire community in order to achieve the most efficient utilization of future research funds, was one of the motivations for the Genome Initiative. Savings in cost and time would clearly be provided by the availability of ordered clone libraries and improved sequencing technologies, for the whole research community as well as DOE projects. In addition much more rapid and economical gene comparison capacities were desired for related DOE programmatic objectives: (a) a capacity to specify on an individual basis susceptibilities to background radiation and energy related and environmental carcinogens; and (b) development of efficient technologies to monitor mutations in human populations.

### I. Focused Research A. Human

#### FY 1987 PROJECTS CLASSIFIED AS FALLING UNDER THE "HUMAN GENOME INITIATIVE"

LLNL	Ordered Clone/Mapping of Chromosome # 19	\$875K
LANL	Ordered Clone/Mapping of Chromosome # 16	670
Columbia U.	Pulse Field Gel Mapping of Chromosome #21	430
Harvard U.	Evaluation of Multiplex Sequencing	150
LLNL	Gene Library Project	735
LANL	Gene Library Project	771
LANL	Computational Support for Mapping	200
LANL	Database Development	200
LANL	New Sequencing Technologies	75
ORNL	Gel Imaging Technology	94
	Total	4,200

Several other proposals are nearing completion of review for possible funding in FY 1987. In addition, there are several planning and information transfer workshops which have and will be supported which are not included in these figures. Within DOE, the above activities are all considered to be focused Genome Initiative tasks.

#### OTHER RELATED RESEARCH

##### II. Basic Research

Research projects which merely have some gene cloning or sequencing component are not considered part of the Genome Initiative. Rather they will be beneficiaries of resources developed in that program. Abstracts of such basic research projects are appended below. In FY1986, \$5,447K was devoted to these projects. In FY1987, it is estimated that over \$6,000K can be so classified.

##### III. Research Infrastructure

NIH	GenBank	200
LLNL	Flow Cytometry Instrumentation	545
LANL	Flow Cytometry Instrumentation	275

#### CURRENT AND FUTURE FUNDING LEVELS

FY86	Basic Research	\$5,447
FY87	Genome Initiative	4,200
	Infrastructure	1,020
	Basic Research	approx. 6,000
FY88	Genome Initiative, Congressional Budget Submission	11,500
	Infrastructure	est. 1,500
	Basic Research	est. 7,000
FY89	Recommendation by HERAC Genome Committee (appended)	40,000
	Infrastructure	rough est. 2,500
	Basic Research	rough est. 8,500

The preparation of the FY 1989 budget is just getting underway and it isn't possible at this time to provide reliable estimates of funding levels.

Technical review panels will be set up to help prioritize proposals under the Human Genome Initiative. These panels will carry out prospective and retrospective review of all projects and proposals whether from DOE Laboratories, academia or industry.

University of California  
San Francisco, California 94143

Laboratory of Radiobiology and  
Environmental Health

428. *Mammalian Cell Culture Systems for Molecular Studies of DNA, Mutagenicity, and Carcinogenicity*  
J.E. Cleaver

\$379,000

Cultured mammalian cells proficient or deficient in various DNA repair and replication pathways are the basic biological systems used to determine the mechanisms by which damage to chromatin (induced by environmental pollutants) eventually cause cell death, mutation, and cancer. Cells from patients deficient in excision repair of ultraviolet light-induced DNA damage and other radiation-sensitive mammalian cells are used as the primary model systems. Biochemical and molecular biological techniques are used (1) to measure various parameters of normal DNA replication and DNA repair, (2) to determine the relationship of chromatin structure to repair and replication, and (3) to identify DNA sequences that code for radiation and carcinogen sensitivity. This project is part of the basic biochemical approach of the laboratory's work to uncover mechanisms of damage induced by energy-related environmental pollutants. [1912]

429. *DNA Replication and Repair*  
R.B. Painter

\$467,000

This project investigates how ionizing radiation and other mutagenic agents affect the organization, replication, and expression of DNA in human and other mammalian cells. Since DNA replication in cells from patients with the autosomal recessive disease ataxia-telangiectasia (A-T) is different from that in normal human cells, and since DNA replication is inhibited much less by radiation in A-T cells than in normal cells, we are using A-T cells as a model system. DNA transfection of human cells and other recombinant DNA techniques are developed to isolate the gene that complements the defect in homozygous A-T cells. Subsequently, the mRNA will be isolated and used both to locate the position of this gene in the genome and to obtain its translation product. To obtain information on regulation of DNA replication, gene amplification is also studied. Current work in other laboratories favors saltatory replication or unequal sister chromatid exchange as mechanisms for gene amplification in mammalian cells, but other explanations may be more likely. We are testing the hypothesis that unequal segregation of DNA, which results from chromosomal aberrations induced by the drug, causes increased copy numbers of genes to occur in some cells. Expression of transfected genes not closely associated with a promoter is studied to determine the effects of translocation and transposition in human cells. A comparison of repair-proficient and repair-deficient cells in these studies should lead to information about the role of repair systems in the expression of rarely expressed genes, such as oncogenes. [4455]

446. *Cytogenetics and Molecular Genetics*  
A.V. Carrano

\$796,000

This project is aimed at: (1) understanding mechanisms by which genetic damage is induced in mammalian systems; (2) developing an animal model to determine the biological significance of such damage; and (3) probing the location of genes in the human genome associated with DNA repair of this damage. Emphasis is on the integration of cytogenetics, mutagenesis, and molecular biology to link initial genetic injury and ultimate consequence for man. Mechanisms of cytogenetic damage are studied in cultured Chinese hamster or human cells. Using Chinese hamster cell mutants deficient in DNA repair, events at the molecular level are related to observed cytogenetic consequences. The significance of cytogenetic and mutation endpoints is fully examined in a mouse model, comparing amount of initial damage and rate of repair for several endpoints, including chromosomal aberrations, sister chromatid exchange, and specific locus mutations at the hprt locus in splenocytes and other tissues. Methods developed can also be applied to humans to estimate exposure to physical and chemical substances and to understand factors that potentially confound interpretation of induced cytogenetic damage. At the molecular level this project is interested in the events that mediate initial DNA damage and the genes that control these processes. Human genes identified as complementary to the repair deficiency in mutant Chinese hamster cells can be isolated to study their function. [3743]

451. *Mutations in Cultured Somatic Cells*  
L.H. Thompson

\$623,000

The objective of this project is understanding, at the molecular level, the complex process of mutagenesis in mammalian cells and the role of DNA repair in the production of genetic alterations by environmental agents. Our approach involves genetic analysis based on studying specific gene mutations that affect DNA repair processes. We have developed in the CHO cell line a collection of mutant strains that are much more amenable to genetic manipulation than the available repair mutants of human cells. The CHO mutants have opened up new ways of identifying and studying the human DNA repair genes. The CHO mutants can be complemented, or corrected, by fusion with human cells or by transfection with human DNA. The genetic similarity of the hamster and human repair genes has allowed us to map the chromosomal locations of the human genes and to begin isolating these human genes by recombinant DNA methods. Recent progress includes: (1) localizing three DNA repair genes to human chromosome 19; (2) constructing DNA libraries for isolating repair genes; (3) finding that two of the excision-repair mutants of CHO complement several genetic groups of xeroderma pigmentosum human mutants; and (4) validating a simple, sensitive, and unique mammalian cell test that appears specific for DNA cross-linking damage. In addition to isolating human repair genes, our immediate objectives include mapping three other repair genes and studying repair by transfecting cells with damaged plasmid DNA molecules. [0261]

468. *Genome Structure and Function*  
R.K. Moyzis

\$174,000

The long-range objective of this project is to determine the molecular mechanisms by which humans organize and express their genetic information. Ultimate applications of these investigations include development of sensitive assays for detection of human genetic diseases, and measuring effects of low-level ionizing radiation and/or carcinogen exposure. A combination of biochemical, biophysical, and recombinant DNA techniques are being used to identify, isolate, and determine the roles of specific DNA sequences mediating chromosome structure and/or involved in regulating differential gene expression. Current efforts focus on determining (1) the organization and function of human repetitive DNA sequences and (2) the factors regulating induction of metallothionein and related genes. These proteins are involved in reducing cytotoxic and carcinogenic effects of heavy metals and alkylating agents. Investigations in the past year have led to the proposal of a novel model for the general organization of human DNA. 75% to 90% of the mass of repetitive sequences present in human DNA have been isolated and characterized. A highly sensitive assay for genetic damage was developed using these cloned repetitive sequences. An initial step in the metallothionein gene studies was completed, resulting in the isolation of the hamster MT-I gene. Future studies will: (1) define and isolate functional DNA repetitive regions; (2) use repetitive sequences to define genomic variability and damage; and (3) determine specific regulatory mechanisms used by the metallothionein genes. Defining

University of Southern California  
Los Angeles, California 90033

552. *Structure and Function of Human Metallothionein Genes and Their Protein Products*

M. Karin

\$26,000

Department of Microbiology  
213-224-7344

Exposure of various organisms to heavy metals can lead to long-term adverse effects (e.g., carcinogenesis and teratogenesis) besides the immediate toxicity induced by active exposure. We are studying the human metallothionein (MT) gene family that encodes heavy-metal-binding proteins responsible for protection against metal toxicity. To gain a better understanding of the protection mechanisms, we have isolated the different human MT genes. We plan to determine the full nucleotide sequence of these genes. The expression and regulation of the individual genes will be studied in various human cell lines and tissues using gene-specific hybridization probes. We also plan to investigate MT gene expression during human embryonic development. Since it is now known why most organisms, especially primates, express multiple forms of MTs we plan to investigate any functional differences between the various genes and proteins in respect to *in vivo* metal-binding capacities and protection against toxicity. Heavy-metal-resistant cell lines whose survival depends on high-level expression of human MT genes carried on plasmid vectors will be constructed. In this way we can study the functional *in vivo* differences between the various human genes. Using the most efficient MT gene(s) identified, we will construct nonpathogenic viral vectors that will allow high-level expression of the inserted human MT gene and as a consequence will lead to increased resistance by cell cultures and experimental animals to heavy-metal contamination. [9314]

520. *Correlation of Chromosome Patterns in Human Leukemic Cells with Exposure to Chemicals and/or Radiation*

J.D. Rowley

\$230,000

Franklin McLean Memorial Research Institute  
312-962-6117

Nonrandom chromosome changes, particularly loss of Nos. 5 and/or 7, are observed in most patients with acute leukemia (ANLL) as a second malignancy (i.e., developed after radio- or chemotherapy for another disease). The frequency of these chromosome abnormalities increases in ANLL *de novo* patients occupationally exposed to chemical solvents or pesticides. One project goal is to correlate data from patients with various myeloid leukemias regarding their genetic background and exposure to occupational, environmental, or medical therapeutic agents that may be mutagenic with (1) the chromosome pattern of the leukemic cells, (2) the morphology of the leukemic cells, and (3) the patient's clinical course. Another goal is to determine if presently available cytogenetic tests can distinguish patients treated with mutagenic agents whose cells are particularly sensitive to the chromosome-damaging effects of these agents. Use of DNA probes to identify the DNA located at the sites of chromosome translocations associated with certain morphologic types of leukemia will allow us to determine the product(s) of the DNA and to investigate the alterations in function of the affected genes. Data on 63 patients with ANLL, as a secondary malignancy reveals a distinctly different chromosome pattern in the leukemic cells compared with patients with ANLL *de novo*. Sixty-one of 63 patients with secondary leukemia had an abnormal karyotype, compared with 50% of ANLL *de novo*. Fifty-five of 61 (90%) patients had consistent loss of chromosomes Nos. 5 and/or 7. We correlated the karyotype and occupation of 74 patients with ANLL *de novo*. Seventy-five percent of patients who had a history of exposure to chemicals or petroleum had an abnormal karyotype compared with 43% of patients not exposed. Losses of chromosomes 5 and/or 7 were much higher in the exposed (67%) than in the nonexposed (20%) population. [8328]

Case Western Reserve University  
Cleveland, Ohio 44106

516. *Repair of DNA Treated with  $\gamma$ -Irradiation and Chemical Carcinogens*

D.A. Goldthwait

\$52,000

Department of Biochemistry  
216-368-3337

Current work is in three areas. The first area concerns oncogenes in tumors of the human central nervous system. We are searching for oncogenes by transfection with CNS tumor DNA of NIH 3T3 cells and of immortalized glial cells. A search for *sis* genome rearrangements and an analysis of *sis* gene expression in cell lines continues. Factors affecting *sis* gene half life, the effect of anti-*sis* RNA, and the expression of a series of other oncogenes are under study. Another area of study involves the effect of carcinogens on CAT gene expression through LTR sequences. We have inserted the CAT gene in a retroviral vector in order to make a mouse cell line suitable for the study of physical and chemical agents on expression of the CAT gene. Finally, we search for a human transposable element as an inactivator of the *gpt* gene inserted into human lung carcinoma cells. We are completing work on the effect of ATP on the structure of chromatin in HeLa cell nuclei. [9239]

481. *Gene Expression in Carcinogenesis*  
F.T. Kenney, K.-L. Lee \$375,000

Project objectives are: (1) to define the molecular mechanisms by which gene expression is regulated by hormones and during differentiation in mammalian cells and (2) to determine how gene control is lost in transition to malignancy. We have shifted away from a focus on a single model gene toward analyses of several. Through molecular cloning we have isolated seven discrete genes whose expression is regulated by glucocorticoids in rat liver. Current work focuses on two of these: tyrosine aminotransferase and a second of (as yet) unknown specificity, provisionally identified as gene 33. Transcription of both genes is accelerated in liver by glucocorticoids, insulin, or cyclic AMP; a similar response to the latter two classical antagonists was thought to be unique to the aminotransferase. Expression of both genes is developmentally activated at about the time of birth during differentiation of the liver. However, the control of their expression is not identical in other aspects (i.e., tissue specificity and hormonal responses in fetal hepatocytes). The structure of gene 33 is being analyzed in detail to compare with the aminotransferase gene; this information should lead to a definitive description of the molecular basis for similarities and differences in their regulation. We are assessing what appear to be *trans* acting regulatory genes controlling differentiation of the liver; these genes have been identified through study of deletion mutants in mice from the Biology Division mouse genetics facility. Initial experiments demonstrate that the usual developmental activation of aminotransferase gene expression does not occur in mutant mice, while that of gene 33 is not affected. These studies will expand understanding of differentiation and of hormonal regulation of gene expression in normal cells, and add to knowledge of the factors responsible for the genomic instability associated with cancer. [1699]

490. *Effects and Uses of Chromosomal Aberrations*  
L.B. Russell, W.M. Generoso \$175,000

This project is concerned with various qualitative properties of heritable mouse chromosome aberrations whose induction is studied and quantitated under other DOE-sponsored research. Major emphasis is on (1) the organismic effects of the aberrations and (2) the events (meiotic pairing and segregation) that play a role in the transmission of aberrations to subsequent generations. The findings are of interest to basic genetics and are needed, in conjunction with data on the frequency of induction, to assess genetic risk from environmental agents. Contrary to earlier belief, even balanced rearrangements can produce dominant phenotypes. Among those currently investigated are neurological disorders associated with a reciprocal autosomal translocation. Another type of heterozygous effects of interest is blockage of spermatogenesis, which is found in certain classes of autosomal translocations and in all sex-linked translocations. Proposed causes of this sterility are under investigation. Mechanisms of chromosome recognition and segregation during meiosis play a major role in how rearrangements are transmitted. In turn, certain types of rearrangement, particularly X-autosome translocations and X-inversions, provide excellent tools for studies of homologous and nonhomologous synapsis and de-synapsis. We are analyzing synaptonemal complexes in pachytene spermatocytes for such studies, which presently emphasize X-Y pairing. Finally, certain chromosome aberrations, particularly genetically characterized deletions and translocations, provide valuable tools for isolating specific DNA regions for studies of fine structure and gene expression. Such studies involving Chromosome-7 loci have been initiated. [0184]

Mammal

494. *Carcinogen-Cell Genome Interaction*  
W.K. Yang \$225,000

The long-range goal of this project is to elucidate the cellular genetic mechanisms related to cancer formation and to understand how energy-related environmental insults may affect these mechanisms in the multi-step process of carcinogenesis. Research has focused on the isolation and characterization of retroviral gene elements inherited in the germ line of the mouse, considering that these potentially transposable genes may interact with proto-oncogenes. Emphasis is on detecting cellular genetic factors that may control oncogene action. Recent major results include (1) discovery of a novel class of mouse endogenous retroviral gene elements that carry the IS-type long terminal repeats and are located in autosomes and specifically in the Y chromosome of the mouse; (2) demonstration that some chemical carcinogen-transformed and tumor-inducing NIH3T3 cells do not contain altered *ras* oncogene expression; and (3) development of a sensitive DNA transfection system for oncogenic DNA detection, including isolation of NIH3T3 cell clones with various morphological properties. Specific aims are to study the expression of various endogenous long-terminal-repeat (LTR)-containing genes in mice before and after carcinogenic exposure and to isolate a dominant genetic factor that can suppress *ras* oncogenes from causing the expression of tumor properties. [1701]

525. *Repair of Lesions and Initiation of DNA Replication in Vertebrate Cells*  
J.H. Taylor \$82,000  
Institute of Molecular Biophysics  
904-644-1421

The DNA of most eukaryotes contains the modified base 5-methyl cytosine (5mC), a postreplication modification at certain CG sites. Higher animals have 2 to 4 percent of the cytosines methylated, while higher plants may have as many as one-third of the DNA cytosines methylated. Some of these modifications play a regulatory role, but they also present a source of mutation due to the spontaneous deamination of 5mC. The mismatched bases are simulated by constructing mismatches with mutants at the unique Hpa II site (CCGG) in the DNA of SV40 virus. These heteroduplexes are used to transfect monkey kidney cells in culture to learn how the cell deals with these lesions. We can methylate the Hpa II site *in vitro* before producing the heteroduplexes. We can also methylate other cytosines in the genome and present the SV40 DNA with and without nicks bracketing the lesion. We have learned that the methylated chain is preferentially retained in mismatch repair, but methylation bracketing the lesion is more effective than methylation within the Hpa II site. Nicks near the lesions tend to override the effects of methylation so that the nicked chain is preferentially repaired. Experiments to explore the repair mechanism are continuing by the use of synthetic polynucleotides to produce pure heteroduplexes. We are also studying the initiation of DNA replication and its effect on repair. A number of dispersed repeats have been cloned and are being tested as possible origins for the regulation of replication and as enhancers for gene regulation. [7780]

# Mammal

## 461. Gene Expression in Mammalian Cells D.L. Grady, R.K. Moyzis

\$179,000 Lf;

The objectives of this project are: (1) to identify factors that control phenotypic expression in mammalian cells; and (2) to determine the structure and organization of a family of genes responsive to heavy-metal induction. The coordinate expression of gene families is critical to normal development, differentiation, and metabolism. Abnormal expression of critical genes can have profound effects on cellular function and is likely to be the basis of most, if not all, pathological states. Heavy-metal induction of the synthesis of metallothioneins (MT) provides an ideal model system for basic mechanistic studies of gene expression and the ways in which altered MT gene expression may be involved in human disease and metal toxicity. A combination of recombinant DNA and two-dimensional gel electrophoretic techniques are used to study gene structure and the control of RNA transcription and processing. A total of 14 kilobases of overlapping DNA clones have been isolated that include the structural genes for Chinese hamster MT-I and MT-II proteins. The complete nucleotide sequence of the MT-II gene uncovered a cryptic splice site in the first intron that if used would code for an altered MT-II protein. Additionally, a search for other metal-regulated proteins was initiated. Future studies will be directed towards the identification of key DNA sequences involved in the control of RNA transcription or processing and the role of DNA-protein interactions in this regulation. Ultimately, an understanding of the mechanisms responsible for the coordinate regulation of this metal-induced gene family will allow realistic assessment of the health effects of energy-related pollutants. [0124]

## 467. Recombinant DNA Repair M.A. MacInnes

\$115,000

This project is developing molecular-genetic understanding of DNA excision repair to characterize major biochemical determinants of repair efficiency *in vitro* and *in vivo*. Emphasis is on identifying repair genes essential for DNA damage recognition and enzymatic incision, which are probable rate-limiting steps of repair for diverse chemical and ultraviolet radiation damages. DNA cotransformation and expression of increased ultraviolet resistance in a Chinese hamster ovary (CHO) repair deficient mutant cell line is used to identify unique hamster and/or human DNA sequences comprising the repair genes. This provides strategies for screening recombinant DNA libraries for functional repair genes. Repair system reconstruction with recombinant repair genes will provide a novel basis for testing specific hypotheses about mammalian repair regulation and identifying repair proteins. We have proved that hamster and human genomic DNAs from repair-competent cells can specifically transform CHO mutant UV-135 to express recombinant repair function. DNA libraries derived from human or hamster DNA-transformed cells will be used to screen for the presence and repair activity of cloned sequences. Active cloned repair genes will provide material for increasing understanding of mechanisms regulating DNA repair in humans. These studies will provide a better empirical basis to risk assessing occupational exposure to carcinogenic chemicals and/or sunlight. [4424]

# Plant

## 415. Genetics and Regulation of the Plant Cell Division Cycle J. Van't Hof

\$250,000

The project objective is to understand the molecular and cellular mechanisms responsible for the temporal order of chromosomal DNA replication and cell division in higher plants. Work concentrates on (1) replicon origins, (2) termination of replication, (3) joining of replicons and replicon clusters to produce mature chromosomal-sized DNA molecules, (4) discontinuity of replication fork movement, and (5) free replicon-sized molecules excised from chromosomal DNA. Knowledge of these basic elements of chromosomal DNA replication is essential to understanding how plant cells distribute genetic information to their progeny and how they handle damaged DNA. [0006]

512. Regulation of Gene Expression: The Sgs-4 Glue Protein Gene of *Drosophila*  
S.K. Beckendorf \$104,000  
Virus Laboratory

The project objective is to study the role of a regulatory gene important in the differentiation and development of *Drosophila melanogaster*. The Sgs-4 locus in *Drosophila melanogaster* codes for a developmentally regulated protein synthesized in large amounts in salivary glands at certain times during development of the fly from an embryo. This protein is not necessary for the viability of *Drosophila*, thus mutants that do not produce it or under-produce it are viable and can be studied genetically. Previous work has demonstrated what appears to be an important regulatory region for the Sgs-4 locus, located 300 to 500 base pairs (bp) upstream or on the 5' side of the Sgs-4 structural gene that codes for Sgs-4 protein. This Sgs-4 regulatory region acts as an enhancer element. We will construct a hybrid gene by recombinant DNA techniques, consisting of the putative Sgs-4 regulatory region fused upstream from the structural gene for alcohol dehydrogenase (Adh) of *Drosophila*. This hybrid gene will be introduced by transformation into embryos of *Drosophila* that are genetically deficient in Adh activity (Adh minus). Transformation will be mediated by special genetic elements (P-transposons) by the P-mediated transformation technique. Successful transformation of the *Drosophila* will be measured by expression of Adh activity. Expression of Adh fused to the Sgs-4 regulatory region will be measured in salivary glands of developing larvae of *Drosophila*. Expression of Adh is equivalent to expression of the Sgs-4 regulatory gene in the cis position. Using the transformed *Drosophila* that carry the Sgs-4/Adh construct, we will select mutations in genes that regulate Sgs-4 expression. [9376]

515. Spermatogenesis in *Drosophila*  
D.L. Lindsley \$35,000  
Department of Biology  
619-452-3109

The project objective is the dissection of the normal gametogenesis network into its separate gene-controlled steps through the study of spermatogenesis and meiosis in *Drosophila melanogaster*. By conducting deficiency mapping and genetic complementation studies of newly induced sterile mutations in regions previously saturated with lethal mutations, we have determined the degree to which lethal mutations fail to complement sterile mutations and thus estimated level of coincidence among loci that can mutate to lethal and those that can mutate to sterile alleles. 8/9 male steriles and 5/7 female sterile mutations are complemented by all the lethals in their vicinity; thus, lethally mutable and sterility mutable loci form largely disjunct populations. We are currently examining the ability of normal germ-cell precursors to become established and produce functional gametes in sterile testes. We have determined the effects of duplications and deficiencies for the base of the X chromosome on male fertility by studying the interaction of a single heterochromatic deficiency with a graded series of free-Xh duplications. In addition, a series of X chromosomes deficient for increasing amounts of heterochromatin is being used to examine the relation between X-Y pairing, as measured by the frequency of nondisjunction, and sperm maturation, as measured by X:Y transmission ratios. In these studies, the effects of heterochromatic constitution on male fertility, meiotic drive, and disjunction of X and Y are highly correlated, suggesting that the same mechanism is involved in all three responses. A series of testis-specific genomic clones has been recovered by means of differential cDNA screens of a genomic library. These clones are being characterized; the corresponding

*Drosophila*

551. Mutagenesis: Alcohol Dehydrogenase in *Drosophila*  
W.H. Sofer \$60,000  
Waksman Institute of Microbiology  
201-932-3052

The project objective is to determine the mechanism(s) of mutagenesis of several chemical mutagens in a higher organism. We study the effects of mutagenesis on the amino acid and nucleotide sequences of alcohol dehydrogenase and the alcohol dehydrogenase gene respectively in *Drosophila melanogaster*. We employ the alcohol dehydrogenase system to study the mechanism of mutagenesis because of its useful properties. First, null mutant flies lacking detectable alcohol dehydrogenase (ADH) activity can be selected by chemical procedures devised in our laboratory. More than 200 ADH-negative mutants now exist that have been selected by this technique. Second, both ADH and ADH-mRNA are present in large quantities in wild-type flies; we estimate that between 1.5 and 2% of the total soluble protein in a six-day-old fly is ADH and that ADH mRNA is also present in relatively large amounts. This abundance of protein allowed us (1) to easily purify the enzyme, (2) to raise antibody against it in rabbits and goats, and (3) to measure and isolate cross-reacting material (CRM). ADH is one of the few well-characterized proteins in *Drosophila* whose gene has been extensively studied. Using recombinant DNA technology, we have isolated an ADH cDNA clone and a series of genomic clones, and sequenced the entire ADH gene. These properties of the ADH gene-enzyme system make it suitable for our approach. In the future, we should be able to identify the full spectrum of changes generated by a given mutagen, to learn how neighboring sequences influence the frequency and type of nucleotide changes, and to deduce the mechanism of mutagenesis in a higher organism from these findings. [8656]

437. Mitotic Recombination and DNA Repair  
M.S. Esposito, J. Hosoda \$94,000

The research objectives are to determine the mechanisms of chromosomal recombination and the molecular defects of yeast strains containing mutant *REC* genes. Recombination occurring spontaneously and after exposure of mitotic cells to recombinogenic treatments is assessed genetically and by *in vitro* studies using plasmids that permit monitoring of Holliday structure resolution. Previously isolated *rec* mutant strains are used for molecular cloning and sequencing of *REC* genes, and are characterized with respect to stages of the recombination process in which they are defective. Previous research has provided strong evidence that both spontaneous and recombination-induced mitotic recombination occur by mechanisms that differ from those occurring in meiotic cells. This project will provide tests of molecular models, and a better understanding of the recombinational pathways of eukaryotes. Understanding the pathways will lead to a critical evaluation of the role of genomic rearrangements in the control of gene expression during development and in differentiated cells. Another goal is to elucidate mechanisms controlling generalized repair pathways including recombinational, inducible, and/or error-prone repair. These pathways are closely related to recombination and replication, and are carried out by multi-protein complexes. Focus is on the protein-protein and protein-DNA interactions within the complex, which are the major factors determining the overall activities of the complexes. [3822]

476. Comparative Genetics  
J.L. Epler, F.W. Larimer, C.E. Nix, L.C. Waters \$400,000

The major objective of this project is to obtain basic information necessary for evaluating the genetic effects of potential chemical hazards through analysis of comparative biological endpoints in eukaryotic systems in order to extrapolate results from short-term assays to the estimate of possible genetic risks to man. Our experimental approach involves the use of (1) *Drosophila melanogaster* for investigations of the molecular mechanisms of genetic control of metabolism of xenobiotics (e.g., the P-450 system) and (2) *Saccharomyces cerevisiae* for molecular studies of DNA repair. In mutagen metabolism studies with *Drosophila*, P-450 has been resolved into two electrophoretically distinct forms (P-450-A and P-450-B). P-450-A is expressed in all strains examined while P-450-B is strain dependent, associated with dimethylnitrosamine demethylase activity, under *trans* regulatory control, and correlated with insecticide resistance. Immunochemical and recombinant DNA techniques along with other biochemical and genetic methodology will be used to isolate the P-450-B gene(s). These DNA clones will be used in experiments to determine the molecular mechanisms of P-450-B expression and the basis of P-450 associated insecticide resistance. In the study of DNA repair in yeast, the *rev1* gene has been isolated from a genomic library. Restriction mapping and subcloning will be used to identify the minimum fragment carrying the *rev1* gene. Future experiments will be designed to identify the gene transcript and gene product, facilitating study of its regulation and mode of action. The approach will lead to a better understanding of (1) the role of P-450-dependent metabolism and DNA repair in mutagenesis, (2) the relationship between these model systems in terms of mutagen metabolism and DNA repair, and (3) extension of similar studies to mammalian systems. [1569]

548. Incision of Ultraviolet-Irradiated DNA in Yeast  
L. Prakash \$30,000  
Department of Radiation Biology and Biophysics  
716-275-2656

We will examine the genetic control and molecular mechanisms of the incision step of excision repair following exposure of yeast cells to ultraviolet light. Several human genetic diseases are associated with defective DNA repair and enhanced neoplastic transformation. Correlations are known to exist between the inability to repair DNA damage and the induction of cancer. DNA repair processes can be easily studied in yeast because of its well-defined and versatile genetic system, and ease of manipulation with recombinant DNA procedures. Our studies will focus on further characterization of two of the four genes we have cloned involving incision of ultraviolet-induced pyrimidine dimers in the budding yeast *Saccharomyces cerevisiae*: *RAD3* and *RAD10*. The inducibility at the transcript and protein levels of these genes in response to DNA damage will be determined. The *RAD3* protein will be purified with the aid of antibodies directed against a *RAD3-lacZ* hybrid protein and the purified protein will be characterized for its DNA-binding properties, ATPase activity, and so forth. The nucleotide sequence of the *RAD3* gene, when compared to that of the other genes in this group, might reveal homologies and consensus sequences. We will also isolate and characterize the *RAD4* gene. A long-term objective is to purify all six proteins encoded by the genes required for dimer incision in *S. cerevisiae*. By reconstituting dimer-incising activity *in vitro*, we will better understand the molecular mechanism of this repair process. We will begin similar studies in the fission yeast *Schizosaccharomyces pombe*, since it may have evolved other pathways for repairing ultraviolet damage and may also differ in the incision step. [2868]

549. Mutation and Structure-Function Relationships of Cytochrome C  
F. Sherman \$65,000  
Department of Radiation Biology and Biophysics  
716-275-2766

Iso-1-cytochrome *c* and iso-2-cytochrome *c* from the yeast *Saccharomyces cerevisiae* are two of the proteins of known primary structure from a microorganism particularly suitable for experimental genetic studies and for manipulation by recombinant DNA procedures. This iso-cytochrome *c* system is used to investigate numerous problems in molecular biology and genetics, including: (1) DNA sequencing of mutations induced by a variety of physical and chemical agents including a number of carcinogens; (2) DNA sequencing of unusual mutations that occur by multiple base-pair changes; (3) DNA sequencing of mutable and immutable sites and of unstable mutations; and (4) DNA sequencing of missense mutations and structure-function relationships of iso-1-cytochrome *c*. Cytochrome *c* genes corresponding to tuna cytochrome *c* and pigeon cytochrome *c* will be synthesized and inserted into the yeast genome. We will investigate the transcription and translation efficiencies of these completely synthetic genes and the degree function of the corresponding cytochrome *c*. The functions of the tuna cytochrome *c* with various mutations and amino acid replacements will be examined. [2869]



# Neurospora

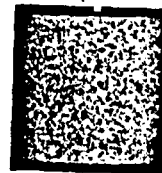
**529. Ribosomal RNA Genes of Neurospora: Heterogeneity and DNA Sequencing of Promoter and Processing Sites**

S.K. Dutta

\$25,000

Department of Botany  
202-636-6942

Ribosomal (r)RNA genes are important for continuity and maintenance of a given cell. These rRNA genes exist in multiple copies in all of the eukaryotes studied, including the lower eukaryote *Neurospora crassa*, as reported earlier. Studies reveal more variety than uniformity in amplification, genomic organization, and interconnected mechanism for regulation of transcriptions of rRNA genes in the developing organism. The primary objective of this project is to increase our understanding of the heterogeneity of rRNAs, promoter sequences, and processing sites of rRNA genes using rDNA clones from *N. crassa* cell types like conidia, germinated conidia, mycelia, normal and abnormal morphological mutants of *N. crassa*, and representative species of sexually incompatible groups of *Neurospora*. Variable regions of rDNA sequences (internal and external spacers) and DNA fragments containing promoter and processing sites will be cloned and sequenced. Since rRNA genes are in multiple copies and that number of copies varies in cell types, it is essential to know whether these copies are identical. The conservation of specific sequences around rRNA genes, which may suggest recognition and/or regulatory sites for initiation, processing, and termination, must be understood in order to know controls of gene expression. [6204]



# Bacteria / Phage

Stanford University  
Stanford, California 94305

553. *Repair of Damaged DNA In Vivo*  
P.C. Hanawalt, A.K. Ganesan \$94,000  
Department of Biological Sciences  
415-497-2424

This project will elucidate the multiple pathways of excision repair and DNA lesion tolerance in *E. coli* and their relationships to biological endpoints such as survival and mutagenesis. Current research objectives include: (1) comparative analysis of the mechanisms and genetic control of inducible long-patch excision repair and daughter strand gap repair in ultraviolet-irradiated bacteria, with focus on effects of mutations that alter normal *recA-lexA* regulatory circuit; (2) analysis of the action of the *denV* gene product of bacteriophage T4, endonuclease V, on defined DNA substrates containing pyrimidine dimers; (3) further characterization of the cloned *denV* gene, an attempt to obtain higher yields of the gene product for use in determining the pyrimidine dimer content of ultraviolet-irradiated DNA and to study the effect of the gene in mammalian cells; (4) development of immunological probes for specific photoproducts in bacterial DNA and their use to analyze disposition of these photoproducts in irradiated cells (with probe reactivity compared with the specificity of other biological probes); and (5) analysis of the spectrum of photoadducts produced in model substrates treated with 8-methoxypsoralen and UVA. The focus is on adducts involving cytosine nucleotides, both in *E. coli* DNA and in oligonucleotides of defined sequence. Our overall rationale is that bacteria continue to serve as the most important model system for guiding exploration of DNA repair schemes in human cells. [6289]

546. *Molecular Mechanisms of Misrepair Mutagenesis*  
C.W. Lawrence \$85,000  
Department of Radiation Biology and Biophysics  
716-275-2948

The project involves two complementary approaches to the study of misrepair mutagenesis in yeast and *E. coli*, which in combination will contribute to the long-term goal of elucidating the molecular mechanisms responsible for error-prone repair. The first approach is to clone some of the more important genes concerned with mutagenesis, with the eventual aim of identifying and isolating their products. The second approach is to construct M13 single-stranded DNA virus derivatives that contain a single defined lesion at a specific site, to be used as tools to investigate *in vivo* replication. The molecules will help to provide precise information concerning the occurrence of translesion synthesis, its error rate, and the induced spectrum of targeted or untargeted mutations.

University of Rochester  
Rochester, New York 14642

544. *Inducible Error-Prone Repair in B. Subtilis*  
R.E. Yasbin \$32,000  
School of Medicine and Dentistry  
716-275-5229

The study of DNA repair in *B. subtilis* is still in its early development. With few exceptions, the work to date has been descriptive in nature. The project objective is to apply molecular biological and biochemical technologies to this areas. Building upon our laboratory's earlier genetic characterization of the inducible SOB system, we propose (1) to investigate the role of mutations on the regulation and control of the SOB phenomena; (2) to clone the genes involved in regulating the SOB system and begin identification of these gene products; (3) to continue the isolation of Din (damage inducible) and Cin (competence inducible) genes, map these genes and use the promoters to begin isolating the cellular repressor(s) associated with the SOB system; (4) to purify the bacteriophage 0105 repressor, using the DNA clone already obtained, and identify the mechanism of inactivation of this repressor when the SOB system is induced; and (5) to study the interaction between the SOB system and the development of competence. Results help in the dissection of the SOB system (both its regulation and its mechanism of action) and the relationship between DNA repair, mutagenesis, and competence development, and how bacteria handle environmental stress situations. [8908]

410. *Nucleic Acid Structure*  
F.W. Studier \$275,000

Bacteriophage T7 and its host *E. coli* provide an excellent system in which to study basic genetic processes. T7 contains a single molecule of linear DNA almost 40,000 base pairs long, and the complete nucleotide sequence is known. During infection, T7 DNA is introduced into the *E. coli* cell where it redirects the metabolism of the host cell to produce new phage particles. Processes such as entry of the DNA into the cell, regulation of gene expression, replication and repair of DNA, gene recombination, and assembly of virus particles can be studied by a combination of genetic and physicochemical techniques. The nucleotide sequence predicts 50 to 55 T7 proteins, and mutations affecting most of them have been found. Specific fragments representing all parts of the T7 DNA molecule have been cloned in bacterial plasmids and are being used for genetic and biochemical analysis of T7. Simple procedures for radioactive labeling, gel electrophoresis, and filter hybridization permit resolution and identification of individual T7 RNAs, proteins, and intermediates in DNA replication and recombination. Thus, all of the T7 genetic processes can be studied in detail at the molecular level. An understanding of basic genetic processes in this simple system should provide useful models for understanding similar processes in more complex plant and animal systems. The T7 system itself, and the techniques developed for studying it, are finding wide application in studies on the detailed biological effects of mutagens, carcinogens, and radiation. [0019]

---

**REPORT ON THE  
HUMAN GENOME INITIATIVE  
for the  
OFFICE OF HEALTH AND  
ENVIRONMENTAL RESEARCH**

---

Prepared by the  
Subcommittee on Human Genome  
of the  
Health and Environmental Research Advisory Committee  
for the  
U.S. Department of Energy  
Office of Energy Research  
Office of Health and Environmental Research

---

April 1987

---



27 April 1987

Dr. Alvin W. Trivelpiece  
Assistant Secretary  
Office of Energy Research  
U.S. Department of Energy  
Washington, DC 20545

Dear Dr. Trivelpiece:

On behalf of the Health and Environmental Research Advisory Committee (HERAC), I am pleased to submit to you the enclosed *Report on the Human Genome Initiative*. This was prepared by a subcommittee under the chairmanship of Dr. Ignacio Tinoco, University of California, Berkeley, and is in response to a charge by you. It has been strongly endorsed by the parent committee.

The report urges DOE and the Nation to commit to a large, multi-year, multidisciplinary, technological undertaking to order and sequence the human genome. This effort will first require significant innovation in general capability to manipulate DNA, major new analytical methods for ordering and sequencing, theoretical developments in computer science and mathematical biology, and great expansions in our ability to store and manipulate the information and to interface it with other large and diverse genetic databases. The actual ordering and sequencing involves the coordinated processing of some 3 billion bases from a reference human genome.

Science is poised on the rudimentary edge of being able to read and understand human genes. A concerted, broadly based, scientific effort to provide new methods of sufficient power and scale should transform this activity from an inefficient one-gene-at-a-time, single laboratory effort into a coordinated, worldwide, comprehensive reading of "the book of man". The effort will be extraordinary in scope and magnitude, but so will be the benefit to biological understanding, new technology and the diagnosis and treatment of human disease.

It may seem audacious to ask DOE to spearhead such a biological revolution, but scientists of many persuasions on the subcommittee and on HERAC agree that DOE alone has the background, structure, and style necessary to coordinate this enormous, highly technical task. When done properly, the effort will be interagency and international in scope; but it must have strong central control, a base akin to the National Laboratories, and flexible ways to access a huge array of university and industrial partners. We believe this can and should be done, and that DOE is the one to do it.

Sincerely,

A handwritten signature in black ink, reading "Mortimer L. Mendelsohn". The signature is fluid and cursive, with a long horizontal flourish extending to the right.

Mortimer L. Mendelsohn, M.D. Ph.D.  
Chairman, Health and Environmental  
Research Advisory Committee

considerably less than half of the coding sequences even if we knew the entire sequence and the locations in the DNA sequence of the primary transcripts. This is an area where focused research could greatly improve the outlook, even without new data. The use of current data with a good expert system (an expert system is a computer program that uses all the information that an expert would have to solve a problem) could significantly increase identification of splice-junction sites. New data will continue to enhance the performance of such programs. Although such programs will never be perfect, they provide predictions that are easily tested. Effective programs would avoid the necessity of sequencing the entire messenger RNA for a protein.

An expert system approach can be used on other patterns as the data emerge. For example, the system used to find splice-junction sites could also be used to identify promoter regions when more data are available to define them. The interaction between computer-aided predictions and experimental results is important. The results will improve the predictions, and the predictions should direct the experiments. An investment begun now in computer applications research will maximize the return, in the short term as well as the long term.

There are many more areas of sequence analysis that will benefit the human genome project. Current search and comparison programs should be made more efficient to handle the enormous size of the data base. We should also do more to understand the biological significance, in contrast to statistical significance, of finding sequence homologies. Research into general pattern identification methods would prove valuable.

Equally important to locating the proteins on the DNA sequence and determining their regulation is to understand their functions. Recent years have produced improvements in our ability to predict protein structures from their sequences. More research is needed to be capable of reliably predicting both structures and functions. That would provide an additional major key to unlock the information of the genome.

Appendix B. The Need for Computer Resources: A Data Bank for the Future

As physical map data are gathered they must be stored in a way that facilitates the cross comparisons required to construct a complete map. Programs that do these functions already exist, but they may be inadequate for this project, because the human genome is about 1000 times as large as the largest current map (E. coli). Inefficiencies that are tolerable on small projects will be major problems on projects the size and complexity of the human genome.

It is particularly important to include in the data base references to other data bases and to facilitate communication between data bases. Specifically, it is necessary to be able to locate physical fragments with respect to any known genetic markers or to restriction fragment length polymorphisms. This is essential for the project to fulfill its promise of facilitating our understanding of human diseases. The entire set of data bases on Genomic Resources (the human gene map at Yale, the mouse gene map at Jackson Laboratories, etc.) which the Howard Hughes Medical Institute is helping to make cross-referenceable, contain data relevant to the human genome project. There are major nucleic acid and protein sequence data banks in the U.S., Europe and Japan which have agreed to collaborate closely. This effort must be supported and further developed. A coordinated effort must be established to maximize the interaction between these data bases, to reduce duplication of effort and to improve speed of data collection. Furthermore, there will undoubtedly be new discoveries, such as the introns which were discovered a decade ago, therefore it is important that the data bases be designed to absorb such changes gracefully.

Sequence Analysis. If the human sequence were magically made available, much of its interpretation would still remain obscure. Research performed now could unlock a substantial amount of the hidden information as the sequence becomes available. For instance, one of the key pieces of information included in the DNA sequence is the protein sequence, but that requires knowledge of the locations of transcription and of the splice junctions. Splice-junction information is usually obtained by sequencing both the genomic DNA and the messenger RNA. This requires substantially more work than would be needed if we could recognize the splice junction from the DNA sequence. However, the best current methods are correct only about 85% of the time in predicting splice junctions in genomic sequences. That means that all the junctions of a three-intron mRNA would be properly recognized only about 40% of the time. If a human gene resembles this example, then it is likely that with current methods we would know

Report on the Human Genome Initiative  
Office of Health and Environmental Research

Prepared for Dr. Alvin W. Trivelpiece  
Director, Office of Energy Research

by a Subcommittee of the  
Health and Environmental Research Advisory Committee (HERAC)

Dr. Ignacio Tinoco, Jr. (Chairman)  
University of California, Berkeley

Dr. George Cahill  
Howard Hughes Medical Institute

Dr. Charles Cantor  
College of Physicians and Surgeons  
Columbia University

Dr. Thomas Caskey  
Baylor College of Medicine

Dr. Renato Dulbecco  
Salk Institute

Dr. Dean L. Engelhardt  
Enzo Biochemicals, Inc.

Dr. Leroy Hood  
California Institute of Technology

Dr. Leonard S. Lerman  
Genetics Institute

Dr. Mortimer L. Mendelsohn  
Lawrence Livermore National Laboratory

Dr. Robert L. Sinsheimer  
University of California, Santa Cruz

Dr. Temple Smith  
Dana/Farber Cancer Institute  
Harvard University

Dr. Dieter Söll  
Yale University

Dr. Gary Stormo  
University of Colorado

Dr. Raymond L. White  
University of Utah Medical Center

Strategy. A substantial effort directed at technology, mapping and pilot-project sequencing can begin immediately. The committee recognizes that implementation of this initiative by DOE has already begun, and it praises the speed and thrust of the effort. \$11.5 million has been requested for fiscal year 1988; an amount double this would be more appropriate. Funds spent early in this project will save money later, because each advance in technology will make all the following steps more efficient and less costly. Support of \$40 million dollars the first year (fiscal year 1989) and increasing linearly to \$200 million dollars by the fifth year (fiscal year 1993) could be used very effectively. We envision three types of grants -- to individual investigators, to centers with 3 to 10 senior investigators and to a few large centers that will include mapping, sequencing and interpreting the human genome. In addition to the principal investigators, each project will involve junior scientists and engineers, and students. A total of 2500 professional people might be working on the initiative by 1993. The professional personnel will include molecular biologists, chemists, engineers, physicists, computer scientists and so forth.

Recommended funding levels are:

FISCAL YEAR	\$ MILLION	TOTAL
1988	20	20
1989	40	60
1990	80	140
1991	120	260
1992	160	420
1993	200	620
1994	200	820
1995	200	1,020

Reasonable goals to attain by the end of seven years of support at the level requested (by the end of 1995 with \$1 billion spent) are:

- 1) The United States should have the capacity to sequence ten million bases per day.
- 2) The complete map of each chromosome and an essentially complete sequence of at least one human chromosome should be finished.

Attainment of these goals will prove that the U.S. has the capabilities to continue the process to obtain all the benefits promised. We assume that equivalent progress will have been made in computer algorithms to analyze the sequences, and to characterize medically important genes.

---

Technologies Required for Sequencing the Human Genome

1. Production of DNA fragments containing 100 to 1000 kilobases
  - a. Chromosome separation
  - b. Sequence-specific chemical and enzymatic scissors (restriction enzymes)
  - c. Separation and purification of large fragments
  - d. Large-insert cloning
2. Automated DNA handling, mapping and sequencing
  - a. DNA preparation
  - b. DNA cloning
  - c. Physical, restriction fragment, and genetic mapping
  - d. Chemical, physical and enzymatic sequencing
3. Data storage and analysis
  - a. Immediate data entry with uniform notation
  - b. Efficient searching with cross-referencing and access to other data banks
  - c. Rapid data distribution
  - d. Parallel or concurrent processing
  - e. New algorithms for analyzing and interpreting DNA and protein sequences
4. Detection and analysis of DNA, RNA and protein at very low levels
  - a. Single molecule analytical methods
  - b. Methods for detecting large numbers of DNA fragments simultaneously (multiplexing)

---

We estimate that the cost of the development of all of these technologies will be about \$500 million dollars. The total cost will be near \$1 billion and completion of the project will take many years. However, each advance in technology will produce immediate benefits to medicine, agriculture and industry.

TABLE OF CONTENTS

REPORT ON THE HUMAN GENOME INITIATIVE	1
RECOMMENDATIONS	2
REPORT	4
A. Concerted efforts in several different areas should be supported.	6
B. DOE can and should organize and administer this initiative.	9
C. Major advances in diagnosis, prevention and treatment of disease will result.	11
D. The process will produce advances in biotechnology	13
E. Fundamental knowledge in biology will result, and young scientists will be trained to be able to make new discoveries.	14
F. Deleterious effects on other programs must be prevented.	15
Appendix A. Analysis of Costs	16
Appendix B. The Need for Computer Resources: A Data Bank for the Future	20



Multiplex sequencing techniques such as those being developed by George Church are still in their infancy. However, their potential attractiveness is so great that a careful evaluation and refinement of such methods is surely warranted before one embarks on large-scale sequencing. Direct physical approaches to sequence determination such as mass spectrometry or scanning tunneling microscopy are speculative, but their potential impact must not be overlooked. Such approaches should be critically tested in the next few years.

Current strategies for using any of the existing sequencing methods are mostly shotgun approaches which sequence random fragments of DNA. These are quite inefficient since they require sequencing the same region many times over. Sequencing of overlapping fragments is needed to determine the order of the fragments; this is called a bottom-up approach. Phased, or top-down approaches, including systematic ordering and mapping, linked library construction, and optimized production of DNA fragments will all result in far less redundancy in the sequencing. These preliminary steps probably represent half of the final cost and require more than half of the skilled labor. Each of these preliminaries to the actual acquisition of sequence data needs full exploration, refinement and optimization. Most of these preliminaries can and should be automated. Very exciting developments, like methods for cloning or purifying large DNA fragments, and schemes for orderly generation of nested sets of DNA pieces are so new that their potential cannot yet be evaluated. However, it is inevitable that some of these methods will have to be incorporated into any effective large scale sequencing effort.

Once the speed, error rate, and cost are appropriate then one can begin the organized and coordinated effort to sequence a reference human genome. The technologies will then be sufficient to sequence other genomes and to examine human polymorphisms. The wide range of technologies that must be developed for this project are outlined below.

General. The total cost of sequencing the human genome will certainly fall in the billion dollar range, although it is important to stress that the actual cost will be very sensitive to the state-of-the-art technologies associated with DNA sequencing, and the related requirements for automation of procedures for cloning, mapping, data handling and data analysis. As an example, compare the current and projected future costs for DNA sequencing and their corresponding implications for sequencing the human genome.

Estimated Cost for Determining the DNA Sequence of a Human Genome (Given Unique Fragments)\*

SOURCE	COST	GENOME COST
Current commercial laboratories	\$1/base	\$6 billion
Japanese sequencing machines	\$0.17/base	\$1 billion
Future cost with automation	\$0.01/base	\$60 million

\*This estimate does not include the cost of isolating and ordering the fragments; it only includes sequencing each DNA strand, or 6 billion bases. Sequencing both strands provides a check on the accuracy of the sequence.

This table illustrates the importance of making substantial initial investments in technology. We emphasize that the above estimates do not include costs for cloning, mapping or data analysis. Thus our proposal for sequencing the human genome would necessarily be staged. The first 5 years would focus on three general objectives: 1) mapping the human genome, 2) development of technology, 3) sequencing of selected chromosomal regions.

Advances in technology are a necessary first step in sequencing the human genome. These advances will make large-scale sequencing and subsequent comparative studies practical and cost-effective. At present the only automated sequencing machines are based on Sanger's method. There are probably distinct advantages to be gained from automating the Maxam-Gilbert method. A detailed comparison of the two approaches should precede a major investment in one of them. Both approaches can also benefit from considerable optimization. A twofold increase in the length of sequence accessible on a single gel lane would cut the cost of sequencing by considerably more than a factor of two. A number of ways to increase this sequencing range, such as pulsed-field techniques, are very promising and need to be tested.

Advances in biology and medicine have reached the stage where it is now possible to acquire a thorough and very detailed understanding of human biology and inheritance at the molecular level. This understanding will require mapping and sequencing of DNA on a massive scale, a task which cannot be accomplished efficiently with current technologies.

Two major tools are needed:

- a) The sequence of a reference human genome
- b) Efficient methods for obtaining and interpreting the large amount of additional sequence data needed for a wide variety of biological and medical studies

Creation of these tools will require a broad interdisciplinary research effort that brings together technologies from the fields of biology, computing, materials science, instrumentation, robotics, physics and chemistry. This special focus on technological development is distinct from the current national effort in human biology and genetics and requires a new initiative.

The Department of Energy, through the Office of Health and Environmental Research, has a mission to understand the health effects of radiation and of other harmful by-products of energy production. The Department has long supported work on human mutations, DNA damage and DNA repair. Now it is clear that the ability to determine quickly and accurately the sequence of a DNA is the most rapid and cost-effective way to assess DNA damage, and to protect the public health. Thus, the Department of Energy is poised for this initiative because of its research support and interest in human genetics, and its experience in developing large scale, long-term interdisciplinary projects. Development of these new technologies will place the United States in a commanding position in the biotechnology of the 21st century.

## RECOMMENDATIONS

1. DOE should fund a major new initiative whose goal is to provide the methods and tools which will lead to an understanding of the human genome. Funding should start in fiscal year 1989 at \$40 million and increase over a five year period to reach a level of \$200 million per year. Appendix A provides details.
2. The early goals (first 5 to 7 years) of this program should be to:
  - a) Make a physical map of the human genome. A physical map consists of a complete set of segments of the DNA, arranged in order.
  - b) Locate genes and other markers on the map.
  - c) Produce and distribute cloned DNA sequences and other materials needed for using and improving the physical map.
  - d) Develop new techniques and improve existing methods for large-scale DNA mapping and sequencing (including applications of automation and robotics).
  - e) Develop new methods for characterizing and locating genes; both computational and cloning techniques are needed.
  - f) Establish computer facilities, and develop computer data bases for the storage, retrieval and dissemination of cloning, mapping, and sequence information (including cross-references to other relevant data bases). Improve and invent algorithms for analyzing DNA sequences, including methods for identifying coding regions, predicting protein structures and functions, and identifying genetic regulatory sites.
3. The major long-term goal is to obtain a base sequence for each of 24 reference human chromosomes, and to make DNA sequencing technology readily available to search for disease-related variations and to make biological comparisons. The improvements in technology listed in Recommendation 2 are necessary to attain this goal.
4. Work on these goals should take place in the National Laboratories, in universities and in industry. Both prospective and retrospective peer review should be used. Cooperation and collaboration among all groups is essential; in particular, all new map and sequence information must be placed promptly in a designated data base. Clones and cell lines must be made available for distribution to other qualified investigators.

## F. Deleterious effects on other programs must be prevented.

A major new initiative, no matter how worthy, must not disrupt or hinder ongoing worthwhile programs. This initiative deserves the highest priority. However, the most efficient progress will occur if research in all aspects of the relation of DNA to RNA to protein and to health are strongly supported. This requires major increases in funding for the human genome.

It is also important that effort not be shifted from current projects on the genetics of other organisms to study the human genome. The human genome is the emphasis of this initiative; it is not its only component. Everyone must realize the similarity among genes and the utility of transferring knowledge from one organism to another. Furthermore, the DOE initiative will involve people from a wide range of disciplines, including biology, chemistry, engineering, physics and mathematics. There is a large pool of scientists and engineers available.

There is some fear that a large influx of money into a field will distort and disrupt current research. However, there is good precedence that this is not necessarily so. The Howard Hughes Institute increased its biomedical funding from \$3 million in 1975 to more than \$200 million in 1986. There has been a significant beneficial effect.

A large and increasing financial commitment should be made to support this initiative. It should be distributed among the National Laboratories, Universities and Research Institutes; industry contracts may be used when appropriate. Both small science and large science projects should be supported. Peer review should be used for initial funding, and continuing funding should require further review. Flexibility and innovation should be fostered. It is particularly important in this rapidly developing field not to start any large, inflexible organizations whose direction would be hard to change. A large part of the challenge of this initiative is to think of new ideas and to develop relevant technology. A wide range of funding mechanisms will be needed and a wide variety of organizations must be supported.

E. Fundamental knowledge in biology will result, and young scientists will be trained to be able to make new discoveries.

The many practical applications of this initiative have been discussed, but we must stress that the most important result will be new knowledge. We cannot predict what new insights we will obtain, but we are certain to learn completely new patterns of biological organization, structure and control. The discovery of large numbers of currently unknown genes will further our knowledge of all biological processes. The human genome sequence will serve as a reference library that will stimulate and coordinate the next century of biological research. The graduate students and other young investigators who work on this initiative will obtain the background and training to attain the goals of 21st century initiatives. Their exciting research findings should also encourage more entering college students to choose the fields of biological and physical sciences and engineering.

5. Two scientific panels should be established immediately. One would develop policy, define overall strategy, and provide continuing oversight. The other would provide scientific review of proposals and programs for their technical merit and feasibility. The initial phase of the program should consist primarily of technological development in the areas of construction of large scale maps, automation, sequencing and the determination and analysis of sequence data. Because of the highly creative nature of this beginning phase, it is essential that the effort be widely distributed. The project should involve single-investigator-initiated proposals as well as multidisciplinary consortia that bring together the development of instrumentation and software, as well as biotechnology.

6. DOE should encourage wide collaboration at the scientific and managerial levels for the human genome project. Cooperation is needed with other agencies within the U.S. and with other countries throughout the world. Results should be open and in the public domain, within the constraints of technology transfer and the promotion of industrial involvements. Information transfer should be emphasized among the cooperating scientists, the scientific community and the public at large.

## REPORT

### THE ULTIMATE GOAL OF THIS INITIATIVE IS TO UNDERSTAND THE HUMAN GENOME

Knowledge of the human genome is as necessary to the continuing progress of medicine and other health sciences as knowledge of human anatomy has been for the present state of medicine. The DNA of the human genome contains complete instructions for construction of each human being, but we know only the crudest features. We each have two sets of 23 chromosomes with a total of about three billion base pairs per set. Each set consists of 22 autosomes plus one sex chromosome; thus there are 24 distinct chromosomes -- one female (X), one male (Y) and 22 autosomes. The chromosomes contain an unknown number of genes with estimates which range from 20,000 to 200,000. Presently only about 500 of these genes have been cloned and characterized. Our knowledge is equivalent to that of 15th century anatomists who knew about the major bones and organs, but knew very little about their functions. The significance of most vital organs, including obvious ones such as the liver and pancreas, or small ones such as the pituitary and the adrenals, was completely unknown. Most important, even the simplest concerted functions of the body, such as provided by the circulatory system, were not mapped.

We are at the same early state of knowledge with respect to the human genome. We do not know within a factor of ten how many genes there are, nor the range of functions performed by the gene products. We have very limited knowledge of how the expression of genes is controlled. What sequences of the DNA turn genes on and off at the right time for correct development and differentiation? We do not understand how the coordinated control of genes is accomplished. We expect that vital elements that exist in the human genome have not even been imagined. The human genome has been called the book of man; it contains the instructions that describe each human. It is time to obtain a copy of the book to begin to understand what the text means.

It should also be clear that understanding the human genome is a very long-range task. Once the gross features of a human genome are mapped, it will be important to identify and localize all the genes. The control elements must be identified which determine when and where each gene is expressed, and thus program our development from a single cell to a complex structure. The study of single-gene defects in humans has already been extremely beneficial for the diagnosis and treatment of some diseases. Although genes may account for only ten percent of the human genome, complicated chromosomal changes and aberrations, which are not simply dependent on DNA sequences in genes, are also heavily implicated in genetic diseases. Thus, Down's syndrome is caused by an extra copy of chromosome 21, Cri-du-chat is caused by a deletion -- a loss of a segment -- in chromosome 5,

### D. The process will produce advances in biotechnology.

The long-range goal of this initiative is to understand the human genome. This will require improved technology in many other fields. It will automatically further fundamental advances in molecular biology. It will encourage correct theories which relate DNA structure and function, RNA structure and function, and protein structure and function. The ability to organize, manipulate, correlate and retrieve large amounts of data must be improved. Fast and accurate robots that can clone, purify and sequence DNA need to be developed. The advances in all these areas will be applicable to the use of biological materials in industry and agriculture. For example, more efficient production of biomass for energy production should result. Important environmental goals that are of major importance to the Department of Energy will be furthered, such as protection of plants by improving their resistance to environmental stress, and neutralization of toxic wastes by using genetically engineered microbes. Development of the new technologies for this initiative in the fields of biology, chemistry, physics, instrumentation, automation and computing will place the U.S. at the forefront of the biotechnology of the 21st century.

New knowledge about the human genome also means new knowledge about all other genomes. Fundamental knowledge about DNA structure applies to all organisms. Even more directly, sequences of some genes are similar from animals to plants to bacteria. Studies on other organisms, where genetic experiments can be done, will help progress in the human genome. Also maps of other species will greatly increase the validity of applying the results of experiments on other organisms to human health problems. Thus, the human genome project will complement all the other biological research being done on humans and other organisms to increase the rate at which we understand human biology. Now is the appropriate time to begin the direct examination of the human genetic system.

genes following birth is well documented in the development of the body's immune defenses. Abnormal alteration (mutation) of genes is responsible for numerous cancers. Thus the knowledge of the human genome -- the genes, their regulation and their abnormal function -- will have the greatest impact on health maintenance yet experienced in medicine. No individual will be untouched by this initiative.

We cannot afford, nor do we have foundation support for, individual and redundant efforts on the 3500 inherited diseases presently known. Many laboratories are presently working in parallel to obtain DNA fragments and sequences near important human genes. Progress has been made on particular diseases because of foundations dedicated to them, but much of the effort has been redundant. Although the gene for Huntington's disease has been localized to a region on the short arm of chromosome 4 for three years, and the gene for cystic fibrosis has been localized to a small region of chromosome 7 for over a year, overlapping DNA fragments which span these regions are yet to be developed. The high cost of these important studies would be markedly reduced by the development of much faster and comprehensive sequencing and mapping studies. A reference sequence would thus provide rapid, and much more economical, discovery and identification of human disease genes.

The development of rapid, cost-effective methods for sequencing may be the greatest benefit. The more efficient technologies that will be developed for the human genome project will be directly applicable to all sequencing problems. It is appropriate to ask whether we can afford not to develop such improved technology given the level of resources already going into sequencing. We do not know what sequence information is the most valuable. It is likely that the most significant applications to medicine cannot be foreseen at the present time, but the ability to determine DNA sequences routinely will allow immediate application of that knowledge.

and many birth defects and congenital defects have a chromosomal basis. The part of the human genome whose function is not yet known or even imagined must be characterized and understood. Searching analysis will continue to be required to discern differences among human genomes that correlate with sickness and health.

Accomplishing these goals obviously requires sequencing a large fraction of the genome. However, some genomic regions, such as long stretches of repetitive DNA, may not need detailed sequencing. As the details of the genome unfold, it should be possible to set priorities and make rational decisions about what should and should not be done.

A. Concerted efforts in several different areas should be supported.

1) A first step is to map the human genome -- to arrange in order large segments of DNA (in size from 100 to 1,000 kilobases); there are 3,000 to 30,000 of these pieces. As a prerequisite to sequencing the human genome, it is necessary to have pure DNA fragments from known locations on the genome. These DNA fragments constitute an ordered clone bank. At present 30 to 50 kilobase fragments of DNA (cosmid clones) can be prepared routinely and partially ordered; these fragments are vital for current progress. However, methods for preparing and separating larger fragments are becoming available. Large DNA fragments can be formed with restriction enzymes or reagents specific for sequences which are eight or more base pairs long. They can be separated by new methods of electrophoresis. Unique identification of these DNA fragments can be obtained with probes or restriction enzymes; the fragments can be characterized by a complete set of restriction sites with known intervals. Practical methods to determine their order must be worked out.

As human map and sequence data accumulate, many investigators will be able to apply this knowledge to problems of medical and biological importance. They will need access to large numbers of biological samples including cloned DNA fragments and human cell lines. Methods for the efficient production and distribution of these materials need to be developed. Effective quality control for the identity and purity of the samples is essential.

2) Genes should be assigned to the fragments as each fragment is identified. There are standard methods available for locating genes whose gene products (a protein or nucleic acid) are known. These include genes whose defects are responsible for blood diseases such as certain hemophilias, alpha and beta thalassemias and sickle cell anemia. Genes for enzymes with known activities are particularly easy to find. There are essential enzymes whose absence causes death, but the deficiency of other enzymes may only lead to illness, or the predisposition to certain diseases. An enzyme deficiency genetic disease is phenylketonuria which causes mental retardation, but can be treated by removing excess phenylalanine from the diet. A defective anti-trypsin gene produces lungs very susceptible to injury and requires extra care with smoke or other lung irritants. When gene products are not known, as in many human diseases, the process is more difficult. Here the methods which have been successful for Huntington's disease, retinoblastoma, cystic fibrosis, and Duchenne muscular dystrophy can be used. A genetically linked marker for the disease must first be found; this is a sequence of DNA which is located near the disease gene and serves to track the inheritance of the gene. Many genes may only be found from analysis of the DNA sequence; identification will lead to the gene product and its function.

C. Major advances in diagnosis, prevention and treatment of disease will result.

Research to understand the human genome is taking place, so why is it necessary to have a new initiative? The answer is that the results of this initiative are so valuable to humanity that it is essential to proceed as fast as possible. Consider diabetes, for example. One in 300 American children take daily insulin injections by age 18. About half of these will have kidney failure within 30 years. Today about half of all people on kidney dialysis (at a cost of about \$1 billion annually) are diabetics. The disease is genetic, associated with factors on chromosome 6, thus children at risk can be identified. Knowledge of the precise genetic basis of the disease by appropriate sequencing may allow reversal of the autoimmune process which leads to diabetes.

The major killers in this country -- cancer, cardiovascular disease, hypertension and stroke -- all have significant genetic components. The ability to respond to these diseases before they strike will save lives. The immune system controls the body's intrinsic defenses and is responsible for autoimmune diseases and other degenerative diseases such as arthritis. Analysis of the genes of the immune system will allow effective stimulation of the defenses and appropriate therapy for the diseases. Detailed sequence information will lead to methods for more exact matching of donor and recipient in transplantations. Monitoring changes in the DNA sequence of one tissue in one person will reveal damage caused by environmental factors. Many more examples could be given. However, the analogy of knowing human anatomy and knowing the human genome is apt. We could not cure heart disease as soon as we understood blood circulation, but it was a necessary first step. It is also well to emphasize that we do not need to recognize and order all genes for success. Each new fragment of DNA sequence can bring human benefits.

It is now practical to locate genes, to sequence DNA, to supplement some of the deficiencies caused by missing or defective genes. A major effort will bring immediate and continuing benefits. Each new gene identified and mapped will allow certain diagnosis of any diseases associated with this gene. Recent examples include Duchenne muscular dystrophy, chronic granulomatous disease, cystic fibrosis, Alzheimer's disease and Huntington's disease. The identification of genetic risk factors for common diseases such as diabetes and premature coronary disease are further examples where genetic map information could lead to methods of risk modification for an entire population. The recent identification of genes which lead to abnormal development emphasizes a relatively unexplored health problem -- birth defects. Alteration of

Americans receive exposure to various mutagens, including mutagens from energy sources such as the combustion of fossil fuels. The exposure levels of paramount importance to society are low, and there is enormous individual heterogeneity in susceptibility to exposure. Rapid and cost-effective methods are needed to assess exposures and risks to large numbers of people. The definitive measure of mutation is the sequence of DNA. The ability to determine quickly and accurately the sequence of any DNA is the ultimate way to assess immediate and cumulative damage by many agents. Thus DOE has unique capabilities to manage this initiative, and the initiative is central to its mission.

Other Federal and private agencies have a major interest in this initiative. The National Institutes of Health, in particular the National Cancer Institute and the National Institute of General Medical Sciences, are already heavily committed to support research on DNA sequence and function. This support deserves to be increased. The Howard Hughes Medical Institute supports an increasing number of projects on human genetic diseases. Important work is also being done in Europe and Japan. It has become clear to everyone that the tools to map and to sequence the human genome can now be developed; what will be accomplished depends on the effort and commitment.

DOE should develop general methods and provide tools useful to all the other molecular biology projects. Instrumentation, automation, computation and other multidisciplinary approaches should be emphasized. DOE should foster cooperation among all the organizations involved, both national and international. However, it should not delay implementation of its plans or defer to some other organization. Thorough communication should ensure that there is no duplication of facilities and waste of resources. We strongly encourage continuing cooperation among the various agencies.

Genes for all the enzymes involved in metabolism, in biosynthesis and in repair need to be localized. Structural proteins, proteins of the immune response, transport proteins and the RNAs of protein synthesis are all important. The genes for hormones, which act in very small amounts, need to be identified and entered in the genome map. The largely unknown control proteins, which orchestrate differentiation, development and senescence, may be the most important to characterize. As more genes are identified and their gene products determined, the polygenic disorders like heart disease, hypertension, diabetes, schizophrenia, manic depression and even some symptoms of aging can be attacked. It will become possible to develop methods for early diagnosis and effective treatment.

There are currently many projects, sponsored primarily by the National Institutes of Health and the Howard Hughes Medical Institute, involved in locating genes on the human chromosomes. This research is extremely valuable and is primarily aimed toward medically important genes. The DOE initiative will facilitate rather than compete with those projects; it will provide a valuable resource for the projects. Even so, the success of all those other projects would locate only a few percent of the total number of human genes. The tools and methods developed through this project will greatly speed the finding and understanding of the total complement of human genes, a task far beyond the scope of any current research efforts.

3) Current methods for mapping and determining base sequences in DNA need to be increased in speed by orders of magnitude and radical new methods should be encouraged. Automation of current methods has begun. There is a Japanese national project which is trying to develop automated equipment based on current sequencing methods to determine sequences at the rate of 300 kilobases to one million bases per day. Even if the Japanese effort is successful, a thorough sequencing of a human genome will not be possible because methods for preparing, purifying and ordering DNA fragments are not available to provide the necessary fragments for sequencing. However, advances in technology are being developed which will allow complete and cost-effective sequencing. These new methods need to be automated to provide a reference sequence and also to provide the ability to make comparative studies both within the human population and between humans and other animals. Close collaboration between engineers and molecular biologists can provide efficient, reliable methods that use the capabilities of automated instrumentation to fullest advantage. It should be possible to reduce the total cost of sequence determination to one tenth, one hundredth, or less of the cost of current manual methods. Appendix A gives details.

4) Computer facilities to organize, disseminate and interpret the sequence of the human genome must be supported. At present there are several organizations which act as repositories for sequence data and human gene information (such as Genbank at Los Alamos, the National Library of Medicine, the Yale human gene map, the European Molecular Biology Organization, the Japanese National Institute of Genetics).



Easy cross-reference and cross-access between data bases must be assured. Algorithms and programs to interpret the data are in very early stages of development. We need programs to identify accurately DNA sequences corresponding to genes and their control. At present we cannot identify unambiguously the signals to start messenger RNA synthesis, to start protein synthesis, to remove introns and thus to provide a protein sequence. When a DNA sequence predicts a protein sequence, we need to be able to predict the protein shape, its function and possible cellular or extracellular sites for its location. Algorithms to identify the control elements for expression and regulation of the genes (enhancers, repressors, etc.) are needed. DNA sequences involved in chromosome organization, recognition and regulation must be understood. Appendix B provides further details.

B. DOE can and should organize and administer this initiative.

The Department of Energy, extending back to its predecessor the Atomic Energy Commission, has successfully managed many long-term and complex technological programs. DOE has a history of coordinating such projects through contracts with industries, universities and its own laboratories. The size, interdisciplinary nature and long-term scale of the human genome project, with the many technologies involved, fits these experiences of DOE well. In addition, within DOE the mission of the Office of Health and Environmental Research (OHER) is to understand the health effects of radiation and other by-products of energy production. This requires fundamental knowledge of the effects of chemical and physical damage to the human genome.

The OHER mission in human genetics has led to the initiation and support of a number of research and technological developments which are closely linked to the human genome mapping and sequencing project. These include basic research on radiation and chemically-induced damage of DNA and on the repair of DNA damage. Risk analyses of the effects of the deleterious agents on cancer and genetic diseases have also been done. DOE-supported studies in genomic mapping, chromosome isolation, and sequence data management and analysis are even more directly related. Thus, this initiative is a natural outgrowth of current DOE-supported research. Furthermore, the initiative will make important contributions to other DOE missions, including environmental waste control, improving energy production, producing and utilizing biomass, and so forth.

The National Laboratories can be an important resource for the genome project. They are currently furthering the goals of the project by providing sorted chromosomes, genetic probes and clone libraries. Genbank at Los Alamos is presently supported by NIH, DOE and other agencies as a computer facility for organizing and disseminating DNA sequence information. The National Laboratories are experienced in providing technical and engineering support for large projects, and for efficient development of technological tools. The completion of a physical map of the human genome, the organization of associated clone libraries and the production of a reference sequence produce a tool. This tool can be the most powerful technological resource available for the understanding of biology and medicine.

The Office of Health and Environmental Research seeks a fundamental understanding of the health effects of radiation and of energy-related chemical toxicants, so as to apply its findings to the protection and improvement of human health. The complete sequence of a human genome provides a reference base against which perturbations induced by the environment will be recognized and measured. A long-term interest has been the monitoring of somatic cell and germ cell damage caused by radiation and by other toxic agents such as chemical mutagens.

## Support for Research Related to the Human Genome

Please provide FY 1986 and FY 1987 budget information for support in the following categories:

<u>Category</u>	<u>Selected Examples</u>
I. Focused Research	Major projects directed toward mapping or sequencing an entire chromosome or genome
A. Human	
B. Non-Human	
II. Basic Research	Fundamental studies directed toward mapping a specific gene and its control elements, or sequencing a DNA fragment of equivalent size
A. Human	
B. Non-Human	
III. Research Infrastructure	
A. Biological Methods Development	Cloning techniques, identification of new restriction enzymes, or DNA separation methodology
B. Instrumentation Development	Automated DNA purification, Southern blot apparatus, or new or improved chromosome or DNA sorting methods such as electron tunneling or mass spectroscopy
C. Informatics	New algorithms for database search, retrieval, and analysis; methods for protein structure or function prediction; or identification of genetic control elements
D. Biological Databases	Databases such as GenBank®, the Protein Identification Resource, or BIONET
E. Biomaterials Resources	Stock and strain collections such as the ATCC Probe and Gene Library Repositories, the Human Mutant Cell Repository, and the Aging Cell Repository

Current and Future Funding Levels

- I. Please provide aggregate budget information on levels of support in FY 1986 and FY 1987 using the following support mechanisms:
  - A. Direct laboratory support (Intramural)
  - B. Grants
  - C. Contracts
  
- II. Please provide aggregate funding level estimates for FY 1988 and FY 1989 in the following categories:
  - A. Budgeted activities
  - B. Activities that may be undertaken if new funds become available

Domestic Policy Council  
Working Group on Biotechnology  
Subcommittee on the Human Genome

May 8, 1987

Minutes

Attendees:

Dr. Wyngaarden (Chairman)  
Dr. Danello  
Ms. Levinson (Exec. Sec.)  
Dr. Kingsbury

Dr. Kirschstein  
Dr. Noonan  
Dr. Smith  
Dr. Wooley

The first meeting of the Domestic Policy Council, Working Group on Biotechnology, Subcommittee on the Human Genome, opened with a decision to name the group as aforesaid. Dr. Wyngaarden, Subcommittee Chairman, described the function of the subcommittee as that of "coordination" of the activities of multiple federal agencies with an involvement in genetic mapping and sequencing. He noted that although the central planning function will remain in each individual agency, the subcommittee will provide a forum in which plans can be constructed based on up-to-date information contributed by each member agency.

The purpose of this first meeting was to discuss: 1) agency definitions of "mapping and sequencing," and 2) information concerning related agency activities to be presented to the Domestic Policy Council. Questions developed by Dr. Kirschstein were used to construct a request for information that will be sent to member agencies as well as the following federal agencies: the National Bureau of Standards; Veterans' Administration; Centers for Disease Control; Alcohol, Drug Abuse, and Mental Health Administration; and Department of Defense.

Members of the subcommittee agreed to respond to this request quickly so that Dr. Kingsbury can present preliminary figures at the May 28 meeting of the Domestic Policy Council.

Dr. Kingsbury will transmit the same request to several research-supporting foundations and institutions including the American Cancer Society, March of Dimes, and Howard Hughes Medical Institute. Dr. Noonan, representing the Office of Management and Budget, approved NSF distribution of the request for information. (See attached)

(Subsequent to this meeting, an earlier decision by the Environmental Protection Agency not to participate in subcommittee activities was reversed per a telephone conversation between Ms. Levinson and Dr. Jack Fowle, who will serve as EPA's representative.)

There was a brief discussion regarding the announced intent of a private party to copyright sequence data and a possible response by the subcommittee.

## Support for Research Related to the Human Genome

Please provide FY 1986 and FY 1987 budget information for support in the following categories:

<u>Category</u>	<u>Selected Examples</u>
I. Focused Research	Major projects directed toward mapping or sequencing an entire chromosome or genome
A. Human	
B. Non-Human	
II. Basic Research	Fundamental studies directed toward mapping a specific gene and its control elements, or sequencing a DNA fragment of equivalent size
A. Human	
B. Non-Human	
III. Research Infrastructure	
A. Biological Methods Development	Cloning techniques, identification of new restriction enzymes, or DNA separation methodology
B. Instrumentation Development	Automated DNA purification, Southern blot apparatus, or new or improved chromosome or DNA sorting methods such as electron tunneling or mass spectroscopy
C. Informatics	New algorithms for database search, retrieval, and analysis; methods for protein structure or function prediction; or identification of genetic control elements
D. Biological Databases	Databases such as GenBank®, the Protein Identification Resource, or BIONET
E. Biomaterials Resources	Stock and strain collections such as the ATCC Probe and Gene Library Repositories, the Human Mutant Cell Repository, and the Aging Cell Repository

Current and Future Funding Levels

- I. Please provide aggregate budget information on levels of support in FY 1986 and FY 1987 using the following support mechanisms:
  - A. Direct laboratory support (Intramural)
  - B. Grants
  - C. Contracts
  
- II. Please provide aggregate funding level estimates for FY 1988 and FY 1989 in the following categories:
  - A. Budgeted activities
  - B. Activities that may be undertaken if new funds become available

Responses should be addressed to:

James B. Wyngaarden, M.D.  
Director, National Institutes of Health  
Shannon Building, Room 124  
9000 Rockville Pike  
Bethesda, MD 20892

If you or your staff have any questions pertaining to this request, they may be referred to Ms. Rachel E. Levinson. Ms. Levinson may be reached by telephone on (301) 496-1454.

Thank you for your attention to this matter.

Sincerely yours,

David T. Kingsbury, Ph.D.  
Associate Director for  
Biological, Behavioral, and  
Social Sciences  
National Science Foundation

Enclosure

## National Institutes of Health

### Activities Related to Mapping and Sequencing the Human Genome

#### I. Focused Research

Nearly every National Institutes of Health (NIH) institute and division is involved in research directed toward the analysis of complex genomes. At this time, however, there are no major projects directed toward mapping or sequencing an entire chromosome or genome. Rather, the bulk of the research effort, consisting of roughly 3,000 projects, falls into the basic research category.

#### II. Basic Research

NIH support of research relevant to genome mapping and sequencing ranges from viral and bacterial genetics through human genetics. Examples of the types of relevant research projects supported by NIH are described in Attachment I.

#### III. Research Infrastructure

##### A. Biological Methods Development

##### B. Instrumentation Development

Because it was not practical to identify individual projects aimed at biological methods development or instrumentation development, estimates of the funds provided for these subcategories were based on a representative sample drawn from the larger basic research category and added to support for known discrete projects.

##### C. Informatics

The NIH has a long-standing commitment to cataloguing biological information in order to maximize its usefulness to the public as well as to the biomedical community. The recent interest in manipulating molecular biology data has provided a new focus for informatics research. Specific projects under way in the NIH intramural program include an initiative developed by the National Library of Medicine to increase ease of access to and interaction with several molecular biology databases.

##### D. Biological Databases

##### E. Biomaterials Resources

NIH has established and continues to fund a number of biological databases and biomaterials resources that support mapping and sequencing efforts including GenBank<sup>®</sup>, BIONET, the Hybridoma



Data Bank, the Protein Identification Resource, the ATCC Cell/Tumor Bank, the Human Genetic Mutant Cell Repository, the Human DNA Probe Repository, and the Human Gene Library. The NIH National Biomaterial Resource Committee has been established in order to coordinate the NIH approach to management and funding of these resources.

F. Additional Related Activities

Other activities include NIH sponsorship of workshops directed toward promoting progress in genomic analysis. For example, in February 1987, the National Institute of General Medical Sciences and the European Molecular Biology Laboratory cosponsored a workshop addressing the future database needs of researchers in the life sciences. On May 18, 1987, the National Institute of Child Health and Human Development convened a number of NICHD investigators working on mapping chromosome 21 together with experts in various aspects of gene mapping to discuss what the Institute might do to foster and coordinate this effort. The purpose of this workshop was not to construct a map of the entire chromosome, but to share recent research methods and results pertaining to particular loci of interest to the Institute. Financial support for these activities was included in the category best describing the purpose of the workshop.

Future Research Plans

Last October, the Advisory Committee to the Director, NIH, was asked to weigh the relevant scientific and policy considerations attendant to mapping and sequencing the human genome. A major concern raised at that meeting was the need for improved information handling techniques. Additional recommendations of the Advisory Committee concerning the science required to promote progress in mapping and sequencing addressed the following needs:

- Better techniques for cloning large DNA fragments;
- Improved chromosome sorting methods; and
- Need for cosmid libraries, more genetic markers and highly specific restriction enzymes.

Following the meeting of the Advisory Committee, an NIH working group was established to examine recommendations put forward at the Advisory Committee meeting. Working group members are: Dr. James B. Wyngaarden, Director, NIH, working group chairman; Dr. Ruth L. Kirschstein, Director, National Institute of General Medical Sciences; Dr. Donald A.B. Lindberg, Director, National Library of Medicine; Dr. Betty Pickett, Director, Division of Research Resources; Dr. Duane F. Alexander, Director, National Institute of Child Health and Human Development; Dr. Jay Moskowitz, Associate Director for Program Planning and Evaluation; Dr. George E. Palade, Yale University School of Medicine, Advisory Committee Member; and Ms. Rachel E. Levinson, Office of Program Planning and Evaluation, Executive Secretary.

The working group has developed two program announcements intended to convey to the scientific community the particular interest at NIH in research related to the analysis of complex genomes. These announcements were published in the May 29, 1987 issue of the NIH Guide For Grants and Contracts.

Program Announcement 1.

New Approaches to the Analysis of Complex Genomes

Examples include but are not limited to:

- Advanced techniques for obtaining pure preparations of individual chromosomes and their fragments;
- Improved methods for cloning large fragments of DNA; and
- Enhanced DNA sequencing technologies, including automation of DNA isolation and purification.

Program Announcement 2.

Computer-based Representation and Analysis of Molecular Biology Data

Directed toward improving computerized manipulation of the data generated by molecular biologists via:

- Improved database design for searching, analyzing, transmitting, and storing biological information;
- Development of software algorithms for predicting structure and/or function based on primary sequences of nucleotides and amino acids;
- Specialized computer hardware for economical and rapid comparisons of large volumes of biological data; and
- Establishment of expert system techniques for automatic generation of annotation information and creation of linkages among related databases via explicit pointers or a common vocabulary.

Applications in response to these announcements will be reviewed in accordance with the usual NIH peer review procedures. Publication will include contacts representing each BID. Although no central pool of funds has been set aside to support these initiatives, individual institutes have established similar program plans and their funding decisions will be made accordingly.

The final budget category contains projected costs for those activities that would be undertaken if new funds became available. These activities include both expansions of existing programs and new initiatives, examples of which are described in Attachment I.

Selected Examples

Current Activities

- The following are examples of research under way that involves either chromosome mapping or DNA sequencing:
  - Past research on Huntington's disease (HD) concentrated on the search for the causative gene in order to predict whether an individual was affected, prior to bearing children who might also be at risk for this lethal neurodegenerative disease. Recently, the HD gene was localized to chromosome 4, and markers were determined that make it possible to identify disease carriers and affected individuals.
  - The NIH supports the National Flow Cytometry and Sorting Resource at Los Alamos National Laboratory. In order to fill the large number of requests for individual chromosomes from investigators all over the country, DNA libraries of flow sorted chromosomes were constructed to increase the supply of the genetic material. Many of these scientists are searching for genetic probes from specific chromosomes, while others have requested chromosomes from a particular cell strain that they are studying to determine sites of viral integration or genetic rearrangement.

Planned New or Expanded Initiatives

- Plans for Fiscal Year 1987 indicate modest expansion of these programs and initiation of new projects including the following:
  - The NIH is developing an experimental online computer system for simultaneous access to several databases containing information on published DNA sequences, the availability of cDNA probes and markers that can be used to locate genes, and other resources. The system is being tested on molecular biologists in the NIH intramural program to make it efficient and easy to use.
  - Since the development of the "super mouse" in 1983, the transgenic mouse model has been applied to a wide variety of studies including normal gene expression, correction of genetic defects through the introduction of normal genes, and genetic control of immune response. The applications of this animal model system to different disease problems are expected to increase rapidly as more laboratories become expert in this new technology.

### Future Opportunities

- If additional new funds become available, several important areas of opportunity will be explored. These include:
  - Recombinant DNA linkage studies of retinoblastoma and inherited retinal degenerative disorders, such as retinitis pigmentosa, have resulted in great progress in defining the genetic loci indicating a predisposition to these conditions. Additional funding would assist in isolation of the genes involved and the determination of the DNA sequences within these genes that may lead to a more defined understanding of the mechanism of the disease process.
  - The NIH currently supports efforts to identify, map, and sequence genes responsible for the etiology of Alzheimer's disease. This activity, which is proceeding at a modest level, could be stimulated with additional funds.
  - The gld gene causes the autoimmune disorder, systemic lupus erythematosus (SLE), in mice. This gene is on a part of the mouse chromosome 1 which corresponds to the long arm of the human chromosome 1. With additional support, the mouse gld gene will be compared with the homologous human gene in normal people and patients with SLE.

National Institutes of Health

Agency Support for Research Related to the Human Genome  
(Dollars in thousands)

Fiscal Year 1986

I. Focused Research

NA

II. Basic Research

Fiscal Year 1986

A. Human	\$91,123	B. Non-Human	\$202,761
----------	----------	--------------	-----------

III. Research Infrastructure

A. Biological Methods Development	\$ 29,388
-----------------------------------	-----------

B. Instrumentation Development	200
--------------------------------	-----

C. Informatics	511
----------------	-----

D. Biological Databases	1,500
-------------------------	-------

E. Biomaterials Resources	9,500
---------------------------	-------

National Institutes of Health

Agency Support for Research Related to the Human Genome  
(Dollars in thousands)

Fiscal Year 1987 (est.)

I.	Focused Research		
	NA		
II.	Basic Research		
	A. Human	\$97,000	B. Non-Human \$215,800
III.	Research Infrastructure		
	A. Biological Methods Development		\$31,280
	B. Instrumentation Development		240
	C. Informatics		600
	D. Biological Databases		1,500
	E. Biomaterials Resources		9,500

National Institutes of Health

Agency Support for Basic Research Related to the Human Genome\*

Current and Future Funding Levels  
(Dollars in thousands)

I. Current Funding Levels by Mechanism

Fiscal Year 1986

A. Direct Laboratory Support (Intramural)	\$ 43,716
B. Grants	180,852
C. Contracts	22,041
D. Training	47,275
Total	293,884

Fiscal Year 1987 (est.)

A. Direct Laboratory Support (Intramural)	\$ 46,607
B. Grants	194,120
C. Contracts	24,711
D. Training	47,362
Total	312,800

II. Future Funding Level Estimates

A. Budgeted Activities

Fiscal Year 1988 \$336,000

Fiscal Year 1989

B. Activities that may be undertaken if new funds become available

Fiscal Year 1988 \$ 45,000

Fiscal Year 1989

\*Research Infrastructure not included

**Organizations to be contacted regarding activities related to mapping and sequencing:**

**Howard Hughes Medical Institute  
Dr. George Cahill**

**American Cancer Society**

**March of Dimes**

**AIRI Association of Independent Research Institutes?**

**Hereditary Disease Foundation  
Dr. Nancy Wexler**

**Cystic Fibrosis Foundation**

**Committee on Huntington's Disease**

**National Huntington's Disease Association**

**Association of Biotechnology Companies  
Dr. Bruce Mackler**

**Industrial Biotechnology Association  
Dr. Alan Goldhammer, Richard Godown**

**Office of Technology Assessment  
Dr. Patricia Hoben**

**National Neurofibromatosis Foundation**

**Retinitis Pigmentosa Foundation**

**National Tay-Sachs and Allied Diseases Association**

**National Association for Sickle Cell Disease**

**National Genetics Foundation**

**National Association for Down Syndrome**

**Myasthenia Gravis Foundation**

**National Hemophilia Foundation**

**Muscular Dystrophy Association**

**Alzheimer's Disease and Related Disorders Association**



Dear :

As you know, recent technological advances in molecular biology have brought about the possibility of embarking on a major effort to map and sequence the entire human genome. As a result, discussions as to the wisdom and feasibility of such an initiative are taking place among the Federal agencies with responsibility for funding and conducting research related to the analysis of complex genomes. These agencies are aware that representatives of the academic research community and the private sector are important participants in such activities, as are the non-profit organizations that support research and educational functions relevant to genetic diseases.

Recently, the United States Domestic Policy Council Working Group on Biotechnology established a Subcommittee on the Human Genome. The Domestic Policy Council is a cabinet-level body charged with advising the President on such matters as biotechnology and economic competitiveness. This interagency Subcommittee, chaired by the Director of the National Institutes of Health, has undertaken a review of current and future support for research related to the mapping and DNA sequencing of complex genomes. Because of the important role of many institutions outside of the Federal Government, we feel that it is critical that this review include information on pertinent research supported by these organizations. Therefore, we would appreciate your participation in this effort by completing the enclosed questionnaire.

**Workshop Report**

**Repository, Data Management, and Quality Assurance Needs  
for the National Gene Library and Genome Ordering Projects**

**26 - 27 August 1987**

**Sponsored by:**

**Office of Health and Environmental Research, U.S. Department of Energy**

**and the**

**National Institutes of Health**

**Workshop Report**

**Repository, Data Management, and Quality Assurance Needs  
for the National Gene Library and Genome Ordering Projects**

**26 - 27 August 1987**

Sponsored by:

Office of Health and Environmental Research, U.S. Department of Energy

and the

National Institutes of Health

## **SUMMARY**

This report is an overview of a meeting on Repository, Data Management, and Quality Assurance Needs for the National Gene Library and Genome Ordering Projects. The meeting was cosponsored by the U.S. Department of Energy and the National Institutes of Health. It was held in Pleasanton, California and hosted by the Biomedical Sciences Division of Lawrence Livermore National Laboratory. Dr. Anthony V. Carrano was the organizer. The meeting consisted of four sessions over 1½ days on 26 - 27 August 1987. The agenda is enclosed as appendix 1. The 27 attendees included representatives from the Department of Energy, National Institutes of Health, Howard Hughes Medical Institute, Office of Technology Assessment, as well as from university and government laboratories. A list of participants is given in appendix 2. The attached report summarizes the discussions and lists the recommendations agreed upon by the majority of those present. In most cases, the agreement was unanimous. A set of cassette tapes are available that cover the last three sessions of the meeting. Unfortunately, background noise reduces their quality. The meeting served as a mechanism to bring the government agencies to date on the status of three DOE-funded efforts on human genome ordering, the National Gene Library Project and the status of the NIH-funded repository at ATCC. The discussions highlighted many of the issues unique to the gene library and genome ordering efforts and provided specific recommendations for future funding.

## **THE NATIONAL GENE LIBRARY PROJECT**

Drs. Larry Deaven (Los Alamos National Laboratory) and Marv van Dilla (Lawrence Livermore National Laboratory) described the current status of the National Gene Library Project. The construction of a complementary set of chromosome-specific small insert libraries in Charon 21A has been completed. Both the Eco RI and Hind III libraries have been deposited in the repository at the American Type Culture Collection (ATCC, Rockville, MD). Both laboratories have concentrated their efforts on developing the methodology to establish large insert libraries for each of the human chromosomes. The plan for this Phase of the project is to construct both lambda and cosmid partial digest libraries for the sorted chromosomes. Charon 40 has been proposed as the lambda vector. Livermore is exploring the use of two cosmid vectors, c2RB and the Lorist series developed by Peter Little (United Kingdom). Los Alamos is exploring the pCos series developed by Hans Lehrach (United Kingdom) and a new vector developed by Glen Evans (Salk Institute).

Molecular weight of the DNA from sorted chromosome and efficiency of cloning into cosmids are the main issues to be resolved. High molecular weight DNA seems to be attainable from chromosomes isolated in a polyamine buffer. The need to do partial digests means that more chromosomes have to be sorted since some of the extra DNA will be used to optimize the digests. Cosmid vectors with double cos sites seem to be the most useful since they eliminate the need to size select the DNA prior to ligation and cloning. Initially, each laboratory (Livermore and Los Alamos) will clone half the human genome. It was estimated that it will take three years to complete this phase of the project.

Both laboratories expressed concern about the lack of cooperation from some sources in obtaining hybrid cell lines for use in the library project. A more serious issue has been the lack of feedback from the user community of the Phase I libraries. Despite requests that are made either by phone, letter, or both, the percentage of users responding has been very low. This might indicate that many of the "users" are collectors. Specific recommendations that would benefit the National Gene Library Project were discussed.

### **RECOMMENDATIONS**

- 1. Funding should be made available to establish a set of stable monochromosomal hybrids for each of the human chromosomes.**
- 2. A repository should be established for such hybrids.**
- 3. A data base should be established for both the libraries in the repository and those still maintained by the originator.**

## **PHYSICAL MAPPING**

The status of the Office of Health and Environmental Research/U.S. Department of Energy (OHER/DOE) physical mapping efforts funded at Columbia University, Los Alamos and Lawrence Livermore National Laboratories were described by Drs. Cassandra Smith, Ed Hildebrand, and Tony Carrano, respectively. The goal of the Columbia effort was to make restriction maps of several genomes using restriction enzymes that cut DNA infrequently (e.g. Not I, Sfi I). Pulsed-field electrophoresis is used to separate the fragments and each fragment is visualized by Southern blotting using unique or repeat sequence probes. The group has essentially completed the restriction map of *E. coli* and *S. pombe*. Work has begun on human chromosome 21. A major limitation for the human chromosome mapping is the lack of or inability to obtain useful chromosome-specific probes to link the large fragments.

The Los Alamos effort focuses on human chromosome 16. Ed Hildebrand discussed the ability to obtain high molecular weight DNA from sorted chromosome 16 and the discrimination of repetitive DNA elements from this chromosome on a transverse-field electrophoresis system. The group plans to pursue the electrophoretic separation of large fragments and to model, and perhaps test, the hybridization approach for cosmid overlap detection suggested by Hans Lehrach. Tony Carrano presented the Livermore effort which focuses on human chromosome 19. Their program has four major research components: computations and modeling, vector construction and cloning, cosmid ordering, and large fragment mapping. The modeling efforts have pointed out the inefficiencies of the lambda or cosmid ordering schemes employed by Maynard Olson's group for yeast and Sidney Brenner's group for the nematode. This might be accomplished more efficiently by generating a large (approx. 70 - 150) number of restriction fragments for each insert. The group has developed a high-resolution fluorescence-based restriction fragment analysis system to discriminate fragments from a four-cutter digest. The approach uses a new chemistry and the Applied Biosystems DNA Sequencer. The computations team is developing new software to facilitate both fragment identification and overlap analysis. New vectors have been constructed to optimize cosmid cloning and to create Not I linking probes for large fragment analysis. Approximately 25 unique sequence probes have been collected for chromosome 19.

There was considerable discussion both during and after the presentations both on points of clarification and to express support for the programs presented.

### **RECOMMENDATIONS**

1. Establish a repository of chromosome-specific DNA probes for mapping the large fragments by transverse-field electrophoresis.
2. Establish sets of chromosome-specific linking probes (e.g. Not I or Sfi I-based) for linking the large fragments together.

## **PHYSICAL REPOSITORY**

The status of the library and probe repository at ATCC was presented by Dr. Bill Nierman. The repository is in its second contract year of a five year contract. The repository receives material, verifies its authenticity, and sends material out to qualified investigators. Lag time from receipt of material to distribution is about 70 days and most of this time is spent in verification. All Phase I libraries are now in the repository. More than 630 library aliquots have been mailed to users. Probes have been solicited from the originators. The response has been slow and inadequate. Of 422 probes solicited, only 118 have been sent to the repository (approx. a 30% response). The presumed reasons for the lack of response include fear of competition and the proprietary nature of industry originated probes. Of the probes received, about 20% were found not to verify with the author's description. The ATCC cost per clone is about \$1500 for the five-year period of the contract. The majority of the cost is associated with verification procedures.

The user community for both the libraries and the probes include universities (most users), government laboratories, and industry (fewest users). The only other repository that exists worldwide is a small one operated in Japan by the Ministry of Health. Two problems have arisen with the user community. Laboratories have reported inconsistent titers with the Phase I libraries. It appears that this is due to the lack of a standard protocol to determine titers. Also many recipients have stored their libraries at -20°C which is known to cause loss in the titer. To minimize titer problems, the ATCC has been mailing a storage and titer protocol with each library.

### **RECOMMENDATIONS**

- 1. Examine ways to automate the verification procedures for received material to minimize the associated costs.**
- 2. Expand the repository to include the physical output (i.e. configs, clones, and probes) that will be derived from the chromosome ordering efforts.**

## **RESEARCH NEEDS GENERATED BY THE LIBRARY AND ORDERING PROJECTS**

Following the above presentations, an open discussion session was chaired by Dr. Cantor. The discussion was lively and constructive. It focused on sources of DNA and repository requirements. The discussion is best summarized with a list of recommendations.

### **RECOMMENDATIONS**

- 1. A mechanism should be established to insure that current recipients of government funding make any hybrid cell lines and probes available that are derived from that funding.**
- 2. Journals should take a stronger position in accepting publications on probes and libraries. Acceptance should be contingent upon deposition of the reported materials in a repository.**
- 3. Any new hybrid cell lines established should have the human donor material available in another form, e.g. as fibroblast or lymphoblast cultures. If at all possible, do not use hybrids for which normal cells or DNA from the donor is not available.**
- 4. Monochromosomal hybrid cell lines should be established from family kindreds, even if only two generations (i.e. parents and children) are available. Perhaps the "Utah" kindreds would be an appropriate source of material.**
- 5. Avoid the use of lymphoblast cell lines for Phase II libraries, if possible. Rearrangements in these lines are thought to occur at high frequency.**
- 6. Establish a repository and data base for all hybrid cell lines.**
- 7. Only contigs above a minimum size should be stored in the ordering repository. No consensus was reached on minimum size requirements but suggestions ranged from 100 to 300 kb. Only a 1X contig coverage should be stored..**
- 8. Any ordered DNA segments to be placed in the repository should be coupled to the genetic map, i.e. have either an RFLP or a gene assigned to it.**
- 9. Southern blots of large fragments from transverse field electrophoresis separations should be mass produced and stored in a repository for other users. This could be a commercial venture.**
- 10. Dot blots of each contig could also be mass produced and stored in a repository. This should decrease the demand on the repository to store and supply the original clones.**
- 11. Establish a set of normalized human cDNA libraries. These would be especially useful for map closure. Consensus was not reached on tissue of origin of the cDNAs.**
- 12. A contract should be considered to support a pilot project for the repository and data base for the ordering efforts. The ordered library of E. coli established by Kohara might serve as test material. The goals would be to develop costs and clone management expertise as a prelude to the creation of a larger repository.**



## **ACCURACY AND QUALITY ASSURANCE**

A general discussion on quality control was chaired by Dr. Keith McKenney. Quality control on the National Gene Library project is restricted both by available funding and by the decision of the Advisory Boards to this project. At present quality control is performed in a number of ways. Cell lines used for sorting are karyotyped by banding to determine whether the banding patterns are normal. A flow karyotype is also performed to determine whether there are any major shifts in DNA content. Both of these methods only insure against large (megabase) rearrangements or deletions. Sorting purity is determined by a mathematical analysis of the histograms from the sorted regions. Further analysis is performed on the constructed libraries by the user community. Probes are picked from the libraries and verified as to chromosome of origin by Southern hybridization against DNA from appropriate reduced hybrids or by *in situ* hybridization.

For the physical repository at ATCC, quality assurance is performed in several steps. Transformation of DNA into host cells is completed, if necessary. A one liter plasmid/phage preparation is established to dispense the DNA for storage and distribution. A restriction fragment analysis is then completed and the depositor is allowed 60 days to verify the analysis. After this time, the material is available for distribution. A sample product sheet distributed to users is enclosed as appendix 3.

Quality assurance for ordered material is a new concept. One suggestion for quality assurance was to develop independently a cosmid as well as a large fragment restriction map. If the ordering were done correctly, the two maps should be consistent. This could be done by placing the ordered cosmid contigs on a previously mapped set of large fragments. It was strongly felt that there should be no formalized quality control program for the ordering efforts. Each laboratory should be responsible for their own quality control.

### **RECOMMENDATIONS**

- 1. Additional funding should be provided for increased quality control of the Phase II effort of the National Gene Library Project. Most of this effort would be required in characterization of the libraries to insure purity.**
- 2. Journals should play a major role in assuring quality control by increasing the stringency of their acceptance criteria for probes and libraries. Proper and sufficient characterization information should be provided.**
- 3. An oversight committee to oversee the global issues related to the national ordering effort should meet frequently to minimize redundancies and insure cooperation.**

## **DATA MANAGEMENT**

Dr. Donna Maglott presented the data management methods for the ATCC probe and library repository. Data are maintained in a relational data base program called "RDB" on a Micro-VAX 2 with the VMS operating system. The data base is accessible to the public via modem and is currently the only public access repository data base. The various fields (relations) in the data base are included as appendix 4. The data base has had minimal usage from the user community as yet but it is anticipated that usage will dramatically increase as more information is stored.

Dr. Ken Kidd described the Yale-New Haven human gene mapping data base funded by the Howard Hughes Medical Institute. The data base uses the "SPIRES" data base management system and currently includes the ability to interrelate five independent data bases. They include a data base for literature references, for map information given as chromosome and band location, for mapped probes, for mapped RFLPs, and for names of people to contact for probes. Appendix 5 contains a sample query. Recently, it has been possible to query the Mendelian Inheritance in Man data base in Baltimore while logged into the New Haven data base. Over 3755 individual clones are listed in the probe data base and the rate of acquisition is more than 100 per month. The data base is considering expanding to include information on contigs.

Dr. Jim Leighton described the computer networking facility emanating from Lawrence Livermore National Laboratory termed the Energy Sciences Network (ESNET). This network was established by DOE to support applications in several areas including health and the environment and is anticipated to be operational in 1-2 years. It has connectivity with other agencies (e.g. NSF, NASA, and DOD) and is international in scope with nodes in West Germany, Japan, and Switzerland. Each node would be supported by a Micro-VAX. Approximately 100 such nodes are being established in the United States. Dr. Walter Goad is the biological sciences representative to the advisory committee for this network. It would be feasible to link laboratories performing ordering and mapping work through this network to provide a rapid means for data sharing.

In the concluding remarks to the meeting, Dr. Cassandra Smith emphasized the data management problems associated with handling the probes, cell lines, restriction fragment data and experiment status reports. She emphasized a need for coordination to establish a uniform data base of information on experimental data and a mechanism to communicate these data to collaborators. The group felt that personal computers and existing software could manage these problems. Dr. Walter Goad indicated that a separate smaller meeting should be held to identify the problems of individual laboratories and to define possible solutions. A separate written document on automated sample management was submitted by the Robotics Group at Los Alamos National Laboratory for inclusion in this report (Appendix 6).

### **RECOMMENDATIONS**

1. **Additional funding should be provided for increased quality control of the Phase II effort of the National Gene Library Project. The increased funding would support an additional FTE in each laboratory (Los Alamos and Livermore) for characterization of the libraries.**
2. **A data base for restriction fragment data on cosmid clones and large fragments should be established.**
3. **A standard nomenclature should be developed for the informatics aspects of the genome ordering effort.**
4. **A smaller workshop should be conducted in the near future to deal specifically with the problems of intra-laboratory data management and inter-laboratory data exchange.**

## Agenda

### DOE-NIH Meeting on Repository, Data Management, and Quality Assurance Needs for the National Gene Library and Genome Ordering Projects

26 - 27 August 1987

Host : Lawrence Livermore National Laboratory

Location: Courtyard Hotel by Marriott, Pleasanton, CA.

*26 August, Wednesday*

8:00 a.m.	Convene/Continental Breakfast	Conference Room A
8:15	Introduction and Presentations	B. Gledhill, Chair B. Barnhart J. Willett A. Carrano
8:30	National Gene Library Project Status - LANL	L. Deaven
9:00	National Gene Library Project Status - LLNL	M. van Dilla
9:30	Status of Physical Mapping Effort at Columbia	C. Smith
10:00	Coffee Break	
10:15	Status of Physical Mapping Effort at LANL	E. Hildebrand
10:45	Status of Physical Mapping Effort at LLNL	A. Carrano
11:15	General Discussion	
11:30	Group Lunch Buffet	Lounge Area

---

1:00 p.m.	Physical Repository Status and Needs	C. Cantor, Chair
1:05	Status of ATCC Library and Probe Repository	W. Nierman
1:35	Open Discussion	All
	Needs generated by Phase II libraries Needs generated by genome ordering efforts Costs	
3:00	Coffee Break	

(over)

<b>3:30</b>	<b>Accuracy and Quality Assurance</b>	<b>K. McKenney, Chair</b>
	<b>Proposed quality control for Phase II libraries.</b>	<b>L. Deaven M. van Dilla</b>
<b>3:45</b>	<b>Discussion of Library Quality Control</b>	<b>All</b>
<b>4:15</b>	<b>Quality Control for Physical Repository</b>	<b>W. Nierman</b>
	<b>What is currently done at the ATCC.</b>	
<b>4:30</b>	<b>Discussion of Repository Quality Control</b>	<b>All</b>
<b>5:15</b>	<b>Discussion of Physical Mapping Quality Control</b>	<b>All</b>
	<b>What should be done? Who should do it? Associated costs?</b>	
<b>5:45</b>	<b>Adjourn for day/ Dinner at local restaurants</b>	



*27 August, Thursday*

<b>8:00 a.m</b>	<b>Convene/Continental Breakfast</b>	<b>Conference Room A</b>
<b>8:30</b>	<b>Data Management Status and Needs</b>	<b>W. Goad, Chair</b>
<b>8:35</b>	<b>ATCC Library and Probe Data Management</b>	<b>D. Maglott</b>
<b>9:05</b>	<b>The New Haven Data Base</b>	<b>K. Kidd</b>
<b>9:35</b>	<b>Networking at the LLNL MFE Computer Center</b>	<b>J. Leighton</b>
<b>10:05</b>	<b>Open Discussion</b>	<b>All</b>
	<b>Data handling within each laboratory Data communication among laboratories - networking Data base coupling</b>	
<b>11:30</b>	<b>Open Discussion/New or Additional Issues</b>	<b>All</b>
<b>12:00</b>	<b>Adjourn Meeting</b>	

## Invited Participants

Dr. Anthony V. Carrano  
Biomedical Sciences Division  
L-452  
Lawrence Livermore Natl. Laboratory  
P.O. Box 5507  
Livermore, CA 94550  
415-422-5698

Dr. Marvin A. van Dilla  
Biomedical Sciences Division  
L-452  
Lawrence Livermore Natl. Laboratory  
P.O. Box 5507  
Livermore, CA 94550  
415-422-5662

Dr. Elbert W. Branscomb  
Biomedical Sciences Division  
L-452  
Lawrence Livermore Natl. Laboratory  
P.O. Box 5507  
Livermore, CA 94550  
415-422-5681

Dr. Larry L. Deaven  
Life Sciences Division  
Los Alamos Natl. Laboratory  
P.O. Box 1663  
Los Alamos, NM 87545  
505-667-3144

Dr. C.E. Hildebrand  
Life Sciences Division  
Los Alamos Natl. Laboratory  
P.O. Box 1663  
Los Alamos, NM 87545  
505-667-2803

Dr. Walter Goad  
Theoretical Biology Group  
Los Alamos Natl. Laboratory  
P.O. Box 1663  
K710  
Los Alamos, NM 78545  
505-667-7511

Dr. Charles R. Cantor  
Columbia University College of P&S  
701 W 168th Street  
New York, NY 10032  
212-305-7915

Dr. Cassandra Smith  
Columbia University College of P&S  
701 W 168th Street  
New York, NY 10032  
212-305-4011

Dr. William Nierman  
American Type Culture Collection  
12301 Parklawn Drive  
Rockville, MD 20852  
301-231-5559

Dr. Donna Maglott  
American Type Culture Collection  
12301 Parklawn Drive  
Rockville, MD 20852  
301-231-5586

Dr. Jacqueline Courteau  
Biological Applications Prog.  
Office of Technology Assessment  
U.S. Congress  
Washington, D.C. 20510-8025  
202-226-2196

Dr. Keith McKenney  
National Bureau of Standards  
Center for Chemical Physics  
Gaithersburg, MD 20899  
301-975-2582

Dr. George Cahill  
Howard Hughes Medical Institute  
7984 Old Georgetown Road  
Bethesda, MD 20814  
301-571-0326

Dr. Kenneth K. Kidd  
Yale University School of Medicine  
333 Cedar Street  
New Haven, CT 06510  
203-785-2654

Dr. Jim D. Willett  
National Institute of Health  
Division of Research Resources  
Building 31, Room 5854  
500 Rockville Pike  
Bethesda, MD 20892  
301-496-5507

Dr. Benjamin J. Barnhart  
Health Effects Research Division  
Office of Health and Environ. Res.  
Office of Energy Res. ER-72  
U.S. Dept. of Energy (GTN)  
Washington, D.C. 20545  
301-353-5468

Dr. Marvin Stodolsky  
Health Effects Research Division  
Office of Health and Env. Res.  
Office of Energy Res. ER-72  
U.S. Dept. of Energy (GTN)  
Washington, D.C. 20545  
301-353-4475

Dr. James F. Leighton  
Natl. MFE Computer Center  
L-561  
Lawrence Livermore Natl. Laboratory  
P.O. Box 808  
Livermore, CA 94550  
415-422-4025

**Dr. Rachel Levinson**  
National Institute of Health  
Bldg 1 Room 224  
9000 Rockville Pike  
Bethesda, MD 20892  
301-496-1454

**Dr. Joel Schindler**  
Genetics and Teratology Branch  
Landon Bldg Room 7C08  
NICHD/NIH  
Bethesda, MD 20892

**Dr. Barton Gledhill**  
Biomedical Sciences Div.,L-452  
Lawrence Livermore Nat'l Lab.  
P.O. Box 5507  
Livermore, CA 94550  
415-422-5758

**Dr. Joe Gray**  
Biomedical Sciences Div.,L-452  
Lawrence Livermore Nat'l Lab.  
P.O. Box 5507  
Livermore, CA 94550  
415-422-5610

**Dr. Robert Mortimer**  
Donner Laboratory  
Lawrence Berkeley Lab.  
1 Cyclotron Road  
Berkeley, CA 94720  
415-642-4131

**Dr. Barbara Harrison**  
National Institute of Health  
Bldg 1 Room 224  
9000 Rockville Pike  
Bethesda, MD 20892  
301-496-1454

**Dr. Mark Guyer**  
NIGMS/NIH  
Westwood Bldg Room 920  
5333 Westbard Ave.  
Bethesda, MD 20816

**Dr. Pieter de Jong**  
Biomedical Sciences Div.,L-452  
Lawrence Livermore Nat'l Lab.  
P.O. Box 5507  
Livermore, CA 94550  
415-423-8145

**Dr. Harvey Mohrenweiser**  
Biomedical Sciences Div.,L-452  
Lawrence Livermore Nat'l Lab.  
P.O. Box 5507  
Livermore, CA 94550  
415-423-0534

REPOSITORY OF HUMAN DNA PROBES AND LIBRARIES



AMERICAN TYPE CULTURE COLLECTION

12301 Parklawn Drive, Rockville, Maryland 20852

PRODUCT SHEET FOR

57592 - Freeze-dried E. coli containing plasmid. Rehydrate with LB medium.  
57593 - Purified plasmid DNA distributed as 5 µg in 50 µl TE.

NAME OF CLONE: pcD-hIL-4

LOCUS OR ANONYMOUS PROBE DESIGNATION: IL4

NAME OF GENE PRODUCT: INTERLEUKIN 4

HGM8 MAP POSITION:

DEPOSITOR: TAKASHI YOKOTA  
DEPARTMENT OF MOLECULAR BIOLOGY  
DNAX RESEARCH INST OF MOLEC & CELL BIOL  
901 CALIFORNIA AVENUE  
PALO ALTO, CA USA 94304-1104

DEPOSITOR VERIFICATION RECEIVED: Y DETECTS RFLP:

\*\*\*\*\*

NAME OF VECTOR: pcD 3123b INSERT SIZE IN KB: .72  
SOURCE OF INSERT DNA: cDNA COPY NUMBER: UNIQUE  
FLANKING ENZYME SITES: MARKERS: ampR

REFERENCES:

PROC NATL ACAD SCI USA 1986;83:5894-98

NOTES:

Jul/23/87 Same as ATCC #67029  
Jul/23/87 RESTRICTION DIGESTS ANALYZED IN 1% AGAROSE SHOW: BamHI--2.9 kb, 0.8 kb; HindIII--3.8 kb; PstI--2.3 kb, 1.6 kb; ClaI--3.8 kb; XhoI--2.8 kb, 0.9 kb  
Aug/11/87 Exact restriction sizes are as follows: BamHI--2.98 kb, 0.86 kb; HindIII--3.84 kb; PstI--2.27 kb, 1.57 kb; ClaI--3.84 kb; XhoI--2.92 kb, 0.92 kb

SEQUENCE INFORMATION

REFERENCE

Insert includes about 100 bp of poly A.  
Contains complete coding sequence, GC tail  
between SV40 promoter and 5' end of insert;  
AT tail between 3' end of insert and poly A.  
Contains internal NheI, EcoRV, PstI, PvuII,  
BglI, and EcoRI sites.

PROC NATL ACAD SCI USA  
1986;83:5894-98

REPOSITORY DATABASE: CURRENT SIZE

Record Name	Occurrences	Bytes Used	Avg Rec in Bytes	% Frag	% Total Space	% Used Space
CHARACTERIZATION_LIB	58	9194	159	0	39	73
CHARACTERIZATION_PROBE	442	68761	156	0	73	91
CLASSES	5	75	15	0	1	2
CLONE	433	62622	145	0	69	90
COUNTRIES	38	892	23	0	15	21
CROSS_REF	265	13475	51	0	57	72
INVESTIGATOR	1109	175626	158	0	76	93
LIBRARY	72	13921	193	0	47	80
NO_POLYMORPHISM	91	11014	121	0	42	75
POLYMORPHISM	832	98345	118	0	63	91
POPULATION	24	595	25	0	10	16
PROBE_REQUEST	804	46984	58	0	72	83
RDB\$FIELDS	154	62062	403	0	78	93
RDB\$INDICES	37	3515	95	0	40	52
RDB\$RELATIONS	27	2376	88	0	40	43
RDB\$RELATION_FIELDS	262	118162	451	0	89	96
RECIPIENT	1486	43869	30	0	71	74
REF	910	26998	30	0	65	72
REFERENCE	573	120062	210	0	50	94
SEQNOTE	239	34008	142	0	68	87
TYPES	11	263	24	0	4	8

8437

948259



RELATIONS IN DATABASE

CHARACTERIZATION\_LIB  
CHARACTERIZATION\_PROBE  
CLASSES  
CLONE  
COUNTRIES  
CROSS\_REF  
INVESTIGATOR  
LIBRARY  
NO\_POLYMORPHISM  
POLYMORPHISM  
POPULATION  
PROBE\_REQUEST  
RECIPIENT  
REF  
REFERENCE  
SEQNOTE  
TYPES

**Fields for relation CHARACTERIZATION\_LIB**

LIB_NO	text size is 10
CAT_NO	signed longword scale 0
PURITY	varying string size is 255
LCOMMENTER	signed longword scale 0
COMMENT_DATE	Date
LIB_COMMENT	varying string size is 240
REF_NO	signed word scale 0

**Fields for relation CHARACTERIZATION\_PROBE**

COMMENT_DATE	Date
PROBE_COMMENT	varying string size is 240
ID	signed longword scale 0
REF_NO	signed word scale 0
B_CAT	signed longword scale 0

**Fields for relation CLASSES**

CLASS	signed word scale 0
CLASS_CODE	text size is 4

### Fields for relation CLONE

PROBE	text size is 18
LOCUS	text size is 10
GENE_NAME	text size is 45
CHROMOSOME	text size is 2
CHROM_ARM	text size is 1
REGION_UPPER	text size is 6
REGION_LOWER	text size is 6
MAP_IN_SITU	text size is 1
MAP_LINKAGE	text size is 1
MAP_HYB_CELL	text size is 1
MAP_DOSAGE	text size is 1
DISEASE	text size is 45
VECTOR	text size is 15
INSERT_SIZE	F floating
INSERT_SITE1	text size is 10
INSERT_SITE2	text size is 10
TYPE_CODE	signed word scale 0
TISSUE	text size is 20
PROMOTER	text size is 1
ENHANCER	text size is 1
POLY_A	text size is 1
SEQUENCED	text size is 1
STATUS	text size is 1
ID	signed longword scale 0
HOST	text size is 20
ID2	signed word scale 0
STAR	text size is 1
MARKER	text size is 16
VERIFIED	text size is 1
MAP_DELETION	text size is 1
ENDS	text size is 16
TOTAL_SIZE	signed word scale -2
ISOL_F_LIB	text size is 10
B_CAT	signed longword scale 0
D_CAT	signed longword scale 0

### Fields for relation COUNTRIES

COUNTRY_CODE	text size is 3
COUNTRY	text size is 20

### Fields for relation CROSS\_REF

GENBANK	text size is 6
MCKUSICK	signed longword scale 0
EC_NO	text size is 10
YALE_PROBE	text size is 14
YALE_RFLP	text size is 14
OTHER_SYMBOLS	varying string size is 80
PROTEIN_GROUP	text size is 40
B_CAT	signed longword scale 0

**Fields for relation INVESTIGATOR**

ID	signed longword scale 0
NAME_LAST	text size is 25
NAME_FIRST	text size is 15
NAME_INIT	text size is 1
DEPARTMENT	text size is 40
INSTITUTION	text size is 40
ADDRESS	text size is 40
CITY	text size is 20
STATE_PROV	text size is 4
COUNTRY_CODE	text size is 3
POSTAL_CODE	text size is 10
PHONE	text size is 20
USE_APPROVAL	text size is 1
DEPOSITOR	text size is 1
ATCC_CUST_NO	signed longword scale 0
CLASS	signed word scale 0
RECORDED	Date
DOCUMENTATION	text size is 1
ADVISOR	text size is 1
INFO_REQ	text size is 1

**Fields for relation LIBRARY**

CAT_NO	signed longword scale 0
LIB_NO	text size is 10
CHROM_SOURCE	text size is 24
CHROMA	text size is 2
CHROMB	text size is 2
ISOL_MTHD	varying string size is 30
VECTOR	text size is 15
REST_ENZ	text size is 10
DIGEST	text size is 1
INSERT_SIZE	F_floating
DATE_CLONED	Date
NON_RECOMB	signed word scale -2
DATE_AMPL	Date
GENOME_EQ	F_floating
DIST_VOL	signed word scale -2
ID	signed longword scale 0
TISSUE_SOURCE	text size is 45
TYPE_CODE	signed word scale 0
COVER	signed longword scale 0
INSI2	F_floating
TITER	F_floating
RECOMB_NO	F_floating
SORTING_QUAL	varying string size is 240
HUMAN_CHROM_CELL	varying string size is 50
INPUT_CHROM_NO	F_floating
STATUS	text size is 1

**Fields for relation NO\_POLYMORPHISM**

PROBE	text size is 18
ENZ_NO_POLYMORPH	varying string size is 256
COMMENT	varying string size is 240
B_CAT	signed longword scale 0

**Fields for relation POLYMORPHISM**

PROBE	text size is 18
LOCUS	text size is 10
REST_ENZ	text size is 10
CONSTANT_BANDS	text size is 25
ALLELE	text size is 6
ALLELE_LEN1	F_floating
ALLELE_LEN2	F_floating
ALLELE_LEN3	F_floating
ALLELE_FREQ	signed word scale -2
PIC_F	signed word scale -2
STD_DEV	F_floating
POP_CODE	signed word scale 0
DATE_UPDATE	Date
COMMENT	varying string size is 240
POP_SIZE	signed longword scale 0
REF_NO	signed word scale 0

**Fields for relation POPULATION**

POP_CODE	signed word scale 0
RACE	text size is 1
COUNTRY_CODE	text size is 3
STATE_PROV	text size is 4
AGE	F_floating
DISEASE	text size is 45
SEX	text size is 1

**Fields for relation PROBE\_REQUEST**

PROBE	text size is 18
ID	signed longword scale 0
DATE_REQ	Date
DATE_REC	Date
LOCUS	text size is 10
SENT	Date
RELEASED	Date
DATE_APPR	Date
REF_NO	signed word scale 0
REPLY	text size is 1
DATE_REPL	Date
PARTIAL	text size is 1
DATE_PARTIAL	Date
B_CAT	signed longword scale 0
DATA_VERIF	Date

**Fields for relation RECIPIENT**

ATCC_CUST_NO	signed longword scale 0
ID	signed longword scale 0
DATE_SENT	Date
ITEM	signed longword scale 0
EXCHANGE	text size is 1

**Fields for relation REF**

PROBE	text size is 18
REF_NO	signed word scale 0
LOCUS	text size is 10
B_CAT	signed longword scale 0

**Fields for relation REFERENCE**

AUTHOR	varying string size is 240
TITLE	varying string size is 240
JOURNAL	text size is 45
REF_NO	signed word scale 0

**Fields for relation SEQNOTE**

SEQ_NOTES	varying string size is 240
REF_NO	signed word scale 0
B_CAT	signed longword scale 0

**Fields for relation TYPES**

TYPE_CODE	signed word scale 0
DNA	text size is 7
COPY_NO	text size is 6

**APPENDIX 5**

**SAMPLE OUTPUT FROM HHMI HUMAN GENE MAPPING DATA BASE.**

**HHMI**

**Human Gene Mapping Library**

**Sample Output from Databases**

## HGML Sample Entries

The following 4 pages present samples of entries that may be found in three of the Human Gene Mapping Library databases -- in this instance for PAH, or phenylalanine hydroxylase. The pages consist of, in order:

1. An entry in the **MAP** database.
2. An entry in **MAP** similar to 1 but including the citations for the cross-references to the **LIT** database.
3. An **excerpt** (one probe-enzyme system) from the PAH entry in the **RFLP** database. System C is described here.
4. Three entries in **PROBE** for the probes references in 3.



Symbol: PAH

Entry from HGM Workshop, last updated 08/13/85:

Marker name: phenylalanine hydroxylase

E.C. number: 1.14.16.1

Status: C

Map location: 12q22-q24.2

Assignment mode: RE,S

MIM number: 26160

References relevant to mapping: G0125, H0481, H0655

Entry by local review committee, last updated 08/19/87:

Background references, possibly relevant to mapping: H2496,  
H2973, H3884

Symbol: PAH

Entry from HGM Workshop, last updated 08/13/85:

Marker name: phenylalanine hydroxylase  
E.C. number: 1.14.16.1  
Status: C  
Map location: 12q22-q24.2  
Assignment mode: RE,S  
MIM number: 26160

References relevant to mapping: G0125, H0481, H0655

Entry by local review committee, last updated 08/19/87:

Background references, possibly relevant to mapping: H2496,  
H2973, H3884

REFERENCES IN LITerature DATABASE.

- G0125 Lidsky, A.; Robson, K. J. H.; Chandra, T.; Barker, P.;  
Ruddle, F. H.; Woo, S. L. C.  
The PKU locus in man is on chromosome 12.  
Am. J. Hum. Genet. 35:201A,1983. (abs. 599)
- H0481 Lidsky, A. S.; Robson, K. J. H.; Thirumalachary, C.;  
Barker, P. E.; Ruddle, F. H.; Woo, S. L. C.  
The PKU locus in man is on chromosome 12.  
Am. J. Hum. Genet. 36:527-533,1984.
- H0655 Woo, S. L. C.; Lidsky, A.; Law, M.; Kao, F. T.  
Regional mapping of the human phenylalanine hydroxylase  
gene and PKU locus to 12q21->qter.  
Am. J. Hum. Genet. 36:210S,1984. (abs. 622)
- H2496 DiLella, A. G.; Kwok, S. C. M.; Ledley, F. D.; Marvit, J.;  
Woo, S. L. C.  
Molecular structure and polymorphic map of the human  
phenylalanine hydroxylase gene.  
Biochem. 25:743-749,1985.
- H2973 DiLella, A. G.; Marvit, J.; Lidsky, A. S.; Guttler, F.;  
Woo, S. L. C.  
Tight linkage between a splicing mutation and a specific  
DNA haplotype in phenylketonuria.  
Nature 322:799-803,1986.
- H3884 Chakraborty, R.; Lidsky, A. S.; Daiger, S. P.; Guttler,  
F.; Sullivan, S.; DiLella, A. G.; Woo, S. L. C.  
Polymorphic DNA haplotypes at the human phenylalanine  
hydroxylase locus and their relationship with  
phenylketonuria.  
Hum. Genet. 76:40-46,1987.

LOCUS SYMBOL, MAP LOCATION, AND BRIEF DESCRIPTION.

RFLP ID number : d0209  
 Locus name : phenylalanine hydroxylase  
 Locus symbol : PAH  
 Map location : 12q22-q24.2  
 Highest PIC : 0.37  
 Comments : Ledley et al., 1986 (H2627) report that RFLP's at this locus were used to distinguish haplotypes and alleles in families with PKU and mild hyperphenylalaninemia.  
 Date : 03/13/87

DEFINITION OF ALLELE SYSTEM Ca

Probe : phPAH247, pPH72, hPH7  
 Enzyme : HindIII  
 Polymorph. type : insertion/deletion  
 Number of bands : 6  
 Allele bands : \* 5.6 5.2 2.7 4.2 4.0 4.4  
                   C1 + + + + - -  
                   C2 + + + - + -  
                   C3 + + + - - +  
 LIT Citation : H0154, H2496, H3884, H1648, H2857  
 XRef. to PROBE : p02252, p01946, p02614  
 System comments : The HindIII polymorphism apparently results from an insertion/deletion of 0.2 kb unit (Chakraborty et al., 1987 (H3884)). The HindIII polymorphic site is located in the 3' flanking region of the gene (DiLella et al., 1986 (H2496)).  
 Date updated : 07/30/87

ALLELE FREQUENCY DATA FOR SYSTEM Ca

Population : Caucasian  
 Number of chroms: 40  
 Frequency PIC : .37  
 Allele frequency: C1 0.71 +/- 0.07  
                   C2 0.24 +/- 0.07  
                   C3 0.06 +/- 0.04  
 LIT Citation : H0154  
 Date updated : 01/14/86

ALLELE FREQUENCY DATA FOR SYSTEM Ca

Population : Caucasian (Danish)  
 Number of chroms: 132  
 Frequency PIC : .37  
 Allele frequency: C1 0.636 +/- 0.04  
                   C2 0.363 +/- 0.04  
                   C3 0.001 +/- 0.003  
 LIT Citation : H3884  
 Date updated : 07/21/87

[p01946] Lab symbol : pPH72  
Alternate symbol : prPAH72  
Sequenced : yes  
Vector : pBR322  
Insertion site : PstI  
Size of insert(kb) : 1.2  
Type of DNA : cDNA  
Contact : woo  
Availability : available  
HGM locus symbol : PAH  
Refer to LIT : H0154, H0481, H0892  
Refer to RFLP : d0209  
Date record added : 01/14/86  
Last updated : 04/16/87

[p02252] Lab symbol : phPAH247  
Sequenced : no  
Vector : pBR322  
Insertion site : EcoRI  
Ends : EcoRI  
Size of insert(kb) : 2.448  
Type of DNA : cDNA  
Contact : woo  
Availability : available  
HGM locus symbol : PAH  
Refer to LIT : H0892, H1648, H2496, H2627, H3884,  
H3894  
Refer to RFLP : d0209  
Comments : From a human liver cDNA library.  
This probe encompasses the entire  
PAH coding region as well as the 3'  
and 5' untranslated regions.  
Date record added : 09/25/86  
Last updated : 10/21/86

[p02614] Lab symbol : hPH7  
Sequenced : no  
Vector : pUC9  
Insertion site : EcoRI  
Ends : EcoRI  
Size of insert(kb) : 2.4  
Type of DNA : cDNA  
Contact : cout, dahl  
Availability : not known  
HGM locus symbol : PAH  
Refer to LIT : H2857  
Comments : This is nearly a full-length cDNA  
which is missing approximately 150  
nucleotides at the 3' untranslated  
end.  
Date record added : 02/18/87  
Last updated : 02/18/87

AUTOMATED SAMPLE MANAGEMENT: ORGANIZING, MANAGING,  
AND ACCESSING THE DNA FRAGMENTS GENERATED IN THE PROCESS  
OF MAPPING AND SEQUENCING THE HUMAN GENOME

Dan Knobeloch  
Tony J. Beugelsdijk

Los Alamos National Laboratory  
Mechanical and Electronic Engineering Division  
Group MEE-9  
Robotics Section  
Phone (505) 667-3186  
667-3169  
Mail Stop E537

As human map and sequence data accumulate, many investigators will be able to apply this knowledge to problems of medical and biological importance. They will need access to large numbers of biological samples including cloned DNA fragments and human cell lines. Methods for the efficient production and distribution of these materials needs to be developed. Automated repository management utilizing current technologies will facilitate the coordination of research data through the use of computer database management activities. In fact, repository management is similar to computer databases in the sense that computers manipulate bits of data while repositories will manipulate bits of DNA. Thus, basic requirements in developing automated DNA repositories emulate some of the basic requirements in construction of computational databases. At Los Alamos National Laboratory researchers are investigating the anticipated growth in DNA map and sequence data as progress is made in the Human Genome Project. (C. Burks, et al.) The following considerations in developing an automated DNA sample repository were based on this research and development effort.

The samples generated in the process of mapping and sequencing the human genome will:

- \* represent several (and perhaps many) different kinds of DNA, and different types of vectors used to contain the DNA.
- \* at peak rates, access rates can be expected to exceed the capacity of current manual DNA storage and distribution systems.
- \* require careful and timely management, with input expected from sources all over the world.

To best serve the expected needs of the scientists generating and analyzing these DNA fragments, the corresponding repository(ies) must be designed and managed so as to allow:

- \* uniform organization of samples

- \* continuous input/output of DNA from scientists at many remote locations (with automatic error-checking and as little manual intervention from sample management staff
- \* enhanced accountability for each sample
- \* timely inventory information
- \* flexibility to accommodate changes in DNA sample size, storage requirements (i.e., liquid nitrogen, stabs, etc.), and storage density.

These requirements are not unique to molecular biology and have been addressed by a variety of material management industries. Repositories have been designed to include high-speed automated sample identification techniques, automatic database management systems, and remote material handling technology. Automated repositories have relied on bar code technology to provide data input for coordinated and consistent operation of the subsystems. Bar code technology provides a unique code that can be generated and applied to each sample with automatic print and apply systems. Thus, individual items can be positively identified and manipulated without the need for manual intervention.

Incorporating bar code technology with automated storage/retrieval systems could be one approach to the development of a distributive sample management system for the human genome mapping effort. Bar code technology would be used to monitor sample identification to permit quick access to important information including the source of the sample (which lab or researcher), type of DNA insert (i.e., which chromosome, type of vector, etc.), characterization data, and other markers that would be used to "fingerprint" the sample. Automated sample repositories would also facilitate the on-going distribution of samples in the existing human chromosome specific libraries developed by Los Alamos and Lawrence Livermore National Laboratories. Keyless data entry would eliminate the tedious operations required to monitor distributive operations like those anticipated in the human genome project.

In order to meet these goals the development of an automated repository should utilize proven automated material handling and storage concepts. The system will be modified to manipulate DNA samples, but rely on existing technology to facilitate development and implementation. A typical system consists of Storage/Retrieval (S/R) machines and a storage system that houses removable racks. The operators will enter transactions on an accountability system which will direct the S/R machines to store or retrieve materials. The S/R machines can be designed to interface with the operator through an opening in the storage system, or through a wall, or utilizing pneumatic transfer devices send the item to a remote location. Process automation equipment and robotics technology are developed to the point that such automated warehousing systems are highly reliable and commonly used throughout industry. Similar automated systems are being used to store and retrieve nuclear material at Los Alamos National Laboratory, Rocky Flats Plant, and Westinghouse-Hanford.