Report of the International Strategy Meeting on Human Genome Sequencing held at the Princess Hotel, Southampton, Bermuda, on 25th-28th February 1996

Aims of the Meeting

To discuss mechanisms to co-ordinate, compare and evaluate different strategies for human genome mapping and sequencing.

To consider the potential role of new technologies in sequencing and informatics and to discuss different scenarios for data release.

Summary

The following principles were endorsed by all participants. These included officers from, and scientists supported by, the Wellcome Trust, the UK Medical Research Council, the NIH NCHGR (National Institute of Health, National Center for Human Genome Research, the DOE (U.S. Department of Energy), the German human Genome Programme, the European Commission, HUGO (Human Genome Organisation) and the Human Genome Project of Japan. It was noted that some centres may find it difficult to implement these principles because of legal constraints and it was, therefore, important that funding agencies were urged to foster these policies.

Primary genomic sequence should be in the public domain.

It was agreed that all human genomic sequence information, generated by centres for large-scale human sequencing, should be freely available and in the public domain in order to encourage further research and development and to maximise its benefit to society.

Primary genomic sequence should be rapidly released.

- Sequence assemblies should be released as soon as possible; in some centres, assemblies of greater than 1 kb would be released automatically on a daily basis.

- Finished annotated sequence should be submitted immediately to the public databases.

It was agreed that these principles should apply to all human genomic sequence generated by large-scale sequencing centres, funded for the public good, in order to prevent such centres establishing a privileged position in the exploitation and control of human sequence information.

Whitehead Institute/MIT Centre for Genomic Research

Eric Lander proposed a similar mapping strategy using BACs (Bacterial Artificial Chromosomes) anchored by STSs and an M13 shotgun sequencing strategy with directed closure. The distinctive focus of the Whitehead Institute's approach would be the high level of automation proposed throughout the process; the Sequatron. The Whitehead had already developed an automated system for the selection of BACs using STSs and their characterisation by fingerprinting. A hands-off assembly process was also planned using informatics-directed closure. The Centre's goal was to produce 5Mb of sequence in the first year.

The Institute for Genomics Research

Craig Venter expressed concern that the estimated cost of developing sequence-ready maps by conventional approaches was likely to be of the order of $100m and that the maps produced were unlikley to be complete. He proposed a different approach which involved sequencing the ends of BAC clones to produce sequence tags every 5kb. He anticipated that this would reduce the cost of producing a sequence-ready map for the whole genome to $10m and would enhance international collaboration by providing unique, uniformly spaced tags across the genome.

University of Washington, Seattle

Lee Hood endorsed the use of BACs for developing sequence-ready maps. He proposed a similar strategy to that of the previous speakers and was planning to collaborate with Craig Venter using the end-sequence approach to produce Sequence-Tagged-Connectors (STCs) every 4kb;assuming a 20-fold coverage of BACs with 150Kb inserts (400,000 BACs). He recognised the difficulties in resolving repetitive elements using this approach particularly long elements (LINEs) which comprised approximately 11% of the genome. However, the approach should be very amenable to automation since it only required two processes; isolating DNA and sequencing DNA. He endorsed the general principle that data should be rapidly released and of high quality. The software that Phil Green had developed at Seattle to assess quality and eliminate manual editing should address both these issues.

Japanese Human Genome Project

Naotake Ogasawara explained that the Japanese Human Genome Project was funded by the Science and Technology Agency, the Ministry of Education, Science, Sports and Culture and the Ministry of Health and Welfare; the project heads would be Ken-ichi Matsubara and Yoshiyuki Sakaki. Japan also had a major interest in the rice genome project. Human sequencing focused on regions on chromosomes 16 (HLA), 14 (IgH), 4 (TCR), 21 (Down's critical region), 22 (IgL). P1 and cosmid maps were being developed for chromosomes 21, 17, 11, 8, 6 and 3. The two key technologies were the use of flow-sorted chromosomes to make chromosome-specific cosmid libraries and the development

Lawrence Livermore National Laboratory

Tony Carrano stated that the primary human genome sequencing target for LLNL would be chromosome 19; a cosmid map of the 50Mb chromosome had been published in December in Nature Genetics. A shotgun strategy with directed closure was proposed for the chromosome which had a high proportion of Alu repeats. The Laboratory would provide clones and map information to other groups. Comparative mouse studies would be done in collaboration with Rick Woychik and Lisa Stubbs at Oak Ridge National Laboratories. The LNLL was also planning to use ESTs available through the IMAGE consortium to pull out BACs in a genome-wide, high throughput hybridisation strategy.

Baylor College of Medicine

Richard Gibbs proposed a similar strategy to previous speakers, focusing on chromosomes 12 (12p1.3) and X (Xq28 and Xp22). The strategy would include some reverse sequencing and other approaches to compensate for the incompleteness of available maps. Richard Gibbs concurred with the general consensus that sequence generated should be of high quality aiming for 99.99 % accuracy.

Los Alamos National Laboratory

Bob Moyzis stated that the LANL had funding from the U.S. Department of Energy (DOE) for a pilot project on sample sequencing. 1 Mb had been completed which included regions with single-pass sequencing and regions with 16-fold redundancy where they related to EST sequences.

The German Human Genome Programme

Hans Lehrach explained the proposed structure of the German Human Genome Programme which would comprise a central resource centre with genomic and cDNA libraries. The sequencing programme would focus on chromosome 21 and the long-arm of the X chromosome, for which the map was 70% complete.

EMBL

Wilhelm Ansorge provided brief details of the sequencing technology used at EMBL; with 2 dyes and 2 lasers, 2000 bases could be read from each sequencing reaction. He estimated that with this technology 3 people could sequence 4Mb per year at 5-fold redundancy. The Laboratory had focused on the EC yeast genome programme to date and did not have an in-house human mapping programme. Any human sequencing programme would therefore be dependent on resources being made available from other laboratories.

The session was divided into two parts; the first part focused on strategies for large-scale human sequencing based on experience with model organisms, and the second part focused on resources available for human sequencing and how they were being used.

Part I: Model Organisms

John Sulston - *C.elegans*

John Sulston summarised progress and lessons learned from the *C.elegans* sequencing project. The physical map covered more than 95 Mb with 7-deep coverage over 80% of the genome and with 7 gaps. Cosmid contigs had been assembled by fingerprinting and the deep coverage had meant that the tiling path had been easy to determine. YACs had been incorporated by hybridisation strategies but with lower confidence levels because of repetitive sequences. To date, 34 Mb had been sequenced by the two centres (WashU and the Sanger Centre), 6000 genes had been identified of which 45% showed database matches, 28% of the sequence represented coding information with an average gene density of 1 per 5Kb (autosomal) and 1 per 7 Kb (X chromosome), 30% of predicted gene sequences could be confirmed with EST/cDNA matches.

The key features of the strategy were the use of multilevel maps (cosmids and YACs) to resolve difficulties and provide a range of resources for use by the scientific community. The high resolution map had resulted in greater sequencing efficiency and, as a general rule, it was more efficient to aim for a high quality product than attempt to resolve mistakes at a later date.

In response to questions about testing the integrity of the genomic sequence, John Sulston stated that, at one point, cosmids had been tested against the genome by PCR. However, this had been very expensive and was not 100% efficient, it was not therefore considered to be worthwhile. Fingerprinting allowed validation to 3 Kb resolution but he conceded that the cosmid sequence had not been truly validated against the genome. There was some discussion about the mutation rate in cosmids and M13 clones but it was generally agreed that these were fairly simple to detect, particularly in the worm since it was homozygous.

David Cox noted that there were both technical problems and efficiency costs inherent in the validation of sequence information and producing a truly complete product.
John Sulston commented that completion of the human genome would be asymptotic and that some features (such as centromeres) may be omitted by design.

Rick Myers

Rick Myers described the approach at the Stanford Human Genome Centre to generate high resolution radiation-hybrid maps and BAC contigs. A protocol to verify map and sequence data with oligo chips was being developed in collaboration with Affymetrix. The 10bp chips were likely to have a useful role in checking fingerprinting but could only verify sequence data to a limited extent. The chip technology was likely to be prohibitively expensive for 100% verification. In addition, there were serious technical problems resulting from oligo folding that meant that some sequences would not be accessible using this technology. The Stanford Human Genome Centre had initially proposed a strategy based largely on a directed sequencing approach using transposon-mediated directed sequencing but it was now likely that their strategy would also involve a random approach.

David Bentley

David Bentley summarized progress on the development of a sequence-ready map for chromosome 22. The key elements of the approach were the use of a radiation hybrid map to pick out large insert clones and then to integrate PACs, BACs, cosmids etc. *via* fingerprinting.

Eric Lander

The Whitehead Institute/MIT Center for Genomic Research had mapped 16,500 STSs to generate a human physical map with 8,000 mapped to Radiation Hybrid addresses and 11,500 to YACs. This included 60% Genethon and CHLC loci. The Center planned to map a total of 20,000 STSs by June 96.

In the mouse, 6,500 STSs had been mapped to generate a genetic map published in Nature, 14th March 1996. Eric Lander suggested that in order to interpret the human genome effectively, it would be worthwhile sequencing 5% of the mouse genome for comparative studies.

The rapid STS mapping technology could be used to pull out YACs, BACs and PACs for the generation of sequence-ready maps and closure of gaps. The sequencing strategy proposed by Eric Lander was to pull out BACs using STSs to generate end-sequence and fingerprint the BACs to check integrity.

In discussion about fidelity checking, it was agreed that deep maps of at least 10-fold coverage would be required to distinguish between polymorphisms and clone rearrangements. This was particularly true of regions where genomic rearrangements were to be expected.

In discussion it was recognised that there was general requirement for technological developments such as oligonucleotide synthesizers to be more readily available and distributed. Whilst there was clearly commercial interest in licensing the technology to sell oligonucleotides, this had not facilitated the distribution of the technology itself, which may require the involvement of contract engineering firms.

SESSION III - LARGE-SCALE SEQUENCING

CHAIR - Tom Caskey

The session began with a discussion on data release and led by John Sulston and Bob Waterston. Drs Sulston and Waterston proposed that sequence data should be released automatically on a daily basis and that there should be no patent protection before release. A number of key issues were raised during the discussion:

Data Release

There was a strong scientific argument for immediate data release in order to facilitate co-ordination and encourage further research and development.

The value of unfinished sequence information was queried by some participants but previous experience (e.g.with the BRCA2 region) had shown that such information could be effectively utilised by both academic and commercial groups.

There was a balance between providing high quality information and avoiding long delays in the release of data. It was agreed that early release of data would be essential to ensure that sequence information was freely available for research and development.

IPR

Primary sequence information of unknown function was unlikely to be patentable.

In the U.S. patents could be filed up to one year after data release whereas this was not the case in Europe. This meant that data release would have a different effect on the ability to patent in different countries.

It was important to ensure that centres funded to generate sequence information in the public domain, on a large scale, did not also establish a privileged position in the control and exploitation of that information.

It was noted that groups in different countries and funded by different agencies may be under various legal and political constraints which would make it difficult for them to adopt these principles. It was agreed that funding agencies should be urged to foster these principles in the public interest.

Richard Gibbs

At Baylor, Richard Gibbs' group was using new dyes which were not covered by existing patents to provide ET-primers. Reverse reads were used as a part of the overall strategy to reduce the need for oligonucleotide synthesis and to verify overlaps. A region of chromosome 12 near CD4 had been sequenced and analysed. All except one of the genes identified using software tools had been verified by EST hits. The Group was also involved in comparative sequencing of regions of Xq28 in man and mouse; high quality data was a prerequisite for reliable comparison of coding regions.

Trevor Hawkins

Trevor Hawkins described the Sequatron; a fully automated front-end system for the preparation of sequencing reactions. The system was currently capable of processing 8,000 samples per day with a projected increase to 12,000 samples per day. The Whitehead/MIT Center used a base-calling software called GRACE as a substitute for the ABI software. To date, all the sequencing at the Center had been ESTs and STSs from man and mouse but the aim was to produce 5 Mb of finished human sequence in the first year rising to 80 Mb in the third year.

Andre Rosenthal

Andre Rosenthal was the co-ordinator for the large-scale human sequencing programme of the German Human Genome Project. At Jena, he had a group of 37 people of whom 25 were involved directly in sequencing. His laboratory currently had 12 ABI 377 sequencers and he hoped to increase this in 1996/97. He hoped to produce 5-10Mb finished sequence by the end of 1997, 10-15 Mb in 1998 and 15-20 Mb in 1999, assuming that the restrictive regulations on consumable expenditure could be overcome. Funding had been provided by various agencies including the EU, the DFG, the BMBF and the German Human Genome Programme. The German BMBF provided most of his funding. Targeted regions included: 3 Mb of Xq28 between the MeCP2 locus and DXS304 (in collaboration with Annemarie Poustka, Michele D'Urso, the Sanger Centre, Richard Gibbs and Ellson Chen), 3 Mb of Xp11.23, 1-2Mb of Xp11.4, 1-2 Mb in PAR 1 of X, 1-2 Mb regions on chromosomes 7, and 11 around disease genes and fragile sites. Andre Rosenthal was also involved in a project proposal to the German Human Genome Programme to sequence 30-40 Mb of chromosome 21. This proposal excluded the Minimal Downs Syndrome Region and the PME region which were being sequenced by two Japa: ese groups and the Stanford Genome Centre. Rosenthal's group was also involved in comparative sequencing studies in human and fugu for genes on the X-chromosome. In order to investigat gene evolution between man and fugu as well as possible synteny between the two species, he hoped to carry out comparative analysis on larger regions of the human X chromosome in fugu and man. His group also hoped to analyse genes from chromosome 21 in fugu although funding for comparative analysis on a large scale was not available in Germany. He was supportive, in principle, of the data release policy proposed earlier but would prefer high quality data to be released. His

SESSION IV - INFORMATICS

CHAIR: David Lipman

In view of the strong consensus on data release achieved the previous day, David Lipman proposed that the session should focus on annotation of data, mechanisms for assessing error rates and new approaches.

Mark Adams

Mark Adams described software developments at TIGR for sample tracking (TRACKER), quality control feedback and assessment of randomness of a particular library.

He also described various ways of annotating sequence data based on similarity searches (e.g. EST hits and BLAST analysis) or gene prediction analysis using GRAIL and other software. Different methods may give different results and it was important that these were annotated appropriately, and any contentions marked, since the average user was unlikely to have sophisticated sequence analysis software or hardware.

It was agreed that detailed annotation submitted to public databases should be definitive and that submitting groups should be responsible for revising any inaccuracies in the sequence or annotation. The NCBI could revise EST hits since it carried out daily comparisons of existing data with new submissions. One of the key issues was the high level of redundancy which was accumulating in the public databases from entries with an increase in the level of experimental and computational information attached which were being considered as independent entries rather than revisions.

LaDeana Hillier

The goal at Washington University Genome Sequencing Center was to provide immediate data release with local annotation of sequence. The units of release were BACs, PACs and cosmids which could be updated into larger contigs.

All features annotated were at a high confidence level; local analysis and annotation were important for assessment of error rates. COP and P-COP were used to automatically compare the consensus sequence against all available raw data to check the consistency of the data. Assemblies were confirmed using mapping data from STSs, restriction digests, fingerprints and overlapping clones. The public availability of raw data also provided an independent checking mechanism..

Polymorphisms in the human genome sequence were annotated in a feature table. These could be distinguished from clone mutations with a high confidence level in the human genome sequence because of the depth of coverage. They were not annotated in ESTs

Richard Durbin

Richard Durbin explained how manual operations in the sequencing process were becoming computerised at the Sanger Centre and the need for human input being reduced. The computerisation involved the introduction of new software as modules rather than revision of a single monolithic structure.

He discussed various ways in which confidence levels and error rates could be attributed to data. It was general practice to assign confidence levels for base-calling on a log scale, via PHRAP, rather than use a binary system. Most uncertainties had very low PHRAP scores and tended to be clustered. In order for external users to make best use of the data available, it was important that centres explained how data had been validated and to ensure that PHRAP values (or other confidence levels) and raw traces were also accessible.

SESSION V - PANEL/OPEN DISCUSSION

CHAIRS: Jim Watson, Bob Waterston and John Sulston

The following key issues were identified to be addressed in the summary session:

- Data release and intellectual property rights
- Co-ordination
- Funding available for large-scale human genome sequencing
- Accuracy; standards and evaluation
- Data release and intellectual property rights

John Sulston and Bob Waterston proposed the following principles for release of human genomic sequence generated by large scale centres:

Release

- Automatic release of sequence assemblies greater than 1 Kb (preferably daily).
- Immediate submission of finished annotated sequence
- Aim to have all sequence freely available and in the public domain for both research and development, in order to maximise its benefit to society.

funding under the RFA would allow at least $15 million to be provided for human sequencing this year. This was expected to increase to $60-$80 million (of a total NCHGR budget of $120 million) in 1999 with the move from mapping to sequencing. Grants awarded under the RFA would be reviewed in December 1997 to decide whether a third year of funding should be provided. The assessment criteria for renewal would include the amount of finished sequence submitted to public databases. Early release of data and a high level of accuracy would also be required.

The NCHGR would be holding a workshop of grantees in April to consider potential mechanisms to determine accuracy and validate data. The NCHGR was also planning a symposium on post-genomic biology which would consider: Future technologies in sequencing and genotyping, whole genome approaches to function, evolutionary biology, human variation and the ethical, legal and social issues (ELSI). The mouse genome and yeast whole genome biology were considered to be likely targets for investment in the near future.

The Wellcome Trust

Michael Morgan summarised the Trust's main activities in genome research; the development of the Wellcome Trust Genome Campus at Hinxton, pilot studies on pathogenic genome sequencing and the Wellcome Trust Centre for Human Genetics, in Oxford. The Trust was planning a scientific frontiers meeting "From Gene to Structure and Function" to address potential strategies for genome interpretation and exploitation. The Trust funding for the Sanger Centre totalled £84 million over the next 7 years of which £60 million was attributable to human sequencing and associated mapping; equivalent to £8.5 million per annum.

The Sanger Centre had orginally proposed to sequence a third of the genome and had been funded, by the Trust, to sequence a sixth. John Sulston considered that the Centre could realistically scale up to do 50% of the genome if funding were available.

European Commission

Manuel Hallem stated that the Fourth Framework programme (1994-1998) provided funding of $17,000 million over 5 years of which $11 million per annum was used to support human genome research. This covered human mapping and sequencing, function, disease determinants, gene therapy and data management. An ad hoc working group was currently considering priorities for the Fifth Framework programme. A key criteria for inclusion was that topics should be complementary to activities supported in member states. The first contracts under the fifth framework would be isued from 1st July 1999.

Discussion focussed on measures that might be required to achieve this level of accuracy and the cost/benefit ratio of the various methods. These included:

- Double-stranded coverage
- "Rule of Three": i.e. two clones including one reverse-read or using orthologous chemistry
- Resolution of all ambiguities
- High level of contiguity

It was noted that some regions may require additional reads to achieve this level of accuracy and others possibly less. The quality of the data could be determined by the ease of assembly and the use of software programmes such as cop and pcop which compared the consensus sequence with the raw data. Other methods of quality control which were discussed include the resequencing of a proportion of clones, independent analysis of trace data, and comparison of assembly data with restriction analysis. It was noted that data quality was likely to vary depending on the base composition of particular regions of the genome. Sampling would therefore have to be quite extensive in order to provide a comprehensive picture.

In considering the level of contiguity that might be achieved, it was noted that sequence "gaps" arose for three main reasons; "biological" cloning gaps, technical gaps arising from dinucleotide repeats or G,C-rich regions, and sizing or mapping gaps. In some instances, it may be necessary to develop further technologies to deal with the problems and it was therefore agred that gaps should only accepted if all exisiting technologies had been exhausted.

Participants were informed that the NIH NCHGR would be convening a workshop of grantees to discuss validation and quality control of data in April.