February 11, 2011

Robert Cook-Deegan, M.D.
Director, Center for Genomic Ethics, Law & Policy
Duke University
Institute for Genome Sciences & Policy
Durham, NC 27708

Re: FOIA Case Number: 11-FOI-00056-NHGRI - 38310

Dear Dr. Cook-Deegan:

This is our final response to your November 23, 2010, Freedom of Information Act (FOIA) request addressed to the National Human Genome Research Institute (NHGRI), National Institutes of Health (NIH). You requested: 1) copies of files and/or professional correspondence documenting or regarding involvement of NIH employees in the planning and execution of the International Strategy Meetings on Human Genome Sequencing, which took place in Bermuda in February 1996 and February-March 1997; 2) a list of individual attendees to the International Strategy meetings in 1996 and 1997; and 3) files maintained within the records of NIH/NHGRI staff who may also have attended. In a phone conversation on January 25, you amended to exclude all but: 1) a list of attendees for the 1996 meeting; 2) a list of attendees for the 1997 meeting; 3) information on Bermuda rules for sequence sharing; and 4) a final report following the meeting.

In response to item 2, enclosed is a list of attendees for the Second International Strategy Meeting on Human Genome Sequencing on February/March 1997 (7 pages); and in response to item 4, enclosed is the Report of the Second International Strategy Meeting on Human Genome Sequencing from the February/March 1997 meeting (33 pages). A total of 40 pages are being released with this response.

In response to items 1 and 3, the NHGRI, Office of the Director, Office of Policy, Communications, and Education searched its files, and no records responsive to that part of your request were located. While we believe that an adequate search of appropriate files was conducted for the records you requested, you have the right to appeal this determination that no records exist which would be responsive to that part of your request. Should you wish to do so, you must send your appeal within 30 days of receipt of this letter to the Deputy Assistant Secretary for Public Affairs (Media), Department of Health and Human Services (DHHS), Room 17-66, Parklawn Building, 5600 Fishers Lane, Rockville MD 20857, following the procedures outlined in Subpart C of the enclosed regulations, 45 CFR Part 5. Please mark both the envelope and the appeal letter “FOIA Appeal.”

In certain circumstances provisions of the FOIA and Department of Health and Human Services FOIA Regulations allow us to recover part of the cost of responding to your request. Because the cost is below the $25 minimum, there is no charge for the enclosed materials.

Thank you for your interest in the National Human Genome Research Institute.

Sincerely,

Christy Cecil
Freedom of Information Specialist, NHGRI

Enclosures - 40 pages
Revised Delegates List
Dr Mark Adams  
The Institute for Genomic Research  
9715 Medical Center Drive  
Rockville MD 20850  
USA  
☎ 301 838 0200  
Fax: 301 838 0208  
E-Mail madams@tigr.org

Dr David Bentley  
The Sanger Centre  
Wellcome Trust Genome Campus  
Hinxton  
Cambridge CB10 1SA  
UK  
☎ 44(0) 1223 832244  
Fax: 44(0) 1223 494919  
E-mail db@sanger.ac.uk

Dr Elbert Branscomb  
Lawrence Livermore National Laboratory  
7000 East Avenue  
L-452 Livermore  
CA 94550  
USA  
☎ 510 422 5681  
Fax: 510 422 2282  
E-mail branscomb1@llnl.gov

Dr Graham Cameron  
The European Bioinformatics Institute  
Wellcome Trust Genome Campus  
Hinxton  
Cambridge  
CB10 1SD  
UK  
☎ 44(0) 1223 494444  
Fax: 44(0) 1233 494468  
E-Mail cameron@ebi.ac.uk

Dr Ellison Chen  
ACGT  
ABD-PE  
850 Lincoln Center Drive  
Foster City CA 94404  
USA  
☎ 415 638 5107  
Fax: 415 638 6177  
E-mail ChenEY@perkin.elmer.com

Dr Francis Collins  
National Institutes of Health  
National Human Genome Research Institute  
31 Center Drive MSC 2152  
Building 31, Room 4B09  
Bethesda MD 20892-2152  
USA  
☎ 301 496 0844  
Fax: 301 402 0837  
E-mail FC23A@nih.gov

Professor David Cox  
Stanford University School of Medicine  
Department of Genetics  
Room M336  
Stanford CA 94305  
USA  
☎ 415 725 8043  
Fax: 415 725 8058  
E-mail ne.mrm@forsythe.stanford.edu

Dr Pieter de Jong  
Rosewell Park Cancer Institute  
Human Genetics Department  
Elm & Carlton Streets  
Buffalo NY 14263  
USA  
☎ 716 845 3168  
Fax: 716 845 8449  
E-mail pieter@dejong.med.buffalo.edu
Dr Glen Evans  
University of Texas Southwestern Medical Center  
Human Growth & Development  
6000 Harry Hines Boulevard  
Dallas TX 75235-8591  
USA  
☎ 214 648 1660  
Fax: 214 648 1666  
E-mail gevens@utsw.swmed.edu

Dr Eric Green  
National Institutes of Health  
National Human Genome Institute  
49 Convent Drive  
Building 9, Room 2A08  
Bethesda MD 20892-4431  
USA  
☎ 301 402 0201  
Fax: 301 402 4735  
E-mail egreen@nhghr.nih.gov

Professor Asao Fujiyama  
National Institute of Genetics  
Division of Human Genetics  
1111 Yata  
Mishima  
Shizuoka 411  
Japan  
☎ 81 559 81 6796  
Fax: 81 559 81 6716  
E-mail afujiyama@lab.nig.ac.jp

Dr Phil Green  
University of Washington  
Department of Molecular Biotechnology  
GJ10  
4909 25th Avenue NE  
Seattle WA 98195  
USA  
☎ 206 685 4341  
Fax: 206 685 7344  
E-mail phg@u.washington.edu

Dr Fiona Francis  
Max-Planck-Institut Für Molekulare Genetik  
Ihnessstrasse 73  
Berlin 14195  
Germany  
☎ 49 30 8413 1612  
Fax: 49 30 8413 1380  
E-mail francis@mping-berlin-dahlem.mpg.de

Dr Richard Gibbs  
Department of Molecular and Human Genetics  
Baylor College of Medicine  
One Baylor Plaza  
BCMS-930  
Houston TX 77030  
USA  
☎ 713 798 6539  
Fax: 713 798 5741  
E-mail agibbs@bcm.tmc.edu

Dr Mark Guyer  
National Institutes of Health  
National Human Genome Research Institute  
38 Library Drive  
MSC 6050  
38A/604  
Bethesda MD 20892-6050  
USA  
☎ 301 402 5407  
Fax: 301 480 2770  
E-mail MG25M@NIH.GOV

Dr Trevor Hawkins  
Whitehead Institute/MIT Center for Genome Research  
One Kendall Square  
Building 300  
Cambridge MA 02139-1561  
USA  
☎ 617 252 1906  
Fax: 617 252 1933  
E-mail trevor-hawkins@gatormail.wi.mit.edu
Dr LaDeana Hillier  
Washington University School of Medicine  
Genome Sequencing Center  
Box 8501  
4444 Forest Park Avenue  
St Louis MD 631  
USA  
☎ 314 286 1803  
Fax: 314 286 1810  
E-mail rw@ncmatode.wustl.edu

Dr Thomas Hudson  
Whitehead Institute/MIT Center for Genome Research  
One Kendall Square  
Building 300  
Cambridge MA 02139  
USA  
☎ 617 252 1912  
Fax: 617 252 1902  
E-mail thudson@genome.wi.mit.edu

Dr Ursula Hurtenbach  
DLR - Projektträger des BMBF  
Sudstrasse 125  
D-53175 Bonn  
Germany  
☎ 49 228 3821 234  
Fax: 49 228 3821 237  
E-mail u.hurtenbach@dfr.de

Dr Elke Jordan  
National Institutes of Health  
National Center for Human Genome Research  
31 Center Drive - MSC 2152  
Building 31, Room 4B09  
Bethesda MD 20892-2152  
USA  
☎ 301 496 0844  
Fax: 301 402 0837  
E-mail EJIOD@NIH.GOV

Dr Jill Kent  
The Wellcome Trust  
183 Euston Road  
London  
NW1 2BE  
☎ 44(0) 611 8253  
Fax: 44(0) 611 6888  
E-mail j.kent@wellcome.ac.uk

Dr Ung-Jin Kim  
California Institute of Technology  
Division of Biology 147-75  
Pasadena CA 91125  
USA  
☎ 818 395 4091  
Fax: 818 796 7066  
E-mail ung@caltech.edu

Dr David Lipman  
National Center for Biotechnology Information  
National Institutes of Health  
National Library of Medicine  
Building 38A, 8N805  
8600 Rockville Pike  
Bethesda MD 20894  
USA  
☎ 301 496 2475  
Fax: 301 480 9241  
E-mail lipman@ncbi.nlm.nih.gov

Dr Richard McCombie  
Cold Spring Harbor Laboratory  
P.O.Box 100  
Cold Spring Harbor  
NY 11724  
USA  
☎ 516 367 8884  
Fax: 516367 8874  
E-mail mccombe@cshl.org
Professor John Mattick  
University of Queensland  
Centre for Molecular and Cellular Biology  
St Lucia  
Brisbane Queensland  
Australia 4072

Tel: 617 3365 4446  
Fax: 617 3365 4388  
E-mail: j.mattick@cmcb.uq.edu.au

Dr Catherine Moody  
Medical Research Council  
20 Park Crescent  
London  
W1N 4AL  
UK

Tel: 44(0) 171 636 5422  
Fax: 44(0) 171 436 6179  
E-mail: catherine.moody@hq.mrc.ac.uk

Dr Michael Morgan  
The Wellcome Trust  
183 Euston Road  
London  
NW1 2BE  
UK

Tel: 44(0) 171 611 8432  
Fax: 44(0) 171 611 8688  
E-mail: 100071.72@compuserve.com

Dr Michael Palazzolo  
Lawrence Berkeley National Laboratory  
One Cyclotron Road  
MS 74-157  
Berkeley CA 94720  
USA

Tel: 510 486 6932  
Fax: 510 486 6816  
E-mail: michaelp@genome.lbl.gov

Dr Ari Patrinos  
Health Effects & Life Sciences Research Division  
US Department of Energy  
19901 Germantown Road  
Germantown MD 20874-1290  
USA

Tel: 301 903 3251  
Fax: 301 903 5051  
E-mail: ari.patrinos@oer.doc.gov

Dr Jane L. Peterson  
National Institutes of Health  
National Center for Human Genome Research  
Room 614, Building 38A  
Bethesda MD 20892  
USA

Tel: 301 496 7531  
Fax: 301 496 2770  
E-mail: jane_peterson@nih.gov

Professor Bruce Roe  
The University of Oklahoma  
Chemistry Department  
620 Parrington Oval, Room 208  
Norman  
Oklahoma 73069  
USA

Tel: 405 325 4912  
Fax: 405 325 7762  
E-mail: broe@ou.edu

Dr Jane Rogers  
The Sanger Centre  
Wellcome Genome Campus  
Hinxton  
Cambridge, CB10 1SA  
UK

Tel: 44(0) 1223 834244  
Fax: 44(0) 1223 494919  
E-mail: jrh@sanger.ac.uk
Professor André Rosenthal  
Institute of Molecular Biotechnology  
Department of Genome Analysis  
Beutenbergstrasse 11  
D-07745 Jena  
Germany

Phone: 49 3641 65 62 40  
Fax: 49 3641 65 62 55  
E-mail arosenth@imb-jena.de

Dr Barbara Skene  
The Wellcome Trust  
183 Euston Road  
London  
NW1 2BE  
UK

Phone: 44(0) 171 611 8232  
Fax: 44(0) 171 611 6866  
E-mail b.skene@wellcome.ac.uk

Mrs Jilly Steward  
The Wellcome Trust  
183 Euston Road  
London NW1 2BE  
UK

Phone: 44(0) 171 611 8899  
Fax: 44(0) 171 611 8237  
E-mail j.steward@wellcome.ac.uk

Dr John Sulston  
The Sanger Centre  
Wellcome Trust Genome Campus  
Hinxton  
Cambridge CB10 1SA  
UK

Phone: 44(0) 1223 834244  
Fax: 44(0) 1223 494919  
E-mail jes@sanger.ac.uk

Professor Gert-Jan Van Ommen  
University of Leiden  
Department of Human Genetics  
P.O.Box 9503  
2300 RA Leiden  
The Netherlands

Phone: 31 71 5276293  
Fax: 31 71 5276075  
E-mail gvanomme@ruy46.medfac.leidenuniv.nl

Dr Craig Venter  
The Institute for Genomic Research  
9712 Medical Center Drive  
Rockville, MD 20850  
USA

Phone: 301 838 3500  
Fax: 301 838 0209  
E-mail jcventer@tigr.org

Ms Susan Wallace  
HUGO Americas  
7986D Old Georgetown Road  
Bethesda MD 20814  
USA

Phone: 301 654 1477  
Fax: 301 652 3368  
E-mail hugo@gdb.org

Dr Robert H. Waterston  
Washington University School of Medicine  
Genome Sequencing Center  
Box 8501  
4444 Forest Park Avenue  
St Louis MO 631  
USA

Phone: 314 286 1803  
Fax: 314 286 1810  
E-mail rw@nematode.wustl.edu
Professor Sir David Weatherall
Institute of Molecular Medicine
John Radcliffe Hospital
Headington
Oxford
OX3 9DU

☎ 44 1865 222359
Fax: 44 1865 222501
E-mail janet.watt@mailgate.jr@ox.ac.uk

Dr Jean Weissenbach
CNRS URA 1922
Genethon
1 rue de l'internationale
BP 60 - 91002 Evry Cedex
France

☎ 33 1 69 47 28 00
Fax: 33 1 60 77 86 98
E-mail jsbach@genethon.fr
CONFIDENTIAL

Report of the Second International Strategy Meeting on Human Genome Sequencing held at the Hamilton Princess Hotel, Bermuda, on 27th February - 2nd March 1997

Summary

- The principles enunciated at the first International Strategy meeting, of rapid data release and public access to the primary genomic sequence, were reaffirmed.

- Scientists and funding agencies should take the necessary steps to ensure that the principles are adhered to by all participating organisations.

Sequence Quality Standards
The following standards were agreed:

- The nucleotide error rate should be 1 error in 10,000 bases or less for most sequence.
- Assemblies should be verified by restriction digest using two or more restriction enzymes.
- Gaps in sequence. The agreed long term goal is no gaps, recognising that this is not yet routine.
- Closing gaps is the responsibility of the original sequencer.

The following proposals were endorsed by the participants:

- It was agreed that a useful trial to assess sequence accuracy would be to perform a data exchange exercise. Raw sequence data would be exchanged among sequencing centres, centres would reassemble the data and identify outright discrepancies or ambiguities with reference to the sequence submitted to the database. These would be resolved by further consultation or resequencing. The same data sets would be sent to two centres which would hopefully engender competition to detect errors.
- All sequence reads should be archived in a retrievable form.
- Sequencing centres should define explicitly how error rates and costs have been calculated.
Sequence Submission and Annotation

Sequence data should be classified simply as “finished” or “unfinished” and should be stored in distinct databases; consideration should be given to establishing a public database for unfinished sequence data.

Sequence annotation should be standardised if possible, and include the following information:

- Error estimation such as PHRED AND PHRAP data.
- Enzymes used to verify assembles, and sizes of fragments produced.
- Exact details on how to assemble adjacent clones, with a minimum of 100 bp of overlapping (preferably unique) sequence between clones for verification.
- Gaps must be sized and the surrounding sequence oriented and ordered. The methods used for sizing, and reasons for not closing the gap should be stated.
- If features such as coding sequence and splice sites are included in the annotation, it should be stated if they were identified experimentally or by computer predictions.
- Unfinished sequence; it should be stated how near the sequence is to completion.

Potential development of a database listing all gaps in ‘finished’ sequence.

Sequence Claims and Etiquette

Mapping investment does not automatically entitle sequencing claims over the same region until a sequence ready map has been generated.

Potential conflicts with other sequencers to be resolved by early communication.

Collaborations with groups with a biological interest in a region should be subject to the same principles of data release and communication.

Investigate whether the Human Sequence Map Index should be relocated to be more closely associated with the other major human sequence databases.

Claims allowed on the Index:

- Duration - maximum 1 year.
- Size of region - minimum 1 Mb; regions to be defined by Genethon markers if possible, other agreed and available markers if not.
- Maximum amount - in the order of three times the sequence released by the centre in the preceding year.
- Sequence claims must span the entire region between, and including, the delimiting markers.

Next Meeting

- To be held at the end of February 1998 in Bermuda (dates to be confirmed)
Aims of the Meeting

To discuss current progress, effectiveness of strategies, quality standards and evaluation of quality, data release and allocation of genomic regions for sequencing.

Introduction

The meeting had been sponsored jointly by the Wellcome Trust, the NIH and the US Department of Energy. Participants were welcomed to the Second International Strategy Meeting by Dr Michael Morgan from the Wellcome Trust who gratefully acknowledged the contributions from the other sponsors.

Session I

Progress, Strategies and Developments

Chair: David Cox

The aim of this session was for each sequencing group to present a progress report addressing the effectiveness of their strategies for constructing sequence-ready maps and producing finished sequence.

John Sulston,
The Sanger Centre, Cambridge

John Sulston summarised the main human sequencing targets at the Sanger Centre. Initial targets had focused on regions of the X chromosome (90 Mb) and chromosome 22 (25 Mb) in collaboration with the Genome Sequencing Centre at St. Louis. Work was now proceeding on chromosomes 1 (300 Mb), 6 (160 Mb) and 20 (80 Mb) with the greatest emphasis on chromosome 6.

The strategy has involved radiation hybrid mapping of STS markers to a defined density (currently 10-20 STSs/Mb); these markers were then used to screen PAC libraries which were assembled into contigs by fingerprinting and verified by STS content analysis. John Sulston emphasised the importance of software in the finishing process and also the potential of YAC sequencing for gap closure. With current funding commitments, the Sanger Centre has set a total human genomic sequencing target of 655 Mb of DNA. To date, 14.6 Mb had been finished and submitted to EMBL/GenBank, an additional 11.9 Mb of unfinished sequence was also available via ftp. The target for this year was to finish 30-40 Mb of human sequence, this target would be raised to 80-100 Mb in subsequent years. The increase in output would be facilitated by the transfer of production capacity from the nematode sequencing project to the human. The cumulative total of finished sequence, including nematode, human, yeasts and TB was 52.2 Mb; 34 Mb of this had been finished in the previous year.
It was reported that bacterial clone coverage was good for chromosome 22, with clones available for 19 Mb of the 25 Mb target region. The X chromosome project was also progressing well but there were a number of persistent gaps in Xq22 where YACs were deleted.

Robert Waterston,
Genome Sequencing Center, St. Louis

Bob Waterston summarised the main sequencing targets at the Genome Sequencing Centre, St. Louis. Regions of the X chromosome and chromosome 22 were being sequenced in collaboration with the Sanger Centre. Most of the sequencing efforts at St. Louis were focused on chromosome 7; the mapping of this chromosome was being performed in close collaboration with Eric Green. To date, 4.8 Mb of human sequence had been finished, of which 2.95 Mb had been submitted to the public databases.

The Genome Sequencing Centre used a similar strategy to that of the Sanger Centre to generate clones for sequencing. Eric Green had developed an STS map for chromosome 7 with an average marker density of 1 STS per 79 Kb. The STSs were used to identify clones, and restriction enzyme analysis was then used to determine overlaps and to pick a minimal tiling path for sequencing. Initial contig assembly and determination of the minimal tiling path had been semi-automated using a Molecular Dynamics Fluorimager together with software developed by the Sanger Centre.

In summary, the mapping and sequencing status of chromosome 7 was, that 600 STSs had been mapped over 50 Mb, and 128 BAC/PAC contigs (average size 250 Kb) had been constructed. 175 clones had been chosen for sequencing and 21 had been finished. The sequencing strategy was a shotgun directed strategy using a mixture of M13 and pUC clones. The software used included PLAN, PHRED and PHRAP developed by Phil Green for the initial shotgun stages followed by FINISH to carry out the initial directed stages automatically. CONSED had also been developed as an interactive editor in collaboration with Phil Green and David Gordon.

Quality control included verification of the sequence using three different restriction digests, reassembly of the sequence using alternative versions of PHRED and PHRAP and annotation of genes to highlight potential errors. Finished sequence was completely continuous with an error rate of less than 1 in 10,000.
Future developments included the production of software to replace human decision making, and the use of a central database to track all clones through the sequencing process. Efforts to automate some aspects of finishing included the development of a robot to re-array clones selected by the FINISH programme. In order to increase throughput, attempts had been made to convert the ABI 377s to run 64-72 lanes; this had required solving a number of technical problems. The Genome Sequencing Center was experimenting with the Amersham dye terminators (these consisted of the same dyes as ABI but with the Amersham enzyme) and had also begun to use the Ty1 transposon technology to disrupt regions which had been difficult to sequence through or to produce mapping information to assist with assembly. The Center was working with Lloyd Smith to develop his cheaper and more accurate sequencer to see if it was suitable for high throughput sequencing. The Center was also assessing the potential of capillary sequencers.

In response to questions about the output of finished sequence, Bob Waterston explained that the Center was in the process of scaling up their human sequencing effort and that until recently clone supply had been a limiting factor. The efficiency of finishing also represented a bottle-neck and they were working to improve this. Current output of finished human sequence was 1 Mb per month and it was anticipated that this would increase to 2 Mb per month over the following year.

Tom Hudson and Trevor Hawkins,
Whitehead Institute/MIT Center for Genome Research, Cambridge.

Tom Hudson summarised the mapping strategy at the Whitehead Institute which is STS-based using a 30,000 marker map with, on average, one marker every 100 Kb. In regions of high marker density (1.4 markers per BAC) initial screening of BAC and PAC libraries with 20-fold coverage have isolated clones which covered 94% of the region. However contigs are still very small with only 2.5 markers per contig. Even with high density markers, the strategy does not produce very large contigs; a high level of BAC-end sequencing and walking will be required to close gaps. In regions of lower marker density, a different strategy is being used. Single STS markers are used to identify BACs (usually 6-12 clones per STS are obtained). These are then validated by fingerprinting and, in some cases, used to select new STSs for walking. The Whitehead Institute have created very high density BAC pools which required only 70 PCRs to screen half a genome equivalent. 2,800 PCRs are required to screen a 20-fold coverage library with each marker. Using the Genomatron, 100 STSs can be screened per day; the rate of clone identification is much higher than can be accommodated with the current sequencing capacity.
Trevor Hawkins summarised current progress in sequence production. Up to 31st January 1997, 2.1 Mb had been finished, with a further megabase due to be released on the Web in March. The sequencing target for the year up to 1st May was 5 Mb. Initial sequence data had been obtained from various regions (particularly 9q34), but the future focus would be on chromosome 17 using BAC clones provided by the Whitehead mapping group. Most of the previous year had been spent in establishing the necessary infrastructure and developing automation to deal with high throughput production sequencing. Quality control systems had been developed to allow rapid identification of individual components that were operating sub-optimally.

Trevor Hawkins identified finishing as the major bottleneck. The Whitehead Institute were trying to develop methods to finish 80-90% of clones via a production line process. This included a system which assigned a numerical value to the status of individual clones relating to how close to “finished” they were.

In response to questions about the number of gaps present in finished sequence, Trevor Hawkins stated that in the current release of 2.1 Mb, there were 11 gaps (i.e. approximately 1 gap per 200 Kb). Trevor Hawkins confirmed that none of the finished sequence had been submitted to GenBank although it was accessible via the Web site. A submission of 1 Mb was planned for the following week.

Mark Adams,
The Institute for Genomic Research, Rockville.

Mark Adams explained that the TIGR human genomic sequencing initiative represented a collaboration between TIGR and Caltech, relying heavily on the Los Alamos STS map developed in Bob Moyzis’ group. The initial strategy involved sequencing 40 non-overlapping BACs which had been isolated from a 4-fold coverage library using 50 STS markers over a 30 Mb region of chromosome 16p. End probes from these BACs were used to select BACs with minimal overlaps for further sequencing. The rationale for this approach was that the STS map would be unlikely to provide a set of minimal clones for sequencing. This strategy therefore represented an alternative solution to the problem of gap closure by walking early on in the process. The selection of minimal overlaps also reduced the total amount of sequencing required to cover a given region.

Genomic Southern blots were used to verify that the DNA from the low coverage BAC library was representative of the human genome. BAC clones chosen for sequencing were also checked with STS markers and FISH mapping.
Currently TIGR sequence output was highest for bacterial genomes followed by the 
Arabidopsis and human genomes. The rate of human sequencing output was 
increasing with 2.6 Mb having been finished and submitted to GenBank in the first 
year. The second year target was 11 Mb. Scaling up production would be facilitated 
by the introduction of a new robot from SAIC (Allekto-DNA System), but this would 
not be available until 1998. Mark Adams identified information processing and 
management as the key issue in scaling up the finishing process. TIGR’s approach to 
the finishing problem focused on software development, particularly in relation to 
quality control. Differences in the results from two assembly programmes, PHRAP 
and TIGR assembler, were used to identify potential errors for further investigation 
by the closure team.

In the 2.6 Mb region of the short arm of chromosome 16 sequenced by TIGR, only 12 
genes had been identified (1 gene per 200 Kb) using five different gene prediction 
programmes including GRAIL and Genefinder. Given that only 5% of the genome 
was likely to be sequenced in the coming year, Mark Adams queried whether there 
should be greater focus on gene-rich regions as initial targets.

Richard Gibbs,
Baylor College of Medicine Human Genome Sequencing Center, Houston

Richard Gibbs reported that, to date, the Human Genome Sequencing Centre at 
Baylor had finished and submitted 3 Mb to GenBank; contigs ranged from 185 Kb to 
350 Kb. Initial objectives had been to reduce redundancy and improve costs. 
Finished sequence now required a total of 16.5 reads per Kb with costs of $1.35 per 
reaction. Cost reductions over the previous year were the product of small 
increments over all elements in the sequencing strategy. The introduction of 
BODIPY dyes, which had been developed at Baylor, for most production sequencing 
had also contributed to cost savings. The sequencing strategy used at Baylor was 
identical to that described by John Sulston at the Sanger Centre.

Richard Gibbs emphasised the value of full length cDNA sequencing for gene 
identification and gene structure determination. Analysis of a 200 Kb region on 
chromosome 12p13 had lead to the identification of twenty genes using experimental 
PCR, gene prediction programmes and comparison with full length cDNAs. Of 
these resources the information from full length cDNA sequences had proved the 
most valuable. The group at Baylor had sequenced 180 full length cDNAs using 
concatenation cDNA sequencing; this involved concatenating up to 70 cDNAs and 
then using shotgun sequencing to build up contigs representative of each cDNA.

The Baylor group was also involved in comparative sequence analysis and had 
sequenced regions of Xq22, Xq28 and chromosome 2 in the mouse genome. This had 
provided interesting data with respect to regulatory sequences but had been less 
informative than cDNA sequence in predicting gene structure.
Systems had been introduced to reduce redundant sequencing in overlapping regions and avoid “double finishing”. There was currently 1.3 fold redundancy of sequencing in overlapping regions.

Initial sequencing targets were Xpter, chromosome 12 (CD4 region) and regions of chromosome 3. Output in the current year would be 2.5 Mb and the target was to produce 15 Mb of finished sequence in the coming year, scaling up to 100 Mb in 1998. All data was immediately available on the Web site, but only sequence submitted to GenBank was cited as finished sequence.

The Baylor Human Genome Sequencing Center had recently established a number of collaborations with the Dallas Center and it was anticipated that the introduction of the SAGIAN robot from Dallas together with the use of Baylor BODIBY dyes would be significant factors in the scale-up process.

David Cox,
Stanford Human Genome Centre

David Cox reported that, to date, the Stanford Human Genome had submitted 100 Kb of finished sequence to GenBank. It had also submitted 1.2 Mb of unfinished sequence ranging in size from 3 Kb to 100 Kb. The primary target of the Centre was chromosome 4 with an overall goal of sequencing 200 Mb by 2005. Initial sequencing targets were 5 Mb on chromosome 4q25, 1.2 Mb in the EP1 region of chromosome 21 together with a smaller more proximal region (DS) of 400 Kb. David Cox considered it unlikely that the Center would meet its original target of 2.5 Mb in the first year but was confident of reaching the second year target of 5 Mb.

David Cox summarised the theoretical and practical utility of different radiation hybrid mapping strategies. He argued that in order to be cost-effective it was important to use high resolution mapping strategies to identify mapping gaps in advance of the sequencing process. The Stanford Center was using a very directed approach based on a high resolution map with markers ordered every 100 Kb. These markers were used to pull out BACs from a low redundancy library which were then fingerprinted to determine the coverage. The BACs were then sheared to 3 Kb to produce a 5-fold redundancy library of ca. 200 clones. These were then end-sequenced to identify minimal overlapping clones. The problem with this strategy was that it was not possible to determine the overall contiguity in the library until the sequencing process was almost complete. In order to assess the quality of libraries in advance, the Stanford group had developed Affymetrix chip technology to determine the minimal tiling path from the BAC sub-clones. The chips consisted of 25 bp oligonucleotides from the end sequences of each of the 200 sub-clones. The chip technology could also be used to check sequence assemblies to 1 Kb resolution and was particularly useful to adjudicate between alternative assemblies generated by different software programmes. The cost of this technology was 1.5 cents per base pair to determine the minimal tiling path and to check the assembly. The anticipated costs to do this for 20 Mb of sequence was estimated at $1 million. In response to questions about the robustness of this strategy, David Cox explained that the underlying strategy of the group was to use different technologies to develop hypotheses that could be further tested; none of the technologies were expected to provide the absolute answer.
Fiona Francis,  
MPI fuer Molekulare Genetik, Berlin

Fiona Francis explained that the Berlin group operated as part of a German consortium with an overall goal of sequencing 40 Mb over the next three years; the Berlin group aimed to produce 6 Mb of sequence in that time. The main sequencing target was chromosome 21 but given other international interests in this chromosome, it was likely that the German consortium would also target regions of the X chromosome and chromosome 17.

Chromosome 21 has a high density of STS markers and these had been used in non-radioactive hybridisation screening of chromosome 21 specific cosmid libraries and whole genome PAC and BAC libraries to build up contigs. RNA probes from the ends of contigs and cDNA probes had been used to select clones for fingerprinting and to construct a minimal tiling path for sequencing. The sequencing strategy used was a standard shotgun sequencing approach but based almost totally on PCR templates derived from pUC. This allowed the group to take advantage of existing colony picking and PCR robots. Sequence assembly and analysis was carried out using software packages provided by other sequencing centres, particularly the Sanger Centre.

The Berlin group had finished and submitted a contiguous sequence of 243 Kb from Xp22 containing the PEX gene. Other projects in progress included regions ranging from 150 Kb to 1 Mb on 21q22.3, Xq28, Xq13, Xq12 and 17p11 (totaling 2.6 Mb).

Fiona Francis described a technique developed by the group to reduce the redundancy in the shotgun sequencing process. PCR generated inserts were arrayed on a membrane and hybridised with short oligonucleotides (octamers) to produce a “barcode” for each shotgun clone. This “barcode” could then be used to identify shotgun clones which were evenly distributed across the insert with a 3-4 fold redundancy. The technique was currently being evaluated by testing a previously sequenced shotgun library.

The Berlin group did not yet have the facility to present the status of their mapping and sequencing data on the Web. They were, however, currently developing links to allow existing “in-house” databases to be displayed on the Web within the next few months.

The efforts to develop technologies to reduce redundancy in the sequencing process were commended but participants queried whether this technology could be generally applicable to human sequencing, particularly in highly repetitive regions.
Jean Weissenbach,  
Genethon, Evry

Jean Weissenbach described progress in the establishment of a French Sequencing Centre which would be sited in Evry near Genethon and was expected to begin work in summer 1997. Weissenbach would be the Director of the Centre which would be funded by the Ministry of Research with an annual budget of $14 million and a staff complement of 110-120. The Centre would be a joint venture between the Ministry of Research and the CNRS together with a third partner which would be a private company. The involvement of a private company was required in order to allow the Centre to employ people outside of the CNRS.

Projects would be evaluated by a Scientific Committee; these would include “in-house” projects and external collaborative projects. The expected ratio between external and internal projects had yet not been decided and may be influenced by a steering committee (comprised of representatives of different research organisations) which would set priorities and make strategic recommendations about projects. The steering committee would also make decisions about data release and protection of intellectual property. Weissenbach envisaged that scientists working on “in-house” projects may be able to release their data according to the principles agreed at the first Bermuda meeting but that academic collaborative projects may be handled differently.

The scope of sequencing projects at the Centre had not yet been determined but these were likely to include human, model organisms, *Arabidopsis*, pathogens and other micro-organisms.

John Mattick,  
University of Queensland, Brisbane.

John Mattick summarised the current status of the Australian initiative to establish the a national genome research facility for high throughput sequencing and genotyping. The Federal Government had voted $8 million to set up such a facility which should be operational by mid 1997. The facility would be based on two sites; one at the WEHI in Melbourne which would focus on high throughput genotyping and mutation detection under the auspices of Simon Foote and Dick Cotton, the other would be based in Queensland and would provide a sequencing facility. The total projected capacity would be 30 ABI machines providing 1500 reads/day and 8 million genotypes/year.
The funding would provide equipment and infrastructure for the facility but individual projects would need to be funded separately. John Mattick envisaged that the sequencing facility would accommodate a range of projects including micro-organisms, plants and mammals, funded either as external contracts or "in-house" projects. Cloning, sequence assembly and annotation would be the responsibility of the originating groups. It was anticipated that projects would be funded via the existing major funding agencies; the Australian Research Council and the National Health and Medical Research Council. However, it was hoped that the Federal Government may consider providing a special fund for genome projects in recognition of the difficulties associated with obtaining support through traditional funding modes.

Andre Rosenthal, Genome Sequencing Centre, Institute of Molecular Biotechnology, Jena.

Andre Rosenthal presented information on the goals and targets of the Genome Sequencing Centre at Jena. In 1996, the Centre had finished 2.6 Mb of sequence of which 1.5 Mb had been submitted to GenBank. In the period 1997-1999, the Centre aimed to complete 37 Mb; this could be divided into annual targets of 6 Mb (1997), 12 Mb (1998) and 19 Mb (1999). The main sequencing targets were on the human X chromosome [Xq28 (3 Mb), Xp11 (2.5 Mb) and PAR1 (1 Mb)], chromosome 21q (28 Mb) and chromosome 7q22 and 7q32 (7.5 Mb); maps and clones for these regions had been provided by both German and international groups. The Centre also planned to sequence regions in the mouse genome with homologous synteny to human Xq28 (3 Mb).

The Centre would be resourced by 20 ABI machines from May 1997 and would be organised into six production groups of four people, one bioinformatics group of five people and one library group of four people. The production groups would also perform the assembly, finishing and annotation of the sequence. The bioinformatics group would be involved mainly in software development. The total funding available from the federal government was DM 17 million over 4 years.

Andre Rosenthal described the German consortium of three groups which would be targeting 40 Mb of chromosome 21 over the next 3 years in collaboration with Sakaki's group in Japan. If the Japanese contribution to this effort increases over the next three years, the German group would transfer resources to regions of the X and chromosome 17 in the third year. The Centre was also pursuing various research interests in disease gene identification, comparative genomic studies and bacterial genome sequencing.
Phil Green,
University of Washington Genome Center, Seattle

Phil Green began his presentation by commenting that sequencing quality criteria were instrumental in determining the sequencing strategy. The criteria should include assessments of fidelity, accuracy and contiguity. At UWGC, sequence quality was assessed by 2-fold validation of all clones to detect small coligations and deletions, insertions (all data must be confirmed by at least one other clone). A base-specific error rate of less than 1 error per 10 Kb are required; error rates are submitted with the sequence data. In addition the assemblies are tested by an independent method, all gap sizes are estimated and sequence contigs are oriented and ordered within the chromosome; this latter being essential for PCR retrieval of genomic fragments across gaps.

Phil Green highlighted particular differences in the strategy adopted by UWGC compared to other sequencing centres. The strategy involved Multiple Complete Digest (MCD) mapping which provided a number of benefits including clone validation, the choice of more efficient tiling paths, more efficient finishing, simplified assembly verification and less redundancy in the sequencing process overall. The estimated costs were $0.05 to $0.12 per bp. Green stressed the value of long reads in improving efficiency by reducing finishing and assembly problems. The Center had also worked to develop software to provide objective finishing criteria.

The Center had been established in May 1996 and was focusing on three main sequencing projects: part of human chromosome 7 (7q31.3 and 7p14) in collaboration with Eric Green, the Human HLA Class I region in collaboration with Dan Geraghty, and the Mouse T-cell receptor alpha region in collaboration with Lee Hood. The Center had submitted 340 Kb of sequence to the databases, another 1.74 Mb had been completed but was undergoing further editing and annotation. The chromosome 7 region had a relatively low gene density but the HLA and T-cell receptor regions were gene-rich and therefore required a great deal of annotation. The sequencing groups also had a strong biological interest in the HLA and T-cell receptor regions. Phil Green was confident that the Centre would meet its first year goal of submitting 2 Mb by May 1997 but the second year goal of 6 Mb would be dependent on funding.

Green briefly described the MCD mapping strategy which involved subcloning from YACs or BACs (at 2-fold depth) into cosmids at 20-30-fold depth. Restriction digests with three enzymes were then used to construct a map of restriction sites and clone ends. Internal accuracy assessment included comparison of the restriction map with the sequence generated; to date, there had been no mapping errors detected in 1.2 Mb of finished sequence. No sequencing errors had been detected in chromosome 7, in the HLA, two sequencing errors had been detected with one attributable to PHRAP and the other to a small (12 bp) insertion or deletion in a cosmid clone.
Technology development focused on improving the MCD mapping procedure to allow automated detection of clone anomalies and also on improving software for sequence assembly and editing.

In discussion, Phil Green was asked about the technical limitations of subcloning directly from YACs to plasmids for sequencing. Green considered that the large clone size presented greater problems in dealing with repeats and gap closure; for BACs and YACs the number of reads to close gaps was much greater than for cosmids.

Ellson Chen,
Applied Biosystems Division of Perkin Elmer Corp., Foster City

Ellson Chen explained that his group represented an independent division of ABI called the Advanced Centre of Genetic Technology. The group consisted of 20 people divided into informatics, cloning, production sequencing and technology development which was resourced with 11 ABI sequencers. Funding was provided by NIH (50%), NSF (5%), industrial contracts (25%) and by Perkin Elmer (20%).

The major project of the group was sequencing of the human X chromosome in collaboration with David Schlessinger's group at Washington University. To date, the group had completed 2.4 Mb and was currently producing sequence at the rate of 0.25 Mb per month (3 Mb per year). Other projects included the completion of the Ureaplasma genome (760 Kb), in collaboration with University of Birmingham, Alabama; the finished sequence would be submitted to GenBank within the next month.

With current funding the group planned to complete 5 Mb of sequence on the X chromosome but if the NIH funding was renewed, the group planned to sequence a further 30 Mb including regions of the X chromosome (17.5 Mb), chromosome 3 (10 Mb) and the mouse (8 Mb).

Ellson Chen described the ordered shotgun sequencing strategy (OSS); BACs were subcloned to produce 10 Kb fragments in 10-fold lambda libraries which were end-sequenced to produce a partial physical map from which a minimal tiling path was chosen. PCR was used to prepare all sequencing templates for end-sequencing and random shotgun sequencing of the lambda inserts.

The BAC clones for the X chromosome were provided by David Schlessinger's group after they had mapped by STS content analysis, fingerprinting and end-sequencing. The main bottleneck in Dr Chen's strategy was in the generation of sub-clones from the 10 Kb Lambda inserts by long-range PCR. However, the advantage of the strategy was in its effective handling of repeats. Dr Chen concurred with David Cox's earlier comments that it was not possible to generate a complete tiling path during the initial stages of the process although longer end-reads could improve this significantly. It was hoped that the new dyes recently developed by ABI would allow average read lengths to increase to 1 Kb which would further increase the efficiency of the Ordered Shotgun Sequencing Strategy.
Asao Fujiyama,
National Institute of Genetics, Shizuoka

Asao Fujiyama described the Japanese Human Genome Sequencing Programme funded by the Japan Science and Technology Corporation (JST). The programme involved four main groups with different sequencing targets; Hideshoto Inoko at Tokai University, Yuske Nakamura and Yoshiyuki Sakaki both at the University of Tokyo, and Nobuyoshi Shimizu at Keio University. Fujiyama focused his talk on the chromosome 21 project which represented a collaboration between four groups: Sakaki, Shimizu, Cox and Rosenthal.

Fujiyama described current progress by Sakaki’s group in which 2.7 Mb had been completed in three contigs ranging from 300 Kb to 1.4 Mb. There were a number of gaps in the sequence partly due to the presence of chimeric clones in the P1 libraries that they were using. Sakaki’s group had a further 2 Mb in sequence-ready contigs available for sequencing in the near future. The main resources for sequencing were derived from chromosome-specific libraries in P1s, cosmids, fosmids, PACs and BACs.

Sakaki’s group used a directed sequencing strategy based on nested deletions. This process was relatively labour intensive and so attempts were being made to reduce these costs by increasing the level of automation. Two of the Japanese groups were involved in the testing of new prototype capillary sequencers from Hitachi; these machines were capable of running 96 capillaries at one time and were being used for both cDNA and genomic sequencing.

The next phase of the Japanese sequencing programme, after 1998, was currently being negotiated and it was hoped that the sequencing output would increase to 30-60 Mb/year with a maximum budget of $50 million/year. It was anticipated that chromosome 21 would be completed by 1999/2000 as an international collaboration and future sequencing targets were likely to comprise mouse regions with homologous synteny to chromosome 21 and other comparative sequencing studies on chromosome 11, that were also under discussion.

Fujiyama drew attention to the World Wide Web home pages set up by the JST (Advanced Life Science Information systems - ALIS) and by Sakaki’s group at the University of Tokyo. The information available on Sakaki’s home page was considered to be extremely useful and his efforts were commended.
Glen Evans,
University of Texas Southwestern Genome Science and Technology Center
(UTSW GESTEC), Dallas

Glen Evans described the three main activities at UTSW; sequencing regions of chromosomes 11 and 15 funded by the NCHGRI, production of a PAC/BAC end sequence data resource funded by the DoE, and technology development in collaboration with commercial companies.

The chromosome 11 sequencing project was based on an existing YAC/STS content map with 905 STSs which had been supplemented with 17,965 “binned” cosmid end sequences. PACs and BACs were isolated from 20x libraries by high density grid hybridization with pooled STS-specific oligonucleotides. The STS content was confirmed by PCR and the PACs were fingerprinted with four restriction enzymes to build up small contigs. PACs and BACs were end-sequenced to generate new STSs and to assist in map assembly. Chimeric PACs were eliminated using FISH. The map generated by this strategy was displayed on the WWW and represented the framework for the sequencing process.

To date the map production team had screened 465 STSs to isolate 3,185 PACs with an average hit rate of 12.45 STSs per PAC (ranging from 2.5 to 24.4 hits per PAC). 467 PACs had been confirmed by PCR and fingerprinted; 216 of these had been analysed by FISH and 1.3% (three clones) were potentially chimeric.

Efforts had been made to improve automation and accuracy in the sequencing strategy. A Sagian/Beckman robot had been developed with a potential capacity of 24,000 reactions per day. Oligonucleotide primers for gap closure and resequencing were synthesised automatically using a MerMade 192-channel oligonucleotide synthesizer. Where necessary accuracy was improved by resequencing to give an average PHRAP score of >40. The Center had submitted 1.6 Mb of sequence to GenBank of which 0.5 Mb had a PHRAP score of > 40 (i.e. an error rate < 10⁻⁴). The largest stretch of continuous sequence was 341,110 bp. The clone end-sequencing project funded by the DoE had generated a further 5 Mb of sequence mostly from chromosome 11.

The Center had developed an automated annotation protocol using a superparallel computer; final assembly and annotation could be carried out in 2 hours and it was anticipated that this could be reduced to 20 minutes once re-coding was complete. The programme annotated the following features: GenBank matches, EST matches, STS matches, end-sequence matches, GRAIL-predicted exons, repetitive sequences, simple sequence repeats and restriction sites.

Maps and sequence data are made available on the WWW; contigs and closed sequence (2.9 Mb) are available although the unassembled raw data is not. The PHRED/PHRAP score for each base along the sequence is also available on the Web to allow interrogation of the sequence accuracy.
Glen Evans described the current automation projects proceeding at the Center; these included the MerMade oligonucleotide synthesizer, the Sagian/Beckman robot and the Astral DNA sequencer. Most of these technologies were likely to be available as commercial production models or via contracted engineering companies.

Dr Evans emphasised the need to agree the boundaries of sequencing targets on the basis of STSs or other precise markers in order to prevent duplication.

Michael Palazzolo,
Lawrence Berkeley National Laboratory, Berkeley

Michael Palazzolo stated that the Lawrence Berkeley National Laboratory had submitted 5 Mb of *Drosophila* genome sequence (funded by the NIH) and 4 Mb of human sequence (funded by the DoE). Current sequence output was 800 Kb per month. The sequencing strategy was similar to that of other groups although Palazzolo considered that finishing did not represent a bottleneck in his strategy. As part of the plans to scale up the sequencing process the DoE Joint Genome Institute had developed a partnership with industry to introduce effective manufacturing practices into sequence production. This had required identification of goals for volume, quality, cycle time and cost. The importance of cycle time and precise goal definition were emphasised during this review process.

A number of metrical tools had been introduced by the industrial partners (Motorola) to evaluate the operations at Berkeley; these included process models, cost models, cost accounting and pick-a-mix. A process model had been developed for the sequencing activities at Berkeley, excluding physical mapping. This had shown that the bottle-neck in the sequencing process was in the loading of agarose gels. Predictive tools could be used to increase efficiency as new bottle-necks were identified with changing work practices. These models allowed rational decisions to be made about the balance between volume, quality and cost objectives and the inter-relationship between these objectives.
Bruce Roe,
University of Oklahoma

Bruce Roe reported that 1.8 Mb of human genome sequence data had been submitted to GenBank in the first year, almost 2 Mb in the second year and an additional 2.2 Mb was now in progress. The project was on a smaller scale than most participants as funding is less than $1 million per annum.

Currently the major target of the group is the region centromeric of the Sanger Centre and St. Louis portion of chromosome 22q; sequencing of the homologous syntenic regions in mouse, particularly the chromosome 16 region, is also being pursued. It was noted that Shimizu was sequencing 1.2 Mb around the immunoglobulin light chain region in q11.21. Regions on chromosome 9 had also been sequenced, and this may become the major focus of the group once chromosome 22 is finished. The group is not involved in mapping and therefore is dependent on clones for sequencing being supplied by collaborators.

Sequencing has been virtually completely automated; sequencing technologies exclusively utilise double-stranded vectors and dye-terminator chemistry, and protocols are made available on the Web site. PHRED and PHRAP are now used almost exclusively. Before sequence is declared finished it must be sequenced three times, twice in one direction once in reverse.

In addition to the human, sequencing of two bacterial genomes (N. gonorrhoeae and S. pyogenes) is underway as is an Aspergillus nidulens EST project. Funding has also recently been obtained from the NIH-Dental Institute to sequence Actinobacillus actinomycetemcomitans (a 2 Mb bacterial genome) and from the NIH-National Genome Research Institute to sequence a total of 4 Mb of the mouse genome. Data, including the human sequence, is immediately available via the Web site. It was reported that there had been enthusiastic feedback on the bacterial data but only minimal response on the human data had been received. From the data produced it was estimated that in the human there are on average two genes per 100 Kb.
# HUMAN SEQUENCE PRODUCTION (MB)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sullston</td>
<td>14.6</td>
<td>35</td>
<td>80</td>
</tr>
<tr>
<td>Waterston</td>
<td>4.8</td>
<td>24</td>
<td>24 + *</td>
</tr>
<tr>
<td>Lander/Hudson/Hawkins</td>
<td>2.1</td>
<td>20</td>
<td>80 *</td>
</tr>
<tr>
<td>Adams</td>
<td>2.7</td>
<td>11</td>
<td>14 + *</td>
</tr>
<tr>
<td>Gibbs</td>
<td>3</td>
<td>12</td>
<td>18 + *</td>
</tr>
<tr>
<td>Cox</td>
<td>0.3</td>
<td>5</td>
<td>?</td>
</tr>
<tr>
<td>Lehrach</td>
<td>0.24</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Weissenbach</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Mattick</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>Rosenthal</td>
<td>1.5</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Bloecker +</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Green/Olson</td>
<td>0.59</td>
<td>6</td>
<td>?</td>
</tr>
<tr>
<td>Chen</td>
<td>2.4</td>
<td>3.5</td>
<td>6.0</td>
</tr>
<tr>
<td>Sakaki +</td>
<td>2.7</td>
<td>3.4</td>
<td>30</td>
</tr>
<tr>
<td>Other Japan efforts</td>
<td>-</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td>Evans</td>
<td>1.6</td>
<td>5</td>
<td>50 *</td>
</tr>
<tr>
<td>Palazzolo/DoE</td>
<td>4</td>
<td>20.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Roe</td>
<td>3.8</td>
<td>5.5</td>
<td>12 *</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>44.33 Mb</strong></td>
<td><strong>172.4 Mb</strong></td>
<td>? Not meaningful to estimate total (384+)</td>
</tr>
</tbody>
</table>

* Production dependent on funding decisions - some centres (Lander, Evans) give numbers based on anticipated ramp up if funding is not an obstacle, others (Waterston, Adams, Gibbs) are more conservative.

+ Not attending meeting, reported by a colleague
Session II
Sequencing Quality, Costs and Data Release
Chair: Francis Collins

The aim of this session was to discuss standards for sequence quality, cost and data release, and the processes by which these could be measured and verified. Such standards must be defined, credible, and based on a scientific rationale to be of the greatest benefit to the scientific community.

Sequence quality is dependent on:
1. Accuracy of the nucleotide sequence
2. Assembly
3. Presence of gaps within clones and within contigs
4. Fidelity to the human sequence

Accuracy of the Nucleotide Sequence
At the first strategy meeting, an acceptable level of nucleotide error was agreed to be 1 error per 10 Kb. In response to a question about the rationale for this choice, it was noted that 1 in 10,000 is ten times lower than the frequency of single nucleotide polymorphisms in the human genome. One attendee argued that the existence of a polymorphism could be very easily verified at a later date with other techniques. However, there remained general agreement that this level was a reasonable goal and that, currently, most centres should be able to produce sequence with an error rate of 1 in 10,000 or less, except in particularly difficult regions of the genome.

Methods for assessing the error frequency were discussed. Approaches to nucleotide error assessment included:
1. PHRED and PHRAP analysis
2. Checking of raw data against the consensus sequence / Reassembly of raw data from another centre
3. Sample sequencing
4. Resequencing by another centre

The NIH-NHGRI discussed its tentative plan to determine the quality of sequence by commissioning the resequencing of BACs at two different centres. Owing to the variation in the ease of sequencing different regions, representative BACs would have to be chosen carefully to ensure that meaningful additional data was obtained. Many participants felt that this was a relatively expensive strategy. The Sanger Centre described data obtained from clones that have been accidentally resequenced ‘in house’ as providing a good indication of the error rate.
The merits of the PHRED and PHRAP software to assess the quality of sequence data were discussed. There was general agreement that these programs should not be relied upon as sole sources of quality estimations. Although the programs in general were considered to be robust, accurate and valid, they tended to give a slightly high estimation of the error rate and the error rates are prone to distortion by high GC content. A plan to recalibrate the program with DNA with >70% GC content was described. Donations of appropriate sequence data for recalibrating were requested.

The differences between various chemistries were highlighted. Dye terminator reactions were thought to yield a higher level of accuracy and are also capable of reading through GC rich regions that dye primer reactions cannot manage.

In Germany, raw data from sequencing centres are accessible by the other human sequencing centres within the consortium to enable comparison with a consensus sequence. This was proposed as a cost effective mechanism for identifying errors. Both poor quality sequence data and finishing errors can be identified by such comparisons. This was also a educational exercise as it allowed the problems experienced in particular centres to be shared.

The importance of clearly defined goals and standards was highlighted. This also required detailed information about the rationale and process behind the calculation of error rates.

The setting of high standards was agreed to be valuable because it helps to drive technological improvements. Francis Collins stated that the quality of the sequence produced at the outset of the project should be rigorously assessed and once the quality had been established more routine monitoring could be considered.

Returning to the subject of resequencing there was a general consensus that data exchange (release of raw data for checking by other centres) was a cost effective and educational method for assessing sequence quality in NIH centres. A plan was therefore proposed to go through the data exchange exercise, reassemble the data and identify outright discrepancies or ambiguities. These would be resolved by further consultation or resequencing. The same data sets would be sent to two centres which would hopefully engender competition to detect errors. Centres should be able to define explicitly the method by which their error rates had been established.
Assembly

Mechanisms for assembling sequence and validating the assemblies were discussed. The use of more than one assembly package or different stringency levels was suggested. This enables areas where the assembly is less robust to be identified. If any orphan clones remain once an assembly has been completed, investigators should be cautious of the validity of the assembly.

There was general agreement that the most reliable and effective method for validation of the assemblies was by restriction enzyme digestion. The use of two or three enzymes, chosen for their predicted digestion pattern, was agreed to be sufficient in most cases. Potential difficulties when long inverted repeats were present in the sequence were highlighted. Restriction enzyme digests are interpretable for cosmids, PACs and BACs but become more difficult for YACs. It was considered valuable to submit information on the enzymes used and the sizes of fragments obtained to databases along with the sequence.

Other techniques such as PCR, comparison with cDNA sequence and forward and reverse sequencing were also proposed as being of value for verifying assembly.

The need for verification with reference to the long range maps was discussed. No one method was thought to be perfect but methods such as STS content analysis, comparison of maps with the human DNA, and fibre FISH were thought to be useful.

Gaps

There was extensive discussion of whether to allow gaps in “finished” sequence. It was agreed that the goal for finished sequence should be zero gaps. At the same time, it was acknowledged that currently, sequencing and cloning difficulties make this impractical in some instances. Specific criteria should be produced as to when a gap was allowable.

Data were provided on the frequency of gaps in various regions of the genome that have been sequenced (see table). To date the frequency of gaps has been highly dependent on the composition of the DNA sequence, CpG islands being a major problem; sequencing reads are much more difficult through sequence of >80% GC content.
### Frequency of gaps in sequence

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Gap Frequency</th>
<th>Reason</th>
<th>Investigator</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1/55 Kb</td>
<td>CpG rich</td>
<td>Bentley</td>
</tr>
<tr>
<td>X</td>
<td>1/750 Kb</td>
<td>gene-poor region</td>
<td>Bentley</td>
</tr>
<tr>
<td>22</td>
<td>1/282 Kb</td>
<td></td>
<td>Bentley</td>
</tr>
<tr>
<td>13</td>
<td>1/121 Kb</td>
<td>CpG rich</td>
<td>Bentley</td>
</tr>
<tr>
<td></td>
<td>1/200 Kb</td>
<td></td>
<td>Hudson</td>
</tr>
<tr>
<td>C. elegans</td>
<td>1/250 Kb</td>
<td></td>
<td>Sulston</td>
</tr>
</tbody>
</table>

It was agreed that if a sequence containing a gap was to be allowed into the databases as finished sequence, the gap and the surrounding regions should be:
1. Oriented
2. Ordered
3. Sized
4. Justified

This information should be included in the sequence submission, along with the methods used for sizing the gap and the reasons for not closing the gap. For difficult sequences the cost/benefit ratio of trying to close the gap in the short term should be considered (and whether new enzymes or technologies would be required to solve the problem).

A database containing information on gaps was suggested. This would enable the community to be aware of the number of gaps being left by different centres and would also be a resource to facilitate collaborations or ‘SWAT’ teams to tackle problem sequences. The closing of gaps would however remain the responsibility of the original sequencers. The information on gaps should be available at subsequent genome co-ordination meetings.

The question of larger gaps arising from unclonable sequence or mapping gaps was also considered. It was felt that fewer sequences were proving to be totally unclonable with the increase in the number of different vectors, and sequence should only be deemed unclonable if all of these had been tried. Gaps should be avoided between contigs, again responsibility lying with the sequencer. On submission to databases a minimum of 100 bp overlap (preferably unique sequence) between clones should be supplied (as well as accurate information on how the different clones relate to each other).

The consensus was that the goal should be zero gaps in finished sequence, but it was recognised that this was not practical at the moment and sequencers should fulfil the conditions outlined above before submitting a sequence containing a gap to the databases.
Fidelity to the Human Sequence

Strategies for ensuring that the DNA sequenced accurately represented human genomic DNA were considered. Potential sources of error were point mutations and rearrangements during cloning. Only limited quantitative data on the stability of clones is available and more data are required.

For older libraries, there are problems in obtaining original source genomic DNA to compare with the cloned DNA. Differences between clones derived from a single source (excepting allelic differences) could be used to estimate the frequency of changes relative to the source. BACs were generally considered more stable than cosmids. It was reported that 5% of BACs were degraded after 100 propagations and therefore it was important how strains were maintained and stored. The DNA sequence also influenced the fidelity of the cloned DNA relative to the source. In one instance DNA 200-300 Kb from the telomere was 30-40 times more likely to rearrange. In C. elegans point mutations were found 1 in $10^6$ bases but in S. Cerevisiae it was as high as 1 in 60 Kb.

In summary it was thought that to assess and maximise fidelity; deep coverage, overlap of sequences and genomic Southern blots are required. Fidelity problems will eventually be resolved by technological developments which will allow the genome to be resequenced directly from genomic DNA.

Costs

The Chair invited Michael Palazzolo to present his approach to determining sequencing costs.

Issues of costing considered:
1. Value
2. Methods
3. Validation

The reasons for performing cost evaluations were discussed. They were needed to estimate the funds that will be required to complete the human genome sequence, to assess how costs might be reduced and also for review purposes.

Methods of cost evaluation were:
1. Cost model extrapolation
2. Cost accounting
3. Cost models
4. Top down cost and finance analysis

Dr Palazzolo described his laboratory’s experience with cost analyses. Both cost model extrapolation and cost accounting had been performed; it was found that the cost model extrapolation significantly underestimated the actual cost due to omission of some peripheral items. Cost accounting was required to track all costs throughout the process and to ensure that all costs were accounted for. A cost model could be used to analyse the individual steps and allow the cost of individual processes to be defined. Bottlenecks in the processes could be identified and therefore targeted for development.
Other centres had been estimating their costs in a more informal way using a funds in/sequence out method. It was generally recognised that without careful analysis, meaningful estimates of the actual sequencing costs, including all overhead costs, could not be obtained. As this was an expensive operation (TIGR employs three full-time accountants to do this) it was agreed that these centres would benefit from professional help in cost analysis. All the centres agreed that this process would be welcomed. The information generated on how to accurately estimate costs could then be disseminated. A request was made to funding agencies for help in supplying the necessary expertise, either by training scientists to carry out this type of analysis or obtaining suitably qualified assistance.

Once the processes to define costs have been established, a meaningful comparison of sequencing costs between centres can be made. This will be extremely important for assessing the future costs: one model discussed assumed that sequencing costs would decay with time; unfortunately it is impossible to make useful predictions without knowing either the half life of costs or the initial cost.

A suggestion for the formation of a consortium between the genome centres to negotiate improved deals with suppliers was suggested. It was recognised, however, that Government agencies have not been able to become involved in such negotiations and that companies prefer to negotiate deals on an individual basis.

Data Release

There was general agreement that the statement released after the first international strategy meeting was workable, useful and credible and should remain unchanged. The early data release policy appeared to have been welcomed by the scientific community and the wider public.

The practice of immediate data release should be maintained, with unfinished data (assemblies over 1 Kb) being accessible immediately through the home World Wide Web site (WWW). No information is available on the number of different groups accessing the sequence, but this would be a useful indication of the interest of the rest of the scientific community in the sequence that was being generated.

In most centres, efforts were being made to release data quickly; in the case of the NHGRI, for example, every centre has proposed a plan to the Institute that includes working toward rapid data release. At the moment some centres were releasing data as infrequently as quarterly; technical difficulties being cited as the main reason. The official NHGRI policy states that grantees should strive for early data release. Their compliance will be considered as part of the review process. It was suggested that early data release should be made an absolute condition of funding, especially for new grants. The DoE is also trying to make its investigators adhere to the principles in the strictest interpretation. Attendees encouraged all of the agencies to enforce the policy in order to increase public confidence in the way in which the policy has been being implemented. The special need to monitor those efforts being made by scientists in centres which have a biological interest in regions that they were sequencing was mentioned.
The conditions imposed on data release in Germany were extensively discussed. The German genome sequencing initiative is partly funded by industry and partly via the BMBF. The BMBF funding is dependent on the demonstrated benefits to industry. Raw data is not released but submitted to a private database for three months to which the industrial funders have exclusive access. At the end of this period, sequence which has generally been finished in this time is released into the public databases. The policy is scheduled for review after one year. Participants at the meeting felt that an official policy of privileged access was completely contrary to the Bermuda agreement and every effort should be made change this policy. There were concerns that continuation to the German policy could both endanger the early data release policies in other countries and also lead to duplicate (and therefore uneconomic) sequencing. It was suggested that a similar problem may be encountered in France and the scientific community should exert its influence to prevent this.

In contrast, in the last year there had been success in encouraging early data release in Japan. Investigators were now able to release their data directly onto their WWW site. Data had to be submitted at least every six months to the Japan Science and Technology Corporation (JST) which acts as a quality control site and submits sequence to the public databases every three months. The efforts of one particular Japanese investigator to embrace the concept of immediate data release were praised.

There was a consensus that pressure must be exerted on the BMBF to change its data release policy. Andre Rosenthal asked that the government funding agencies meet with BMBF to help persuade them to change their policy.

The importance of considering the DNA sequence itself as precompetitive and discourage patenting was reiterated. Data release prevented patents being filed on sequence in Europe but not the US, and therefore it would be contrary to the spirit of the agreement to file patents on the data once it had been released. The NIH asserted that although it could not prevent its researchers from filing patents, the grantees are required by law to inform the agency of any patents filed.
Data Submission

David Lipman outlined the different types of data currently being released into the public domain.
1. Raw data published on the local WWW site
2. Unannotated sequence containing gaps
3. Finished sequence

The database providers were keen to make the sequence data as accessible as possible. To this end they described plans to mirror sequencing centres' ftp sites. The usefulness of a database division, distinct from that containing the finished sequence, where unfinished sequence would be located was reiterated. This would mean that the scientific community would only need to search two databases, one of finished and one of unfinished data, to cover all the sequence in the public domain.

The system of assigning “levels” to describe the status of the sequence was unanimously rejected. Such descriptions are meaningless as they are not being applied consistently by all groups. A better alternative was considered to be a distinction simply between finished and unfinished, with data being located in the appropriate database division. A comment field could be included to describe how near the unfinished data was to completion. It was confirmed that sequence from clones that had been dropped from a sequencing strategy would be removed from the databases.

There was a request from the database providers for more interaction with the sequencing community to help improve the sequence databases. It was also requested that centres be meticulous about the information provided on how adjacent clones overlapped. Data on similarities to other sequences was updated daily as new sequence was submitted.
Session III
Allocation of Regions / Etiquette for sharing
Chairs: John Sulston and Bob Waterston

Bilateral claims were considered first.
Issues identified were:
1. Mapping
2. Sequencing
3. Limits - maximum and minimum
4. Communication
5. Conflicts / resolution

The system of posting sequencing intentions on the Human Sequence Map Index at the HUGO WWW site appeared to be working reasonably. It was enabling the general scientific community to be aware of what was happening as well as the sequencers themselves.

It was agreed that, at the mapping level, a certain degree of redundancy was inevitable and could be useful. Multiple resources are often required to generate large contigs and a variety of different approaches helped to determine the quality of the map. The major difficulty was deciding whether significant investment in mapping a region gave a group the right to sequence the region. In general, mapping investment was not considered sufficient to claim sequencing rights over a region, especially with groups generating chromosome wide maps. David Bentley, however, argued that long term investment and therefore commitment was required for efficient whole chromosome mapping. It was agreed that if a group had generated a sequence-ready map of the region it would be considered bad etiquette for another group to try and claim the region. In general, the group in the best position to begin sequencing a region should be allowed to do so. To avoid more than one group making a large investment in a region there should be a mechanism of publicizing long term aims. While in itself this would not be a claim it would allow for early communication and collaboration.

Sequencing claims allowable on the WWW site were discussed. Claims would be allowable for one year before proof of the product was required, i.e. sequence in the public domain. Only a realistic amount of sequence should be claimed, it was suggested that this might be up to three times the amount of sequence produced in the preceding year, but this may vary depending on the size of the sequencing operation. The current large designations corresponding to chromosome bands were thought to be too large and too vague. Regions should be defined by agreed markers. The most universal markers, which should be used if possible, were the Genethon markers. If no Genethon marker was available in the region of interest, other agreed and widely available markers could be used, as was the case on chromosome 11. It would then be the responsibility of the sequencers to sequence up to and including these markers. The smallest region that could be claimed was agreed to be 1 Mb. It was hoped that groups would try to sequence large contiguous regions rather than claiming many smaller regions.
The potential problems associated with interactions with the wider scientific community were considered. Two related issues were identified, whether groups should target regions of particular biological interest such as disease regions for early sequencing or whether announcing such an intention may have detrimental effects on groups with funding to identify a gene or genes in that region.

John Sulston discussed the pressures from the scientific community to target certain regions. This had not been the case with *C. elegans* where no custom sequencing had been done. For the human, several projects had been undertaken by setting up collaborations with gene hunting groups, but the immediate release of data was an absolute condition. Before any firm commitment was made sequencing interests of any other parties should also be considered, particularly those engaged in systematic mapping and sequencing of a region containing the area of interest. It was felt that some custom sequencing could be useful, as it increased the immediate benefit to the scientific community which could lead to wider benefits including increased support for the sequencing project. This did not mean that the genome should be sequenced in a piece-meal fashion, with regions of greater biological interest being sequenced outwith the systematic sequencing projects. Collaborative projects should be no more expensive than sequencing any other region of the genome.

Sequencers should be aware that sequencing a disease associated region may jeopardise the funding of groups involved in gene identification in these regions. The best way to tackle this issue was to set up collaborations with these groups to share resources, but no privileged access should be allowed to the data. It was also recognised that some gene hunting groups will be sequencing significant amounts of DNA, if the sequence is of sufficient quality it could be incorporated into the genome sequence as opposed to resequencing.

Although the Human Sequence Map Index had been found to be significantly useful, the Single Chromosome Workshops had also helped to maintain co-ordination on a more local level, especially for chromosomes being sequenced by many groups. Susan Wallace from HUGO Americas summarised the developments that had been made on the WWW page. The site was nearly complete but there was room for improvement.

The project needs were:
1. A clear mandate/action plan
2. A small advisory group to provide guidance
3. Funding for a half or possibly full time curator
4. Clear international support
5. Computer server space (currently donated by GDB)

It was envisaged that a curator could take on a more active role in monitoring local sites; including checking that claims were still current and even ensuring that the sequencing groups were releasing appropriate data. It was agreed that the site was extremely useful and needed little extra development. The possibility of expanding it to cover other organisms was raised.
At the moment funds were available until June, when HUGO would be reorganised. The participants were very concerned that the site would suffer as a consequence of the reorganisation. The advantage of having the site managed by an international and neutral organisation was recognised. The possibility of a relocation to the sequence database providers (NCBI, EBI and DDJB) was suggested as this would maintain the international aspect. This would have to be considered by the HUGO Council and at the international database meeting. It was noted that the WWW site was set up at the request of the participants at the First International Strategy meeting and it was important that the WWW site served the interests of those protagonists rather than the interests of a single organisation.
Session IV
Interpretation
Chair David Bentley

Types of data which were considered important in the interpretation of the human sequence were:
1. Completed sequence
2. Full length sequence and ESTs
3. Expression data
4. Comparisons with model organisms i.e. the mouse

Both computer and experimental techniques could be used to define features of the human genomic sequence which would be used to annotate the sequence. These included CpG islands, open reading frames (ORFs), splice sites, the 5' end of a gene, exon connections, and pseudogenes. Some level of analysis by the sequencing groups was thought to be valuable. This would allow early identification of any discrepancies in the sequence: this included frameshifts in ORFs, gaps in important regions such as CpG islands, non consensus splice sites and wrongly oriented contigs with disrupted exon connections. The discrepancies in the sequence could be immediately checked with reference to the original traces.

There should be certain standards set for annotation, including a standard format for labelling particular features; currently, groups are even annotating Alu repeats differently. PHRED and PHRAP values should be included in the annotation to demonstrate the level of confidence in the sequence. Experimental and computer data should be labelled as such. There was no quantitative data on how accurate the computer predictions can be, and there was disagreement on the ease with which these analyses could be carried out. Particular problems were identified in analysis of GC-rich sequence and regions that were evolving rapidly. It was felt that many of the gene prediction programs were not accurate in predicting exons, but they did give an indication of sequences likely to contain genes. Multiple programs were being used although it was not possible to combine their predictions. It was felt that analyses should not be pushed too far as this would lead to more inaccurate predictions, and predictive annotations should be labelled as such. More detailed annotation could be left to outside groups with a biological interest in the region, although there was a place for annotation by sequencers as the programs and expertise were not always accessible to outside groups. The potential danger of circular interpretation was raised, i.e. that prediction programs should derive the parameters for predictions from experimental data and not from other predictions. The similarities between sequences and protein families in the databases are being recalculated on a regular basis as new data is added; this information should be accessible in the near future.
The handling of queries concerning the sequence itself was discussed. All traces should be archived by the sequencing groups in a form that could be retrieved in response to queries. It was thought to be impractical to house the data on a central server. At the moment CD-ROM might be considered. In the long term it might be possible to establish a central repository for the data and queries could be charged on a cost recovery basis.

In addition to sequencing, many groups were pursuing some biological investigations, which were funded separately; these were often exon trapping or expression analyses. The full length sequencing of cDNA clones was felt to be within the remit of the genome programme as it would lead to valuable information on the location of gene sequence.

It was reported that the EST sequencing project was still progressing at Washington University and is funded until the end of the year. NCI, Merck, Genethon, and Bristol Myers Squibb were currently funding the sequencing of 8000 ESTs per week. The NCI had contracted LifeTech and Stratagene to produce 20 libraries each. These would be subtracted against 1500 known sequences; this process had reduced the abundance of these sequences in libraries by four fold in pilot experiments. Different source tissues and better libraries meant that new sequences were still being identified, although the number of singletons was rising more slowly than the number of clusters. Mapping of ESTs was valuable as they were useful in marker-poor regions for sequencing purposes as well as for association studies to identify disease genes.

Since the first report of the EST mapping consortium in October, an additional 17,000 ESTs had been mapped (see table)

<table>
<thead>
<tr>
<th>Number of ESTs mapped</th>
<th>Centre</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,000</td>
<td>Sanger Centre</td>
</tr>
<tr>
<td>6,000</td>
<td>Genethon</td>
</tr>
<tr>
<td>3,000</td>
<td>Whitehead</td>
</tr>
<tr>
<td>2,000</td>
<td>Stanford</td>
</tr>
<tr>
<td>17,000</td>
<td>Total</td>
</tr>
</tbody>
</table>

The EST-map WWW site at NCBI was due to be updated in June with a second edition of the transcript map using data from RHdb (Radiation Hybrid database). It was requested that any new data should be submitted to RHdb by then. It was suggested that updates should be more frequent. At the moment only the minimum amount of work was being done on RHdb as no funding was allocated to it.
Mouse

The mouse genomic sequence was thought to be of considerable value both for the interpretation of the human sequence and as a biological model.

So far, there had been only a few anecdotal comparisons of mouse and human sequence. The participants outlined their current activities in this area (see table). It was thought that the same data release policy and a similar level of accuracy was required for the mouse as for the human. It was proposed that 10% of the mouse genome should be sequenced, in gene-rich regions, for comparative studies with human but it was considered unlikely that funding agencies would make a commitment to do this until it was clear that there were sufficient funds available internationally to complete the human sequence. It was reported that Howard Hughes was now funding the sequencing of mouse ESTs at the rate of 4000 (3000 submitted to the database) a week. Funds were also committed to cDNA sequencing with the aim of sequencing 30,000 full clones in the next two years.

Mouse Pilots: Underway / Proposed

<table>
<thead>
<tr>
<th>Human syntenic region</th>
<th>Size of region (Mb)</th>
<th>Investigator</th>
</tr>
</thead>
<tbody>
<tr>
<td>11p23</td>
<td>1</td>
<td>Rosenthal</td>
</tr>
<tr>
<td>12p13 (CD4)</td>
<td>0.2</td>
<td>Gibbs</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>Sanger</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>Hudson</td>
</tr>
<tr>
<td>22</td>
<td>&gt;= 0.7</td>
<td>Roe</td>
</tr>
<tr>
<td>Xp22 (PGK)</td>
<td>0.1</td>
<td>Gibbs</td>
</tr>
<tr>
<td>Xp27</td>
<td>0.2</td>
<td>Gibbs</td>
</tr>
<tr>
<td>Xq28</td>
<td>3</td>
<td>Rosenthal</td>
</tr>
<tr>
<td>Xq28</td>
<td>2.5-3</td>
<td>Brown (Oxford)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>~10 Mb</td>
<td></td>
</tr>
</tbody>
</table>

Currently 160,000 5' end sequences are available; the funding from Howard Hughes should be sufficient to increase this to 400,000. The mouse libraries are richer in diversity than the human, probably reflecting the wider range of tissues available. The decision to sequence 5' ends was to obtain sequence in coding regions to enable cross-species homologies to be detected. Some mouse ESTs are being mapped in Europe using an RH panel generated by Peter Goodfellow's laboratory. The resolution of this panel had not been established, but the value of developing and using a higher resolution panel was raised.
Session V
Future Meetings and Public Statement
Chair: Michael Morgan

There was a consensus that there should be a continuation of the meetings. Although Bermuda was a somewhat inconvenient location for some delegates, its neutrality and isolation from other distractions was thought to be more important. A provisional date was agreed for the same weekend in 1998. It was hoped that a free afternoon session could be accommodated if an evening session was scheduled.

A statement for public release was read out which reiterated the genome sequencing community’s commitment to early data release. The conditions in certain countries hindering this was alluded to and exclusion of these countries would be considered if their policies were not reversed. This latter point was strongly contested by the German delegates. There was general agreement that it was essential that the international concern with the German policy was made known in the strongest possible terms. It agreed that Dr Morgan should liaise with the German participants to produce a mutually acceptable statement.