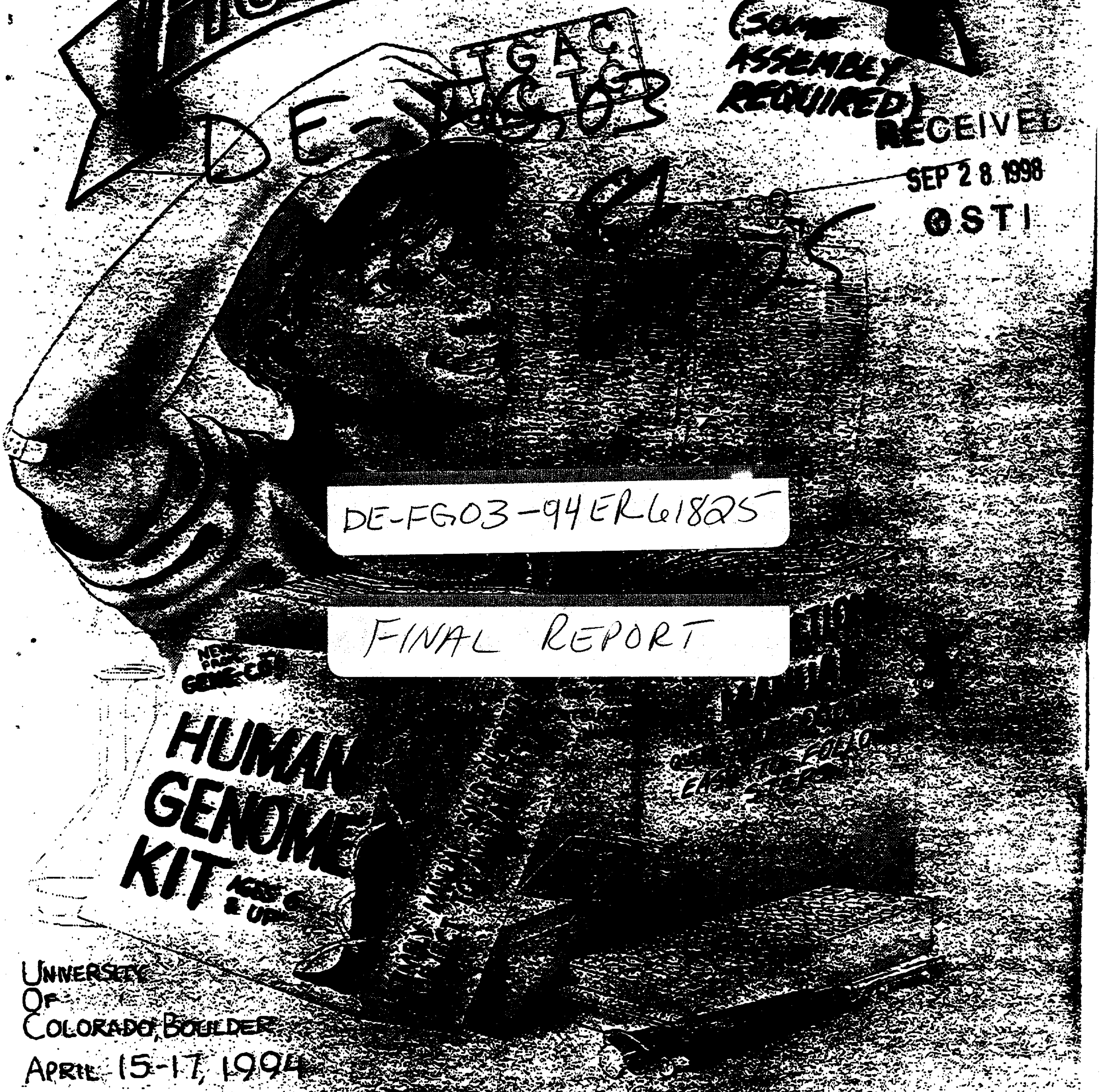
CONF-9404338-Summ.

THE 1994 MCDB GRADUATE STUDENT SYMPOSIUM

HUMAN GENOME

(SOME ASSEMBLY REQUIRED)

DE-



DE-FG03-94ER61825

FINAL REPORT

UNIVERSITY
OF
COLORADO, BOULDER

APRIL 15-17, 1994

THE METHODS, GOALS & IMPLICATIONS OF THE HUMAN GENOME PROJECT

Kay final

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

1994 Graduate Student Symposium
The Human Genome: Some Assembly Required
The Methods, Goals and Implications of the Human Genome Project

Friday, April 15th – Sunday, April 17th

Friday April 15

4:00 PM	Registration, Porter BioSciences First Floor Lobby
5:00 PM	Dinner, Porter BioSciences
7:00 PM	**Keynote Address: Leroy Hood**, Univ. of Washington School of Medicine *Perspectives on the Human Genome Project*

Session I: Finding the Parts – Large Scale Sequencing Technology

8:00 PM	**Bob Waterston**, Genetics Dept., Washington Univ. School of Medicine *The* C. elegans *Genome Project: Lessons*
9:00 PM	Reception and Dessert Party, Koenig Alumni Center

Saturday April 16

8:00 AM	Breakfast, Porter BioSciences
8:30 AM	**Leroy Hood**, University of Washington School of Medicine *Large Scale DNA Sequencing*
9:30 AM	**Stephen Fodor**, Affymetrix, Santa Clara, CA *Oligonucleotide Arrays and Sequence Analysis by Hybridization*

Session II: Assembly Instructions – Analysis of Genomic Sequence Data

10:30 AM	**David Searls**, Department of Genetics, Univ. of Penn School of Medicine *Genome Linguistics*
11:30 AM	Lunch, Porter BioSciences
12:30 PM	**Richard Mural**, Biology Division, Oak Ridge National Laboratory *Combining Neural Networks and Expert Systems to Identify Features in DNA Sequences*
1:30 PM	**Phil Green**, Genetics Department, Washington Univ. School of Medicine *Ancient Conserved Regions: Implications for Gene Identification*

Session III: Trouble Shooting – Understanding Human Genetic Disease

2:30 PM	**Katheleen Gardiner**, Eleanor Roosevelt Institute *Chromosome 21: Its Associated Genetic Diseases and Its Place in the Human Genome Project*
3:30 PM	Free Time / Group Outing – Hiking in the Mountains
5:00 PM	Dinner, Restaurants in Boulder
7:00 PM	Coffee, Porter BioSciences
7:30 PM	**Charles Laird**, Department of Medicine, University of Washington *Triplet Repeat Disease and Genomic Imprinting*
8:30 PM	**Mary-Claire King**, University of California, Berkeley *Mapping Genetic Disorders*
9:30 PM	Symposium Party, Porter BioSciences, Room 121

Sunday April 17

Session IV: Disclaimers – Ethical, Legal and Social Issues

8:00 AM	Breakfast, Porter BioSciences
9:00 AM	**Kenneth Kidd**, Department of Genetics, Yale University Medical School *Diverse Human Genomes*
10:00 AM	**Dean Hamer**, National Cancer Institute, National Institutes of Health *Genetics and Sexual Orientation*
11:00 AM	**Michael Yesley**, Los Alamos National Laboratory *The NIH–DOE ELSI Program*
12:00 PM	Closing Remarks
12:15 PM	Lunch, Porter BioSciences

All talks will be held in Ramaley C250

The Human Genome: Some Assembly Required

The Human Genome Project promises to be one of the most rewarding endeavors in modern biology. The cost and the ethical and social implications, however, have made this project the source of considerable debate both in the scientific community and in the public at large. In the 1994 Graduate Student Symposium, we would like to address the scientific merits of the project, the technical issues involved in accomplishing the task, as well as the medical and social issues which stem from the wealth of knowledge which the Human Genome Project will help create.

To this end, we have brought together speakers who represent the diverse areas of expertise characteristic of this multidisciplinary project. The keynote speaker will address the project's motivations and goals in the larger context of biological and medical sciences. The first two sessions will address relevant technical issues, data collection with a focus on high–throughput sequencing methods and data analysis with an emphasis on identification of coding sequences. The third session will discuss recent advances in the understanding of genetic diseases and possible routes to treatment. Finally, the last session will address some of the ethical, social and legal issues which will undoubtedly arise from having a detailed knowledge of the human genome.

1994 Graduate Student Symposium Committee
Tim Nickles
Eric Snyder
Jack Tabaska
Daniel Weaver

Keynote Address: Perspectives on the Human Genome Project

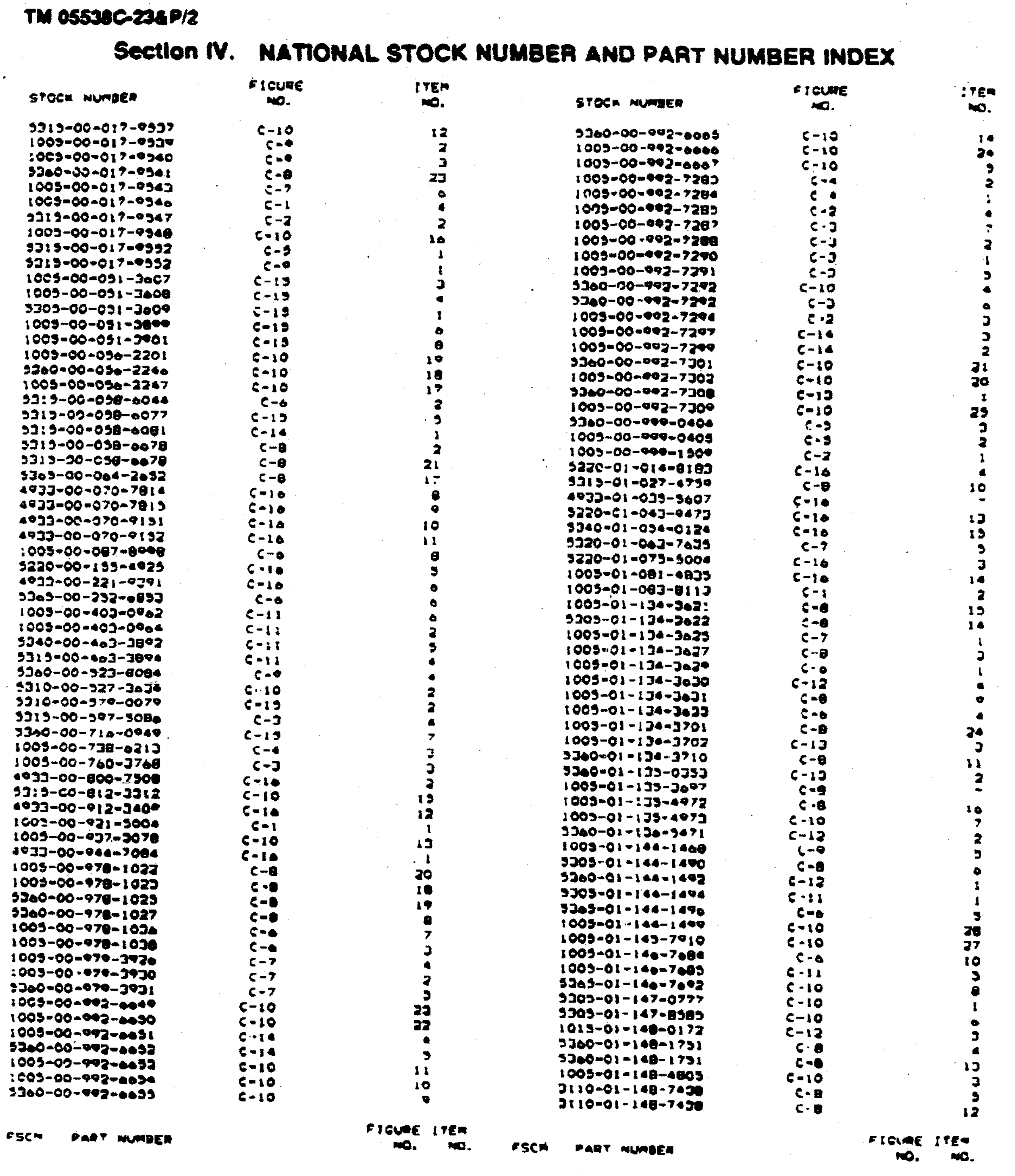
Leroy Hood, University of Washington School of Medicine

The human genome project proposes to decipher the human and model organism genomes over a 15 year period, starting in 1990. This will lead to three types of chromosomal maps: genetic, physical, and sequence. These maps are providing powerful new tools for biology and medicine. This project poses striking technical challenges for developing large scale mapping and sequencing tools, as well as creating the informatics software necessary to delineate the digital information 3.7 billion years of evolution has embedded in the chromosomes of living organisms.

I will discuss these general sequencing and computational challenges and consider some of the contemporary applications of genomics, the application of genome tools, to fundamental problems in developmental biology and immunology.

SESSION I:

FINDING THE PARTS – LARGE SCALE SEQUENCING TECHNOLOGY



TM 05538C-23&P/2

Section IV. NATIONAL STOCK NUMBER AND PART NUMBER INDEX

STOCK NUMBER	FIGURE NO.	ITEM NO.	STOCK NUMBER	FIGURE NO.	ITEM NO.
5315-00-017-9537	C-10	12	5360-00-992-6665	C-10	14
1005-00-017-9539	C-9	2	1005-00-992-6666	C-10	26
1005-00-017-9540	C-9	3	1005-00-992-6667	C-10	5
5360-00-017-9541	C-8	23	1005-00-992-7283	C-4	2
1005-00-017-9543	C-7	6	1005-00-992-7284	C 4	1
1005-00-017-9546	C-1	4	1005-00-992-7285	C-2	4
5315-00-017-9547	C-2	2	1005-00-992-7287	C-3	7
1005-00-017-9548	C-10	16	1005-00-992-7288	C-3	2
5315-00-017-9552	C-5	1	1005-00-992-7290	C-3	1
5315-00-017-9552	C-9	1	1005-00-992-7291	C-3	5
1005-00-051-3607	C-15	3	5360-00-992-7292	C-10	4
1005-00-051-3608	C-15	4	5360-00-992-7292	C-3	6
5305-00-051-3609	C-15	1	1005-00-992-7294	C-2	3
1005-00-051-3899	C-15	6	1005-00-992-7297	C-14	3
1005-00-051-3901	C-15	8	1005-00-992-7299	C-14	2
1005-00-056-2201	C-10	19	5360-00-992-7301	C-10	21
5360-00-056-2246	C-10	18	1005-00-992-7302	C-10	20
1005-00-056-2247	C-10	17	5360-00-992-7308	C-13	1
5315-00-058-6044	C-6	2	1005-00-992-7309	C-10	25
5315-00-058-6077	C-15	5	5360-00-999-0404	C-5	3
5315-00-058-6081	C-14	1	1005-00-999-0405	C-5	2
5315-00-058-6078	C-8	2	1005-00-999-1909	C-2	1
5315-00-058-6078	C-8	21	5220-01-014-8183	C-16	4
5365-00-064-2652	C-8	17	5315-01-027-4759	C-8	10
4933-00-070-7814	C-16	8	4933-01-035-5607	C-16	-
4933-00-070-7815	C-16	9	5220-01-043-9473	C-16	13
4933-00-070-9151	C-16	10	5340-01-054-0124	C-16	15
4933-00-070-9152	C-16	11	5220-01-063-7635	C-7	5
1005-00-087-8998	C-6	8	5220-01-075-5004	C-16	3
5220-00-155-4925	C-16	5	1005-01-081-4835	C-16	14
4933-00-221-9391	C-16	6	1005-01-083-8113	C-1	2
5365-00-252-6853	C-6	6	1005-01-134-3621	C-8	15
1005-00-403-0962	C-11	6	5305-01-134-3622	C-8	14
1005-00-403-0964	C-11	2	1005-01-134-3625	C-7	1
5340-00-463-3892	C-11	5	1005-01-134-3627	C-8	3
5315-00-463-3894	C-11	4	1005-01-134-3629	C-6	1
5360-00-523-8084	C-9	4	1005-01-134-3630	C-12	4
5310-00-527-3634	C-10	2	1005-01-134-3631	C-8	6
5310-00-579-0079	C-15	2	1005-01-134-3633	C-6	4
5315-00-597-5086	C-3	4	1005-01-134-3701	C-8	24
5360-00-716-0949	C-15	7	1005-01-134-3707	C-13	3
1005-00-738-6213	C-4	3	5360-01-134-3710	C-8	11
1005-00-760-3768	C-3	3	5360-01-135-0353	C-13	2
4933-00-800-7508	C-16	2	1005-01-135-3697	C-9	-
5315-00-812-3312	C-10	15	1005-01-135-4972	C-8	16
4933-00-912-3409	C-16	12	1005-01-135-4973	C-10	7
1005-00-921-5004	C-1	1	5360-01-136-5471	C-12	2
1005-00-937-3078	C-10	13	1005-01-144-1468	C-9	5
4933-00-944-7084	C-16	1	5305-01-144-1490	C-8	6
1005-00-978-1022	C-8	20	5360-01-144-1492	C-12	1
1005-00-978-1023	C-8	18	5305-01-144-1494	C-11	1
5360-00-978-1025	C-8	19	5365-01-144-1496	C-6	5
5360-00-978-1027	C-8	8	1005-01-144-1499	C-10	28
1005-00-978-1036	C-6	7	1005-01-145-7910	C-10	27
1005-00-978-1038	C-6	3	1005-01-146-7684	C-6	10
1005-00-979-3926	C-7	4	1005-01-146-7685	C-11	5
1005-00-979-3930	C-7	2	5365-01-146-7692	C-10	8
5360-00-979-3931	C-7	5	5305-01-147-0777	C-10	1
1005-00-992-6649	C-10	23	5305-01-147-8585	C-10	6
1005-00-992-6650	C-10	22	1015-01-148-0172	C-12	3
1005-00-992-6651	C-14	4	5360-01-148-1751	C-8	4
5360-00-992-6652	C-14	5	5360-01-148-1751	C-8	13
1005-00-992-6653	C-10	11	1005-01-148-4805	C-10	3
1005-00-992-6654	C-10	10	3110-01-148-7438	C-8	5
5360-00-992-6655	C-10	9	3110-01-148-7438	C-8	12

FSCM PART NUMBER FIGURE NO. ITEM NO. FSCM PART NUMBER FIGURE NO. ITEM NO.

The C. elegans *Genome Project: Lessons*

R. Waterston, Genetics Department, Washington University

C. elegans is a widely used experimental animal in studies of development, cell biology and neurobiology. In support of these activities and as part of the international Human Genome Project, our laboratories in St. Louis and Cambridge are pursuing sequence analysis of the entire C. elegans genome. An initial pilot phase was devoted to methods development, with gradual scaling up of production. With feasibility demonstrated, both groups have begun a larger effort that should yield the complete genome sequence by 1998.

At present, we have completed more that 2.1 Mb of contiguous sequence and several hundred kb of additional cosmid–sized segments from the central portion of chromosome III. The sequence is gene–rich, and about 30% of the predicted genes have significant database matches. The remainder of chromosome III and part of chromosome II are currently in progress.

Fire, A. and Waterston, R. Proper expression of myosin genes in transgenic nematodes. 1989. *EMBO J* 8: 3429–3436.

Waterston, R. H. The minor myosin heavy chain, mhcA, of *Caenorhabditis elegans* is necessary for the initiation of thick filament assembly. 1989. *EMBO J* 8: 3429–3436.

Benian, G.M., Kiff, J.E., Neckelmann, N., Moerman, D.G. and Waterston, R.H. 1989. Sequence of an unusually large protein implicated in regulation of myosin activity in *Caenorhabditis elegans*. *Nature* 342: 45–50.

Mori, J., Moerman, D.G. and Waterston, R.H. 1990. Interstrain crosses enhance excision of Tc1 transposable elements in *Caenorhabditis elegans*. *Mol. Gen. Genet.* 220: 251–255.

Kondo, . Makovec, B., Waters[illegible] R.H. and Hodgkin, J. 1990. Genetic and molecular analysis of eight tRNAtrp amber suppressors in *Caenorhabditis elegans*. *J. Mol Biol* 215: 7–19.

Burglin, T.R., Ruvkin, G. Coulson, A., Hawkins, N.C., McGhee, J.D., Schaller, D., Wittman, C., Muller, F. and Waterston, R.H. 1991. Nematode homeobox cluster. *Nature* 351: 703.

Barstead, R.J. and Waterston, R.H. 1991. Vinculin is essential for muscle formation in the nematode. *J. Cell Biol.* 114: 715–724.

Coulson, A., Kozono, Y., Shownkeen, R. and Waterston, R. 1991. The isolation of insert–terminal YAC fragments by genomic sequencing. *Technique–A J. of Meth. in Cell and Molec. Biol.* 3: 17–23.

Coulson, A., Kozono, Y. Lutterbach, G., Shownkeen, R., Sulston, J. and Waterston, R. 1991. YACs and the C. *elegans* genome. *Bioessays* 13: 413–417.

Green, E.D. and Waterston, R.H. 1991. The human genome project: prospects and implications for clinical medicine. *JAMA* 266: 1966–1975.

Sulston, J., Ainscough, R., Berks, M., Coulson, A., Craxton, M., Dear, S., Du., Z, Durbin, R., Gleeson, T., Green, P., Halloran, N., Hawkins, T., Hillier, L., Metzstein, M., Qiu, L., Staden, R., Thierry-Mieg, J.. Thomas, K., Wilson, R. and Waterston, R. 1992. The *C. elegans* genome sequencing project: a beginning. *Nature* 356: 37-41.

Waterston, R., Martin, C., Craxton, M., Huynh, Cl. Coulson, A., Hillier., L. Durbin, R., Green, P., Shownkeen, R. Halloran, N., Metzstein, M., Hawkins, T. Wilson, R., Berks, M., Du, Z., Thomas, K., Thierry-Mieg, J. and Sulston, J., 1992. A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genetics* 1: 114-123.

Williams, B., Shrank, B., Huynh, C., Shownkeen, R. and Waterston, R.H., 1992. A genetic mapping system in *Caenorhabditis elegans* based on polymorphic sequence-tagged sites. *Genetics* 131: 609-624.

Green, P., Lipman, D., Hillier, L., Waterston, R., States, D. and Claverie, J-M. 1993. Ancient conserved regions in new gene sequences and the protein databases. *Science* 259: 1711-1716.

Waddle, J.A., Cooper, J.A. and Waterston, R.H. 1993. The α and β subunits of nematode actin capping protein function in yeast. *Mol Biol. Cell* 4: 907-917.

Wilson, R., *et al.* 1994. The *C. elegans* genome project: contiguous nucleotide sequence of over two megabases from chromosome III. *Nature* 368: 32-38.

Williams, B.D., and Waterson, R.H. (in press) Genes critical for muscle development and function in *Caenorhabditis elegans* identified through lethal mutations. *J.Cell Biol.*

Hresko, M.C., Williams, B.D., and Waterston, R.H. (in press) Assembly of body wall muscle and muscle cell attachment structures in *Caenorhabditis elegans*. *J. Cell Biol.*

Large Scale DNA Sequencing

Leroy Hood, University of Washington School of Medicine

Large scale DNA sequencing analysis poses challenges in instrumentation, chemistry, applied physics, computational analysis, systems integration and automation. Today a single DNA sequencing machine may analyze up to 36,000 base pairs of DNA per day. Yet, we will need to increase the throughput of DNA sequence analysis by at least 100 fold before the complete analysis of the human genome will be feasible. I will discuss the contemporary state of DNA sequencing and indicate the direction of future opportunities. I will also discuss out efforts to analyze the T-cell receptor families of human and mouse. These data have provided new insights into the biology, regulation, and evolution of these multigene families.

Hunkapiller T, Kaiser RJ, Koop BF, Hood L (1991) Large-scale and automated DNA sequence determination Science 254(5028):59-67.

Drmanac R, Drmanac S, Strezoska Z, Paunesku T, Labat I, Zeremski M, Snoddy J, Funkhouser WK, Koop B, Hood L et al (1993) DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing Science 260(5114):1649-52.

Strauss EC, Kobori JA, Siu G, Hood LE (1986) Specific-primer-directed DNA sequencing. Anal Biochem 154(1):353-60

Koop BF, Wilson RK, Chen C, Halloran N, Sciammis R, Hood L, Lindelien JW (1990) Sequencing reactions in microtiter plates Biotechniques 9(1):32, 34-7.

Wilson RK, Chen C, Hood L (1990) Optimization of asymmetric polymerase chain reaction for rapid fluorescent DNA sequencing Biotechniques 8(2):184-9.

Wilson RK, Yuen AS, Clark SM, Spence C,, Arakelian P Hood LE (1988) Automation of dideoxynucleotide DNA sequencing reactions using a robotic workstation Biotechniques 6(8):776-7, 781-7.

Koop BF, Rowan L, Chen WQ, Deshpande P, Lee H, Hood L (1993) Sequence length and error analysis of Sequenase and automated Taq cycle sequencing methods Biotechniques 14(3):442-7.

Wettenhall RE, Aebersold RH, Hood LE (1991) Solid-phase sequencing of 32P-labeled phosphopeptides at picomole and subpicomole levels Methods Enzymol 201:186-99.

Kaiser R, Hunkapiller T, Heiner C, Hood L (1993) Specific primer-directed DNA sequence analysis using automated fluorescence detection and labeled primers Methods Enzymol 218:122-53.

Du Z Hood L Wilson RK (1993) Automated fluorescent DNA sequencing of polymerase chain reaction products Methods Enzymol 218:104–21.

Kaiser RJ MacKellar SL Vinayak RS Sanders JZ Saavedra RA Hood LE (1989) Specific–primer–directed DNA sequencing using automated fluorescence detection Nucleic Acids Res 17(15):6087–102.

Landegren U Kaiser R Caskey CT Hood L (1988) DNA diagnostics—molecular techniques and automation Science 242(4876):229–37.

Wilson RK Chen C Avdalovic N Burns J Hood L (1990) Development of an automated procedure for fluorescent DNA sequencing Genomics 6(4):626–34.

Oligonucleotide Arrays and Sequence Analysis by Hybridization

Stephen P.A. Fodor, Affymetrix, Santa Clara, CA

Light-directed chemical synthesis has been used to generate miniaturized, high-density arrays of oligonucleotide probes. These probe arrays, or DNA chips are then used for parallel DNA hybridization analysis, directly yielding sequence information from genomic DNA. Successful implementation of the DNA chip technology requires development of methods for fabrication of the probe arrays, detection of target hybridization, and algorithms to analyze hybridization and reconstruct target sequence. The results of recent experiments addressing each of these methods will be discussed. Application specific oligonucleotide probe array designs will be presented. These designed arrays have been used to demonstrate direct reading of genomic sequence. This method is proving to be a powerful tool for rapid investigations in DNA sequencing, human genetics and diagnostics, pathogen detection, and DNA molecular recognition.

Gordon, EM, Gallop, MA Barrett, R.W., and Fodor, S.P.A.Applications of combinatorial technologies to drug discovery J Medicinal Chemistry, in press.

Jacobs, JW and Fodor, SPA. Combinatorial chemistry-applications of light directed chemical synthesis. Trends in Biotechnology, in press.

Pease, AC, Solas, D., Sullivan, EJ, Cronin, MT, Holmes, CP, and Fodor, SPA. Light generated oligonucleotide arrays for rapid DNA sequence analysis. PNAS, in press.

Fodor, SPA, Lipshutz, RJ, Huang, X. DNA Sequnce by Hybridization. The Welch Foundation 37th Conference on Chemical Research. Houston TX, October 25-26, 1993.

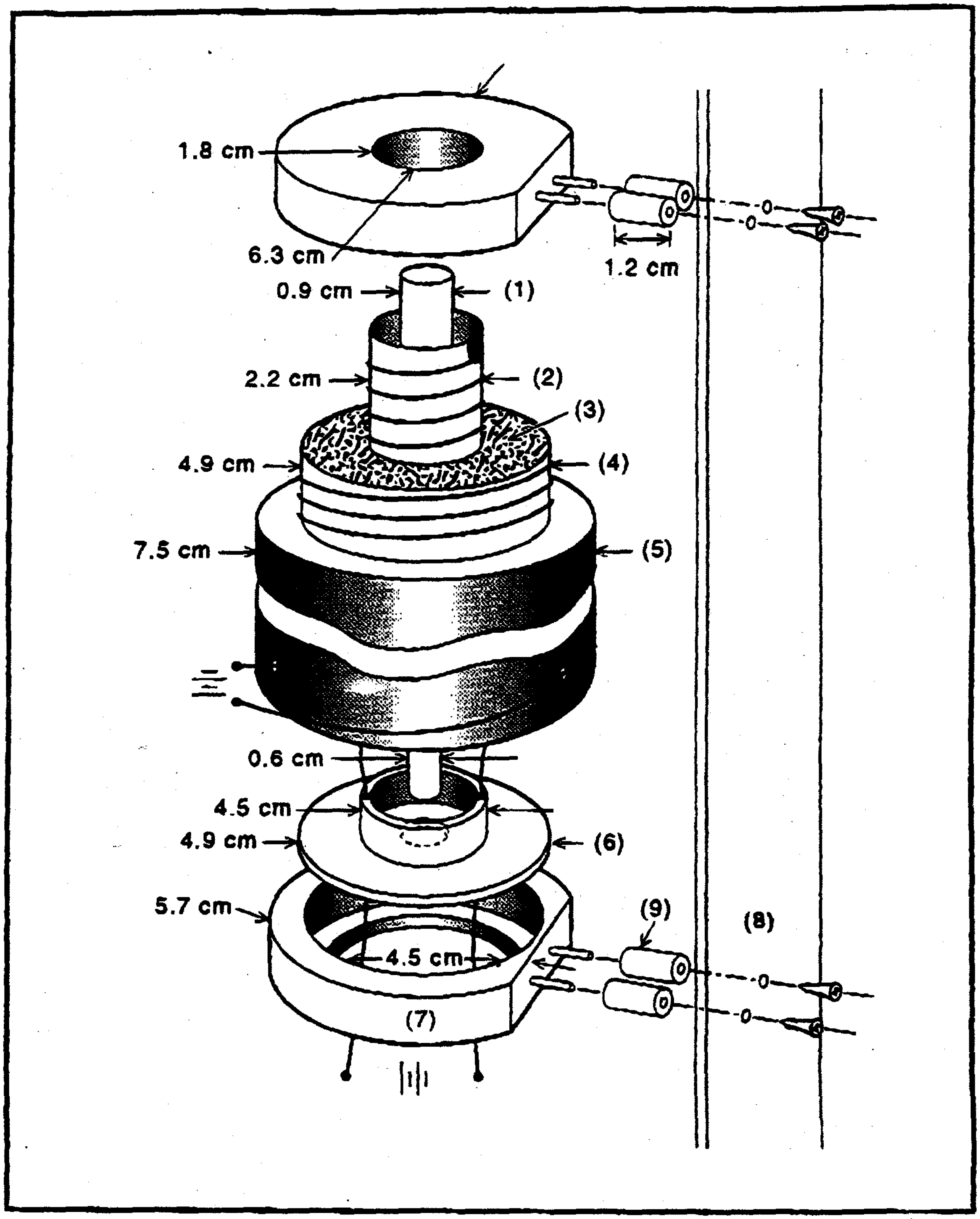
Cho,CY, Moran,EJ, Cherry, SR, Stephans, JC, Fodor, SPA, Adams, CL, Sindoram, A, Jacobs JW and Schultz, PG.(1993) An unnatural biopolymer. Science, 261, 1303-1305.

Fodor, SPA, Rava, RP, Huang, XC, Pease, AC, Holmes, CP and Adams, CL. Multiplexed biochemical assays with biological chips. (1993) Nature, 364, 555-556.

Roznyai, LF, Benson, DR, Fodor, SPA and Schultz, PG. Photolithographic immobilzation of biopolymers on solid supports. (1992) Agnew. Chem Int.Engl. Ed.., 31, 759-761.

SESSION II:

ASSEMBLY INSTRUCTIONS – ANALYSIS OF GENOMIC SEQUENCE DATA



Genome Linguistics

David Searls, Departments of Genetics and Computer & Information Science
University of Pennsylvania.

The metaphor of DNA as language can be extended to the analysis of the genome, by taking advantage of the wide array of tools and techniques that mathematicians and computer scientists have developed for dealing with languages. We have created a grammar formalism for describing higher-order features of sequence data, including gene structure, which can then be used by a general-purpose program called a parser to search for regions satisfying the specifications of a particular grammar. Aspects of DNA make it linguistically complex, in terms of a classification called the Chomsky hierarchy of languages; this new grammar and parser system is uniquely suited to the challenges of the DNA domain. Recent results will be presented in finding tRNA and protein-encoding genes, with results comparable to special-purpose gene-finders such as GRAIL.

Searls, D.B.: The Linguistics of DNA. American Scientist 80:579-591, 1992.

Searls, D.B.: The Computational Linguistics of Biological Sequences. In Artificial Intelligence and Molecular Biology (L. Hunter, ed.), AAAI Press, chapter 2, 47-120, 1993.

Searls, D.B. and Dong, S.: A Syntactic Pattern Recognition System for DNA Sequences. In Proceedings of the Second International Conference on Bioinformatics, Supercomputing, and Complex Genome Analysis (H.A. Lim, J. Fickett, C.R. Cantor, and R.J. Robbins, eds.), World Scientific Publishing Co., 89-101, 1993.

Combining Neural Networks and Expert Systems to Identify Features in DNA Sequences

R.J. Mural, Biology Division, Oak Ridge National Laboratory

Before we begin to assemble the Human Genome it is helpful to be able to recognize the pieces. We have used an approach which combines neural networks and expert systems to locate a variety of features in human DNA sequences. One of our goals has been to develop a system which accurately identifies features of biological interest in DNA sequences. The first tool we developed was GRAIL, an on–line e–mail service, which locates the protein coding regions of DNA sequences. This interface utilizes a multiple sensor–neural network to find protein coding regions and a rule based interpreter to reduce this output to a table which includes other useful information such as the identity of the coding strand and the preferred reading frame of the hypothetical coding exon.

We have recently developed a client–server version of GRAIL which allows the user to interactively explore many features of genomic DNA sequence. This tool assembles predicted coding regions, within a user specified interval, into gene models, allows for database searches of the translation of gene models and locates a variety of sequence features such as potential poly–A addition sites and various classes of repetitive DNA elements, providing an environment to facilitate the annotation of new genome sequence.

Brunack. S., Englebrecht, Jl and Knudsen, S. 1990. Neural Network Detects Errors in the Assignment of mRNA Splice Sites. *Nucleic Acids Res.* 18: 4797–4801.

Ficket, J.W. 1982. Recognition of Protein coding Regions in DNA Sequences. *Nucleic Acid Res.* 10: 5303–5318.

Fields, C., and C.A. Soderlund. 1990. Gm: A Practical Tool for Automating DNA Sequence Analysis. *CANIOS* 6: 263–270.

Gelfand, M. S. 1990. Computer Prediction of the Exon–Intron Structure. *J. Mol. Biol.* 226: 141–157.

Guigo, R., Knudsen, S., Drake, N. and Smith, T. 1992. Prediction of Gene Structure. *J. Mol. Biol.* 226: 141–157.

Hutchinson, G.B. and Hayden. M.R. 1992. The prediction of Exons Through an Analysis of Spliceable Open Reading Frames. *Nucleic Acids Res.* 20: 3453–3462.

Mural, R.J., Einstein, J.R., Guan, X., Mann, R.C. and Uberbacher, E.C. 1992. An Artificial Intelligence Approach to DNA Sequence Feature Recognition. *Trends in Biotechnology* 10: 67–69.

Snyder, E.E. and Stormo, G. D. 1993. Identification of coding regions in genomic DNA sequences: and application of dynamic programming and neural networks. *Nucleic Acids Res.* 21: 607–613.

Uberbacher, E.C. and Mural R.,J. 1991. Locating Protein–Coding Regions in Human DNA Sequences by a Multiple Sensor–Neural Network Approach. *Proc. Natl. Acad. Sci., USA* 88: 11261–11265.

Ancient Conserved Regions: Implications for Gene Identification

Phil Green, Genetics Dept.,Washington University School of Medicine, St. Louis, Missouri.

Over 2/3 of the genes being discovered in current genome sequencing projects are not similar to anything in the sequence databases. It has commonly been assumed that this reflects the relative incompleteness of the databases, and that when the genome sequences are complete most genes will have readily identifiable homologues in other organisms. However an alternative possibility is that the majority of genes are evolving too rapidly to retain detectable similarities over long evolutionary periods.

To investigate this, sets of unselected gene sequences from nematode (1472 ESTs, and 234 predicted genes from > 1 Mb of genomic sequence; C. elegans sequencing consortium), yeast (182 ORFs on S. cerevisiae chromosome 3; S. Oliver et al.), and human (2644 brain ESTs; C. Venter et al.), were compared to each other and to a set of 1916 E. coli genes, in order to detect ancient evolutionarily conserved regions (ACRs) in the encoded proteins. Most of the ACRs so identified — 98% (79/81) of the "prokaryote–eukaryote" ACRs, and 83% (24/29) of the eukaryote–specific ACRs — were found to be homologous to sequences in the protein sequence databases. This suggests that currently known proteins already include representatives of most ACRs, which in turn implies that sequences not similar to any current database sequence are unlikely to contain ACRs. Our analyses also indicate that 2/3 of the ACRs currently represented in SWISSPROT correspond to known conserved regions catalogued in BLOCKS, and yield estimates of 730 for the number of ACRs in SWISSPROT, and 860 for the total number of ACRs. In C. elegans each ACR is, on average, represented in about 7 different genes. Analyses using the worm ESTs suggest that moderately expressed genes are more likely to contain ACRs (and in general are more highly conserved) than rarely expressed genes.

Thus it appears that the majority (60 %) of genes are either phylum–specific or are evolving relatively rapidly, so that functional homologues in distantly related organisms may not be identifiable on the basis of sequence alone. It may be necessary to sequence fairly closely related model organisms (e.g. mouse) to identify homologues for many human genes.

Green P, Lipman D, Hillier L, Waterston R, States D, Claverie J. Ancient conserved regions in new gene sequences and the protein databases. Science 259: 1711–1716 (1993).

Sulston J, Du Z, Thomas K Wilson R, Hillier L, Staden R, Halloran N, Green P, Thierry–Mieg J, Qiu L, Dear S, Coulson A, Craxton M, Durbin R Berks M, Metzstein M, Hawkins T, Ainscough R, Waterston R: The C. elegans genome sequencing project: A beginning. Nature 356: 37–41 (1992).

Waterston R, Martin C, Craxton M, Huynh C, Coulson A, Hillier L, Durbin R, Green P, Shownkeen R, Halloran N, Metzstein M, Hawkins T, Wilson R, Berks M, Du Z, Thomas K Thierry–Mieg J, Sulston J: A survey of expressed genes in Caenorhabditis elegans. Nature Genetics 1, 114–123 (1992).

SESSION III:

TROUBLE SHOOTING – UNDERSTANDING HUMAN GENETIC DISEASE

Section III. INTERMEDIATE TROUBLESHOOTING

3-6. GENERAL.

a. This section contains intermediate troubleshooting information for locating and correcting most of the operating troubles which may develop

Each malfunction for the individual component, unit, or system is followed by a list of tests or inspections which will help you to determine the corrective actions to take. You should perform the tests/inspections and corrective actions in the order listed.

b. This manual cannot list all malfunctions that may occur, nor all tests or inspections and corrective actions. If a malfunction is not listed or is not corrected by listed corrective actions, see individual repair sections for maintenance instructions on each major assembly

3-7. TROUBLESHOOTING PROCEDURES.

Refer to troubleshooting table for malfunctions, tests, and corrective actions. The symptom index is provided for a quick reference of symptoms covered in the table.

Chromosome 21: Its Associated Genetic Diseases And Its Place In The Human Genome Project

Katheleen Gardiner, Eleanor Roosevelt Institute

Chromosome 21 has been of interest in the human genome project for several reasons:
i) It harbors genes associated with the developmental anomalies seen in Down Syndrome.
ii) it contains genes associated with Familial Alzheimers Disease, Familial Amyotrophic Lateral Sclerosis, Progressive Myoclonus Epilepsy, leukemia rearrangements, and the human homologue of the mouse Weaver mutation.
iii) it serves as a prototype in many mapping efforts (including YAC contigs and transcriptional maps) because of its small size and abundant resources, and
iv) it provides unique opportunities for investigation of genome organizational features.

The association with Down Syndrome (DS) is of particular interest because of the complexity of the phenotype, its developmental nature and its origin in the extra copy of perfectly normal genes. While most often associated with mental retardation, important phenotypic aspects also include heart defects, increased risk of leukemia, immune system deficiencies, and defects in the urogenital system.

The development of genotype-phenotype correlations in Down Syndrome, and the identification of other chromosome 21 disease genes would be facilitated by a high resolution physical map. Much progress towards this goal has been made. Currently, a long range NotI restriction map is essentially complete, and a complete YAC contig is being verified. In addition, detailed pulsed field analysis has been carried out in several regions. Correlations of these data with the cytogenetic map, with base composition, and with CpG island and gene densities are possible, and lead to insights into general features of human genome organization. Efforts towards construction of a complete transcriptional map have also begun.

Development of the physical map, and how it has and is affecting identification of chromosome 21 associated disease genes will be discussed.

Chumakov I, Rigault P, Guillou S, Ougen P et al. (1992) A continuum of overlapping clones spanning the entire human chromosome 21q. Nature 359: 380–387.

Epstein CJ, Korenberg JR, Anneren G, Antonarakis SE et al. (1991) Protocols to establish genotype-phenotype correlations in Down Syndrome. Am. J. Hum Genet. 49: 207–235.

Gardiner K, Horisberger M, Kraus J, Tantravahi U et al. (1990) Analysis of human chromosome 21: correlation of physical and cytogenetic maps; gene and CpG island distributions. EMBO J 9: 25–34.

Goate A, Chartier-Harlan MC, Mullan M, Brown J et al. (1991) Segregation of a missense mutation in the amyloid precursor protein gene with familiar Alzheimer's disease. Nature 349: 704–706.

Ichikawa HJ, Hosoda F, Arai Y, Kimiko S, Ohira M and Ohki M. (1993) A notI restriction map of the entire long arm of human chromosome 21. Nature Genetics 4:361–365.

Lehesjoki AE, Koskiniemi M, Sistonen P, Miao J et al. (1991) Localization of a gene for progressive myoclonus epilepsy to chromosome 21q22. PNAS 88: 3696–3699.

Levy E, Carman MD, Fernandez-Madrid IJ, Power MD et al. (1990) Mutation of the Alzheimer's disease amyloid gene in hereditary cerebral hemorrage, Dutch type. Science 248: 1124–1126.

Rosen DR, Siddique T, Patterson D, Figlewicz DA et al. (1993) Mutations in Cu/Zn superoxide dismutase gene are associated with familiar amyotrophic lateral sclerosis. Nature 362:59–62.

Siddique T, Figlewicz DA, Pericak-Vance MA et al. (1991) Linkage of a gene causing familial amyotrophic lateral sclerosis to chromosome 21 and evidence of genetic-locus heterogeneity. N. Engl. J. Med. 324(20): 1318–1384.

Tassone F, Cheng S, and Gardiner K. (1992) Analysis of chromosome 21 yeast artificial chromosome (YAC) clones. Am. J. Hum. Genet. 51: 1251–1264.

Triplet Repeat Diseases and Genomic Imprinting

Charles D. Laird, Department of Medicine, University of Washington

Two current themes in the study of human disease are genomic imprinting and triplet repeat expansions. Genomic imprinting has been observed in many organisms. In contrast, triplet repeat expansions have only been described in humans as a major mutational mechanism. The co-occurrence of these phenomena in several human diseases, most notably fragile-X syndrome, raises the possibility that they are causally related. Genes for two of the triplet repeat diseases correspond to chromosomal fragile sites, which may serve as predisposing breakpoints for chromosome evolution in primates. The replication properties of DNA at these sites offer clues to the possible connection between imprinting and DNA instability. These two current themes highlight astonishingly dynamic aspects of the human genome.

Caskey, C.T., Pizzuti, A., Fu, Y-H., Fenwick, R.G. Jr., and Nelson, D.L. 1993. Triplet repeat mutations in human disease. *Sci.* 256: 784–789.

Hansen, R.S., Canfield, T.K., Lamb, M.M., Gartler, S.M., and Laird.C.D. 1993. Association of fragile-X syndrome with delayed replication of the FMR1 gene. *Cell* 73: 1403–1409.

Laird, C.D., Jaffe, E., Karpen, G., Lamb, M., and Nelson, R. 1987. Fragile sites in human chromosomes as regions of late-replicating DNA. *Trends in Genet.* 3: 274–281.

Laird, C.D. 1987. Proposed mechanism of inheritance and expression of the human fragile-X syndrome of mental retardation. *Genetics* 117: 587–599.

Miro, R., Clemente, E.C., Fuster, C., and Egozcue, J. 1987. Fragile sites, chromosome evolution, and human neoplasia. *Hum. Genet.* 75: 345–349.

Mapping Genetic Diseases

Mary-Claire King, University of California at Berkeley

Cancer is genetic, in the sense that it is caused by DNA alterations at the cellular level. On the other hand, the most important risk factors for the common cancers are environmental: cigarette smoking, environmental pollution, occupational exposures, poor diet, and so on. These two observations are not in conflict: the DNA alterations that lead to cancer are very likely to be caused by environmental mutagens. It would be valuable to know exactly what genes are altered to cause a specific cancer, because the effects of these alterations might then be reversible before cancer has a chance to develop. A key to identifying these cancer genes may lie with rare families at extremely high risk of a specific cancer. Unlike most cancer patients, members of these families may inherit an alteration that confers increased susceptibility to cancer. In these rare instances, cancer is a genetic disease at the level of the family, as well as at the level of the cell. Therefore, in these families, genes predisposing to cancer can be mapped in the same way as genes for purely genetic diseases like sickle cell anaemia, cystic fibrosis, and Huntington's disease. The hypothesis that underlies the mapping of cancer genes in families is that the genes inherited in altered form in these rare families are the same genes that are altered in somatic cells of individuals without a remarkable family history of cancer. This hypothesis has proved correct for retinoblastoma. Genes responsible for other rare cancers have been mapped in families as well: neurofibromatosis, multiple endocrine neoplasia, Wilms' tumour, and colon cancer following familial adenomatous polyps, among others. Genes responsible for common cancers, most notably breast cancer and colon cancer, are also being defined by genetic analysis.

(Abstract adapted from Cancer Survey (1990) 9(3):417–35)

Hall JM, Friedman L, Guenther C, Lee MK, Weber JL, Black DM, King MC Closing in on a breast cancer gene on chromosome 17q. In: Am J Hum Genet (1992 Jun) 50(6):1235–42

Hall JM, Zuppan PJ, Anderson LA, Huey B, Carter C, King MC Oncogenes and human breast cancer. In: Am J Hum Genet (1989 Apr) 44(4):577–84

Shen FM, Lee MK, Gong HM, Cai XQ, King MC, Complex segregation analysis of primary hepatocellular carcinoma in Chinese families: interaction of inherited susceptibility and hepatitis B viral infection. In: Am J Hum Genet (1991 Jul) 49(1):88–93

Zuppan P, Hall JM, Lee MK, Ponglikitmongkol M, King MC Possible linkage of the estrogen receptor gene to breast cancer in a family with late-onset disease. In: Am J Hum Genet (1991 Jun) 48(6):1065–8

King MC, Rowell S, Love SM, Inherited breast and ovarian cancer. What are the risks? What are the choices? In: JAMA (1993 Apr 21) 269(15):1975–80

Ottman R, Pike MC, King MC, Henderson BE, Practical guide for estimating risk for familial breast cancer. In: Lancet (1983 Sep 3) 2(8349):556–8

Leon PE, Raventos H, Lynch E, Morrow J, King MC The gene for an inherited form of deafness maps to chromosome 5q31. In: Proc Natl Acad Sci U S A (1992 Jun 1) 89(11):5181–4

Newman B, Austin MA, Lee M, King MC Inheritance of human breast cancer: evidence for autosomal dominant transmission in high–risk families. In: Proc Natl Acad Sci U S A (1988 May) 85(9):3044–8

Cavalli–Sforza LL, Wilson AC, Cantor CR, Cook–Deegan RM, King MC Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the Human Genome Project. In: Genomics (1991 Oct) 11(2):490–1

King MC Genetic And Epidemiological Analysis Of Cancer In Families: Breast Cancer As An Example In: Inheritance of Susceptibility to Cancer in Man. Bodmer WF, ed. New York, Oxford University Press, 1983. (1983):33–46

Ottman R, Pike MC, King MC, Casagrande JT, Henderson BE, Familial breast cancer in a population–based series. In: Am J Epidemiol (1986 Jan) 123(1):15–21

SESSION IV: DISCLAIMERS – ETHICAL, LEGAL AND SOCIAL ISSUES

Diverse Human Genomes

Kenneth K. Kidd, Department of Genetics, Yale University Medical School

In the mid 1960s gel electrophoresis of various proteins extended to most species the conclusion, previously known clearly for only a few species, that genetic variation was extensive in normal populations and there was no such thing as a "wild type." Extensive normal genetic, and hence DNA sequence, variation is an aspect of virtually all species and certainly is a characteristic of humans. Thus, the genetic constitution of a species – its "genome" – is not just the DNA sequence of a single copy of each chromosome but also includes the amount, the nature, and the distribution among individuals of DNA sequence variation.

This talk will review some recent studies of nuclear DNA variation in human populations from around the world and discuss both what these "pilot" studies have shown and what we can expect to learn from more extensive studies in the future. Finally, the developing plans for a coordinated, collaborative international project – the Human Genome Diversity Project – will be discussed.

Barr, C. and Kidd, K.K. 1993. Population frequencies of the A1 allele at the dopamine D2 receptor locus. *Biological Psychiatry* 34: 204–209.

Bowcock, A.M., Rulz–Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R., and Cavalli–Sforza, L.L 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368: 455–457.

Bowcock, A.M., Kidd, J.R., Mountain, J.L., Hebert, J.M., Carotenuto, L., Kidd, K.K., and Cavalli–Sforza, L.L. 1991. Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proc. Nat. Acad. Sci., USA* 88: 839–943.

Cavalli–Sforza, L.L., Kidd, J.R., Kidd, K.K., Bucci, C., Bowcock, A.M., Hewlett, B., and Friedlaender, J.S. 1986. DNA markers and genetic variation in the human species. *Cold Spring Harbor Symposia on Quantitative Biology*, Vol. LI: Molecular Biology of Homo sapiens, pg. 411–417.

Cavalli–Sforza, L.L., Wilson, A.C., Cantor, C.R., Cook–Deegan, R.M., and King, M.C. 1991. Call for a world-wide survey of human genetic diversity: a vanishing opportunity for the human genome project. *Genomics* 11: 490–491.

Cavalli–Sforza, L.L. and Piazza, A. 1993. Human genomic diversity in Europe: a summary of recent research and prospects for the future. *European J. of Human Genetics* 1: 3–18.

Gillis, A.M., 1994. Getting a picture of human diversity. *BioScience* 44: 8–11.

Giuffra, L.A., Lichter, P., Wu, J., Kennedy, J.L., Pakstis, A.J., Rogers, J., Kidd, J.R., Harley, H., Jenkins, T., Ward, D.C., and Kidd, K.K. 1990. Genetic and physical mapping and population studies of a fibronectin receptor β–subunit–like sequence on human chromosome 19. *Genomics* 8: 340–346.

Kidd, J.R., Black, F.L., Weiss, K.M., Balazs, I., and Kidd, K.K. 1991. Studies of three Amerindian populations using nuclear DNA polymorphisms. *Human Biology* 63: 775–795.

Kidd, J.R., kidd, K.K., and Weiss, K.M.. 1993. Human genome diversity initiative. *Human Biology* 65: 1–6.

Mountain, J.L, Lin, A.A., Bowcock, A.M., and Cavalli-Sforza, L.L. 1992. Evolution of modern humans: evidence from nuclear DNA polymorphisms. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences* 337(1280): 159–165.

Weiss, K.M., Kidd, K.K., and Kidd, J.R. 1992. Human genome diversity project. *Evolutionary Anthropology* 1(3): 80–82.

Genetics and Sexual Orientation

Dean Hamer, National Cancer Institute, National Institutes of Health

Is human sexuality influenced by the genome? To approach this question, we have been studying the role of genes in sexual orientation by pedigree segregation analysis, candidate gene scanning, and DNA linkage studies. In men, we found increased rates of same-sex orientation in the brothers and maternal male relatives of homosexual male probands, but not in their fathers or paternal relatives, suggesting the possibility of sex-linked transmission in a portion of the population. DNA linkage analysis of selected families in which there were two gay brothers and no indication of nonmaternal transmission revealed a statistically significant correlation between homosexual orientation and the inheritance of polymorphic markers on chromosome region Xq28. We are currently attempting to replicate and refine this linkage and to identify the relevant gene(s). In women, we have observed a significant familiality of same-sex orientation. Whether the hypothetical genes act directly, for example by influencing sexually dimorphic brain circuits, or indirectly through personality or temperamental factors, in unknown.

What are the ethical, legal, and social implications of scientific research on human sexuality? This has been the subject of considerable, and healthy, debate. I believe that is would be fundamentally unethical to attempt to genetically asses or alter sexual orientation. At the same time, it is important that we not allow unwarranted fears and speculation to hinder further research in this area. The AIDS epidemic has taught us, all too bitterly, that we have more to fear from our ignorance than from new knowledge about human sexuality.

D. Hamer, S.Hu, V. Magnuson, N. Hu, and A. Pattatucci. 1993. A linkage between DNA markers on the X chromosome and male sexual orientation. *Sci.* **261**: 321-327.

J. Macke, N. Hu, S. Hu, M. Bailey, V. King, T. Brown, D. Hamer, and J. Nathans. 1993. Sequence variation in the androgen receptor gene is not a common determinant of male sexual orientation. *Am. J. Hum. Genet.* 53: 844-852.

S. LeVay and D. Hamer. 1994. Evidence for a biological influence on male homosexuality. *Scientific American*, in press.

The NIH–DOE ELSI Program

Micheal Yesley, Los Alamos National Laboratory

Congress has allocated a portion of the budget for the Human Genome Project at NIH and DOE to research and education efforts related to the ethical, legal and social issues ("ELSI") raised by advances in genetic knowledge and technology. These issues include privacy and discrimination concerns, and the challenges of integrating genetic advances into health care.

Protecting genetic privacy is a complex matter because the broader category of personal medical information is not well protected and because many third parties, e.g., relatives, health care providers and reimbursers, employers, insurers, public health authorities and researchers, may have valid claims to identifiable genetic information under certain circumstances. Further, it is not clear how much genetic privacy is desired by individuals.

In addition to protecting the privacy of genetic information, policy measures, including laws, regulations and professional guidelines, may prohibit improper uses of the information. The federal government and many states have adopted, or are considering, a variety of measures to bar genetic discrimination.

As researchers identify the genes that cause or predispose to many disorders, the pressures grow to screen general populations. Severely limited counseling resources, equivocal information in the form of probabilities, and predictions of late–onset disorders that cannot be treated or prevented are a few of the many implementation issues.

The talk will describe the ELSI program and activities, and provide an overview of the above issues.