THE WELLCOME TRUST

183 Euston Road
London NW1 2BE

Telephone: 0171 611 8888 / Direct: 8899
Fax: 0171 611 8545 / Direct: 8237

JS/SO'D/LET/892

20 May 1997

Mtg. file
2/27 - 3/2/97

Dr Francis Collins
National Institutes of Health
National Centre for Human Genome Research
21 Center Drive MSC 2152
Bethesda
MD 20892-2152
USA

Dear Dr Collins

Attached please find copy of final report following the 2nd International Strategy Meeting on Human Genome Sequencing held in Bermuda in February of this year.

With kind regards.

Jilly Steward
Meetings and Travel Manager

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.

Mtg. file – Bermuda 2/97

CC: EJ
MG 4-7-97
Jane P

Mark & Jane, can you review carefully and return comments to J.K. See my minor edits below
FC

To: [redacted]
cc: [redacted]
From: [redacted]
Date: [redacted]
Subject: Bermuda report 1

Dear Dr Collins

I'm sending two email messages containing the reports of the Bermuda meeting. (Hard copy to follow by fax)
This one contains the report on the session you chaired and the next one is a summary of the meeting.
If you have any comments or amendments can you send them to me by the end of April.
Many Thanks
Jill

Dr Jill Kent
Scientific Officer - Genetics
The Wellcome Trust
[redacted]ad
[redacted]

CONFIDENTIAL

Second International Strategy Meeting on Human Genome Sequencing

Session II
Sequencing Quality, ~~and~~ Costs, and Data Release
Chair: Francis Collins

The aim of this session was to discuss current and desired standards for sequence quality, cost and data release, and also the processes by which these could be measured and verified. These standards had to be clearly defined, credible, and based on scientific principles to be of the greatest benefit to the scientific community.

There was a real need to undertake a realistic assessment of the cost of sequencing. This should take into account ALL sources of funding so as to judge whether sufficient funds are likely to be available to finish the human genome sequence by the target date of 2005 or earlier.

Sequence quality is dependent on:
1. Accuracy of the nucleotide sequence
2. Assembly
3. Presence of gaps
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 in every 10kb. This had been suggested as it was ten

times lower than the frequency of polymorphisms. It was pointed out however that the existence of a polymorphism could be very easily verified at a later date with other techniques. 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.

The ways of assessing the error frequency were discussed and a list of nucleotide error assessment strategies was proposed.
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-NCHGRI was tentatively planning to determine the quality of sequence by commissioning the resequencing of BACs at two different centres. Many participants felt that this was a relatively expensive strategy. At the Sanger Centre clones that are accidentally resequenced æin houseÆ provided a good indication of the error rate. Owing to the variation in the ease of sequencing different regions, representative BACs would have to be chosen carefully to ensure that meaningful data was obtained.

The merits of the PHRED and PHRAP software to assess the quality of the sequence data were discussed. It was felt unwise to rely solely on these programs for quality estimations, although they tended to give a slightly high estimation of the error rate. Although the programs in general were robust, accurate and valid, the error rates were prone to distortion by high GC content. There was a plan to recalibrate the program with DNA with >70% GC content but this type of sequence is currently relatively rare. 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 were also capable of reading through GC rich regions that dye primer reactions could not manage.

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

Michael Palazzolo highlighted the problems encountered when reviewing sequencing centres. He asked that the sequencing community should not only clearly iterate its goals and standards but also the rationale behind them. If the process behind the calculation of error rates etc. was clearly defined this would facilitate meaningful comparison between centres.

The setting of high standards was a valuable exercise because it helped to drive technological improvements. Francis Collins believed that the quality of the sequence produced at the outset of the project should be rigorously assessed and once the quality had been established slightly less monitoring could be considered.

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. 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

The most effective mechanisms for assembling sequence and validating the assemblies were discussed. The use of more than one assembly package or different stringency levels was recommended. This enabled areas where the assembly was less robust to be identified. If any orphan clones remained once an assembly had 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 digest. The use of two or three enzymes, chosen for their predicted digestion pattern, was agreed to be sufficient in most cases. Potential difficulties were highlighted when long inverted repeats were present in the sequence. Restriction enzyme digests were interpretable for cosmids, PACs and BACs but became 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 complementary techniques such as PCR, comparison with cDNA sequence and forward and reverse sequencing were also thought to be of value for verification of the assembly.

The need for verification with reference to the long range maps was considered. 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 as to if, and when, gaps in finished sequence could be allowed. The goal for finished sequence should be zero gaps but, currently, sequencing and cloning difficulties made this ~~unobtainable.~~ impractical in some instances. Specific criteria should be produced as to when a gap was allowable.

Data was provided on the frequency of gaps in various regions of the genome that have been sequenced (see table). The frequency of gaps was highly dependent on the composition of the DNA sequence, CpG islands being a major problem - with sequencing reads being ~~unobtainable~~ much more difficult through sequence of >80% GC content.

Frequency of gaps in sequence

Chromosome	Gap Frequency	Reason	Investigator
4	1/55kb	CpG rich	Bentley
X	1/750kb	gene-poor region	Bentley
22	1/282kb		Bentley
13	1/121kb	CpG rich	Bentley
	1/222kb		Hudson

C. elegans 1/250kb Sulston

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 would have to be:

1. Oriented
2. Ordered
3. Sized

Justified

This information should be included in the sequence submission as well as the methods used for sizing the gap and reasons for not closing the gap. For difficult sequence 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 æSWATÆ 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. Currently an acceptable number of gaps was agreed as 1 in 250kb, but centres should aim for 1 in 1Mb. These figures were to act as a guide as different sequences had different levels of difficulty.

Did we agree to that?

The question of larger gaps was considered. These may arise from unclonable sequence or mapping gaps. 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 lay with the sequencers. On submission to databases ~~approximately~~ a minimum of 100bp 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~~ practical gaps in finished sequence, but it was recognised that this was not possible at the moment and sequencers should fulfil the conditions outlined above before submitting a sequence containing a gap to the databases. It was also recognised that certain sequencing approaches such as BAC-end sequencing would initially lead to a large number of gaps that would then have to be closed.

space

Fidelity to the human sequence

Strategies for ensuring that the DNA sequenced accurately represented human genomic DNA were considered. Major potential sources of error were point mutations and rearrangements. Only limited quantitative data on the stability of clones was available and more data was required.

For older libraries there were problems in obtaining original source genomic DNA to compare with the library. 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 200-300kb 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 60kb.

???

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

The Chair invited Michael Palazzolo to present his data on sequencing costs.

Issues of costing considered:
1. Value/Danger
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

Michael Palazzolo described what had been learnt by employing an MBA to perform cost analyses. Cost model extrapolation and cost accounting had been performed; it was found that the cost model extrapolation had significantly underestimated the actual cost due to omission of some peripheral items. Cost accounting was required to track all costs throughout the process and 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 money in sequence out method. It was felt 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 a full audit should be performed on a few centres. 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 had been established, a meaningful comparison of sequencing costs between centres could be made. This was extremely important for assessing the future costs: it could be assumed that sequencing costs would decay with time, unfortunately it was impossible to make 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. Potential problems were the inability of government agencies to become involved and the desire of companies 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 requirements for immediate data release should be maintained, with unfinished data (assemblies over 1kb) being accessible immediately through the home World Wide Web site (WWW). No data was 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. The NIH had asked their centres to make known their data release policy. At the moment some centres were releasing data as infrequently as quarterly; technical difficulties being cited as the main reason. The official N~~C~~HGRI policy states that grantees should strive towards early data release. Their compliance would 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 was also trying to make its investigators adhere to the principles in the strictest interpretation.
There should be an effort 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 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 this 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 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. It was also proposed that there should be lobbying to put the data release policy on the agenda for the next G7 summit.

The need to consider 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 filing patents the grantees were required to inform them 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 were starting to mirror sequencing centresÆ ftp sites. The setting up of a database distinct from that containing the finished sequence where unfinished sequence would be located was discussed. 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 current system of assigning levels to describe the status of the sequence was unanimously rejected. The levels were meaningless as they were not being applied consistently between the groups. A better alternative was considered to be a classification into finished or unfinished, with data being located in the appropriate database. 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.

THE WELLCOME TRUST

183 Euston Road
London NW1 2BE
[redacted]

Our Ref: JK/PK

Dr Francis Collins
National Institutes of Health
[redacted]ch Institute

More....

4th April 1997 cc: EJ, MG, JP

Dear Dr Collins

Second International Strategy Meeting on Human Genome Sequencing
held in Bermuda on 27 February - 2 March 1997

Please find enclosed a summary of the matters agreed on at the above meeting, and the report of the session which you chaired. I would be most grateful if you could look through these and let me know if you have any amendments.

If it is possible could you inform me if the documents have your approval, or if there are any changes that you would like made, before the end of April.

Yours sincerely

Jill Kent

Dr Jill Kent
Scientific Officer - Genetics

Enc.

JKENT/040497/PK.1

CONFIDENTIAL

Summary of the Second International Strategy Meeting on Human Genome Sequencing

- 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 and Costs

- 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 goal was set at zero, currently only 1 gap in 250kb of 'finished' sequence is allowable with the aim to reduce this to 1 in 1Mb or better
  - Closing gaps is the responsibility of the original sequencers
- The following proposals were endorsed by the participants
  - It was agreed that the best way to assess sequence costs was to perform a data exchange exercise. Raw sequence data would be exchanged between sequencing centres, centres would reassemble the data and identify outright discrepancies or ambiguities with reference to the consensus sequence. 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
  - An audit should be undertaken of selected sequencing centres to determine the true sequencing costs
  - Sequencing centres should define explicitly how error rates and costs have been calculated

Sequence Submission and Annotation

- Finished and unfinished sequence should be submitted to separate databases
- Sequence annotation should be standardised if possible, and include the following information:
  - PHRED and PHRAP data → Not necessarily; not all centers use this
  - Enzymes used to verify assembles, and sizes of fragments produced
  - Exact details on how to assemble adjacent clones, and about 100bp or more if available (helps define polymorphisms) of overlapping (preferably unique) sequence between clones for verification
  - Gaps and surrounding sequence, should be orient~~at~~ed, ordered and sized and the methods used for sizing, and reasons for not closing the gap 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 stated how near the sequence is to completion
  - Levels system to be discontinued
- 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 1Mb
  - Maximum amount - in the order of three times the sequence released by the centre in the preceding year
  - Regions defined by Genethon markers if possible, other agreed and available markers if not
- Sequence must include the limiting markers

Doesn't quite catch the
principle - must do the entire
segment from Genethon marker
to
Genethon
marker

Next Meeting

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

DEPARTMENT OF HEALTH & HUMAN SERVICES

Public Health Service

National Institutes of Health
National Human Genome
Research Institute
31 Center Drive MSC 2152
Building 31, Room 4B09
Bethesda, MD 20982-2152
(301) 496-0844
FAX (301) 402-0837

March 21, 1997

Dr. Frank Laplace
Federal Ministry for Research & Technology
Heinemannstrasse 2
D-53175 Bonn
Germany

Dear Dr. Laplace:

At the Second International Strategy Meeting on Large-Scale DNA Sequencing, held February 28 - March 1, 1997 in Bermuda, attendees reaffirmed how critical it is to the integrity and success of the international Human Genome Project that human genomic sequence data be rapidly released, without prior exclusive access to it on the part of anyone. We are writing to confirm that the National Human Genome Research Institute and the Human Genome Program of the Department of Energy agree with this principle and consider it to be critical to continued support for the Project in the United States. Accordingly, the NHGRI and the DOE have adopted policies to implement appropriate rapid data release practices for all of the large-scale human DNA sequencing projects we are supporting.

In our view, the key purpose of the rapid release policy is to ensure that the small number of laboratories (funded by public and private charitable sources) that have emerged with the capability to generate large amounts of human DNA sequence data do not take unfair advantage of that capability to gain privileged access to potentially valuable information. The human DNA sequence is a scientific resource of unprecedented importance to all of humanity. By studying it and coming to understand it in much greater depth than we do now, the human DNA sequence will be the basis of many discoveries and developments that will improve human health through new therapeutic and preventative approaches. We enthusiastically support the roles that the biotechnology and pharmaceutical industries will play in realizing the promise of human genomics through the development of new and important products. However, we also believe that the unusually large amounts of funding for DNA sequencing from both public and private charitable sources was motivated by a desire to make the sequence publicly available. We do not think it is appropriate for selected groups to gain a competitive advantage simply by virtue of having privileged access to the data.

For this reason, we are disturbed that the policy of the German Human Genome Program and the BMBF, as we understand from its description at the International Strategy Meeting, allows German industry restricted access to the prefinished sequence data for a three month period before the finished data are released to the public nucleotide sequence databases. We are convinced that the support enjoyed by the Human Genome Project in the United States, and elsewhere in the world, is predicated on the assumption that no one will have access to the sequence until it is publicly released for all to work with. We are concerned that the BMBF decision to limit access to the sequence produced by the German Genome Program may lead to erosion of that support and potentially to subsequent calls for protection of the sequence produced in this country, and perhaps elsewhere.

It is essential that all of the participants in the international Human Genome Program have the same policy with regard to the release of human DNA sequence data. We urge the BMBF to reconsider its decision and bring its policies into line with those of the other participants. At the International Strategy Meeting, it was argued that the privileged access of German industry to the sequence data produced in Germany was required in order to make German participation in the Human Genome Program possible. It is our opinion that, by definition, continuation of such restrictions on the immediate availability of the sequence data would mean that the German program is not, in fact, participating in the Human Genome Project as it is defined and practiced in the rest of the world.

Sincerely yours,

Francis S. Collins, M.D., Ph.D.
Director
National Human Genome Research Institute
National Institutes of Health

Aristedes A. N. Patrinos, Ph.D.
Associate Director
Office of Health and Environmental Research
Department of Energy

FSC/phf

Fakunding, Patricia

[redacted] [redacted]

Subject: Letter for FC signature

The following is a letter that I have sent by e-mail in Francis' and Ari Patrinos' names (after they both approved it). Could you do a paper version for their real signatures and send it? Thanks.
****

Dr. Frank Laplace
Federal Ministry for Research & Technology
Heinemannstrasse 2
D-53175 Bonn
Germany

Dear Dr. Laplace:

At the Second International Strategy Meeting on Large-Scale DNA Sequencing, held February 28 - March 1, 1997 in Bermuda, attendees reaffirmed how critical it is to the integrity and success of the international Human Genome Project that human genomic sequence data be rapidly released, without prior exclusive access to it on the part of anyone. We are writing to confirm that the National Human Genome Research Institute and the Human Genome Program of the Department of Energy agree with this principle and consider it to be critical to continued support for the Project in the United States. Accordingly, the NHGRI and the DOE have adopted policies to implement appropriate rapid data release practices for all of the large-scale human DNA sequencing projects we are supporting.

In our view, the key purpose of the rapid release policy is to ensure that the small number of laboratories (funded by public and private charitable sources) that have emerged with the capability to generate large amounts of human DNA sequence data do not take unfair advantage of that capability to gain privileged access to potentially valuable information. The human DNA sequence is a scientific resource of unprecedented importance to all of humanity. By studying it and coming to understand it in much greater depth than we do now, the human DNA sequence will be the basis of many discoveries and developments that will improve human health through new therapeutic and preventative approaches. We enthusiastically support the roles that the biotechnology and pharmaceutical industries will play in realizing the promise of human genomics through the development of new and important products. However, we also believe that the unusually large amounts of funding for DNA sequencing from both public and private charitable sources was motivated by a desire to make the sequence publicly available. We do not think it is appropriate for selected groups to gain a competitive advantage simply by virtue of having privileged access to the data.

For this reason, we are disturbed that the policy of the German Human Genome Program and the BMBF, as we understand from its description at the International Strategy Meeting, allows German industry restricted access to the prefinished sequence data for a three month period before the finished data are released to the public nucleotide sequence databases. We are convinced that the support enjoyed by the Human Genome Project in the United States, and elsewhere in the world, is predicated on the assumption that no one will have access to the sequence until it is publicly released for all to work with. We are concerned that the BMBF decision to limit access to the sequence produced by the German Genome Program may lead to erosion of that support and potentially to subsequent calls for protection of the sequence produced in this country, and perhaps elsewhere.

It is essential that all of the participants in the international Human Genome Program have the same policy with regard to the release of human DNA sequence data. We urge the BMBF to reconsider its decision and bring its policies into line with those of the other participants. At the International Strategy Meeting, it was argued that the privileged access of German industry to the sequence data produced in Germany was required in order to make German participation in the Human Genome Program possible. It is our opinion that, by definition, continuation of such restrictions on the immediate availability of the sequence data would mean that the German program is not, in fact, participating in the Human Genome Project as it is defined and practiced in the rest of the world.

/s/
Francis S. Collins, M.D., Ph.D.
Director
National Human Genome Research Institute
National Institutes of Health

/s/
Aristedes N. Patrinos, Ph.D.
Associate Director
Office of Health and Environmental Research
Department of Energy

The Second International Strategy Meeting on Human Genome Sequencing, Bermuda, February 1997

Scientists representing most of the publicly funded large-scale human sequencing centres around the world, together with officers from the funding agencies, recently met to review progress and consider new developments and strategies for sequencing the human genome.

The meeting agreed to a number of proposals designed to improve co-ordination and standardisation of world-wide sequencing efforts and re-affirmed the so called 'Bermuda' principles governing the release of data, such principles being reproduced below:-

"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 1kb 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."

Participants were deeply concerned to learn that not all national human genome programmes have adopted data release policies that are in concordence with the above principles. For example, one programme is funded by a consortium of the Government and industry and will provide a th~~ree mont~~h period of private, privileged access to human sequence data generated by that national programme before it is released into the public domain.

The sponsoring funding agencies (The Wellcome Trust, The National Centre for Human Genome Research Institute, NIH, and the Office of Health & Environmental Research, U.S. Department of Energy) have with the deepest concern agreed to issue the following statement:-

"Participants in the International collaborative programme to sequence the human genome have agreed to adhere to the 'Bermuda' principles on data release. Groups, who for whatever reason, are unable to agree to these principles will not be recognised as participants in the International Collaboration to sequence the human genome and will not be invited to participate in further International Strategy Meetings, until they are able to confirm adherence to the principles."

It is hoped that this statement will be helpful to scientific colleagues engaged in large-scale human sequencing who find themselves in conflict with the data release policy of their government or funding source.

We welcome participation by all groups who agree to the principles adopted by the international collaborative programme.

To: [redacted]
cc:

From: Sulston) @ INTERNET
Date: 03/10/97 11:26:43 AM
Subject: Andre Rosenthal, sequence index

File: Bermuda
mtg 2/98

Dear Michael,

Thanks very much for the information from Andre. This means that removal of the delay in German data release is vital to our continued association with him, because otherwise our position with regard to industrial interests elsewhere will be untenable.

I strongly support your statement that "the web site should be under neutral international control at the sequence databases". The ideal way of implementing this in my opinion is that the project, small as it is, should be jointly funded by the WT, the NIH and the DOE. The primary site could be either at EBI or NCBI, and mirrored by the other (as well as in Japan): these organisations have an excellent track record of working together, and will provide the necessary stability. HUGO can still be involved if it is in a position to do so.

I very much hope that something along these lines can be worked out. In addition to its functional role the sequence index has become an important symbol of cooperation on the human genome, and it needs to be seen as international in every way.

Best wishes

John

To: [redacted] [redacted] [redacted] [redacted] [redacted]
cc: [redacted] [redacted]
From: [redacted] [redacted] [redacted] [redacted] [redacted] [redacted]
Date: [redacted]
Subject: Human Sequence Index

In case this didn't get through on the other attempts:

Dear Francis:

I just returned from Toronto where I attended the HUGO Council meeting for the first time. The recent 'Bermuda' decision on the Human Sequence Index site came under discussion, and it is obvious that there are various forces trying to reverse the steps taken in Bermuda. While it might be appropriate to have HUGO's name attached to the Index in some way, I feel it is important to turn the Index over to the EBI/NCBI/DDBJ to ensure that the job is done right. Obviously, various details must be worked out and it is similarly important that the NHGRI/DOE/WT work together to get this underway and if necessary provide joint funding, so that it is truly internationally based regardless of what country personnel might be located.

I hope I am needlessly worrying about this, and if so I apologize for troubling you. If not, I hope you, Ari and Michael can keep this on track.

See you in Boston.

Bob Waterston

To: [redacted]

cc: [redacted]

From: [redacted]

Date: [redacted]

Subject: from Andre Rosenthal

Dear All,

I thought you should see Andre's response as soon as I received it. I would be very pleased to receive your views. On another Bermuda matter, I am afraid that (in my own personal view) HUGO is close to making a total 'pig's ear' out of the 'Bermuda' sequence index database.

Apparently they believe its development should in essence be the responsibility of GDB. From my perspective this would not be acceptable to the sequencing community, and would damage their effort and confidence in the collaboration. I believe the web site should be under neutral international control at the sequence databases. I would like to hear from you.

Best regards
Michael

---------- Forwarded Message ----------

[redacted]

[redacted]

[redacted]

Subject: from Andre Rosenthal

Dear Michael,

thank you for your mail. Your information is correct.

Schering has asked me to head their new Institute for Genome Research which is located on the campus of the Max Planck Institute of Molecular Genetics in Berlin-Dahlem next to Hans Lehrach.

The institute with the name "metaGen - Gesellschaft fur Genomforschung mbH" is a 100% daughter company of the Schering AG.

Although the building already exist it needs major renovation which is scheduled for the next five months. The official opening of the institute is expected by September this year.

The major objective of the new institute is to perform research on multifactorial diseases and to identify potential new candidate genes for spontaneous forms of prostate and breast cancer by using a combination of new methods. At a later point the interest of the new institute might shift to other multifactorial diseases. The institute is planning close collaborations with several academic groups from the Max Planck Society in Berlin-Dahlem or with! the Max Delbrueck Centre of Molecular Medicine in Berlin-Buch.

As you know I am a full professor at the Friedrich Schiller University in Jena and I will continue to work as a professor in Jena and as head of the Department of Genome Analysis at the IMB in Jena for the years to come. The president of the Jena University and the State of Thuringia have agreed and permitted that I will also work as a head of this institute for the Schering AG at 50% of my time.

These arrangements will not affect my duties and responsibilities as a professor in Jena and head of department at IMB. The "Genome Sequencing Centre at IMB" will continue to work in the framework of the international Human Genome Project as outlined by me last week at the 2nd Bermuda Meeting.

As you know I am personally committed to meet all the goals and agreements of the 1st and 2nd Bermuda Meeting espcially the instant release of genomic sequence data to public databases with no delay. I was the only scientist currently funded by the German BMBF who repeatedly critisized the intention of the German BMBF to submit genomic sequence data into a primary database which is accessible to a selection of German Pharmaceutical Companies called "Foerderverein" for a three months period. This critizism let to a substantial modification of the original proposal set forward by the German BMBF.

I know that the present situation is unsatisfactory to Britain and the United States. I am willing to continue to fight the agreement of the German BMBF but as i said in Bermuda it is important to find the right diplomatic way. In my view it will be more efficient if you and Francis Collins or representatives of the British and US government will get in close contacts with the appropriate officials of the German BMBF. The minister of the German BMBF is Mr. Ruetgers. I am confident that with your and the help of Francis Collins and with the support of all the colleagues at the Bermuda meeting we will be able to persuade the German BMBF to change their intention and adopt the same data release policies agreed by the Sanger Centre and the NIH and DOE funded sequencing centres in the States.

At the last dinner John and I briefly discussed the option to write a letter of correspondence to Nature about this issue signed by John, Bob, yourself, Francis, Jean and myself (and if necessary by all the other colleagues). I would be very willing to do so if you and others feel that this is a smart move.

You might ask how can I separate my work in Jena from the work I will be doing for the Schering AG? The answer is very easy. My research work as a professor and head of the Department of Genome Analysis at IMB in Jena is completely separated from my work for the Schering AG. Their is no overlap and no conflict of interest. I can assure you that this topic was carefully looked at by the State of Thuringia and by the Schering AG before gran! ting the permission. There is no "research collaboration" or any other contract between the IMB and Schering which will give Schering any right or access to genomic sequence data produced in my department.

If you agree I will also inform John Sulston and Bob Waterston about the new situation with Schering.

I hope I could give you all necessary information you wanted to know and I could clarify any doubt you might have had. If you have any questions or need more information please let me know.

Best regards,

Andre Rosenthal

THE WELLCOME TRUST

Mtg file
2/27 - 3/2/97

183 Euston Road
London NW1 2BE
Telephone: 0171 611 8888/Direct:
Fax: 0171 611 8545/Direct:

FACSIMILE TRANSMISSION

TO:	Dr Mark Guyer
FROM:	Dr Barbara Skene
YOUR FAX NO:	[redacted]
OUR FAX NO:	[redacted]
E-MAIL:	[redacted]
TEL:	[redacted]
NUMBER OF PAGES: (inclusive)	
DATE:	16 May 1997
OUR REF:	

Please see attached.

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 EPM1 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 organisiations) 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 bottle-neck 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 $60 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. gonorrohoeae* 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)

Investigator	Cumulative Finished Sequence	Predicted 1/3/97 - 28/2/98	1/3/98 - 28/2/99
Sulston	14.6	35	80
Waterston	4.8	24	24 + *
Lander/Hudson/Hawkins	2.1	20	80 *
Adams	2.7	11	14 + *
Gibbs	3	12	18 + *
Cox	0.3	5	?
Lehrach	0.24	1	2
Weissenbach	0	2	4
Mattick	0	0	?
Rosenthal	1.5	6	12
Bloecker +	0	1	2
Green/Olson	0.59	6	?
Chen	2.4	3.5	6.0
Sakaki +	2.7	3.4	30
Other Japan efforts	-	12	30
Evans	1.6	5	50 *
Palazzolo/DoE	4	20.0	50.0
Roe	3.8	5.5	12 *
TOTAL	**44.33 Mb**	**172.4 Mb**	**? Not meaningful to estimate total (384+)**

* 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

Chromosome	Gap Frequency	Reason	Investigator
4	1/55 Kb	CpG rich	Bentley
X	1/750 Kb	gene-poor region	Bentley
22	1/282 Kb		Bentley
13	1/121 Kb	CpG rich	Bentley
	1/200 Kb		Hudson
C. elegans	1/250 Kb		Sulston

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. There should be an effort 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)

<u>EST mapping</u>

Number of ESTs mapped	Centre
6,000	Sanger Centre
6,000	Genethon
3,000	Whitehead
2,000	Stanford
17,000	**Total**

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

Human syntenic region	**Size of region (Mb)**	**Investigator**
11p23	1	Rosenthal
12p13 (CD4)	0.2	Gibbs
13	1	Sanger
17	1	Hudson
22	>= 0.7	Roe
Xp22 (PGK)	0.1	Gibbs
Xp27	0.2	Gibbs
Xq28	3	Rosenthal
Xq28	2.5-3	Brown (Oxford)
Total	**~10 Mb**	

Currently 400,000 5′ end sequences are available: 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.

2/28/97

Intl. Sequencing Mtg. - Bermuda

I. Introduction - Morgan

Hinxton - Pathogen sequencing facility
Area for courses
Conf. facility

II. Presentations - Cox, chair

GB4, use PACs for ordering

A. Sulston

Chr. $\underbrace{22, X}_{1/2}$, 20, 6, 1

665 Mb

Strategy RH → PAC screening → Fingerprinting /STS analysis
→ GAP closure
→ Sequencing M13, some pUC
Phrap

Need to focus on finishing - software, some hardware

May have to do YACs - Xfer into windows

Status 10-20 markers/Mb
97 Mb covered in clones

Seq: 14.6 out in GenBank] publicly available
(Mb) 11.9 unfinished
17.5 ready to go

Total output 1996: 34 Mb
Ever: 52 Mb

Chr. 22 - substantial coverage now
X - less covered. Example on Xq22

Plan 30-40 Mb this year, 80 next yr.,
100 Mb/yr after that

B. Waterston: 7, 22, X

Finished 1.85 mb

Submitted 2.95 mb

In Finishing 11.6 mb   Bottleneck? Clones were limiting

In Shotgun 15.1 mb

In Library 3.5 mb

Strategy — STSs on chr. 7 (1/79 kb)

→ Bacterial clones (BACs by EG, Hyb. to PACs by Wash U)

→ MD fluorimager of BACs, Hind III

Use Sanger software

Minimal end-walking

600 STSs across 50 mb

128 BAC/PAC contigs

~250 kb avg size → 32 mb

175 clones underway, 21 finished

Shotgun/directed → M13, pUCs → phred, phrap → finish → consed

Q.C. Restriction digest ×3

Reassemble c̄ alt. versions of phred/phrap

Complete contiguity

Annotate

1:10,000 error rate sought   Anderson

Tech — Want 64-72 lanes/377. Gel loaders. Dye term. Ty transposons

Future - 96 lanes, pipetting station, Lloyd Smith sequencer

Projection 1 Mb/month now → aim for >2

C. Hudson

Screen 20X BACs, PACs c̄ STSs q 100 kb

Contigs ~350 kb

Areas of contigs → will need to be walked across

Use BAC end sequencing

Isolated STS → 12 clones

Fingerprint - pick one where every band is represented

in other clones HindIII, EcoRI

Know which go furthest → STS, screen, walk

? end sequence

Re-made RG pools

Use Genomatron - 100 STSs/day

Hawkins - Seq. — 11 gaps (size known by PCR)

2.1 Mb finished — mostly chr 17 BACs

0.65 in finishing

Aim for 5Mb by 5/97, 20mb the next year, then 80

Sequatron II ↓ automate front end

QC/QA - Goal to gel fluctuations, track to well

Finishing! The bottleneck - Needs to be a production line

Human - mouse system

France has joined them. GDB is decimated

Assembles using "Alewife" - overlapping segments

of 25-mers. Also provides a std.

Almost none of this is in GenBank!

C. Adams

Seq. non-overlapping BACs from 60 STSs on 16p

Using BACs as probes against human genome DNA

Arabidopsis scaling to 10,000 reads/month

Same team is doing human

Goal was 2.7Mb → have 2.6 mb, submitted to GenBank

735kb in closure

1.9 mb ready for random seq. (<5% E.coli, <5% pUC)

Library Team, Random Team, Closure Team

Scale up of finishing is a challenge – 90% of it is a software issue

Uses phrap & TIGR Assembler

Goal is 11 mb for next year "very ambitious"

Robot will help

12 genes per 2,363,073

1 gene/196 kb

There are 200kb BACs c̄ no ESTs, no GRAIL hits

D. Gibbs

Progress – 3mb in GenBank (1.2mb in previous 4 mos)

ABI → BODIPY x̄ for walking (very small Fx)

Power of full length cDNA seq.

Concatenation

Have done 180 F.L. cDNAs

One expt 78 cDNAs → 100kb concatemer

Human vs. mouse also very helpful

Want to reach 15mb next year

Xpter, chr.12

Expand Dilks collab.

E. Cox – Goal is to do 200mb of chr.4 by 2005

Last year target: 2.5mb – won't make that

chr.21 EPM1 1.2 mb

D3 0.3 mb

chr.4 4q25 5mb

In GenBank 100kb finished

1.2mb of clones > 3kb

Whole genome radiation hybrid maps – G3, in press
Can map to 240kb theoretically (300-500 more realistic)
GB4 1mb (1.2-1.5)
But if coverage isn't random, won't know there are gaps c̄ larger bin sizes
Transposon method – vulnerable to bad libraries ←
Chip: 140 × 25bp standard design
<1% false ⊖ <2% false ⊕
* Need to understand this
PCR up STSs, 800 at a time, hyb. to chip
Use to determine tiling path, check assembly
Cost is 1.5¢/bp

200 3kb clones — end sequence, then design chip to get tiling

F. Fiona Francis (Lehrach)
Planning 6Mb over 3 years 1-2-3
3 groups – Rosenthal, Lehrach, Max Planck
chr.21 – seq, ready maps
Hyb screening c̄ cosmid lib & PAC (BAC later)
Some FISH, Restr digest like Wash U for minimal tiling path
Shotgun into pUC (no picking robot), Phred/Phrap
Xp22 PAX 243kb, 9 cosmids
In progress 21q, Xq, 17p
Using oligos to preselect shotgun clones (8-mers)
by bar-coding → more even spreading
No data

G. Weissenbach - Hasn't started yet. Announced by Minister of Research,

$14M/yr.

In Evry, near Genethon   30-35 people from Genethon will move over   joint venture c̄ CNRS, private Co (tech. xfer) to allow hiring

Start summer 1997

Sign lease in a couple of weeks -

office bldg, will need 4-5 mos. to renovate

Projects - In house

Collaborative - eval. by scientific committee. Academic

Ratio?

There is also "Steering Committee" which could change priority

Data release & I.P. will be decided by Steering Com.

In house → release more likely

Collabs → different

Will do some Arabidopsis, probably some microorganisms

Also tetraodon

TGS is Gen Set's private facility (5'ends of cDNAs, 30-50K)

H. Mattick Australia

$8M/yr. Voted by Fed. govt.

Facility to begin function mid year

Melbourne - Simon Foote Dick Cotton

Genotyping, mutation data. 8M genotypes/yr.

Queensland - sequencing, Mattick

Expect ~30 ABIs   1500 templates

Have $ for infrastructure, not projects -

will need to draw on other sources - a problem - funding agencies [illegible] +

Service sequencing? - ESTs for plants

In house? Pathogens. Human [illegible] - clones to be provided by suppliers

I. Rosenthal

1.5 mb in GenBank now

Targets – Xq28 3Mb

Xp11 2.5mb

X-PABA 1 mb

21q 25mb

7q 7q22 7mb Scherer/Tsui

7q32 0.5mb

Mouse syntenic region of 3mb – Xq28

1300 reads/day → 3000 by 5/97?

20 ABI's (16 bought by industry) Bloecker

German Human Genome Project – work c̄

IMB Rosenthal	Lehrach	Broecker	
4	1	1	← start 5/97
9	2	2	
15	3	3	

Have 6 mb available

Doing comparative Seq. in Fugu; Disease gene; rhizobium

Zebrafish – no organized effort

Very interested in methylation

J. Green (Olson)

Fidelity – 2x validation of all sequence-ready clones, using methods adequate to detect small (<1 kb) Coligations, deletions, Xposon

→ use as start point

Accuracy: <1/10kb

Submit base-specific error probs.

Independent test of assembly accuracy

Contiguity - All gap sizes estimated, all contigs oriented
and ordered within the chromosome

MCD mapping

Chr. 7 2mb mapped 7q31.3

HLA " 700kb seq

Mouse TCRα

340kb submitted

Bottleneck in editing

Expect to meet 2mb goal for year 1

Doesn't state 2nd year goal - waiting for $

Discrepancies -

Chr. 7 0 in 2x38802bp

HLA 2 in 2x43084 bp

1 was a phrap error

1 cosmid mutation 12bp ins/del

K. Chen - ACGT, div. of ABI - collab. c̄ Schlessinger

20 people, 4 groups - ~~10~~ 11 ABIs

New institute in Shanghai (ABI, Sequana)

55% of budget is govt. grants

X 2.4mb, at a rate of 3mb/yr → 0.5Mb in GenBank

Micro - Ureaplasma 760kb, 99% done

Arabidopsis 0.4mb/yr 3 in 1998

Ordered shotgun 30 by 2000

→ See NAR

10kb clones (λ) 0.5mb/tech/yr

Mapping done by Schlessinger → BACs

New dye primers - lower background (better spectral sep.),
equal mobilities. 4 mos.

Sakaki not here - broke shoulder skiing

L. Fujiyama - Japan

4 groups - JICSD → JSC

Nakamura - chr. 3, 8, 9

Sakaki - chr. 21

Shimizu - chr. 21/22

Fujiyama -

21q

Sakaki

2.7 mb finished in 3 contigs

500 kb to be finished by end of March

Next FY 3.4 mb in 4 regions - have contigs of part

Directed deletion method

Testing Hitachi capillary sequencer (96)

Not sure if it will be commercial

Expanded facility - scientists agree, govt. slow to respond.

Start FY98?

Mb: 15, 30, 60 → (7 yrs)

(98) (99) (00)

chr. 21 h21/m21 m21/h11

m = mouse syntenic region

Budget - economic decline is affecting

$60M will be severely cut ($20M?)

Data release by JST

900 kb available

Sakaki has his own Web site

M. Evans Chr 11, 15

Chr 11 - 905 STSs

17,965 end sequences from cosmids

Chr. 15 - harder, less well mapped

High density grid hyb. c̄ pooled STS-specific oligos

4 restriction enzyme fingerprints of each PAC

Chr. 11   11p → PAC contig of >3.5 Mb

2465 STSs, screened against 465

3185 PACs

467 fingerprints

216 PACs → 3 c̄ mixed signals (1.3%)

Seq. strategy

Sagian

Auto-finishing - use phred/phrap output and high capacity oligo synthesizer

Accuracy - want Phrap scores > 40

Phase I ⎤ 29 → >1 Kb ordered (II) or unordered (I)

II ⎦

III Closed - 10^{-3} to 10^{-4} ⎤ → GenBank

IV Uniform $>10^{-4}$ accuracy ⎦

Annotation

155 Kb 11p13.3 color coded output, showing overlaps

Available on Web

Per base sequence displayed

Automation

<10¢/nt Small scale

MerMade oligo 96/192 300/day → Avantec, Inc., c̄ no-cost license from UT

Sagian robot (Beckman purchased) - 3 m rail

DNA seq'r - Astral. 7 months. A lot like ABI. Uses hyperspectral imaging

Chr. 11 - needs coordination, desperate by STS, not band

N. Palazzolo

Won't present JGI

800 Kb/month

Physical map - random light shotgun, build paths, transposon

Quality - all double stranded

Hardware – New space needed

colony picker, oligos ...

Partnership c̄ Motorola: Chicago group designs their 95 factories, does their tech transfer

Volume, quality, cycle time, cost

Need precise goal definition – we don't have it

Peer review is impossible

Benchmarking – metrical tools

Process model – must have predictive value. Looks only at volume

Bottleneck analysis – predicts where to put R&D

Cost model (Motorola paid)

Cost accounting

Pick-a-mix – cost models.

Predicts effects of changing volume on a spreadsheet

Did an LBNL review – cost $250K, 3 mos.

Chr. 22

O. Roe – Chr. 22 3.8Mb in GenBank

He doesn't do mapping

Chr. 9 (bcr-abl) → Rowley collab.

Interested

Aspergillus

N. gonorrhoeae 2.2mb

Strep. pyogenes 1.9 mb

] 95% on website

Sees 2genes/100Kb

"40% of the human genome is sequenced" – The Alens

III. Data Quality

Day 2

IV. Cost – Palazzolo, chair

Value/Danger

Methods

Validation

Need to collect data in a serious way

Methods – separate out R&D ?

1) Cost model extrapolations – easiest, but prone to error

Ex oligo synthesizer, miss cost of reagents that had to be thrown out

2) Cost accounting

Separate budgets for each activity

Estimates turned out to be 2-3X low

3) Cost models

Define product, establish process flow model, fixed protocols, databases on cost [materials, equipment, stock solns., labor

→ Identify & manage R&D opportunities

Genome Cooperative Purchasing Group ?

Govt. can't take a leadership role

4) Output-based

NHGRI to take a role ?

Do audits in a couple of places

Then send around an MBA to instruct the rest

Rosenthal – unhappy c̄ generality

Aim for 30¢/bp

"Game of liar's poker" – MP

	$ in/out	other $
Gibbs	50¢	60¢

V. Data Release

VI. Etiquette – John/Bob

Mapping
Sequencing] Claims may be different

Mapping doesn't entitle Sequencing

Sanger Center has gotten into conflict on chr. 1 c̄ TIGR

Their mapping strategy focusses on whole chromosome

X chromosome – different mapping resources were very helpful

Mapping can be redundant, Sequencing shouldn't be

Sequencing – claim no more than a year

HUGO site

HSM Index – Flat text file

Don't need to make this link explicit

FC proposes giving it to NCBI

Lipman: GenBank postdoc could curate

Cameron: EBI could support too – be careful about not calling it GenBank

NCBI/EBI/DDBJ – May Advisory Mtg. Put in other organisms too?

HUGO Council will meet next week

Genethon markers as the boundaries

Minimum size – Megabase? Between Genethon markers agreed to

Concern that ~~#~~ small scale efforts not be injured by claims

↳ Is this happening?

Maximum - a year's worth

No more than 3-5X what you did last year

Sequence-ready map is a significant investment -

it's tacky for someone else to move in on it

Specific issue of chr. 1

Should Sanger be expected to turn over maps?

To TIGR?

Ex: Chr. 11 Peter Little wanted to do 11p13 and 11p15

Overlap c̄ Evans?

End up c̄ 2 sequence-ready maps

VII. Annotation

Standards? What should be submitted?

"Electronic BSE"

Can look at 1° data to check for Ψgene; frameshifts -

producing centers are in a better position to

do this than users

Should all traces be made available on the internet?

Storage of traces? Tape → optical disk

* John Spouge, NCBI MD PhD

Sen. Sci. - assist c̄ data exchange

plan *

What about non in-silico methods?

Software: What option is best? Algorithm to synthesize?

Lipman agrees its database better, in fact, shouldn't

even report unless you have real exptl. data

"Suspected gene" is helpful - exon structure isn't reliable

ESTs → through end of 1997 from NCI, Merck, Genentech, BMS

8000/week being asked by NCI

3' ends + sizing

Subtracted libs / normalized libs?

Lifetech, Stratagene → 20 libs each

Soares → 15,000 used to subtract a pool of libs →

4x ↓ in those clones

Cluster algorithm to find all > 1 rep.

singletons is rising at a slower rate

than clusters now (28% → 21%)

Mapping

Cox urges high resolution panels

MIT 3000

Sanger 6,000

Genethon 6,000

Stanford 2,000

} most on GB4.

17,000 more by June!

Update Web then → no, sooner!

→ RHdb

Can get data now but have to go to 4 webs

Schuler

↓

WWW/NCBI

Full clone seq.? NCI will fund ~15,000

Mouse: 1-2 mb comparisons beginning to appear

1 Mb of chr. 11 in Germany (won't say where)

Xq28 2.5-3 mb Steve Brown IDS

1 mb Rosenthal

Mouse IDS /

12p13 (CD4) ~~PGK~~

Xq (PGK)

} Gibbs each ~0.2 Mb

MIT - 1 mb mouse nu / human 17

Roe - 500kb " [illegible] 2 BACs chr. 22 → ~1 mb Reeves + grant

Sanger 1-2mb BRCA2

Bruce – useful to find genes missed by ESTs

10% of human! (André)

FC – no more than that!

Mouse ESTs

Aim for 30,000 full clone seqs. in next 2 yrs.

EC consortium to map mouse ESTs to

Goodfellow RH panel

Oxford ESTs

Genethon will do 3000 – ? TOTAL?

RH panel is low resolution

Not much enthusiasm for higher resolution because it wouldn't coalesce

Feb. 27-28 – March 1

Evening session? Free afternoon?

Statement:

Needs more explanation of rationale?

And moderation of statement re Germany

Michael will work c̄ Ursula to re-word

Assume cost/bp is following an exponential decay rate with half-life $t_{1/2}$

Then $$n = \frac{\ln\left[\frac{kc_oM}{y} + 1\right]}{k}$$

Where

$$k = \frac{\ln 2}{t_{1/2}}$$

c_o = cost at time zero

M = total sequence that must be done (in Mb)

y = $M/year available for sequence production

n = number of years to finish

EVERYONE – PLEASE EDIT THIS TABLE FOR YOUR CENTER, AND RETURN TO FRANCIS COLLINS BY FRI. AFTERNOON COFFEE/TEA.

Human Sequence Production

Investigator	Cumulative Finished Sequence	Predicted 3/1/97 – 2/28/98	3/1/98 – 2/28/99
✓ Sulston	14.6	35	80
✓ Waterston	~~1.9~~ 4.8	~~12~~ 24	24+
✓ Lander/Hudson/Hawkins	2.1	10	80
✓ Adams	~~3.6~~ 2.7	11	14+ (?50)
✓ Gibbs	3	~~15~~ 12	~~100~~ 18 (?100)
✓ Cox	~~0.1~~ 0.3	5	?
✓ Lehrach	0.24	1	2
Weissenbach	0	?	?
✓ Mattick	0	0	?
✓ Rosenthal	1.5	6	12
✓ Blaecker*		1	2
✓ Green/Olson	~~0.34~~ 0.59	6	?
✓ Chen	2.4	3.5	6.0
Sakaki*	2.7	3.4	} 30
Other Japanese efforts		12	
✓ Evans	1.6	5	50
Palazzolo	4		
✓ Roe	3.8	5-6	12?
	44.33 = 1.5%		

[PLEASE FILL IN ANY GAPS! – AND DON'T BE OFFENDED AT ERRORS!]

* Not present, but reported on by others

2/97

HUMAN SEQUENCE PRODUCTION (mb)

Investigator	Cumulative Finished Sequence	Predicted 3/1/97 - 2/28/98	 3/1/98 - 2/28/99
Sulston	14.6	35	80
Waterston	4.8	24	24+*
Lander/Hudson/Hawkins	2.1	20	80*
Adams	2.7	11	14+*
Gibbs	3	12	18+*
Cox	0.3	5	?
Lehrach	0.24	1	2
Weissenbach	0	2	4
Mattrick	0	0	?
Rosenthal	1.5	6	12
Bloecker+	0	1	2
Green/Olson	0.59	6	?
Chen	2.4	3.5	6.0
Sakaki+	2.7	3.4	} ~~30~~ 15+*
Other Japan efforts	~~1~~3.7	~~12~~ 3.7	
Evans	1.6	5	50*
Palazzolo/DoE	4	20.0	50.0
Roe	3.8	5.5	12*
TOTAL	44.33 mb	172.4 mb	? Not meaningful to estimate ~~total~~ (384+)

* Production dependent on funding decisions - some centers (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

HUMAN SEQUENCE PRODUCTION (mb)

	Cumulative	Predicted	
Investigator	Finished Sequence	3/1/97 - 2/28/98	3/1/98 - 2/28/99
Sulston	14.6	35	80
Waterston	4.8	24	24+ *
Lander/Hudson/Hawkins	2.1	20	80 *
Adams	2.7	11	14+ *
Gibbs	3	12	18+ *
Cox	0.3	5	?
Lehrach	0.24	1	2
Weissenbach	0	2	4
Mattrick	0	0	?
Rosenthal	1.5	6	12
Bloecker +	0	1	2
Green/Olson	0.59	6	?
Chen	2.4	3.5	6.0
Sakaki +	2.1	3.4	} 30
Other Japan efforts	-	12	
Evans	1.6	5	50*
Palazzolo/DoE	4	20.0	50.0
Roe	3.8	5.5	12*
TOTAL	44.33 mb	172.4 Mb	? Not meaningful to estimate - total (384+)

* Production dependent on funding decisions - some centers (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

F/u to Bermuda

DoE - Aravinda to visit Ari.

San Antonio Gene Expression mtg?

Santa Fe Contractors workshop?

Evaluation of quality - write up conclusions

Convene a working group? → CSH

Talk to NCBI person?

Cost - Auditors to 2-3 places?

Do it soon - educational, not punitive

FC visit TIGR

McCombie

MOUSE PILOTS

UNDERWAY / PROPOSED

Human chr. syntenic region		
11	1 Mb	?
11.23	1 Mb	?
12p13	0.2 Mb	Gibbs / Baylor
13	1 Mb	Sanger
17	1 Mb	Hudson / Whitehead
22	≥0.7 Mb	Roe
Xq22	0.1 Mb	Gibbs
Xq27	0.2 Mb	Gibbs
Xq28	3 Mb	Rosenthal
Xq28	2.5-3 Mb	Brown / Oxford
	~ 11 Mb	

France

Jean Weissenbach stated that France was currently considering the development of a French genome sequencing programme but nothing had yet been agreed.

Germany

Frank Laplace (Federal Ministry of Research and Technology; BMBF) informed participants that a Scientific Advisory Board for the German Genome Programme would convene shortly to initiate the programme. The BMBF would be providing funding of DM 40m-50m per annum which would include support for two resource centres to be directed by Hans Lehrach and AnneMarie Poustka. The Deutsches Forschung Gemeinschaft would be providing an additional DM 5m-10m for genome studies focussed on the identification of disease genes. It was hoped that additional funds would be provided *via* investment from industrial partners and discussions were currently in progress with this aim. Industrial participants were requesting privileged access to data for three months prior to publication but this was currently the subject of further negotiations. Notwithstanding industrial sponsorship, Frank Laplace endorsed the principle that work funded with public money should be in the public domain.

U.S. Department of Energy

David Smith stated that the DoE budget for the human genome programme in 1996 was $70m per annum of which $10m was attributable to human and mouse sequencing and $15m to development of new sequencing technologies. In addition to this funding, the DoE also provided $4m per annum in support of microbial genome sequencing.

U.K. Medical Research Council

Sohaila Rastan stated that the MRC currently provided support for the C.elegans sequencing programme at the Sanger Centre at the level of £13.1m over 5 years (1993-1998). In addition, a further £10m would be available over 5 years from 1995 for genome research at the Sanger Centre; £2m of which would be used to ramp up and complete the C.elegans genome sequencing project. The remainder would go towards the human sequencing programme.

Accuracy

It was agreed that sequencing centres should aim to achieve 99.99% accuracy.

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.

THE AUSTRALIAN GENOME RESEARCH FACILITY (AGRF)

— funded at $A10m ($US8m) for equipment only (project funding to be obtained separately)*

— Two DIVISIONS :

(1) DNA SEQUENCING at the Centre for Molecular and Cellular Biology, University of Queensland, Brisbane

(2) DNA GENOTYPING at the Walter and Eliza Hall Institute of Medical Research, Melbourne

— currently in final stages of planning and equipment acquisition, due to begin operations mid-1997 (~ 30 × 377s + assoc. equipment, robotics)

DNA SEQUENCING (University of Queensland)

Current status: 4 × 373s ~ 800 templates/week

Projected: ~ 15 × 377s (+ existing 373s)
~ 1500 - 2,000 templates/day

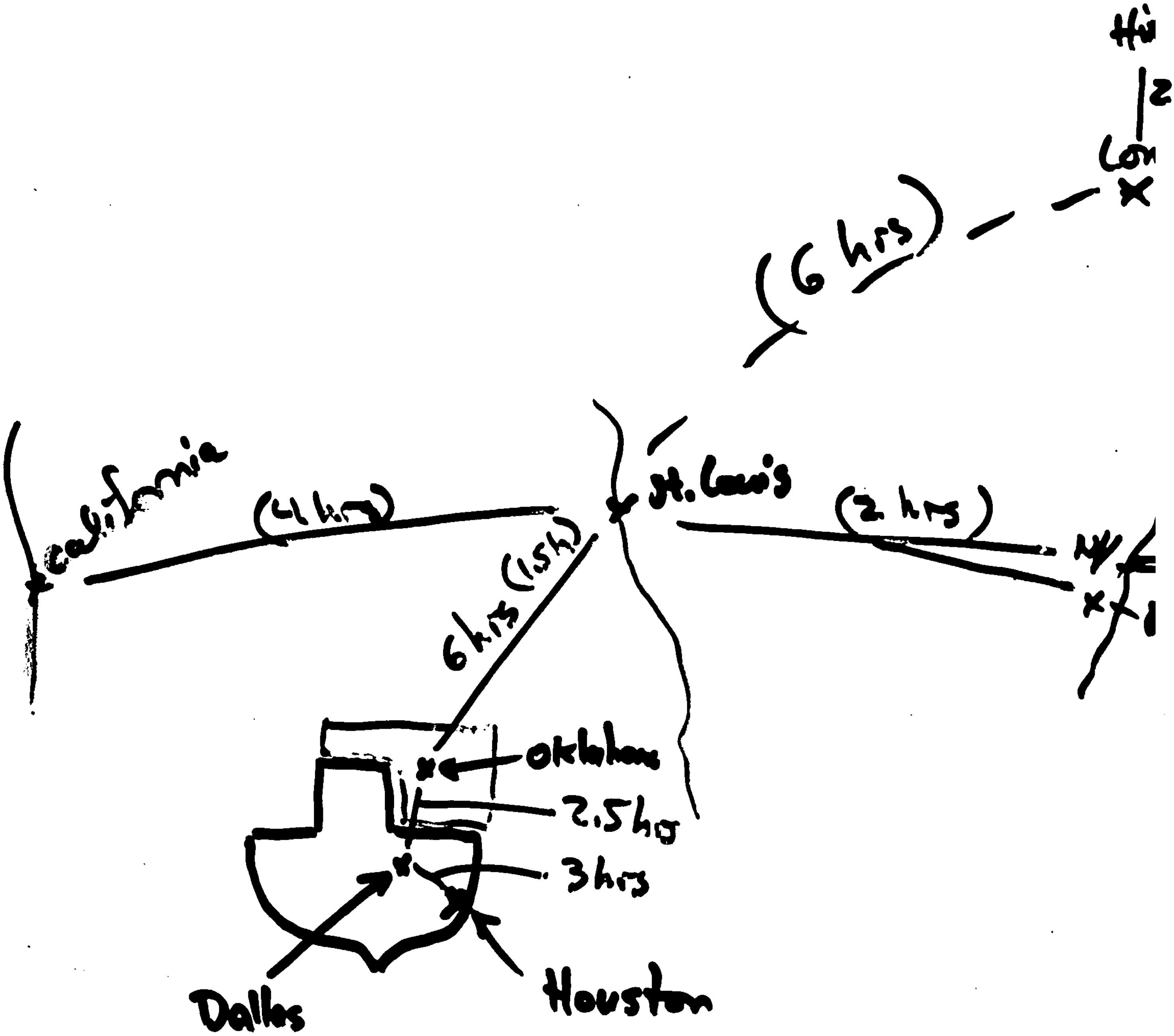
Housing: Proposed new Institute $A50m ($US40m)
— have obtained $A30m from University of Queensland and State Government
— attempting to raise $A20m from Federal Government and other sources
— construction 1997 - 1999 with temporary housing for facility in the interim

- OPERATIONAL

  - AGRF will be a generic high-throughput sequencing facility, not restricted to particular projects, available to Australian and regional research community.
  - no operational funds (yet) voted to the facility. These are intended to be derived by participating groups from granting agencies.
  - two modes of operation:
    - (a) contract/service — on behalf of client-groups who will supply funds and who will take primary responsibility for cloning, library construction, sequence assembly and annotation (using own facilities, supported by services provided by AGRF and ANGIS - Australian National Genome Information Service)
    - (b) bid for specific funds to undertake large projects in-house, and construction of teams for this

FUNDING*

- from existing granting agencies
- working to convince Australian Government to set up a specific fund (~ $A20m/year) to support genome-scale projects, including an Australian participation in the human genome sequencing project.


Hi
2
Con
(6 hrs)
California
(4 hrs)
St. Louis
(2 hrs)
NY
6 hrs (1.5 h)
Oklahoma
2.5 hrs
3 hrs
Dallas
Houston

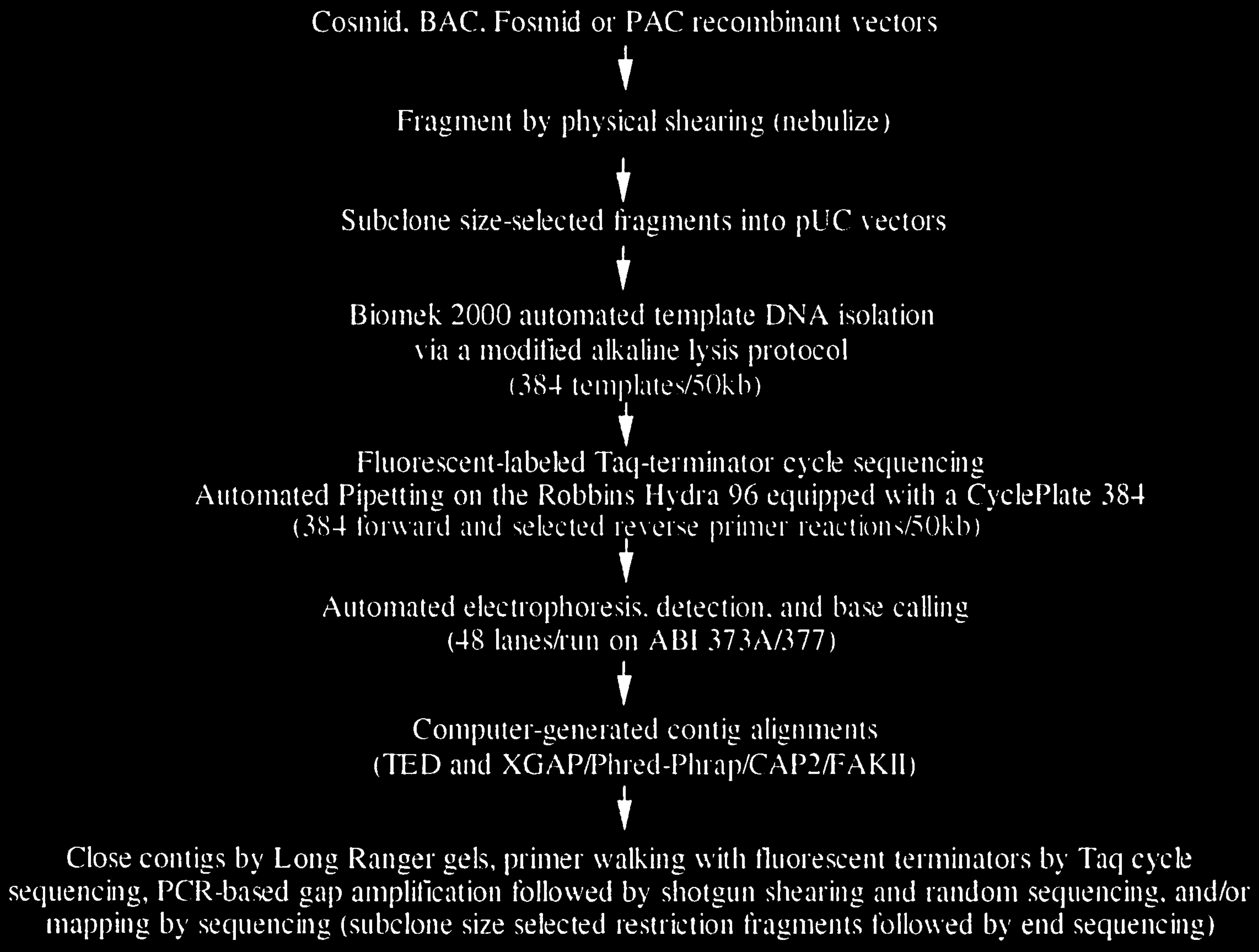
Total sequence data submitted to GenBank

8-31-95-9-1-96	1,846,870 bp
9-2-96 - 11-15-96	1,997,137 bp
Total additional in progress	2,154,832 bp
Total	5,998,839 bp

4(2) 377's - Hu/Mo

2(2) 377's - Bact.

Shotgun Cloning, Automated DNA Isolation, Fluorescent-Based DNA Sequencing, and Closure



Cosmid. BAC. Fosmid or PAC recombinant vectors

Fragment by physical shearing (nebulize)

Subclone size-selected fragments into pUC vectors

Biomek 2000 automated template DNA isolation
via a modified alkaline lysis protocol
(384 templates/50kb)

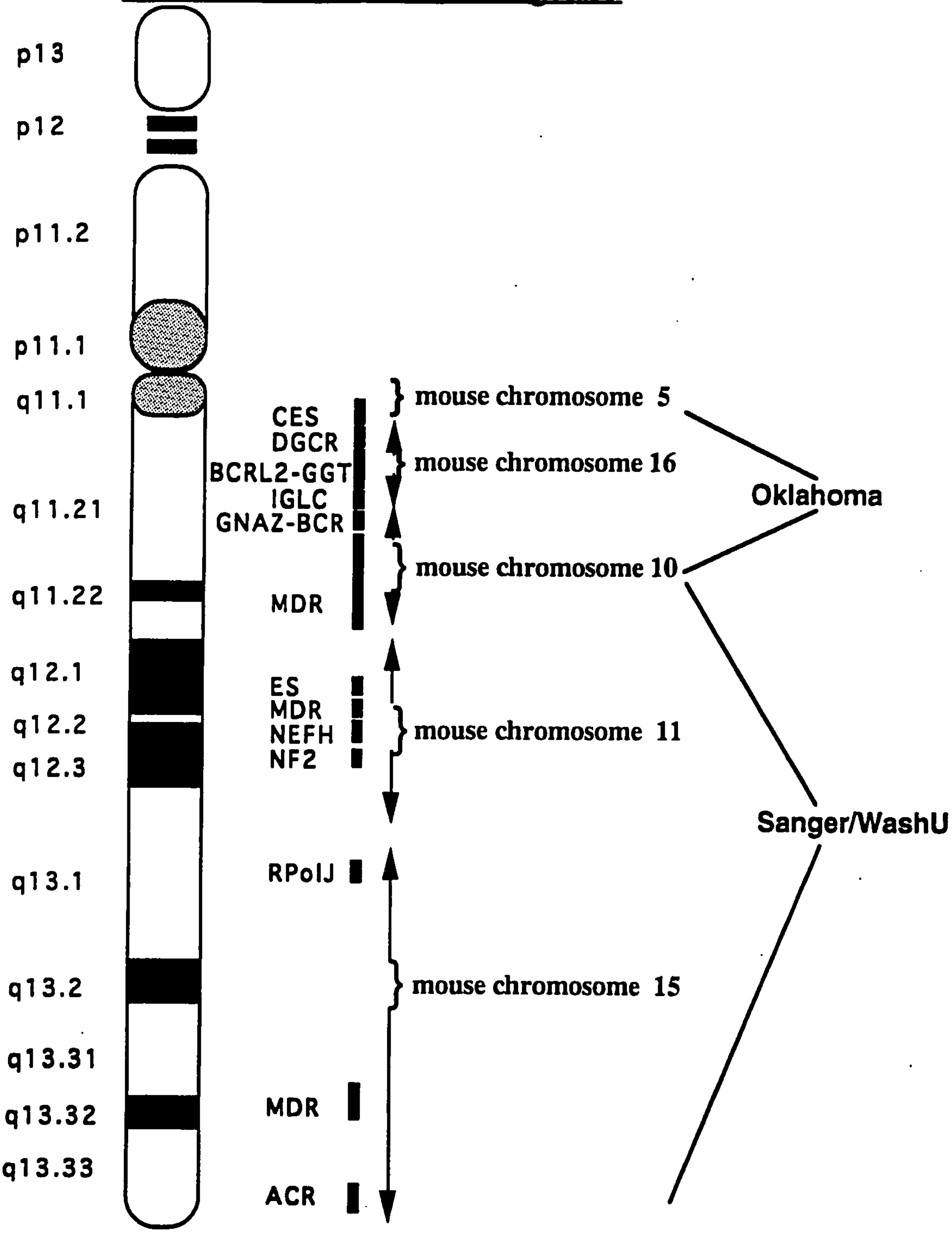
Fluorescent-labeled Taq-terminator cycle sequencing
Automated Pipetting on the Robbins Hydra 96 equipped with a CyclePlate 384
(384 forward and selected reverse primer reactions/50kb)

Automated electrophoresis. detection. and base calling
(48 lanes/run on ABI 373A/377)

Computer-generated contig alignments
(TED and XGAP/Phred-Phrap/CAP2/FAKII)

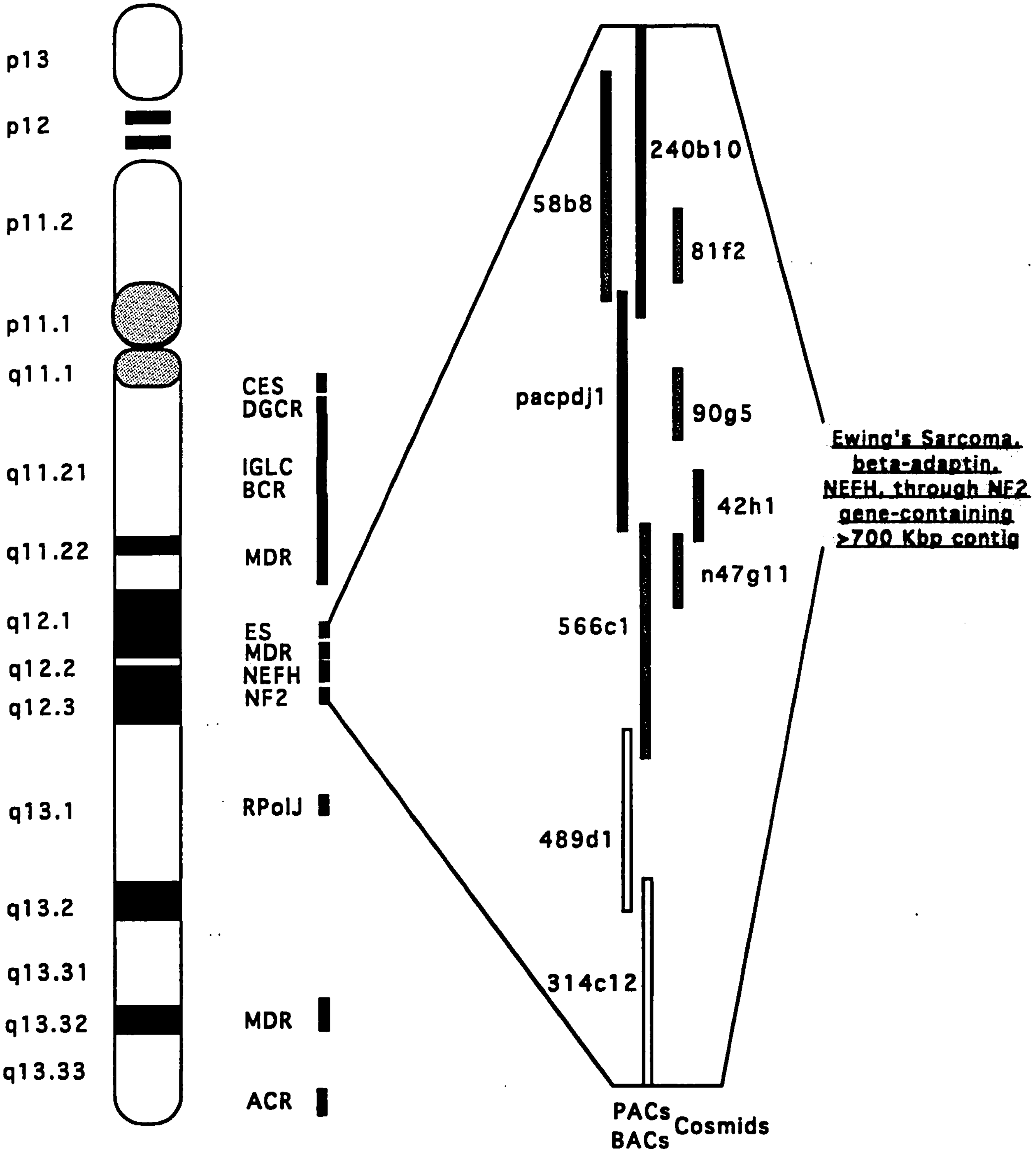
Close contigs by Long Ranger gels, primer walking with fluorescent terminators by Taq cycle sequencing, PCR-based gap amplification followed by shotgun shearing and random sequencing. and/or mapping by sequencing (subclone size selected restriction fragments followed by end sequencing)

Human Chromosome 22 and Syntenic Mouse Chromosomal Regions



CES = Cat Eye Syndrome Region
DGCR = DiGeorge Syndrome Critical Region
IGCL = Immuoglobulin Light Chain Region
GNAZ =Guanine Nucleotide Binding Protein
BCR = Breakpoint Cluster Region
MDR = Meningioma Deletion Regions
ES = Ewing's Sarcoma
NEFH = Neurofilament Heavy Subunit
NF2 = Neurofibroblastoma Region 2
RPolJ = RNA Polymerase II subunit J
ACR = Acrosin

Cosmid, BAC, and PAC clones in the Ewing's Sarcoma through NF2 Regions of Human Chromosome 22

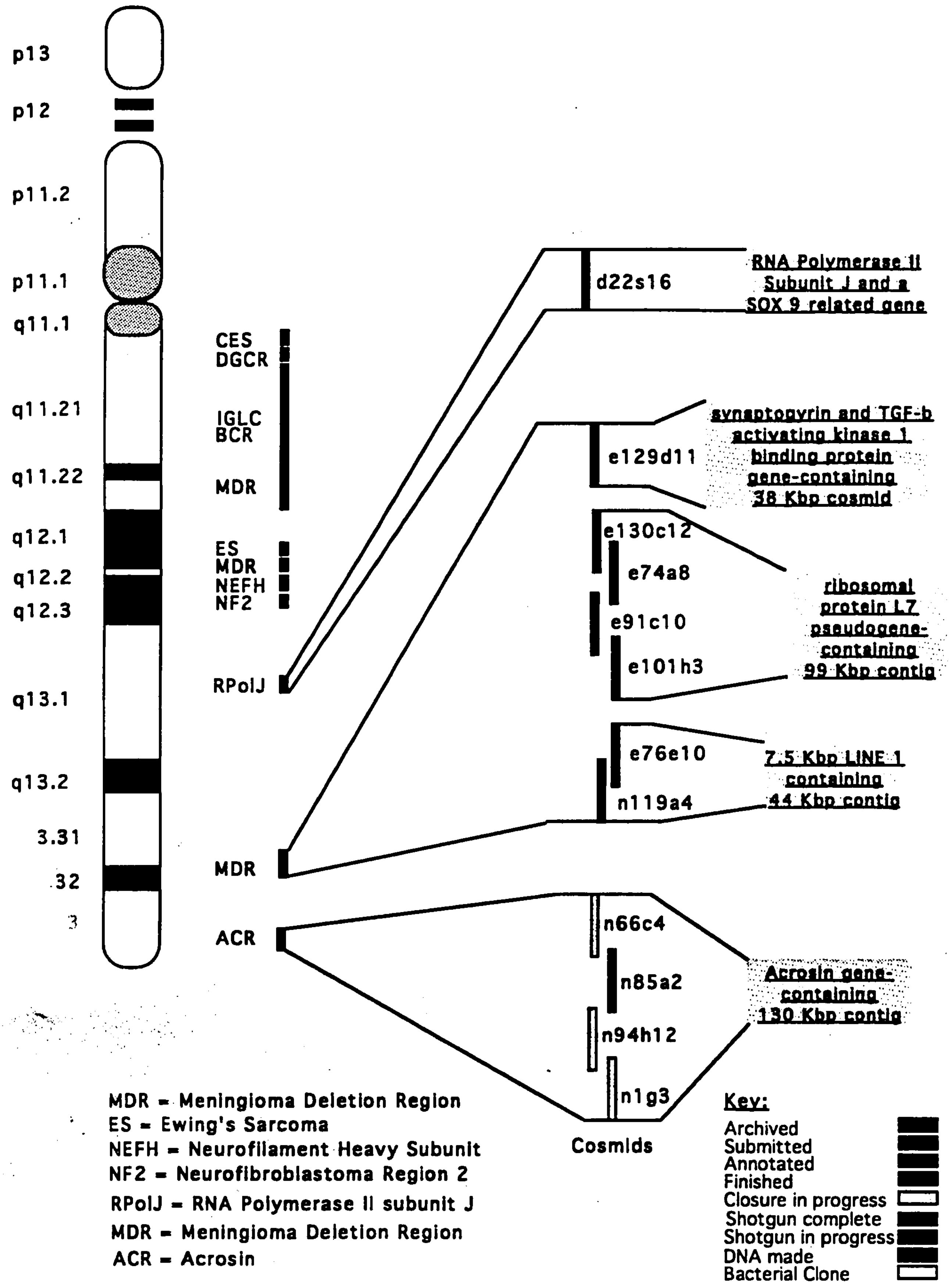


MDR = Meningioma Deletion Region
ES = Ewing's Sarcoma
NEFH = Neurofilament Heavy Subunit
NF2 = Neurofibroblastoma Region 2
RPolJ = RNA Polymerase II subunit J
MDR = Meningioma Deletion Region
ACR = Acrosin

Key:

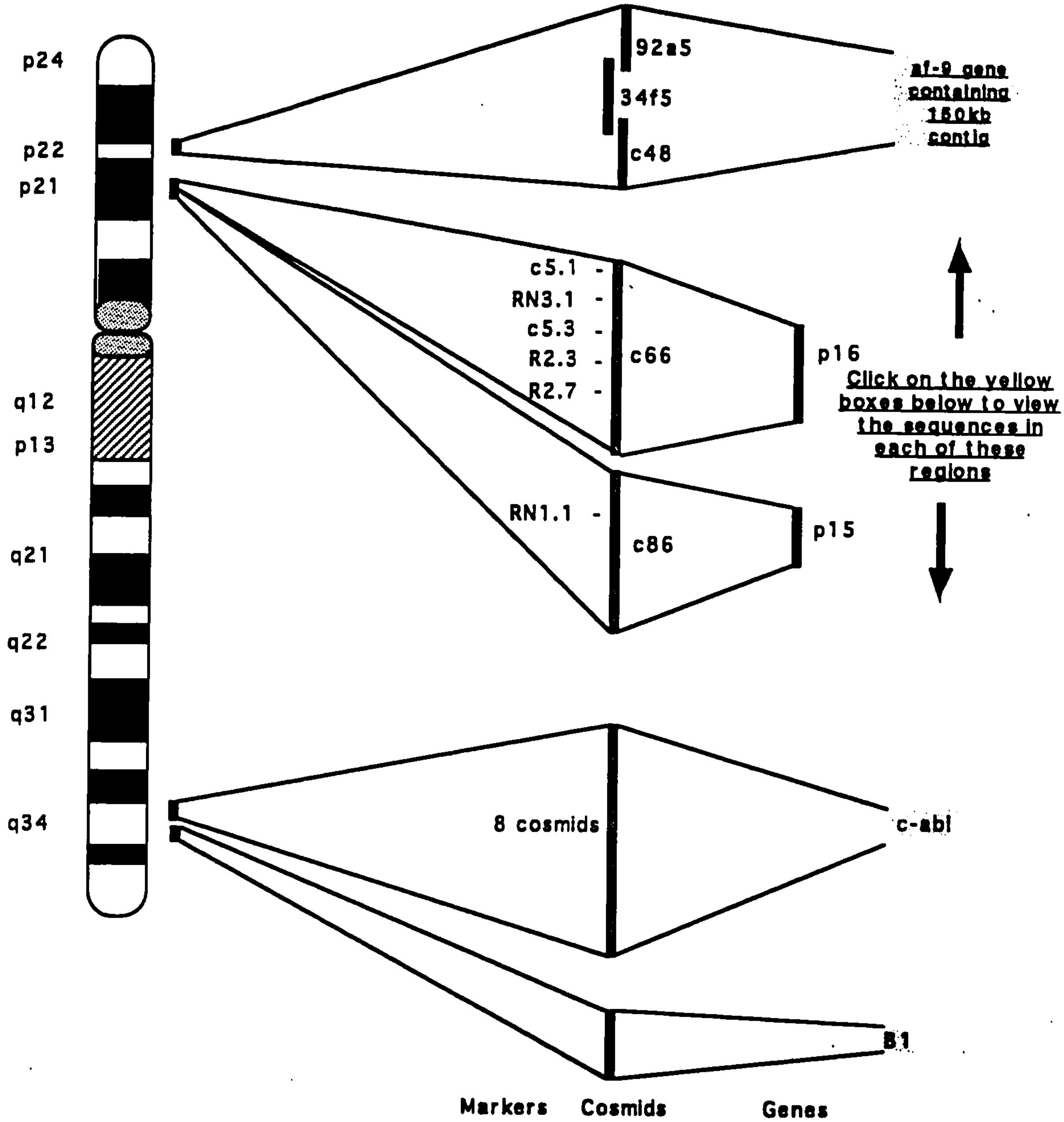
Archived	■
Submitted	■
Annotated	■
Finished	■
Closure in progress	▨
Shotgun complete	■
Shotgun in progress	■
DNA made	■
Bacterial Clone	□

Regions Sequenced at the University of Oklahoma from Clones that Map to the Lower Half of Human Chromosome 22



MDR = Meningioma Deletion Region
ES = Ewing's Sarcoma
NEFH = Neurofilament Heavy Subunit
NF2 = Neurofibroblastoma Region 2
RPolJ = RNA Polymerase II subunit J
MDR = Meningioma Deletion Region
ACR = Acrosin

Cosmids, and P1's Implicated in Leukemia, Melanoma, and Other Cancers from Human Chromosome 9



Click on the yellow boxes below to view the sequences in each of these regions

Notes:

C48 encodes the portion of the af-9 gene involved in leukemogenic t(9:11) translocations. At least six breakpoints have been mapped to C48.

C66 and C86 encode all.of p16 (CDK-INK4) and p15 (CDK-INK4b), respectively, which, when deleted, are involved in melanomas and other cancers.

Key:

Archived	■
Submitted	■
Annotated	■
Finished	■
Closure in progress	▨
Shotgun complete	■
Shotgun in progress	■
DNA made	■
Bacterial Clone	□

Bacterial Genomes and A. nidulans EST Sequencing Projects

- The initial shotgun sequencing phase of the *Neisseria gonorrohoeae* 2.2 Mbp and *Streptococcus pyogenes 1.9 Mbp* genomes is complete and in closure.

- 95% of each genome is now publicly available on our website. http://www.genome.ou.edu

- *Aspergillus nidulans* EST project is underway.

- Indicates Data Not Available

Level 0 = In Shotgun   Level 1 = Unordered Contigs   Level 2 = Ordered Contigs
Level 3 = Completely Finished (3-x coverage and fewer than 1 ambiguity/10,000 bases)

Notes Regarding Sequencing Progress:

Maps showing the location of the clones sequenced or in progress are available along with our protocols on our web site:
http://www.genome.ou.edu

All the clones with GenBank accession numbers AC000067 through AC000095 have no gaps and a sequence ambiguity of approximately 5/10,000 bases due mainly to the lack of "rule of three" coverage. These regions presently are being finished by a combination of long gel reads and sequencing off pcr-generated templates prior to declaring that they are at level 3.

It should be noted that to date we have generated:

Total sequence data submitted for 8-31-95 - 9-1-96:	**1,846,870**	**bp**	
Total submitted 9-2-96 - 11-15-96:	**1,997,137**	**bp**	
Total additional in progress:	**2,135,627**	**bp**	*
Total:	**5,979,634**	**bp**	*

*** = changed since November 15th submission**

Genome
(oral or poster)

Chen

PRODUCTION SEQUENCING OF MAMMALIAN DNA BY ORDERED SHOTGUN SEQUENCING (OSS) STRATEGY

[1]Peter Ma, [1]Chun-Nan Chen, [1]Ying Su, [1]Primo Baybayan, [1]Aleli Siruno, [1]Jeanette Evans, [2]Richard Mazzarella, [2]David Schlessinger and [1]Ellson Chen
[1]Advanced Center for Genetic Technology, Applied Biosystems Division of Perkin Elmer Corp., 850 Lincoln Center Drive, Foster City, CA 94404, and [2]Department of Molecular Microbiology, Washington University School of Medicine, St. Louis MO 63110.

Ordered shotgun sequencing (OSS) has been successfully carried out to sequence over 2.3 megabases DNA (>20 large-insert clones) from human X-chromosome isochores with different GC levels. The approach combines mapping and sequencing of YACs, BACs, or PACs with a hierarchical strategy that incorporates a feedback loop [Chen, E. et al., Genomics 17, 651-656 (1993); Chen,C et al., Nucleic Acids Res, 24, 4034-4041 (1996)]. Clones are recovered by STS-based screening of clones (see Williams et al., these ABSTRACTS). The method starts by randomly fragmenting a BAC, YAC or PAC to 8-12 kb pieces and subcloning those into lambda phage. Insert-ends of these clones are sequenced and overlapped to create a partial map. Complete sequencing is then done on a minimal tiling path of selected subclones.

OSS is currently delivering sequence at a cost comparable to methods that have been established far longer. Automation is facilitated by adapting PCR to prepare all sequencing templates, along with further improvements in sequencing technology and informatics. The approach also provides considerable flexibility in the choice of sequencing substrates. For example, subclones containing contaminating DNA can be recognized and ignored with minimal sequencing effort; regions overlapping a neighboring clone already sequenced need not be redone; and segments containing tandem repeats or long repetitive sequences can be spotted early on for targeted handling.

The encouraging results have led to an expanded goal of increasingly cost-effective genomic sequencing of 35 megabases, initiated on portions of Xq26 (1.5 Mb), Xq27 (1.5 Mb), Xp11.2 (1 Mb), Xq 12 - q21 (17.5 Mb), Xq21.3 (4.5 Mb); chromosome 3 (10 Mb, primariily in 3p21); and comparative sequencing of 8 Mb of mouse DNA, including the t-complex In1 and In2 regions (and corresponding human 6q24-q27), and segments homologous to Xp11.2 and Xq13 DNA already in process.

PROSPECTS

Feb, 97

(Production sequencing at PE-ABD/WU Genome Center)

Short-term (in 1997), 3Mb finished sequences (in addition to 2 Mb finished as of 12/31/96) on portions of:

- 1 Mb in Xp11.2, from DXS1008E to DXS423E.
- 1 Mb in Xq13.2, from DXS227 to DXS7025E
- 1 Mb in Xq26, from GPC3 to DXS8033.
- 1.5 Mb in Xq27, from F9 to DXS984.

Long-term (by 2000), >30 Mb sequences on:

- 13 Mb region in Xq11.2-q13.2 from DXS1 to DXS441 (about 2 Mb of which is being sequenced so far),
- 4.5 Mb Xq21.3 XY homology region, from DXS1217/DXYS1X to DXS3.
- 10 Mb of chromosome 3p21 and selected BACs from 3q23 and 3q29.
- 8 Mb of mouse DNA, including the xce locus and inversion regions In1 and In2 of the t-complex (as well as the corresponding human 6q24-q27).

Status of Human X Chromosome Sequencing at ACGT

	Locus	"MB" Index	Clone type	Insert size (Kb)	ABD Project #	Sequence Region (Marker Limits)	Status	Kb done	Remarks	STS Content: sWXDs
1	q28	158	9 cosmids	220	*	CV-G6PD	completed	220		
2	q24-25	118	yWXD703	135	A&B	ANT2	completed	135		
3	q13	67	bWXD161	220	C	DXS227 -7025E	1 gap	210		
4	q26	131	bWXD8	165	D	GPC3-DXS8033	completed	165		2271, 1455
5	q13	67	bWXD3	95	E	DXS227 -7025E	completed	95	submit	
6	q26	131-132	bWXD9	250	F	GPC3-DXS8033	1 gap	240		
7	q25	124-136	pWXD6	110	G	DXS100	1 gap	132		
8	q25	124-136	pWXD1	100	H	DXS7831	3 gaps	95	too many gaps	
9	q13.2	72-74	bWXD27	135	J	DXS227 to 7025E	1 gap	135		905,3679,3678,15, 981, 599
10	q13.2	72-74	bWXD40	103	K	DXS227 to 7025E	final checking	138		515
11	q13.2	72-74	bWXD42	110	L	DXS227 to 7025E	final checking	99		1254, 1253, 870
12	q13.2	72-74	bWXD14	112	M	DXS8066 - DXS1221	2 contigs	81		1255, 2891
13	q13.2	72-74	bWXD20	102	N*	DXS1679 ?	14 contigs	100		2870
14	q13.2	72-74	bWXD36	177	O	DXS8066 - DXS1221	2 contigs	151		1875
15	p11.2	54-56	bWXD142	140	P	OATL2-CEN	2 contigs	89		1995, 3675, 2559
16	p11.2	54-56	bWXD111	109	Q	OATL2-CEN	2 contigs	70		570, 3676, 2106, 3527, 3525, 1977, 2183, 2894, 2107, 1118
17	p11.2	54-56	bWXD137	158	R	OATL2-CEN	starting			2560, 1977
18	q27	139-140	bWXD90	77	S	DXS1192-DXS119	2 contigs	57		1445
19	q27	139-140	bWXD100	152	T	DXS1192-DXS119	2 contigs	64	Tough to PCR	2623
20	q27	139-140	bWXD105	124	U	DXS1192-DXS119	1 contig	65	Tough to PCR	2462
21	q26	131-132	bWXD168	121	V	GPC3-DXS8033	on hold			415, 27
22	q26	131-132	bWXD171	160	W	GPC3-DXS8033	on hold		33% coli?	791, 2457
23	q26	131-132	bWXD173	179	X	GPC3-DXS8033				791, 1319
24	q26	131-132	bWXD180	160	Y	GPC3-DXS8033				1307, 1334, 415
25	q26	131-132	bWXD181	160	Z	GPC3-DXS8033	Cancel?		overlap with D	1151, 1455, 2271, 2863
26	q26	131-132	bWXD200	240	AA	GPC3-DXS8033				1182, 1928, 2457
27	q26	131-132	bWXD177	92	AB*	GPC3-DXS8033	by shotgun			1151, 385
			total	3906				2341		

Total finished 2,341 Kb on 2/12/97

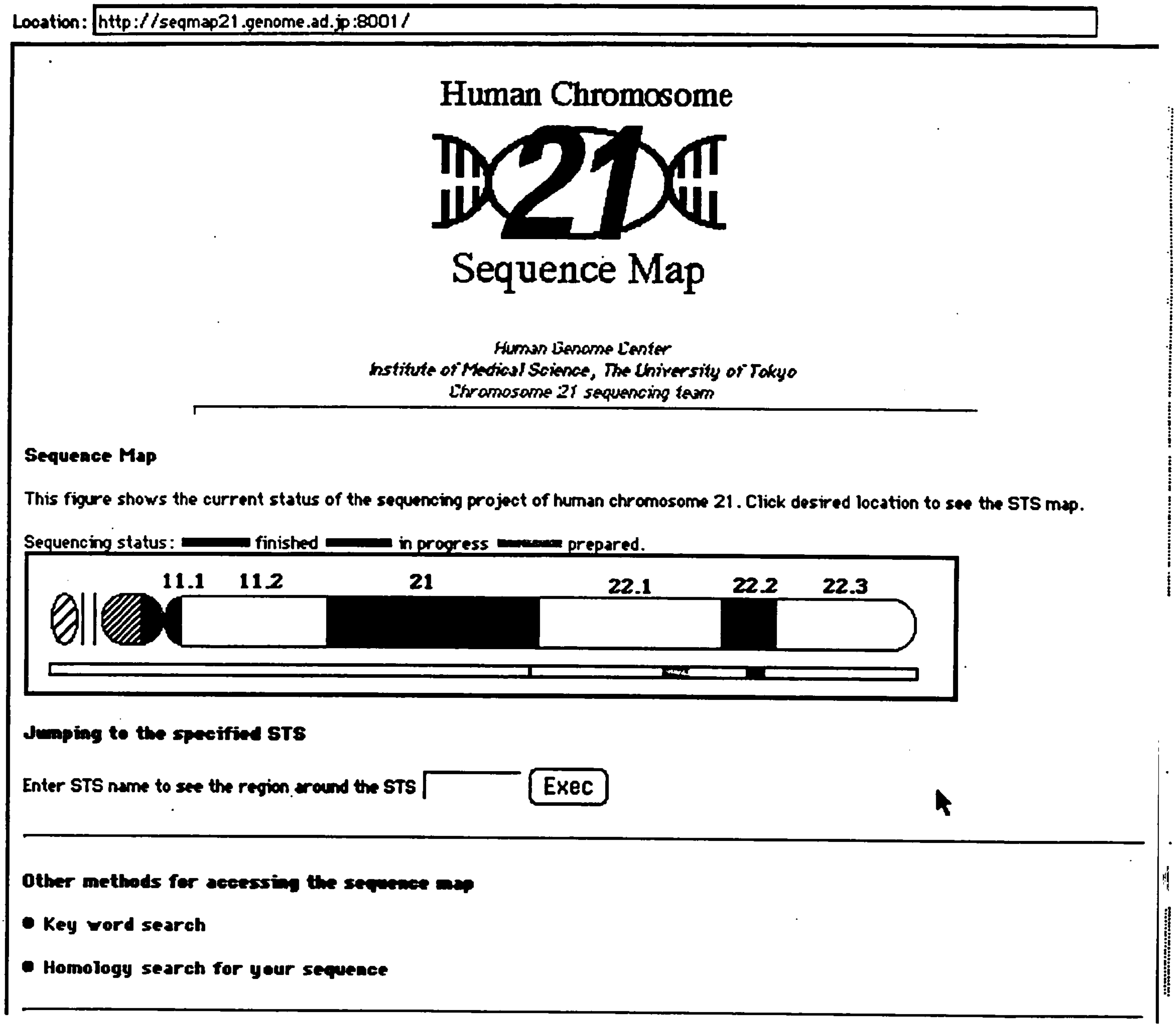
* by shotgun sequencing

(y;YAC: p;PAC: b; BAC)

? May be second site for STS

The locus is defined in cytogenetic bands; the clones to be sequenced are localized in an interval defined by "MB index on the complete X map (Nagaraja et al., 1977), and by bracketing STS markers; "status" indicates the degree of completion of the project, given either as growing contigs with one or a few remaining gaps, in final checking or completed.

All sequencing are done by OSS approach, except those labeled with * (which were done by random shotgun sequencing).



Location: http://seqmap21.genome.ad.jp:8001/

Human Chromosome 21 Sequence Map

Human Genome Center
Institute of Medical Science, The University of Tokyo
Chromosome 21 sequencing team

Sequence Map

This figure shows the current status of the sequencing project of human chromosome 21. Click desired location to see the STS map.

Sequencing status: ▬ finished ▬ in progress ▬ prepared.

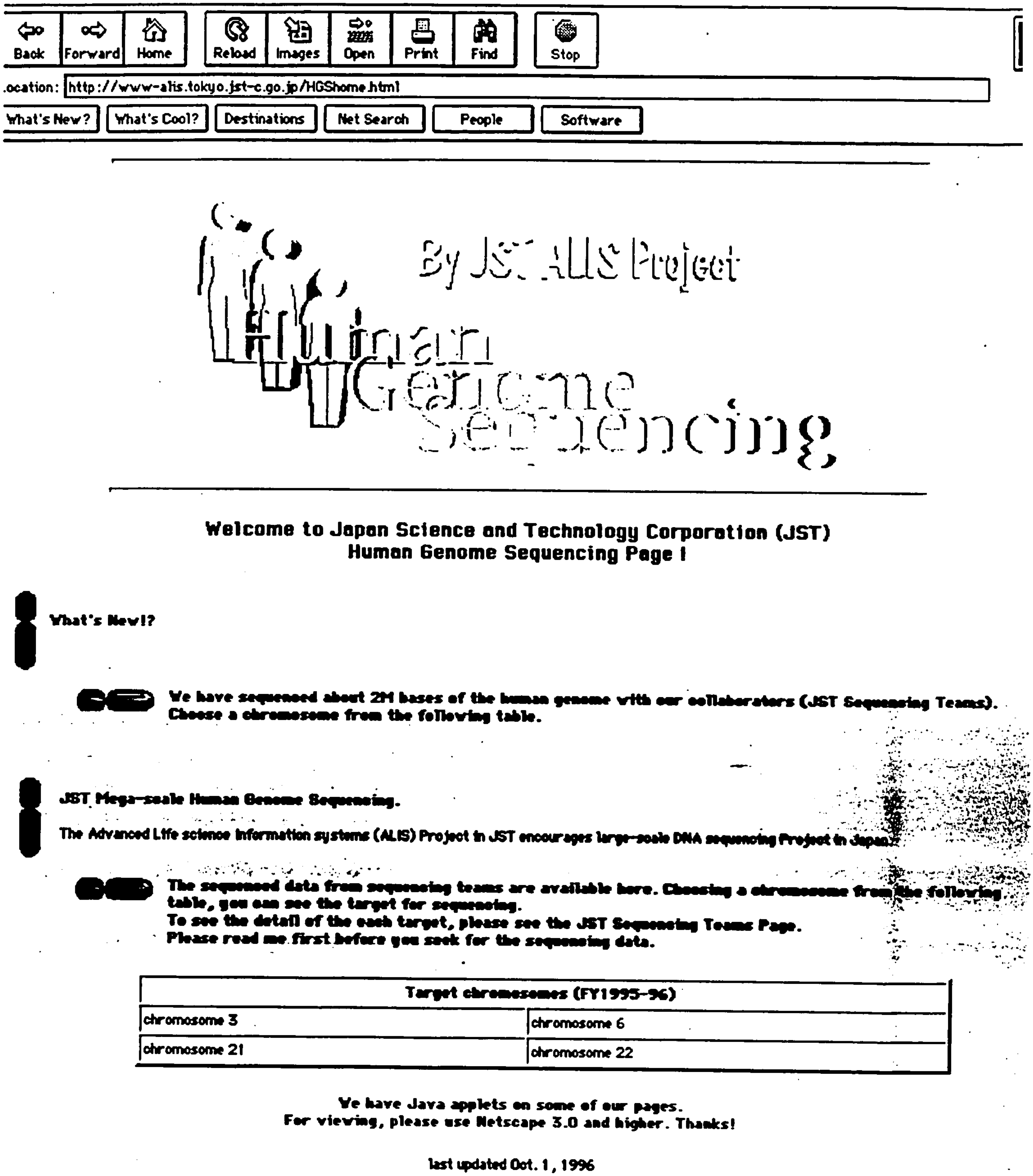
Jumping to the specified STS

Enter STS name to see the region around the STS [ ] Exec

Other methods for accessing the sequence map

- Key word search
- Homology search for your sequence

Japan Science and Technology Corporation



.ocation: http://www-alis.tokyo.jst-c.go.jp/HGShome.html

Welcome to Japan Science and Technology Corporation (JST)
Human Genome Sequencing Page !

What's New!?

We have sequenced about 2M bases of the human genome with our collaborators (JST Sequencing Teams).
Choose a chromosome from the following table.

JST Mega-scale Human Genome Sequencing.

The Advanced Life science Information systems (ALIS) Project in JST encourages large-scale DNA sequencing Project in Japan.

The sequenced data from sequencing teams are available here. Choosing a chromosome from the following table, you can see the target for sequencing.
To see the detail of the each target, please see the JST Sequencing Teams Page.
Please read me first before you seek for the sequencing data.

Target chromosomes (FY1995-96)	
chromosome 3	chromosome 6
chromosome 21	chromosome 22

We have Java applets on some of our pages.
For viewing, please use Netscape 3.0 and higher. Thanks!

last updated Oct. 1, 1996

● **Sequencing Schedule**

Target¥FY	1995	1996	1997	1998	Total
3p21.3	1,000kbp	-	-	-	1,000kbp
8p11.2	300kbp	1,000kbp	1,200kbp	-	2,500kbp
8p21.3-p22	-	-	200kbp	800kbp	1,000kbp
9q32	-	700kbp	300kbp	-	1,000kbp
Total	1,300kbp	1,700kbp	1,700kbp	800kbp	5,500kbp

This plan may be altered by annual budgeting.

● **Sequencing Schedule**

Target¥FY	1995	1996	1997	1998	Total
6p21.3	150kbp	400kbp	450kbp	200kbp	1,200kbp
Total	150kbp	400kbp	450kbp	200kbp	1,200kbp

This plan may be altered by annual budgeting.

● **Sequencing Schedule**

Target¥FY	1995	1996	1997	1998	Total
21q22.2	400kbp	800kbp	-	-	1,200kbp
21q22.1	-	1,000kbp	1,000kbp	-	2,000kbp
21q22.3	-	500kbp	2,000kbp	2,000kbp	4,500kbp
Total	400kbp	2,300kbp	3,000kbp	2,000kbp	7,700kbp

This plan may be altered by annual budgeting.

● **Sequencing Schedule**

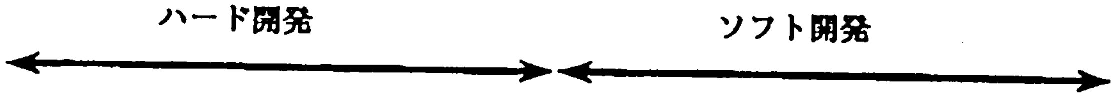
Target¥FY	1995	1996	1997	1998	Total
21q22.2	100kbp	400kbp	-	-	500kbp
21q22.3	-	100kbp	500kbp	300kbp	900kbp
22q11.2	500kbp	800kbp	-	-	1,300kbp
22q11	-	-	800kbp	400kbp	1,200kbp
Total	600kbp	1,300kbp	1,300kbp	700kbp	3,900kbp

This plan may be altered by annual budgeting.

年次計画と所要経費

年　度	98 (H10)	99 (H11)	2000 (H12)	01 (H13)	02 (H14)	03 (H15)	04 (H16)	05 (H17)
データ生産能力	15Mb	30Mb	60Mb	60Mb	60Mb	60Mb	60Mb	60Mb
解析対象	h21	h21／m21	m21／h11	h11	h11／m11	m11	h／m	h／m
人員* リソース	6(4)人	12(8)	15(12)	15(12)	15(12)	15(12)	15(12)	15(12)
シークエンス	12(10)人	24(20)	40(36)	40(36)	40(36)	40(36)	40(36)	40(36)
データ処理	4(3)人	5(4)	10(8)	10(8)	10(8)	10(8)	10(8)	10(8)
技術開発	1人	3(2)	6(4)	6(4)	6(4)	6(4)	6(4)	6(4)
事務部門	2(1)人	4(2)	8(4)	8(4)	8(4)	8(4)	8(4)	8(4)
計	26(18)人	48(36)	79(64)	79(64)	79(64)	79(64)	79(64)	79(64)
運営経費	20億円	30億円	60億円	60億円	60億円	60億円	60億円	60億円

*（　）は人材派遣で可な人数



Princess Hotels 3-1-97

Dear –

I

In

Sep

clea

on

dat

pul

rele

I s

now

shotgu

agree

domai

Al

Project is different, however, because of your position, others may not interpret this as we have –

Regards –

Bruce

The Princess
BERMUDA

RHW

m lebco

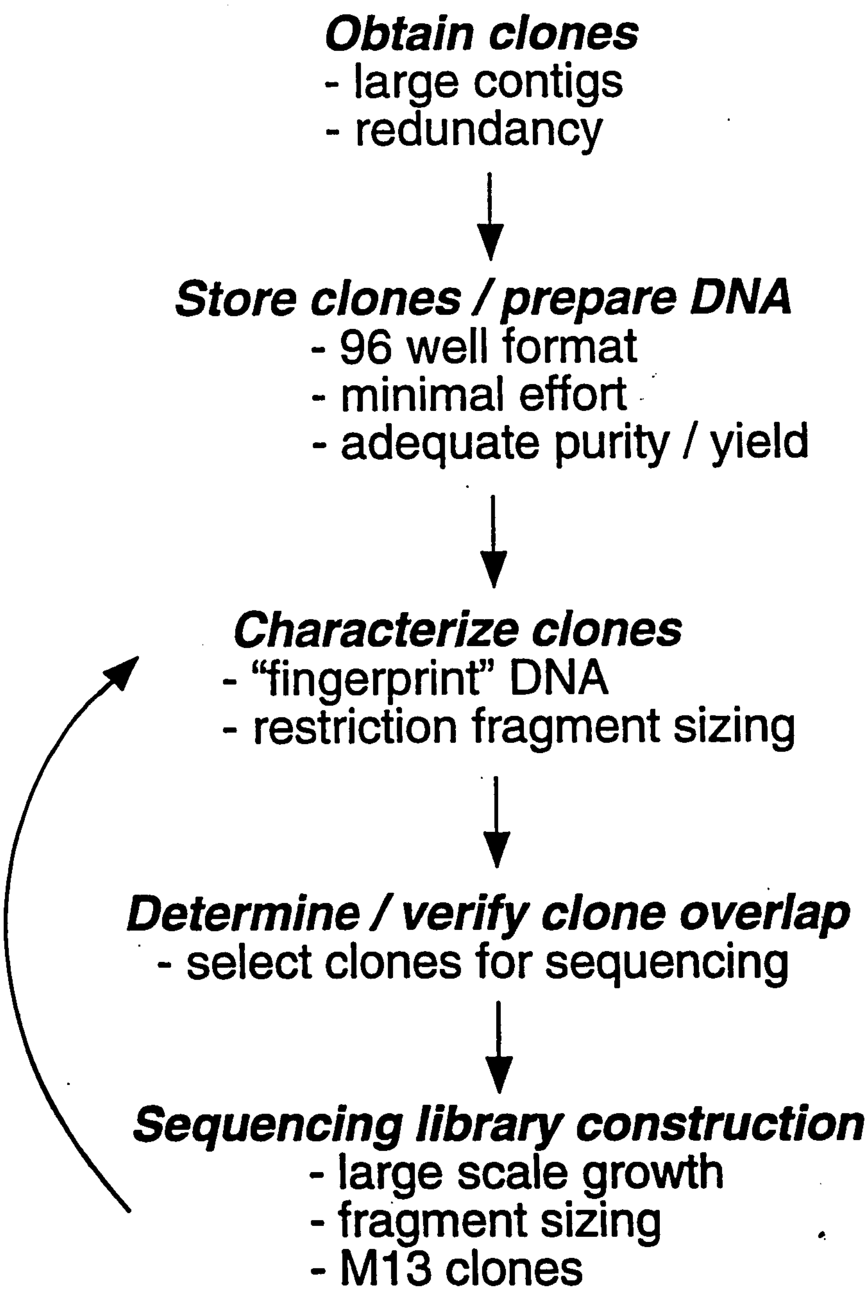
m Sho

m Fw

Fmsh

Subm

7,



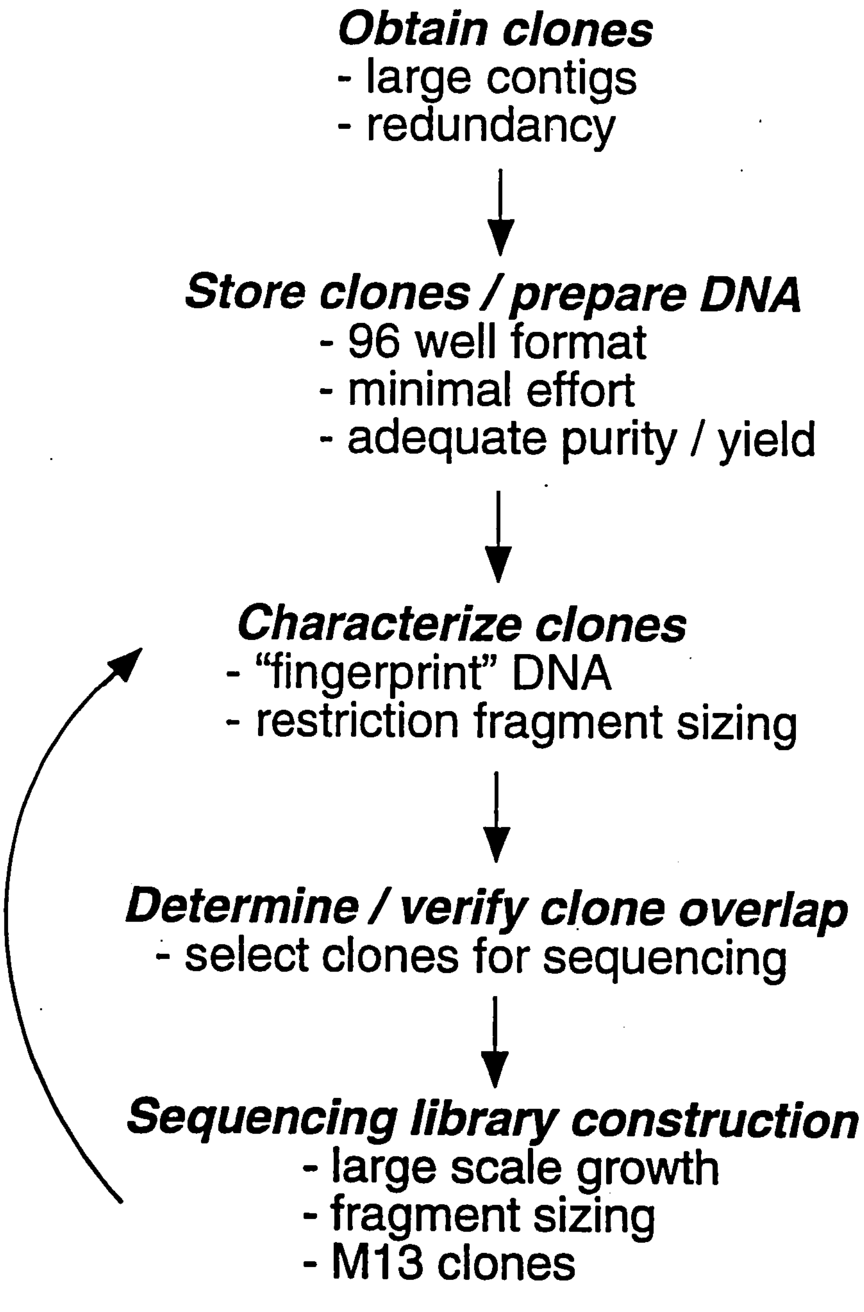
Obtain clones
- large contigs
- redundancy

Store clones / prepare DNA
- 96 well format
- minimal effort
- adequate purity / yield

Characterize clones
- “fingerprint” DNA
- restriction fragment sizing

Determine / verify clone overlap
- select clones for sequencing

Sequencing library construction
- large scale growth
- fragment sizing
- M13 clones



Obtain clones
- large contigs
- redundancy

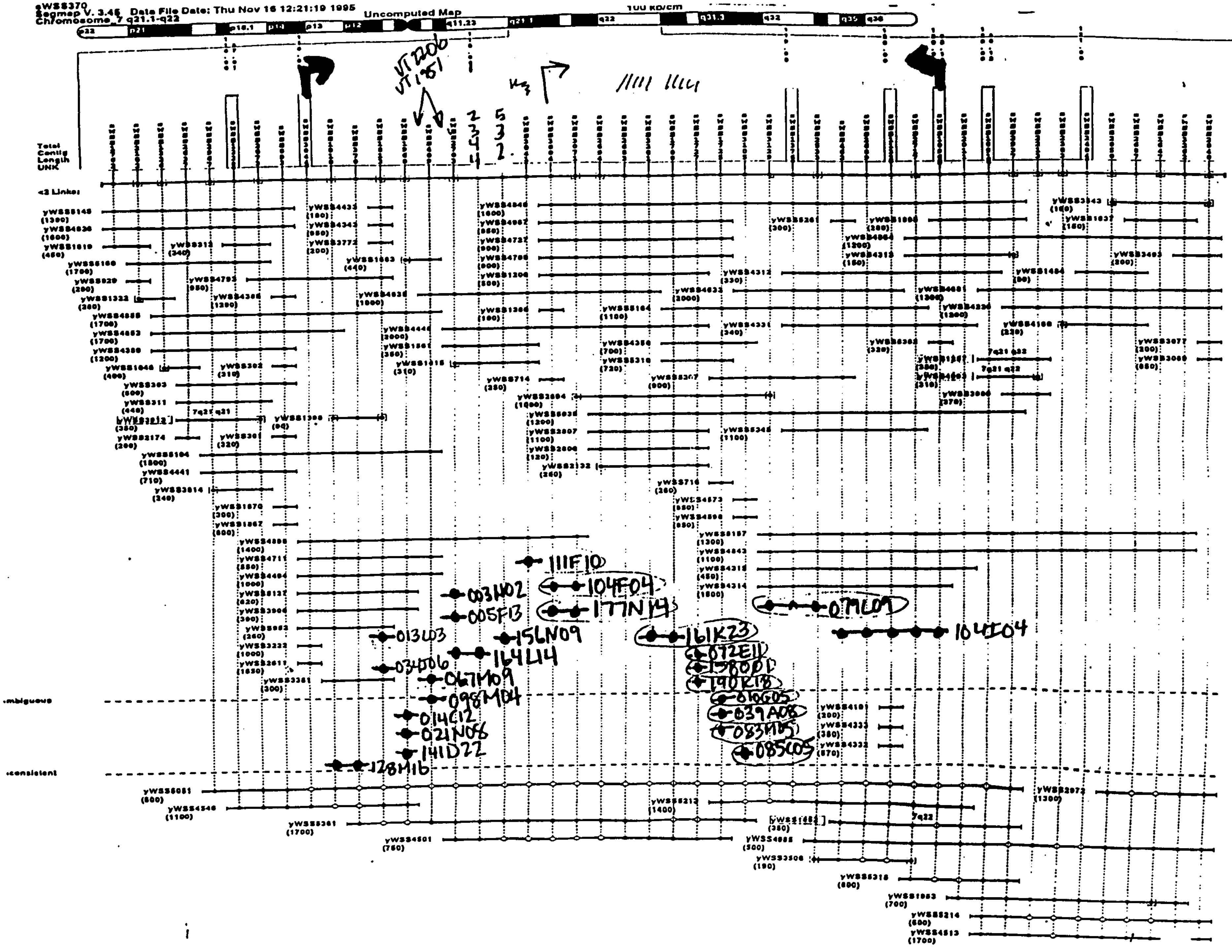
Store clones / prepare DNA
- 96 well format
- minimal effort
- adequate purity / yield

Characterize clones
- "fingerprint" DNA
- restriction fragment sizing

Determine / verify clone overlap
- select clones for sequencing

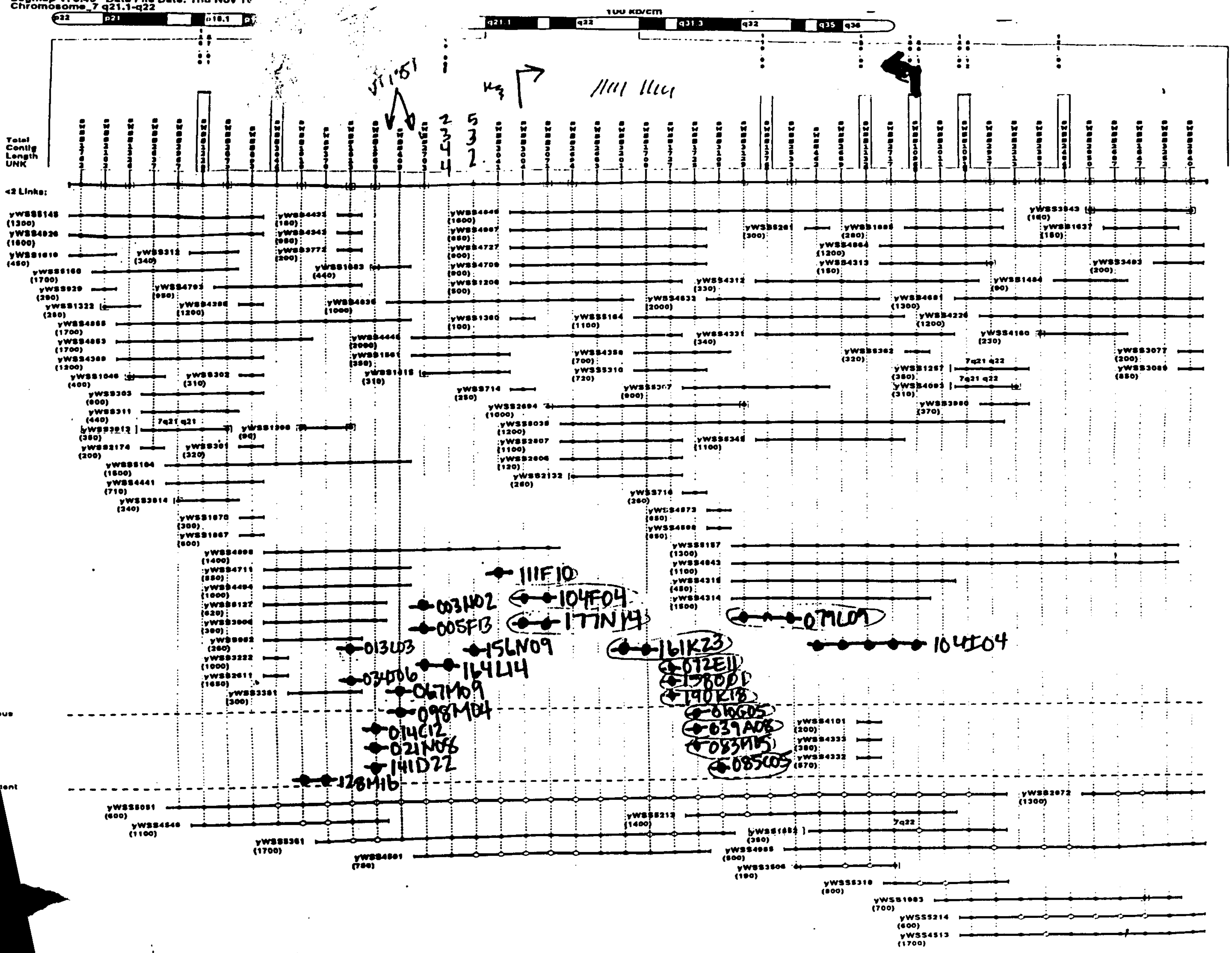
Sequencing library construction
- large scale growth
- fragment sizing
- M13 clones

sWSS370
Segmap V. 3.45 Date File Date: Thu Nov 16 12:21:19 1995
Chromosome_7 q21.1-q22

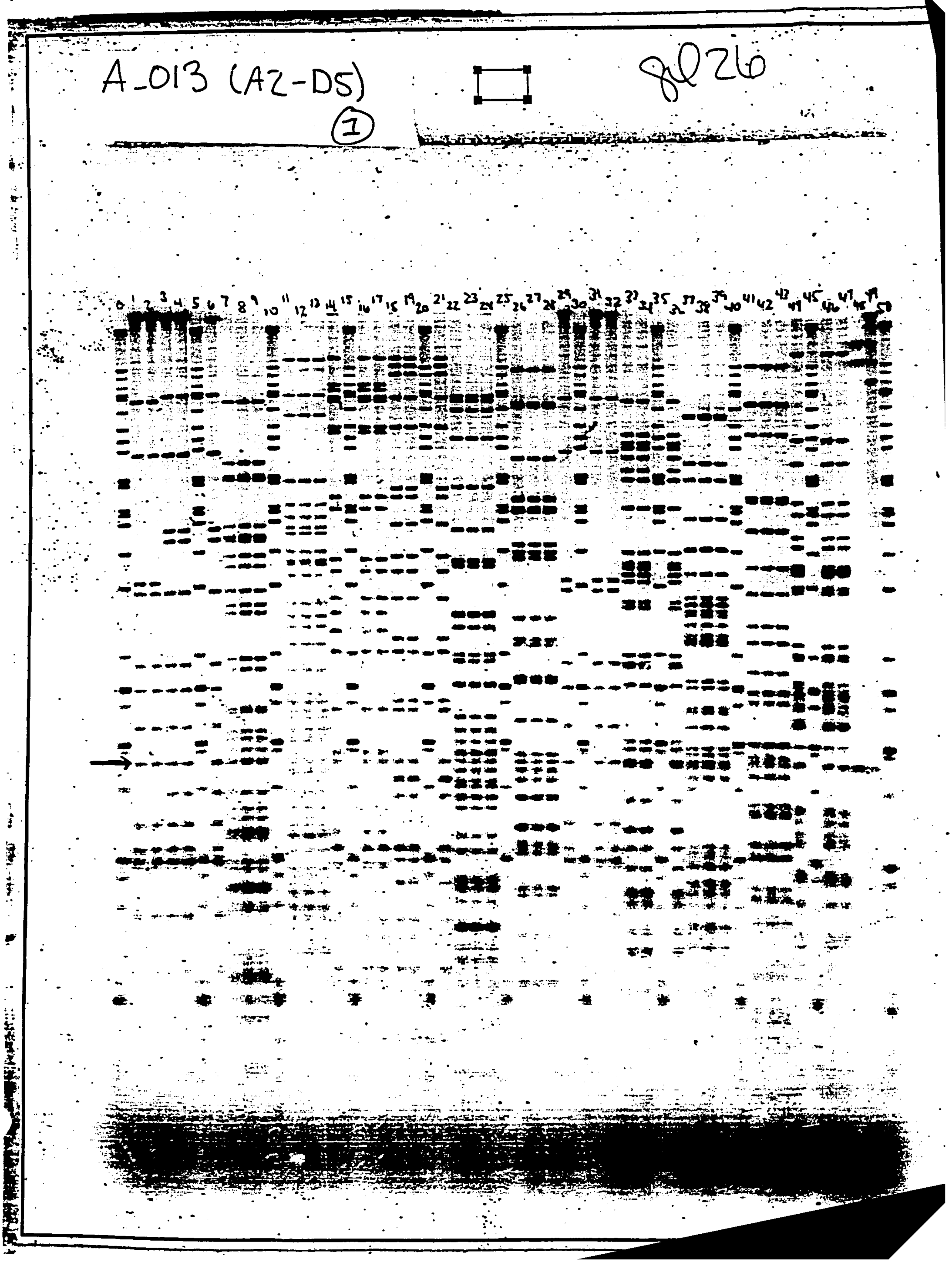


CCM1 region

sWSS370
Segmap V. 3.45 Data File Date: Thu Nov 1[illegible]
Chromosome_7 q21.1-q22



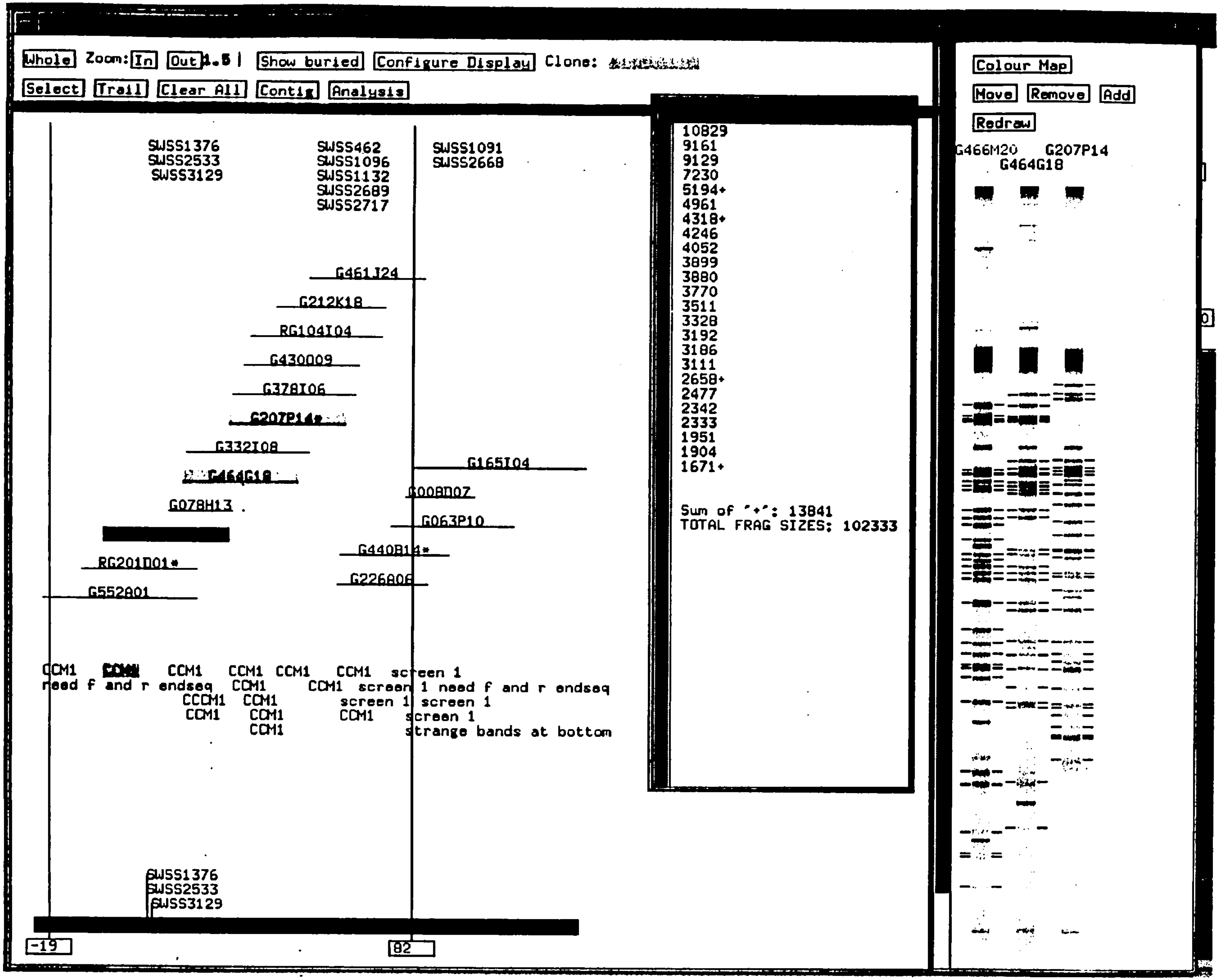
~~e:\m~~m61107h.ds , Range = 100.3-28503 rfu, 105%



A_013 (A2-D5)

①

gel26



[redacted]

```
10829
9161
9129
7230
5194+
4961
4318+
4246
4052
3899
3880
3770
3511
3328
3192
3186
3111
2658+
2477
2342
2333
1951
1904
1671+

Sum of '+': 13841
TOTAL FRAG SIZES: 102333
```

```
CCM1    CCM1    CCM1    CCM1  CCM1    CCM1   screen 1
need f and r endseq   CCM1      CCM1  screen 1 need f and r endseq
             CCCM1  CCM1         screen 1 screen 1
              CCM1    CCM1       CCM1    screen 1
                      CCM1               strange bands at bottom
```

RHW

Map status Summary

600 STS's spanning 50 Mb

128 BAC/PAC contigs

~250 kb average size

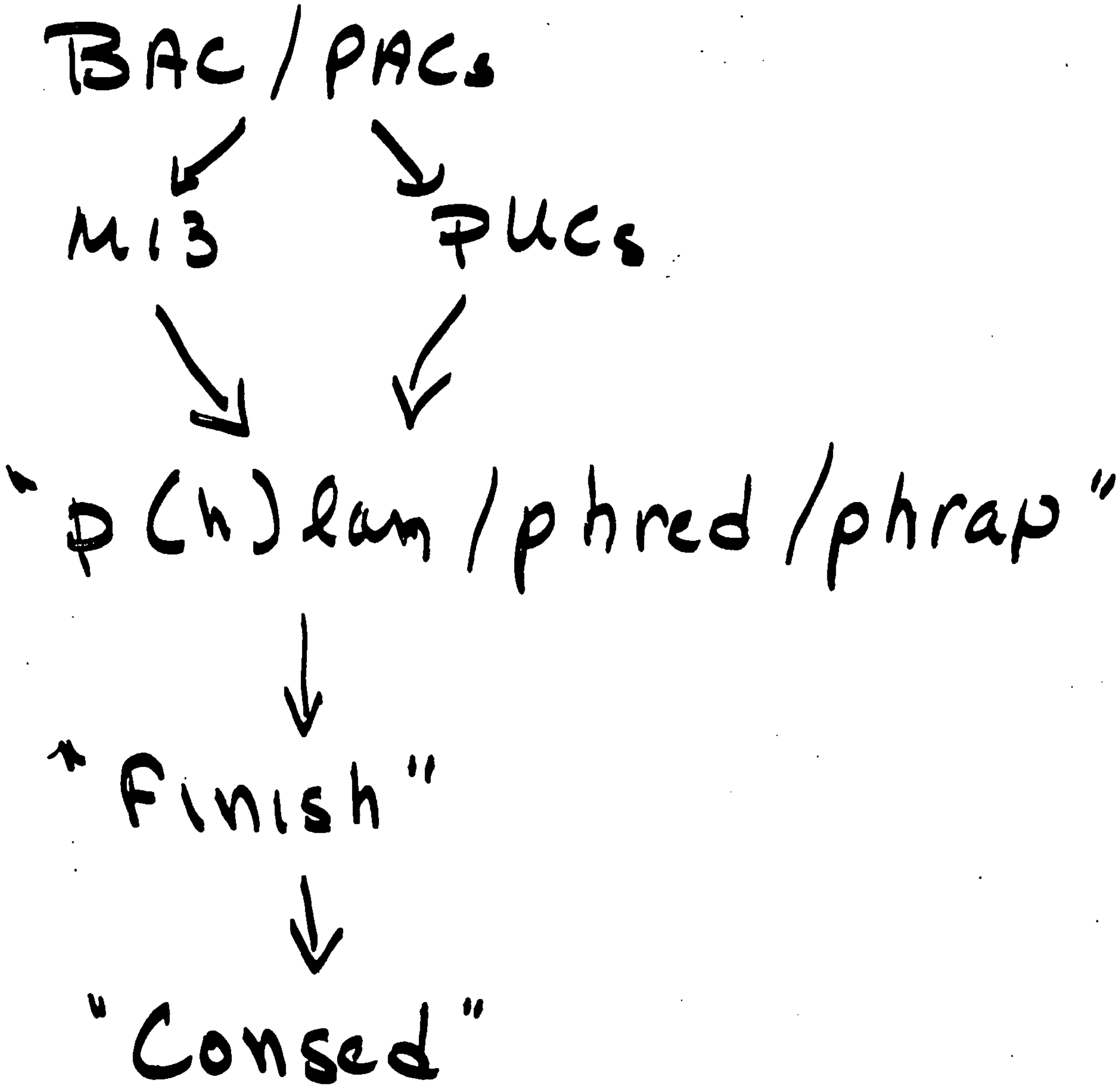
32 Mb TOTAL

175 clones underway

21 finished

Producing Sequence

Shotgun / directed



BAC / PACs

M13

PUCs

"p(h)lam / phred / phrap"

"Finish"

"Consed"

RHW

Quality control

- restriction digests x 3
- reassembly with alternate versions of phred/phrap
- complete continuity
- annotation
- 1/10,000 error rate

RHW

Software for human decision making.

data tracking -
central database
bar coding

getlanes / plan / phred / phrap

finish - rearraying

RHW

Technologies

Present -

64-72 lanes on 377

gel loaders

Amersham dye terminators

Transposons

Future

96 lanes on 373, 377

pipetting station

U.W. sequencer

R [illegible]

1

technology

1997 20 ABI 377 < 4 (IMB)
8 (BMBF-BEO)
8 (BMBF-DLR)

(96' 1.300 reads/day
97' 3.000 reads/day)

6 production groups (1 Postdoc
3 technicans)

1 bioinformatics group (5 people)

1 library group (1 Postdoc
3 technicans)

production (6 groups: 1 Postdoc, 3 TA's)

- picking, preping, sequencing, loading, data transfer, assembly, finishing, annotation

1997	1998	1999
6 x 1 Mb = 6 Mb	6 x 2 Mb = 12 Mb	6 x 3 Mb = 18

funding

- Land Thuringia (renting lab space
lab furniture)
- federal government BMBF-BEO 13 Mill
1995-2000
- federal government BMBF-DLR 14 Mill
(May 97 - April 2.000) (30 cents/

1 Resources

cosmids, PAC's, BAC's

targets	(1997-2000)		maps
X	Xq 28	3 Mb	- Nelson/Gibbs - Poustka Kioschis (Heidelberg
	Xp 11.23 Xp 11.4	2.5 Mb	- Meindl (Munich)
	PAP1	1 Mb	- Rappold (Heidelberg)
21q		28 Mb	- Yaspo (Berlin) - internat. chr. 21 consort
7	7q 22	7 Mb	- Scherer Tsui (Toronto)
	7q 32	0.5 Mb	
mouse	syntenic to Xq 28	3 Mb	

Genome Sequencing Centre at IMB, Jena (Germa

1996 2.5 Mb completed → 1.5 Mb Genbank

→ 1 Mb annotati
phase

1997 - 2000 (April)

Σ 40 Mb

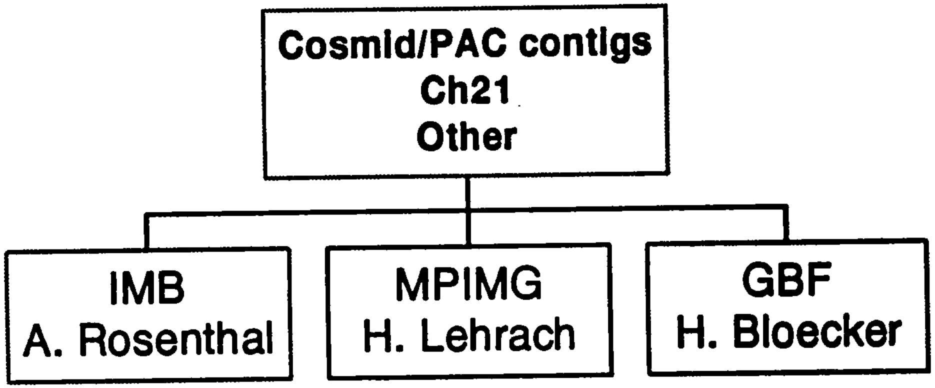
→ 28 Mb (BMBF-DLR
State Thuringia)

→ 12 Mb (BMBF-BEO
State Thuringia)

1997	6 Mb	(4 + 2)
1998	12 Mb	(9 + 3)
1999-	19 Mb	(15 + 4)
2000 (Jan-April)	3 Mb	

German Human Genome Project

Genomic sequence analysis of human chromosome 21 and selected regions of the human genome



Cosmid/PAC contigs
Ch21
Other

IMB
A. Rosenthal

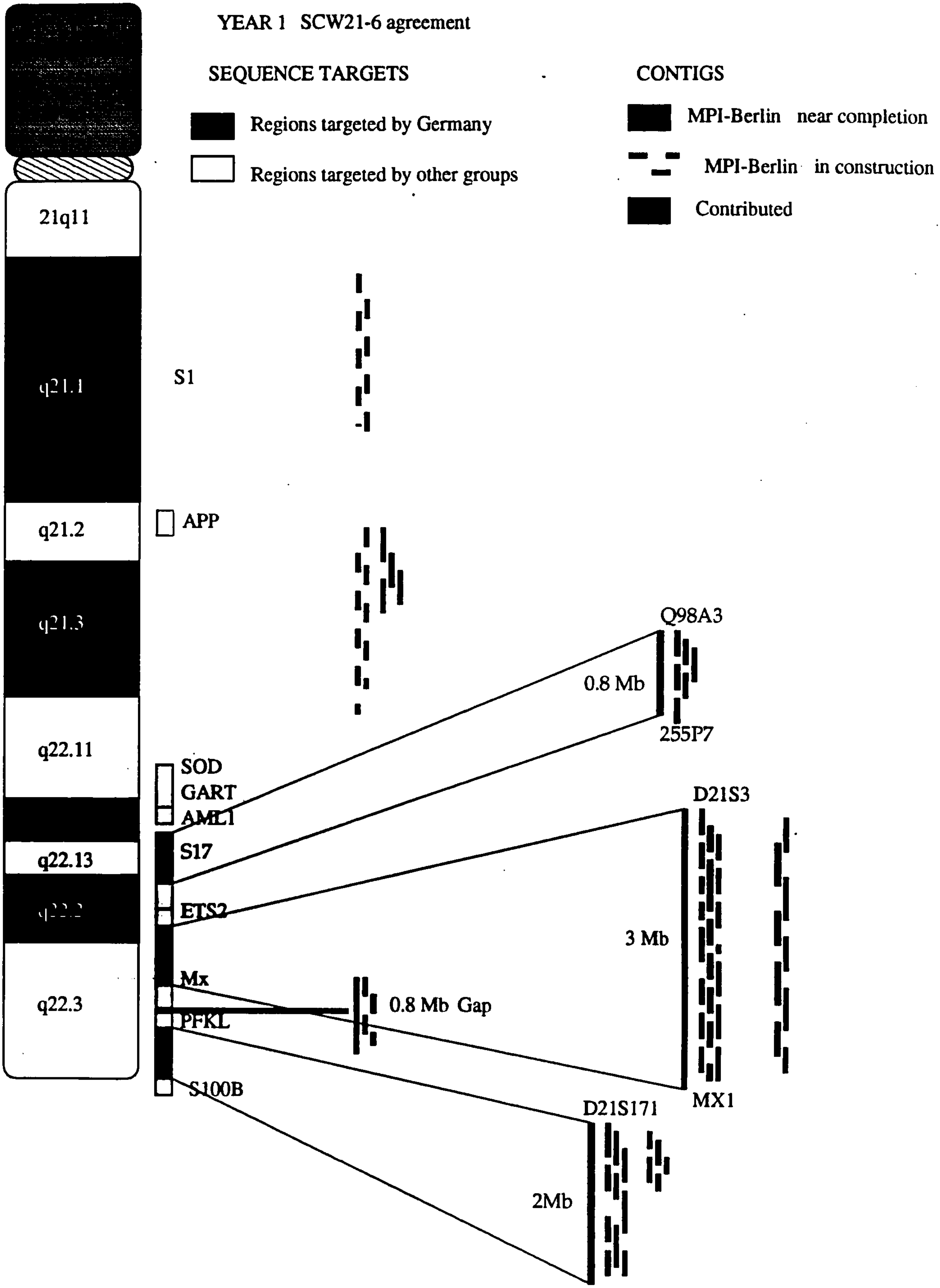
MPIMG
H. Lehrach

GBF
H. Bloecker

Coordinator

year 1	4Mb	1Mb	1Mb
year 2	9Mb	2Mb	2Mb
year 3	15Mb	3Mb	3Mb

YEAR 1 SCW21-6 agreement



UTSW Genome Science and Technology Center

Ongoing Projects:

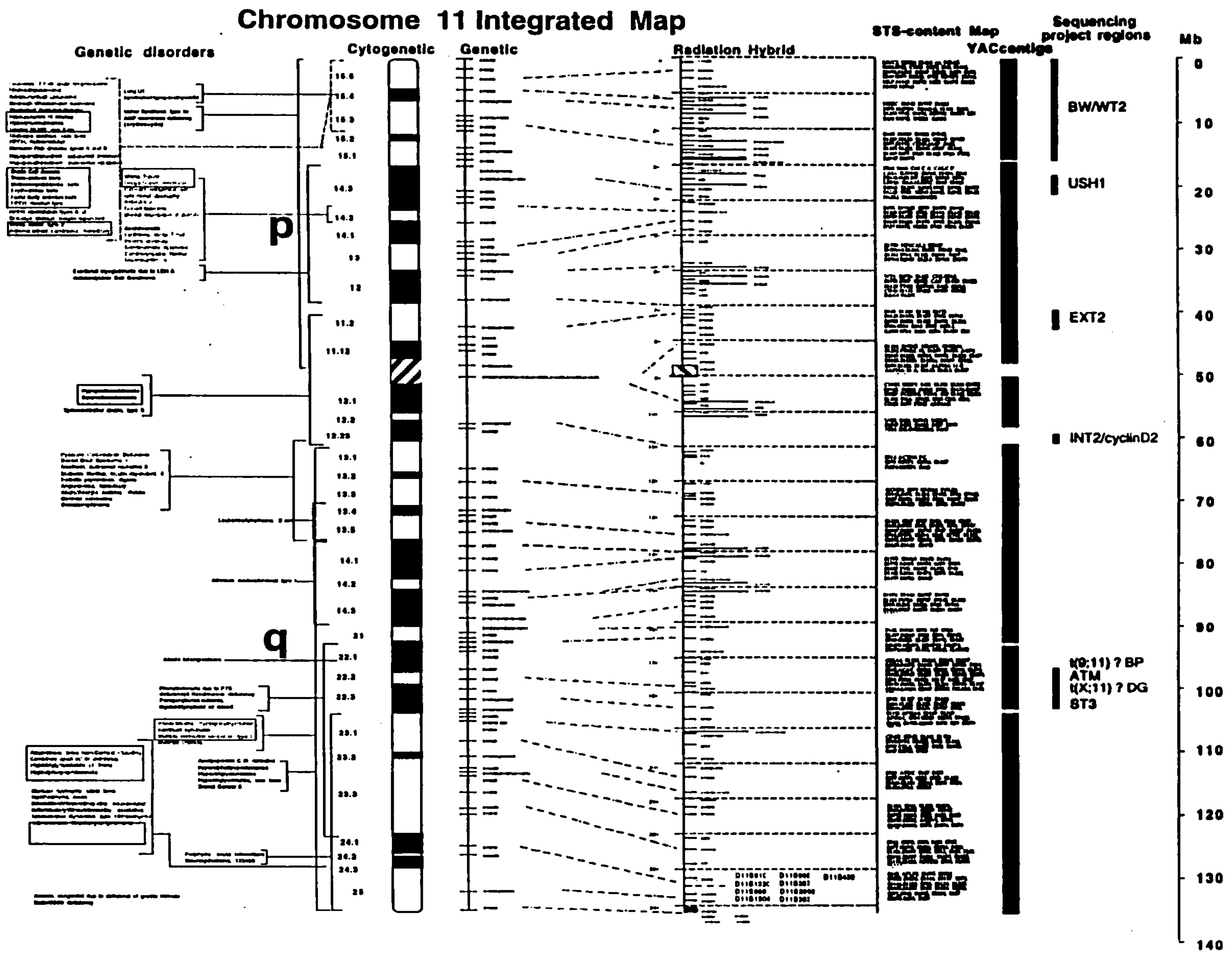
- **NCHGR Genome Science and Technology Center - Sequencing portions of chromosome 11, 15**

- **Department of Energy - PAC/BAC end-sequence data resource for sequencing the human genome (consortium with RPCI, Cedars-Sinai Medical Center)**

- **Collaborations with Hewlett Packard/Convex, Beckman Instruments/Sagian, Texas Instruments, Nanogen.**

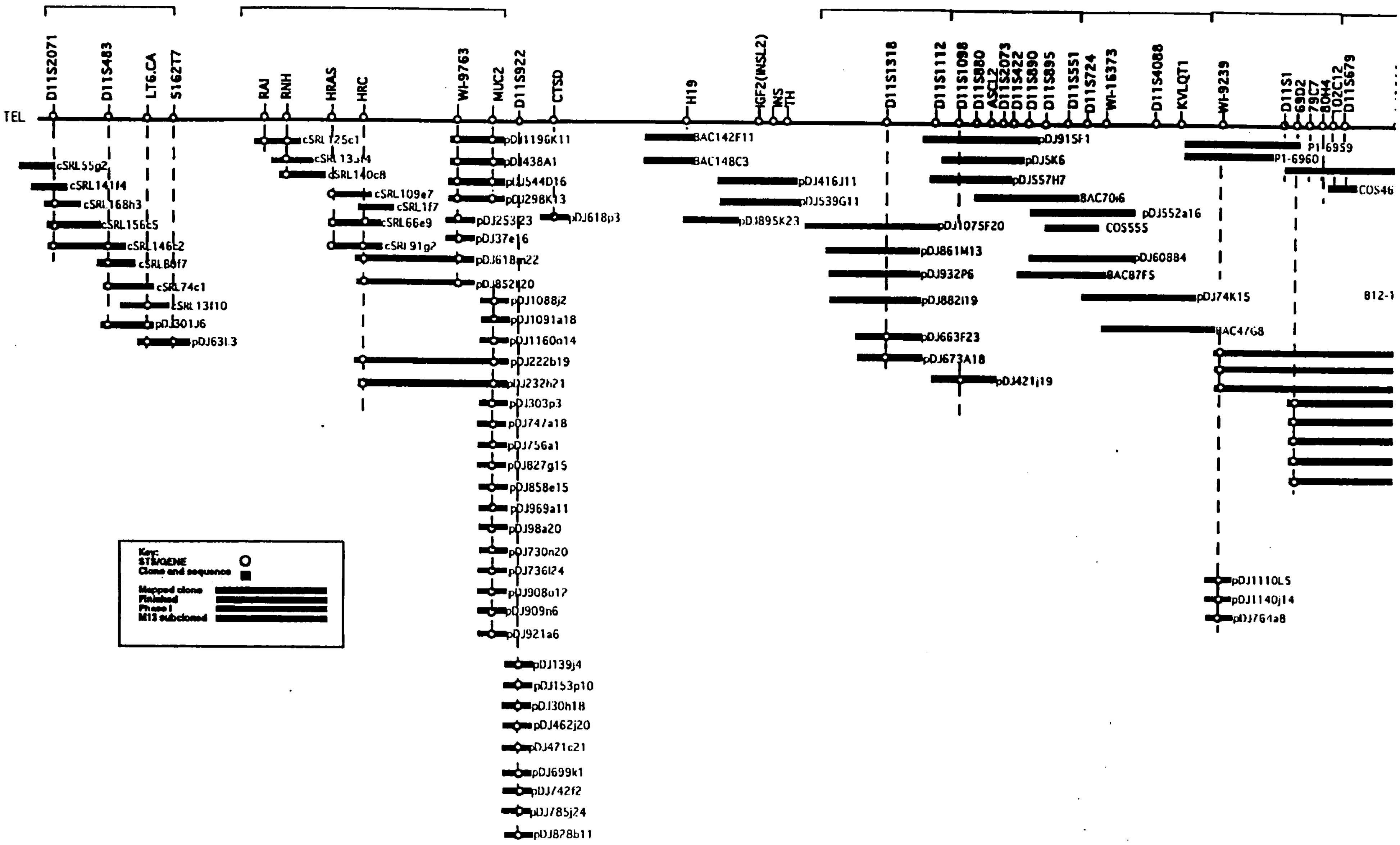
Map Construction

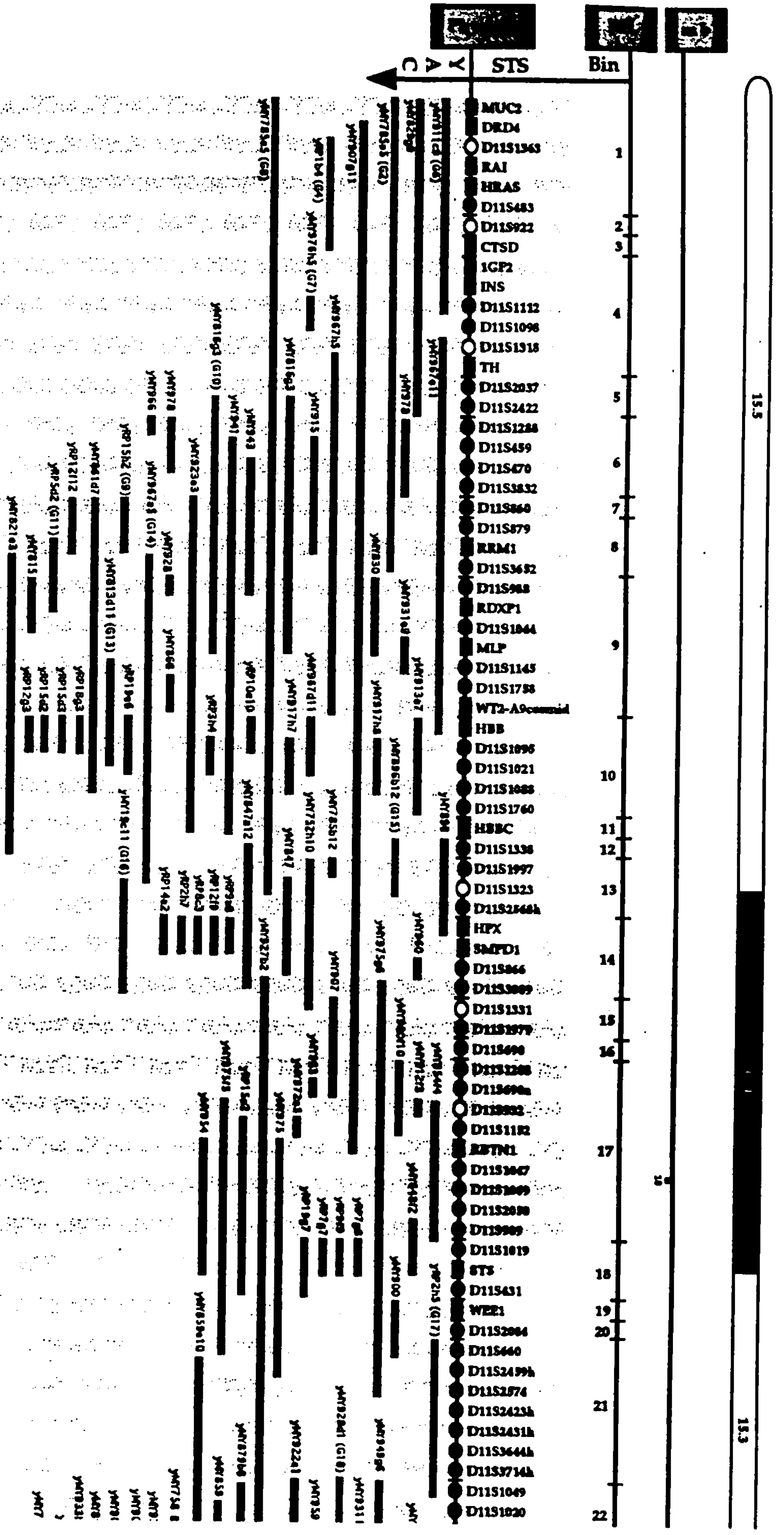
- Based on YAC/STS content map (905 STSs) supplemented with 17,965 "binned" cosmid end-sequences (chr 11), FACS sorted M13 sequences (chr 15)
- Conversion to PAC/BAC map
- PAC/BAC isolation by high density grid hybridization with pooled STS-specific oligonucleotides (20X)
- Confirmation by PCR with STSs (5X)
- Four restriction enzyme fingerprints of each PAC
- PAC/BAC end-sequencing of all clones to detect overlaps, generate additional "gap-filling" STSs and assemble map
- All PACs FISH confirmed to eliminate chimeras (<2%)
- Map becomes the display feature of sequence presentation on WWW

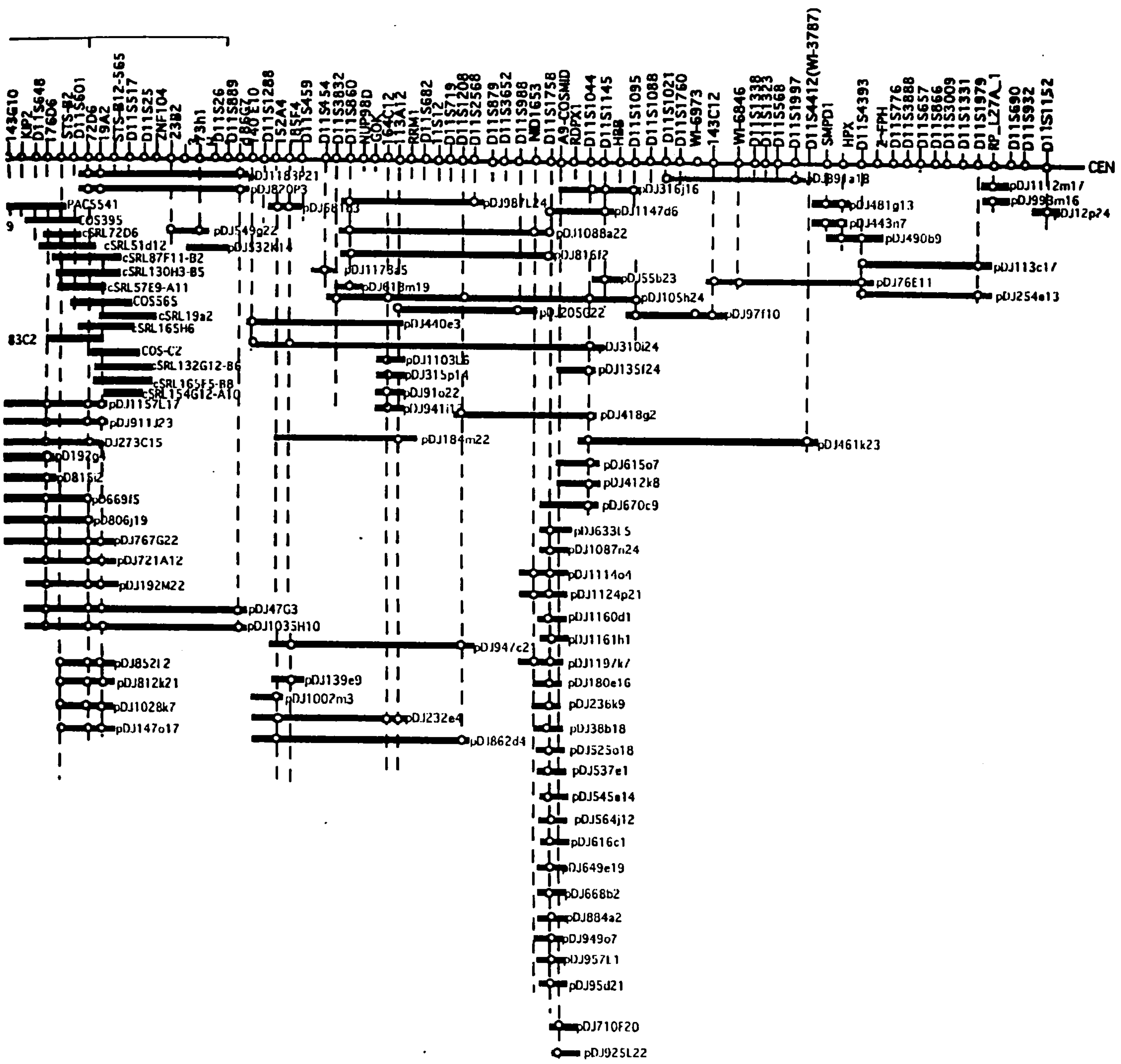


Chromosome 11 Integrated Map

SEQUENCING MAP - CHROMOSOME 11p15.5 - RH BINS 1- 16




C
A
Y
STS
Bin
MUC2
DRD4
D11S1363
RAI
HRAS
D11S483
D11S922
CTSD
IGF2
INS
D11S1112
D11S1098
D11S1318
TH
D11S2037
D11S2422
D11S1288
D11S459
D11S470
D11S3832
D11S860
D11S879
RRM1
D11S3652
D11S988
RDXP1
D11S1044
MLP
D11S1145
D11S1758
WT2-A9cosmid
HBB
D11S1096
D11S1021
D11S1088
D11S1760
HBBC
D11S1338
D11S1997
D11S1323
D11S2566h
HPX
SMPD1
D11S866
D11S3009
D11S1331
D11S1979
D11S690
D11S1208
D11S690a
D11S902
D11S1182
RBTN1
D11S1047
D11S1009
D11S2080
D11S909
D11S1019
STS
D11S431
WEE1
D11S2084
D11S660
D11S2499h
D11S2574
D11S2423h
D11S2431h
D11S3644h
D11S3714h
D11S1049
D11S1020
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
15.5
15.3
yWZ811a8 (G5)
yWZ785a5 (G2)
yWZ783a5 (G6)
yWZ807g11
yWZ892h2 (G17)


143G10
KIP2
D11S648
176D6
STS-B2
D11S601
72D6
19A2
STS-B12-S65
D11S517
D11S25
ZNF104
23B2
3
73h1
D11S26
D11S889
186G7
140E10
D11S1288
152A4
85F4
D11S459
D11S454
D11S3832
D11S860
NUP98D
GOK
164C12
113A12
RRM1
D11S682
11S12
D11S719
D11S1208
D11S2568
D11S879
D11S3652
D11S988
NID1653
D11S1758
A9-COSMID
RDPX1
D11S1044
D11S1145
HBB
D11S1095
D11S1088
D11S1021
D11S1760
WI-6973
143C12
WI-6846
D11S1338
D11S1323
D11S568
D11S1997
D11S4412(WI-3787)
SMPD1
HPX
D11S4393
2-FPH
D11S776
D11S3888
D11S657
D11S866
D11S3009
D11S1331
D11S1979
RP_L27A_1
D11S690
D11S932
D11S1152
CEN
pDJ1183P21
pDJ870P3
PACS541
9
COS395
cSRL72D6
cSRL51d12
cSRL87F11-B2
cSRL130H3-B5
cSRL57E9-A11
COS565
cSRL19a2
cSRL165H6
83C2
COS-C2
cSRL132G12-B6
cSRL165F5-B8
cSRL154G12-A10
pDJ1157L17
pDJ911J23
pDJ273C15
pDJ192g4
pDJ815i2
pDJ669I5
pDJ806j19
pDJ767G22
pDJ721A12
pDJ192M22
pDJ47G3
pDJ1035H10
pDJ852I2
pDJ812k21
pDJ1028k7
pDJ147o17
pDJ549g22
pDJ532k14
pDJ68103
pDJ1173a5
pDJ818m19
pDJ440e3
pDJ1103L6
pDJ315p14
pDJ91o22
pDJ941i13
pDJ184m22
pDJ947c21
pDJ139e9
pDJ1007m3
pDJ232e4
pDJ862d4
pDJ987L24
pDJ1088a22
pDJ816f2
pDJ205G22
pDJ1147d6
pDJ316j16
pDJ55b23
pDJ105h24
pDJ97f10
pDJ310i24
pDJ135I24
pDJ418g2
pDJ461k23
pDJ615o7
pDJ412k8
pDJ670c9
pDJ633L5
pDJ1087n24
pDJ1114o4
pDJ1124p21
pDJ1160d1
pDJ1161h1
pDJ1197k7
pDJ180e16
pDJ236k9
pDJ38b18
pDJ525o18
pDJ537e1
pDJ545e14
pDJ564j12
pDJ616c1
pDJ649e19
pDJ668b2
pDJ884a2
pDJ949o7
pDJ957L1
pDJ95d21
pDJ710F20
pDJ925L22
pDJ891a18
pDJ481g13
pDJ443n7
pDJ490b9
pDJ113c17
pDJ76E11
pDJ254a13
pDJ1112m17
pDJ993m16
pDJ12p24

UTSW GESTEC
Map Production

STSs screened (RH bins 1-85)	465
PACs isolated by hyb	3,185
"Hit" rate (av/range)	12.45 (2.5-24.4)
PACs confirmed by PCR	467
Clones fingerprinted	467

UTSW GESTEC
Resource Lab FISH analysis

PACs analyzed by FISH	216
unique signal	213
chimeric signal	3
% putative chimeras	1.3%
band assignment	192
band analysis	142

Sequencing Strategy

- ○ M13/plasmid shotgun library of entire PAC < 6X coverage.
- ○ Automated reaction assembly using Sagian/Beckman robot, currently 3,000/day with capacity of 24,000/day.
- ○ Initial 75%primer/25% terminator chemistry and automated assembly using Phred/Phrap.
- ○ Automated synthesis of oligonucleotide primers from initial assembly using Primo software and MerMade 192-channel synthesizers (300/day) for closing and accuracy improvement.
- ○ Finishing using alternate strand reads, long reads, oligo gap closing "auto-finishing" and primer production using Primo, etc.
- ○ Accuracy assessment and additional reads to generate average Phrap score of >40 over entire sequence.

Sequence levels and estimated accuracy

- Phase I Assembled contigs > 1 kb, unordered

- [illegible]

- Phase III Closed contig, no gaps, no resequencing for accuracy improvement, estimated accuracy 10^{-3} to 10^{-4} Genbank acceptable

- Phase IV QualPlot analyzed, accuracy improved by resequencing to 10^{-4} based on average Phred/Phrap score > 40

DNA Sequence Production

Level	Type	No.	bp
Data collection	raw data	18	2,160,000
Phase I/II	contigs	41	2,902,496
Phase III	closed	33	1,137,005
Phase IV	$<10^{-4}$ accuracy	4	482,752
Genbank	closed + 10^{-4}	34	1,619,757
Largest contig			341,110

Clone End-Sequencing Project

End-sequence files generated:

Chromosome 11 cosmids	17,965
Giardia lamblia cosmids	2,590
Chromosome 11 PACs	546
Whole Human Genome PACs	1,523

End-sequence database of 5,636,750 bp

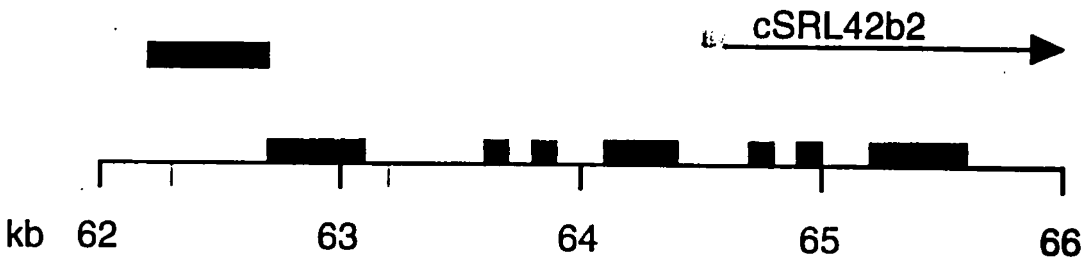
Annotation Protocol

- Final assembly and annotation carried out on HP/Convex Exemplar superparallel computer (8 hrs --> 2 hrs --> 20 minutes)

- Sequence annotated for:
  - Genbank matches
  - EST matches
  - STS matches (map confirmation)
  - End-sequence matches (determination of clone overlap)
  - Grail-predicted exons
  - Repetitive sequence
  - Simple sequence repeats
  - Restriction sites (comparison with fingerprint to confirm assembly)
  - Other features

- QualPlot output - accuracy estimation

Automated sequence annotation



■ Human repetitive element

Simple sequence

■ EST Genbank

■ Non-EST Genbank

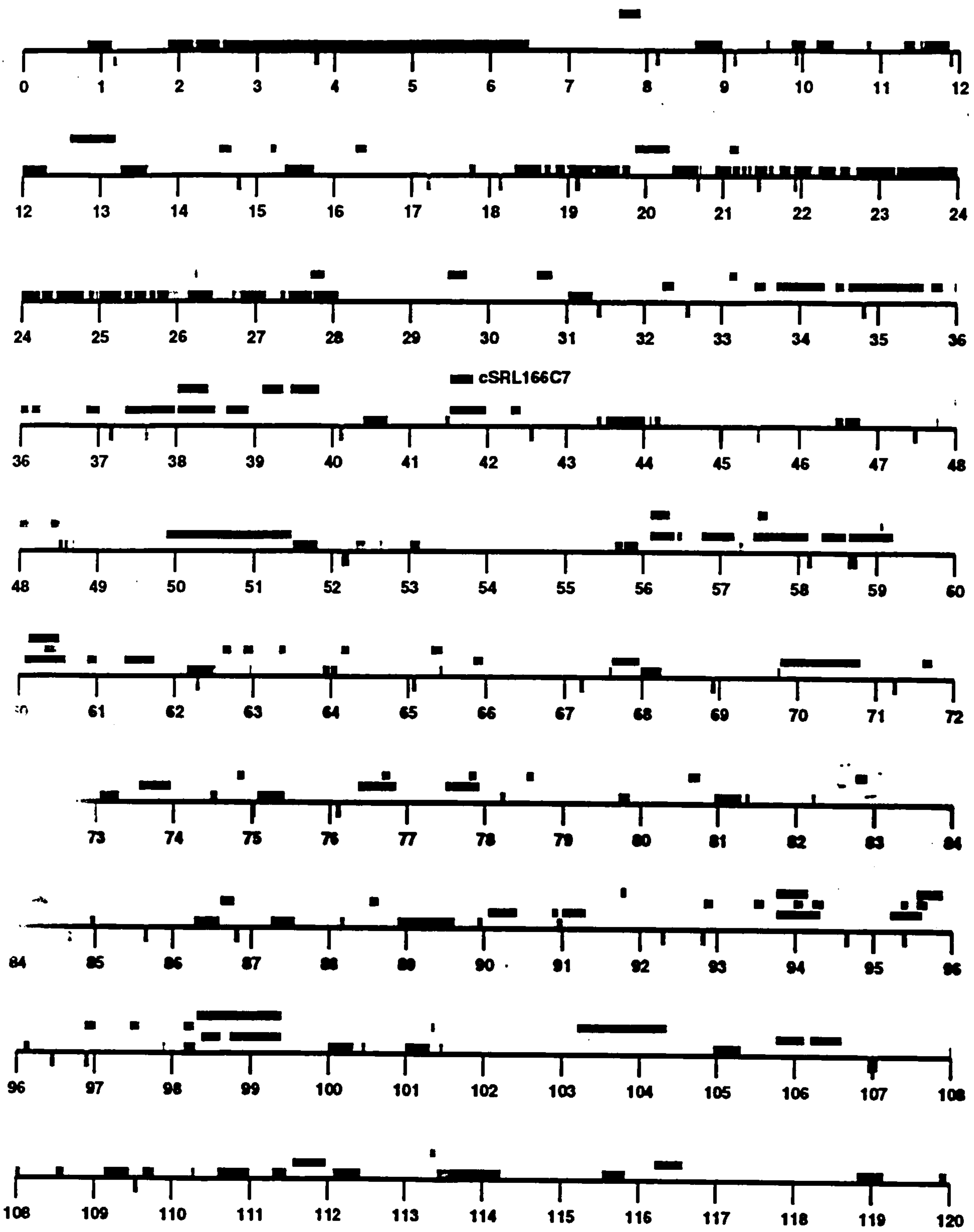
Grail-predicted exon

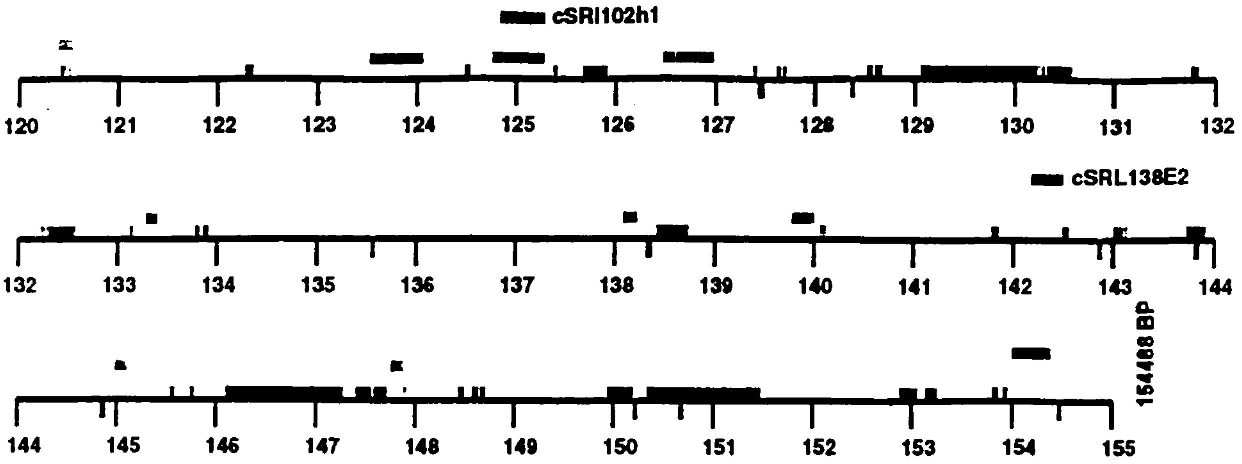
End sequence (overlapping clone)

| EcoR1, BamH1 sites (QC)

Sequence Features from a 155KB Contig of 11p14.3

Tue Dec 17 08:21:01 CST 1996





LEGEND:

Human Repetitive Element

Simple Sequence

Non-EST GenBank

Predicted Exon

EST GenBank

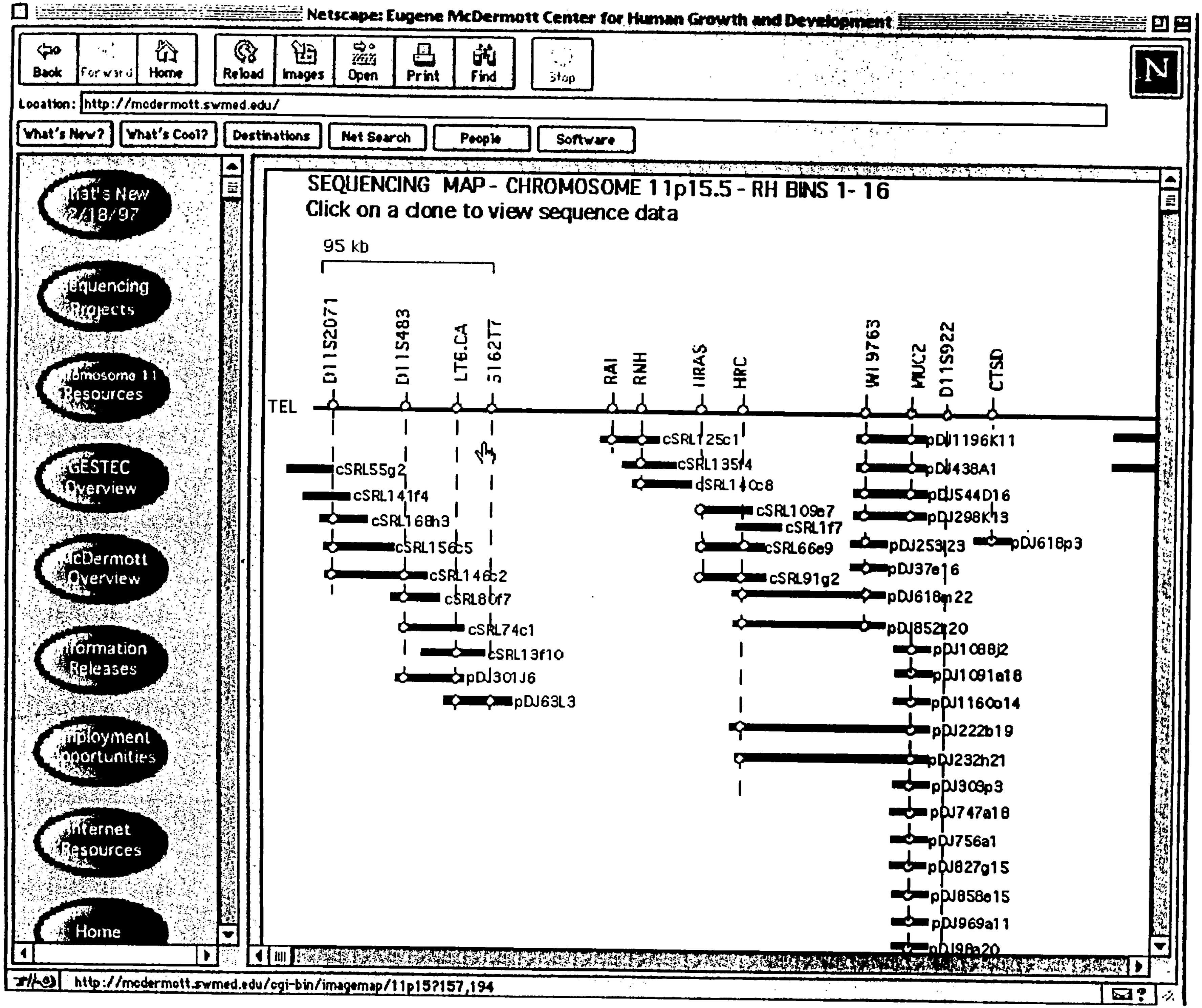
End Sequence

BamHI: 13.37.67,516,521,1045.1062,2352.2446,2493,2578,3757,4012,4329,4384
4384.4679.5381,5480,5486,5736,5929,6643,8501,9631.9955.10866.13083,15581.17925
17925

EcoRi: 75.470.774.919.979.1092.1188.1233.1987.2252.2259.2493.2542.2777.2798
2798.3804.3851.5321.5363.6097.7183.7299.7361.7948.8564.8993.10062.10654.16818
16818.21332

Data Distribution

- Maps and sequence available at http//:mcdermott.swmed.edu/ - updated weekly
- Phase I (contigs) and Phase II (closed) made available; unassembled raw data is not made available
- Phase III and Phase IV submitted to Genbank when completed
- WWW display includes:
  - Map of sequenced region including clones in progress
  - Graphic features display of each clone
  - Complete features tables
  - QualPlot (accuracy estimate) output



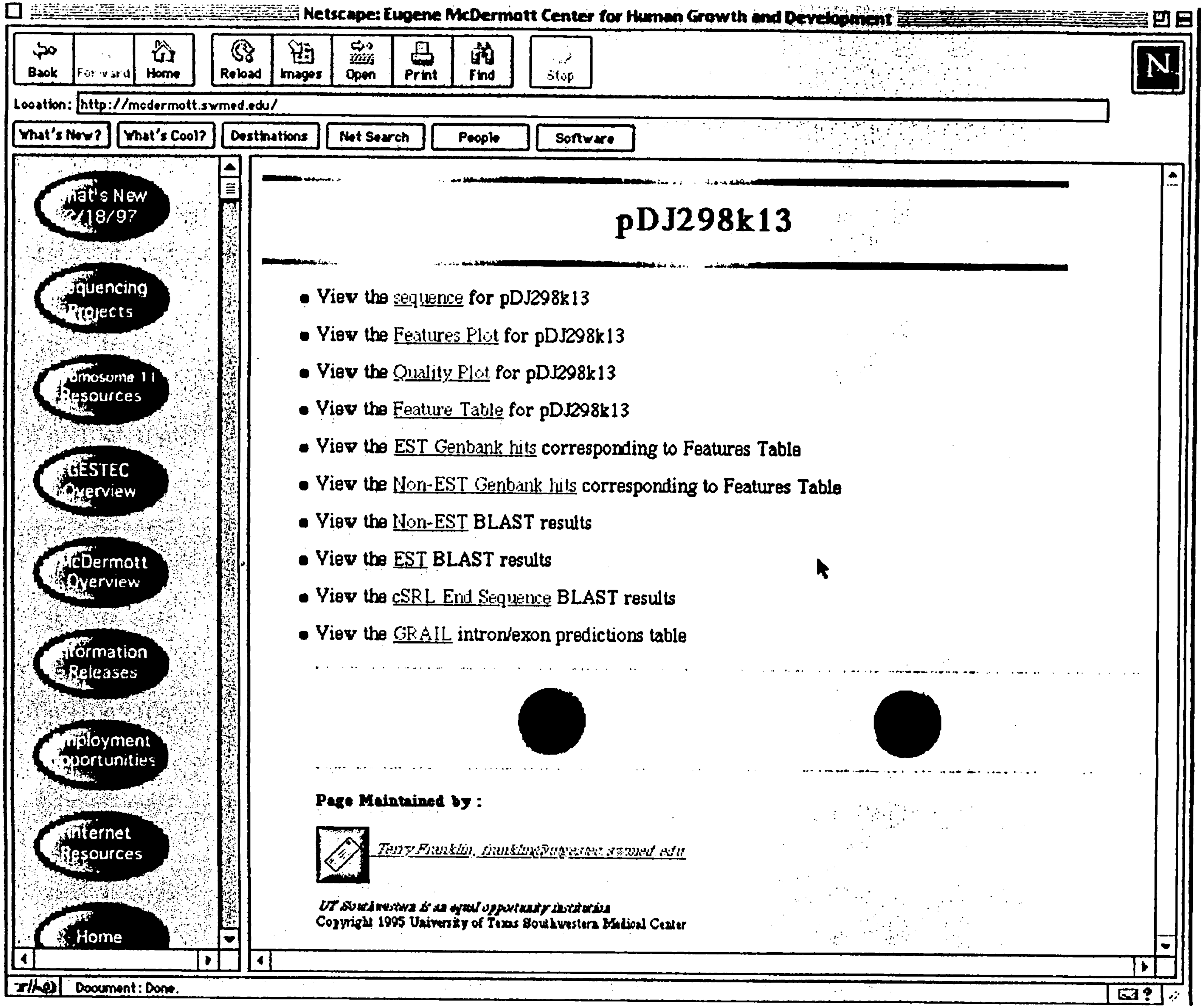
Netscape: Eugene McDermott Center for Human Growth and Development

Location: http://mcdermott.swmed.edu/

SEQUENCING MAP - CHROMOSOME 11p15.5 - RH BINS 1-16

Click on a clone to view sequence data

http://mcdermott.swmed.edu/cgi-bin/imagemap/11p15?157,194



pDJ298k13

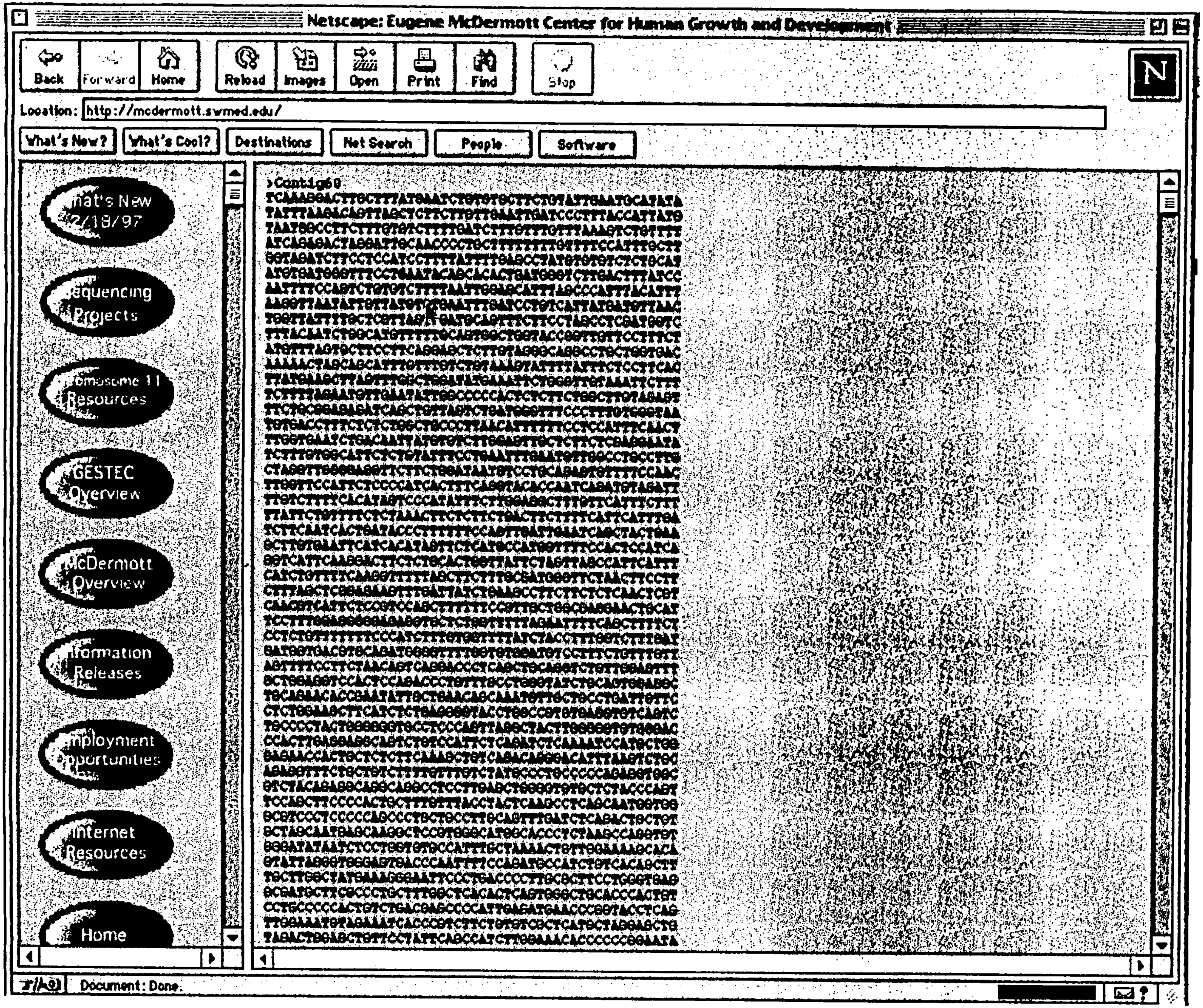
- View the sequence for pDJ298k13
- View the Features Plot for pDJ298k13
- View the Quality Plot for pDJ298k13
- View the Feature Table for pDJ298k13
- View the EST Genbank hits corresponding to Features Table
- View the Non-EST Genbank hits corresponding to Features Table
- View the Non-EST BLAST results
- View the EST BLAST results
- View the cSRL End Sequence BLAST results
- View the GRAIL intron/exon predictions table

Page Maintained by :

Terry Franklin, franklin@genetics.swmed.edu

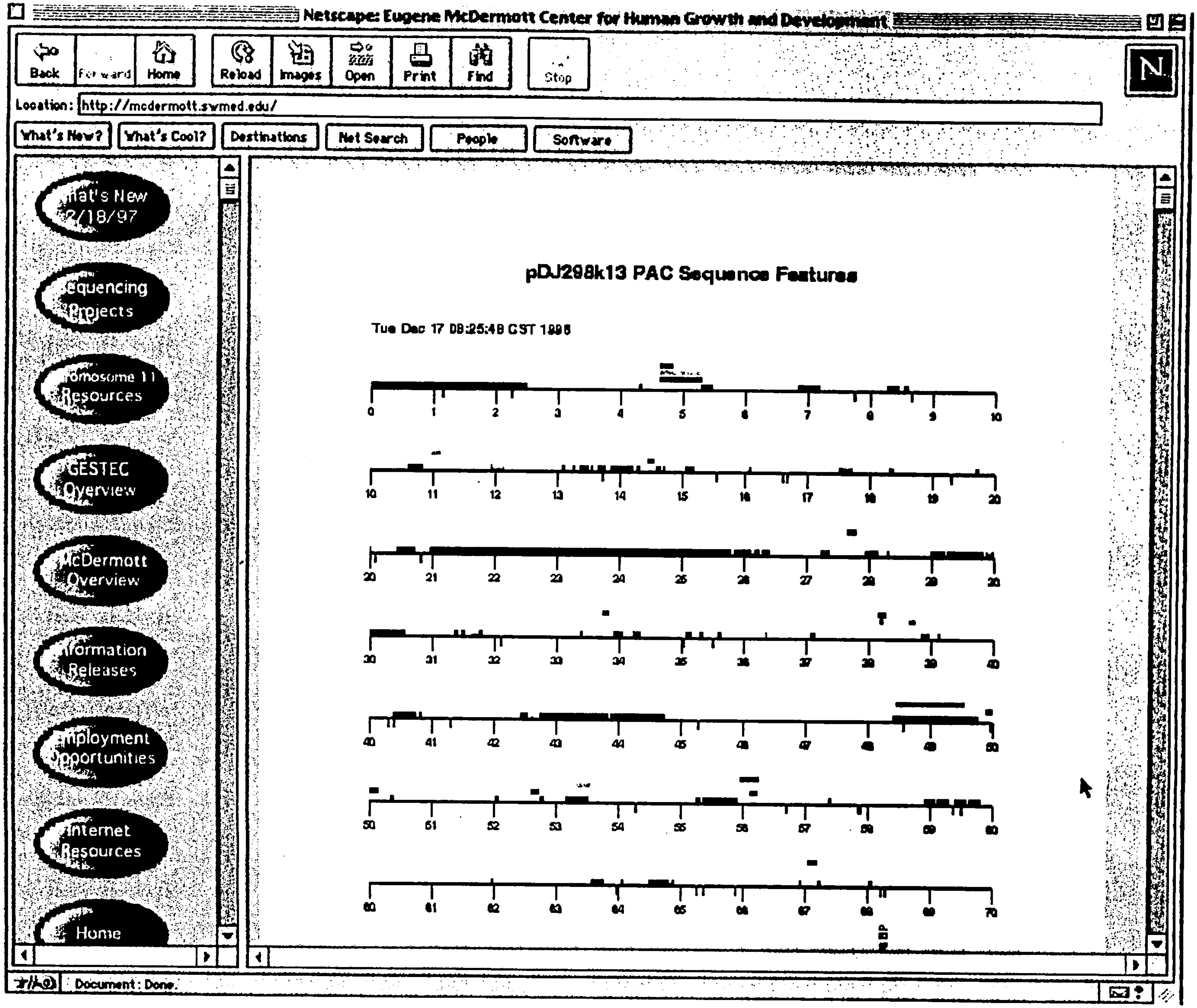
UT Southwestern is an equal opportunity institution

Copyright 1995 University of Texas Southwestern Medical Center



Location: http://mcdermott.swmed.edu/

```
>Contig60
TCAAAGGACTTGCTTTATGAATCTGTGTGCTTCTGTATTGAATGCATATA
TATTTAAGACAGTTAGCTCTTCTTGTTGAATTGATCCCTTTACCATTATG
TAATGGCCTTCTTTGTGTCTTTTGATCTTTGTTTGTTTAAAGTCTGTTTT
ATCAGAGACTAGGATTGCAACCCCTGCTTTTTTTTGTTTTCCATTTGCTT
GGTAGATCTTCCTCCATCCTTTTATTTTGAGCCTATGTGTGTCTCTGCAT
ATGTGATGGGTTTCCTGAATACAGCACACTGATGGGTCTTGACTTTATCC
AATTTTCCAGTCTGTGTCTTTTAATTGGAGCATTTAGCCCATTTACATTT
AAGGTTAATATTGTTATGTGTGAATTTGATCCTGTCATTATGATGTTAAC
TGGTTATTTTGCTCGTTAGATGATGCAGTTTCTTCCTAGCCTCGATGGTC
TTTACAATCTGGCATGTTTTTGCAGTGGCTGGTACCGGTTGTTCCTTTCT
ATGTTTAGTGCTTCCTTCAGGAGCTCTTGTAGGGCAGGCCTGCTGGTGAC
AAAAACTAGCAGCATTTGTTTGTCTGTAAAGTATTTTATTTCTCCTTCAG
TTATGAAGCTTAGTTTGGCTGGATATGAAATTCTGGGTTGTAAATTCTTT
TCTTTTAGAATGTTGAATATTGGCCCCCACTCTCTTCTGGCTTGTAGAGT
TTCTGCGGAGAGATCAGCTGTTAGTCTGATGGGTTTCCCTTTGTGGGTAA
TGTGACCTTTCTCTCTGGCTGCCCTTAACATTTTTTCCTCCATTTCAACT
TTGGTGAATCTGACAATTATGTGTCTTGGAGTTGCTCTTCTCGAGGAATA
TCTTTGTGGCATTCTCTGTATTTCCTGAATTTGAATGTTGGCCTGCCTTG
CTAGGTTGGGGAGGTTCTTCTGGATAATGTCCTGCAGAGTGTTTTCCAAC
TTGGTTCCATTCTCCCCATCACTTTCAGGTACACCAATCAGATGTAGATT
TTGTCTTTTCACATAGTCCCATATTTCTTGGAGGCTTTGTTCATTTCTTT
TTATTCTGTTTTTCTCTAAACTTCTCTTCTGACTTCTTTTCATTCATTTGA
TCTTCAATCACTGATACCCTTTTTTTCCAGTTGATTGAATCAGCTACTGAA
GCTTGTGAATTCATCACATAGTTCTCATGCCATGGTTTTCCACTCCATCA
GGTCATTCAAGGACTTCTCTGCACTGGTTATTCTAGTTAGCCATTCATTT
CATCTGTTTTCAAGGTTTTTAGCTTCTTTGCGATGGGTTCTAACTTCCTT
CTTTAGCTCGGAGAAGTTTGATTATCTGAAGCCTTCTTCTCTCAACTCGT
CAACGTCATTCTCCGTCCAGCTTTTTTCCGTTGCTGGCGAGGAACTGCAT
TCCTTTGGAGGGGAGAGGTGCTCTGGTTTTTAGAATTTTCAGCTTTTCT
CCTCTGTTTTTTTCCCATCTTTGTGGTTTTATCTACCTTTGGTCTTTGAT
GATGGTGACGTGCAGATGGGGTTTTGGTGTGGATGTCCTTTCTGTTTGTT
AGTTTTCCTTCTAACAGTCAGGACCCTCAGCTGCAGGTCTGTTGGAGTTT
GCTGGAGGTCCACTCCAGACCCTGTTTGCCTGGGTATCTGCAGTGGAGGC
TGCAGAACACCGAATATTGCTGAACAGCAAATGTTGCTGCCTGATTGTTC
CTCTGGAAGCTTCATCTCTGAGGGGTACCTGGCCGTGTGAGGTGTCAGTC
TGCCCCTACTGGGGGGTGCCTCCCAGTTAGGCTACTTGGGGGTGTGGGAC
CCACTTGAGGAGGCAGTCTGTCCATTCTCAGATCTCAAAATCCATGCTGG
GAGAACCACTGCTCTCTTCAAAGCTGTCAGACAGGGACATTTAAGTCTGC
AGAGGTTTCTGCTGTCTTTTGTTTGTCTATGCCCTGCCCCCAGAGGTGGC
GTCTACAGAGGCAGGCAGGCCTCCTTGAGCTGGGGTGTGCTCTACCCAGT
TCCAGCTTCCCCACTGCTTTGTTTACCTACTCAAGCCTCAGCAATGGTGG
GCGTCCCTCCCCCAGCCCTGCTGCCTTGCAGTTTGATCTCAGACTGCTGT
GCTAGCAATGAGCAAGGCTCCGTGGGCATGGCACCCTCTAAGCCAGGTGT
GGGATATAATCTCCTGGTGTGCCATTTGCTAAAACTGTTGGAAAAGCACA
GTATTAGGGTGGGAGTGACCCAATTTTCCAGATGCCATCTGTCACAGCTT
TGCTTGGCTATGAAAGGGAATTCCCTGACCCCTTGCGCTTCCTGGGTGAG
GCGATGCTTCGCCCTGCTTTGGCTCACACTCAGTGGGCTGCACCCACTGT
CCTGCCCCCACTGTCTGACGAGCCCCATTGAGATGAACCCGGTACCTCAG
TTGGAAAATGTAGAAAATCACCCGTCTTCTGTGTCGCTCATGCTAGGAGCTG
TAGACTGGAGCTGTTCCTATTCAGCCATCTTGGAAACACCCCCCGGAATA
```



pDJ298k13 PAC Sequence Features

Tue Dec 17 08:25:48 CST 1996

Current Automation Projects

- MerMade oligonucleotide synthesizer
- Sagian/Beckman ^{3}S robot
- Astral DNA sequencer

MerMade

96/192 channel automated oligonucleotide synthesizer

Programmed by Phred/Phrap assemblies using Primo oligonucleotide design software

24 hour/day unattended operation

MerMade I and II in operation, MerMade III constructed, MerMade IV ordered from commercial supplier (Avantec, Inc.)

Cost <$0.10/nucleotide

Available to non-commercial genome centers by no-cost license from University of Texas and contract for construction to Avantec, Inc.

Sagian/Beckman S^3 robot

3 meter rail robot

8 MJ research 96-well PCR thermal cyclers

4 Robbins Hydra 96 channel pipettors

Automated refrigerator storage

Multiple grippers

Currently used for all primer/terminator chemistry sequence assemblies in GESTEC

Current capacity 3,000 to 24,000 samples/day

Custom driver software

Developed at UTSW in collaboration with Sagian - available as a commercial product from Beckman Instruments

Astral DNA Sequencer

Gel based

High lane density (48, 96, 144, 384)

Multiple dye chemistry without altering hardware
(ABI, ET, Bodipy, other)

Higher sensitivity for dyes

Allows a 5th fluorescent dye for auto lane tracking

Spectral decomposition for increased data quality

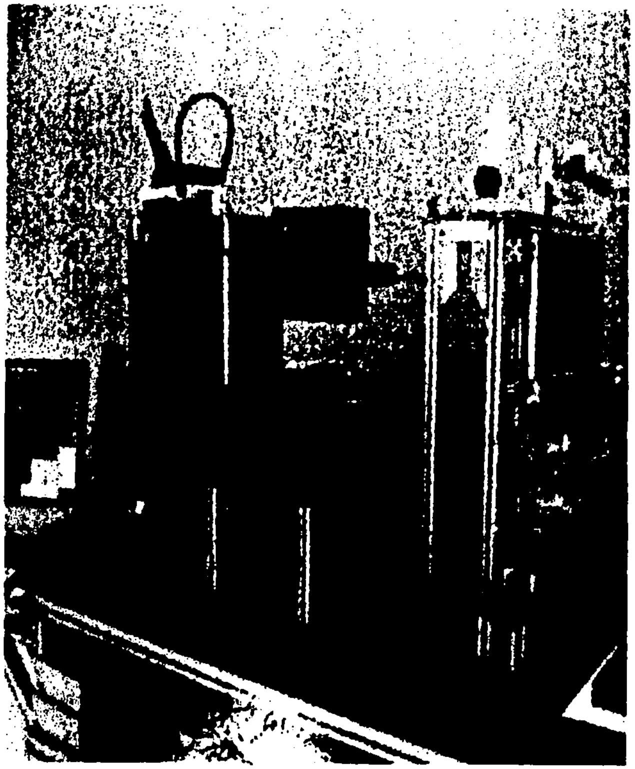
Software for data conversion to industry standard

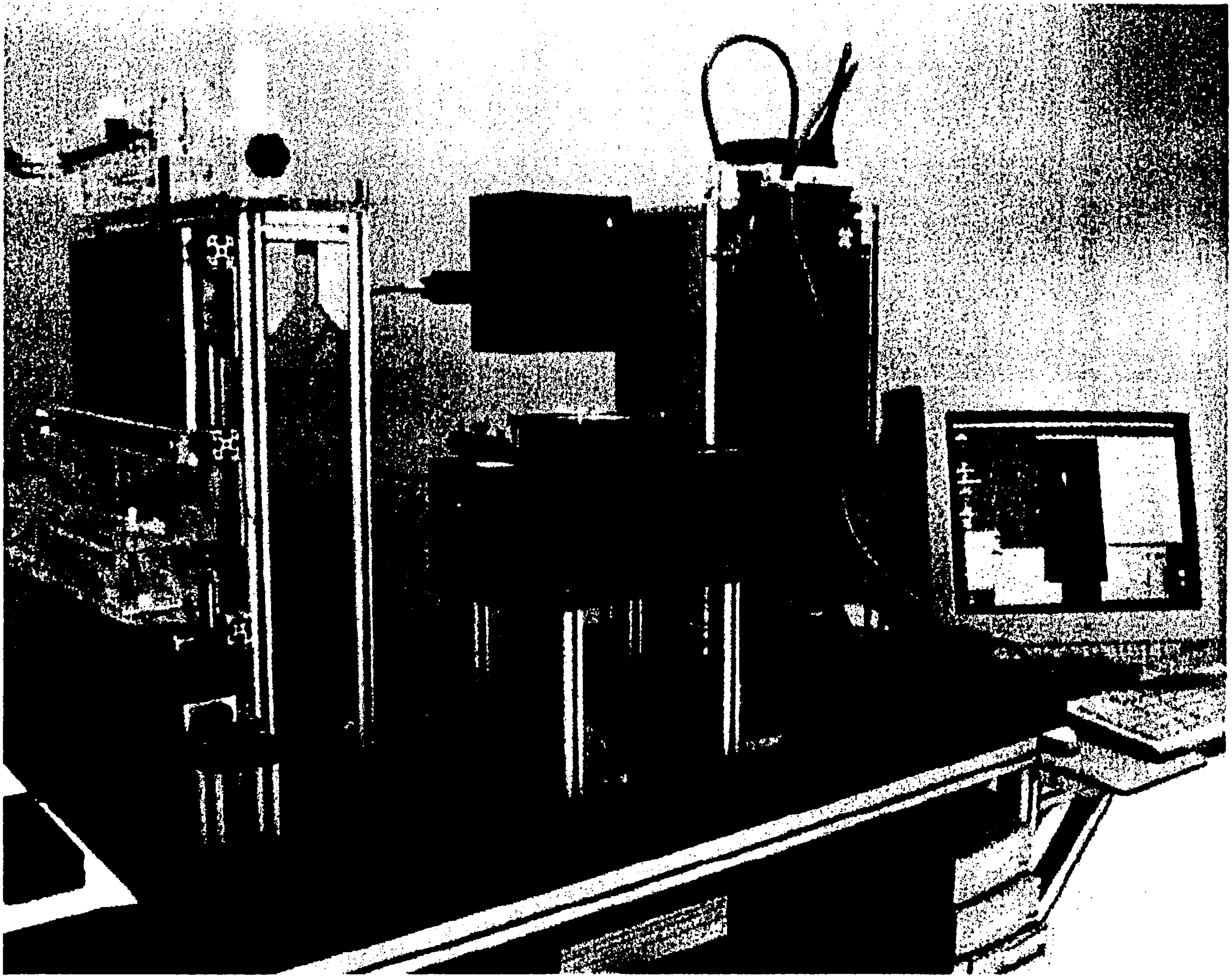
No moving parts for higher reliability

Distribution mechanism under development

ASTRAL, like MISTI employs Hyperspectral Imaging

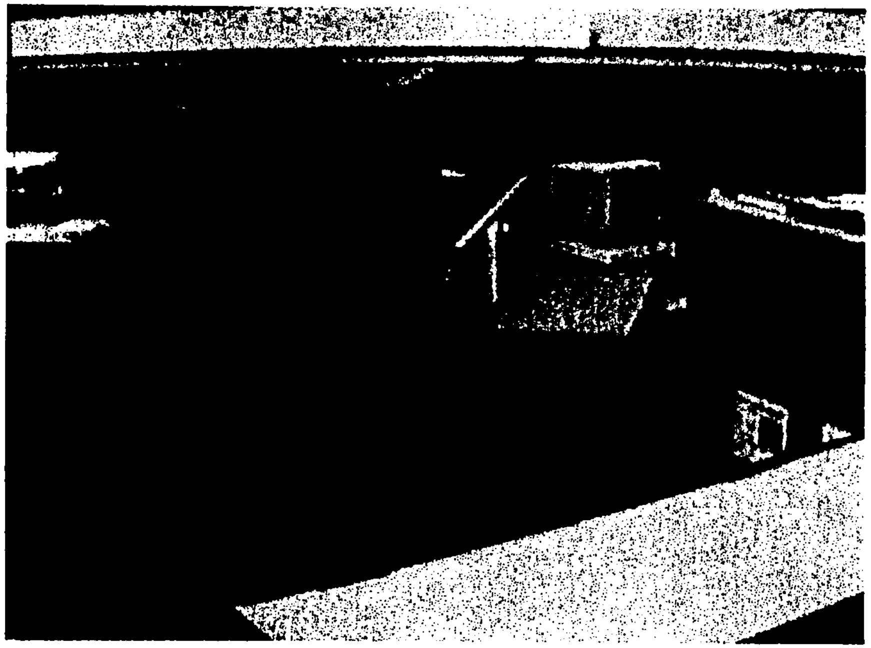
Earth orbit imaging to DNA sequencing





The HP/Convex Parallel Supercomputer

Parallel processing and a large shared memory speed data analysis



8 RISC processers
0.5 GB shared RAM
28 GB Hard Drives
2 Workstations
UNIX, parallelizers
Compilers and optimizers

Groups Sequencing Chromosome 11

Region	Markers		Group	Contact
11p15	11pter	D11S932	UTSW	Evans
11p14	RBTN1	CALCA	GGP	Zabel
11p14	D11S1228	D11S1944E	UTSW	Evans
11p12	D11S1944E	D11S981E	SANGER	Little/Sulston
11p11.2	D11S981E	D11S2399	UTSW	Evans
11q12.2	D11S1368	D11S678	UTSW	Evans
11q13.1	D11S987	D11S3866	UTSW	Evans
11q23	D11S2058	D11S2085	UTSW	Evans

HAWKINS

Finishing Focus

- **Lab/Automation**
  - **Biochemical 'tool box' of methods**
  - **Learning Process**
  - **Finishatron automation**
- **Computer**
  - **Automated workflow**
    - **TaskMaster LIMS**
    - **Trout signal processing/base calling**
    - **Alewife assembler**
    - **Autoeditor**
    - **List generator**
  - **Post Sequencing Varification**
    - **Big Brother**
    - **Restriction enzyme/forward-reverse path checking**

HAWKINS

Production Finishing

- Finishing should be a production line
  - 80-90% of clones must be treated within the system for optimal throughput
  - Set-up ‘swat team’ for completion of more unusual clones
- Finishing by Numbers
  - Set of methods and landmarks for progress and automation streamlining

Hawkins

Whitehead Institute/MIT Genome Sequencing Project

View by Progress ***Last updated January 31st 1997***

Total Finished	2175Kb
Total In Finishing	659Kb
Total	2834Kb

Clone name	Internal Name	Clone Type	Size (kb)	Location	Status	Gaps	Completed
L196C8	L3	Cosmid	39	Human9q34	Finished	0	Sequence
L2C9F1	L5	Cosmid	39	Human9q34	Finished	0	Sequence
S30E11	L6	Cosmid	38	Human9q34	Finished	0	Sequence
L124D6	L15	Cosmid	40	Human9q34	Finished	0	Sequence
S272C1	L16	Cosmid	33	Human9q34	Finished	0	Sequence
S63C9	L19	Cosmid	40	Human Y	Finished	0	Sequence
5195	L22	P1	79	Mouse 19	Finished	0	Sequence
B287E5	L24	BAC	140	Mouse 9	Finished	0	Sequence
1204	L36	Cosmid	42	Mouse 11	Finished	0	Sequence
46A6	L43	Cosmid	44	Human Y	Finished	0	Sequence
L101D11	L27	Cosmid	46	Human9q34	Finished	0	Sequence
-	L18	Cosmid	29	Mouse 11	Finished	0	Sequence
182E3	L8	Cosmid	46	Human9q34	Finished	0	Sequence
152F5	L10	Cosmid	49	Human9q34	Finished	0	Sequence
44J6	L107	BAC	136	Human 17	Finished	0	Sequence
OC401	L53	PAC	107	Human13	Finished	0	Sequence
320L17	L26	BAC	146	Mouse 9	Finished	0	Sequence

HAWKINS

The Learning Curve

April '96

- Infrastructure
  - Computing
  - Team of 20 people, 6 ABIs
  - Team Leaders
- Development
  - Procedures that scale
  - New electrophoresis conditions/devices
- Library Construction
  - Skills/early QC
- Production Sequencing
  - Sequatron Systems
  - WorkFlow
- QC/QA
  - Reagents
  - Gel to gel
  - Projects
  - Auto trend detection

Dec '96

HAWKINS

Current Issues

- Finishing
  - Lab issues
  - Computer issues
  - Automation
- Interpretation
  - Human-Mouse synteny
  - Computational methods
    - » Ken Fasman

Current

Near Future

I. BERKELEY

Drosophila (NIH)
LBNL HGC (DOE)

II. DOE JOINT GENOME INSTITUTE

Livermore
Los Alamos
Berkeley

- Producing Finished sequence

Totals 5 Mb Drosophila
4 Mb HUMAN
RATE 800 KB / MONTH

IV. STRATEGY

- PHYSICAL MAP
- RANDOM LIGHT SHOTGUN
- BUILD PATHS
- TRANSPOSON-FACILITATED

QUALITY - ALL DOUBLE STRANDED
. REDUNDANCY FOR ASSEMBLY
. 1 in 10,000

V. SOFTWARE

PATH-BUILDING SUITE
SPACE (ASSEMBLY, ED., MAP)

VI. HARDWARE

COLONY PICKER
LIBRARY POOLING + REPLICATION
OLIGOSYNTHESIZER
AGAROSE GEL IMAGING
AGAROSE GEL LOADER
DNA PREPARATION ROBOT

VII. PARTNERSHIP WITH INDUSTRY AND JGI DOE

GOALS (VOLUME, QUALITY, CYCLE TIME, COST)

PRECISE GOAL DEFINITION

Benchmarking

METRICAL TOOLS

ROADMAPS FOR Operation + TECH.

IMPLEMENTATION + EVOLUTION

NEW SPACE

VIII. METRICAL TOOLS

PROCESS MODEL

COST MODEL

COST ACCOUNTING

PICK-A-MIX

IX. PROCESS MODEL

A. BUILD

SPACE, EQUIPMENT,
LABOR, PROCESS FLOW

B. VALIDATION

PEOPLE AGREE

MODEL OUTPUT MATCHES
FACILITY OUTPUT

PREDICTIVE VALUE

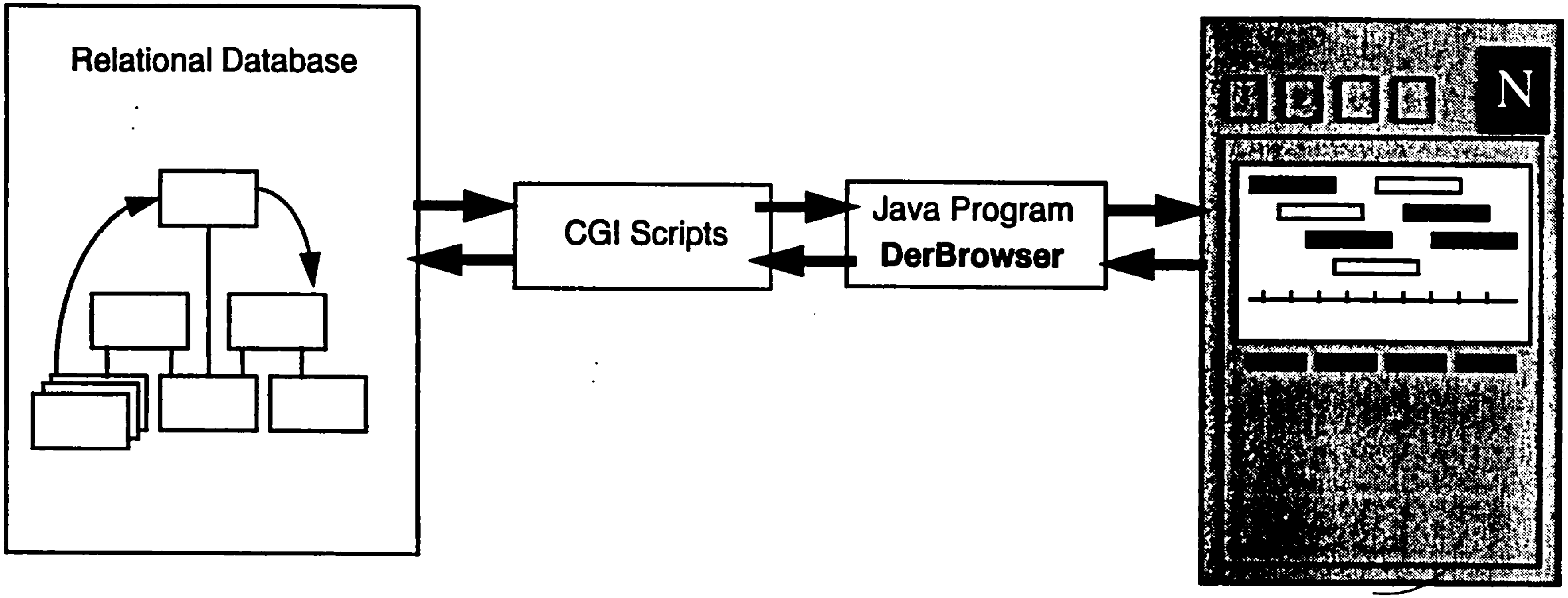
C. UTILITIES

PERSONNEL ACTIVITY RATE
EQUIPMENT UTILIZATION
BOTTLENECK ANALYSIS
LAY OUT ALTERNATIVES
PROCESS ALTERNATIVES

D. BOTTLENECKS

X. PICK-A-MIX / XI CIM

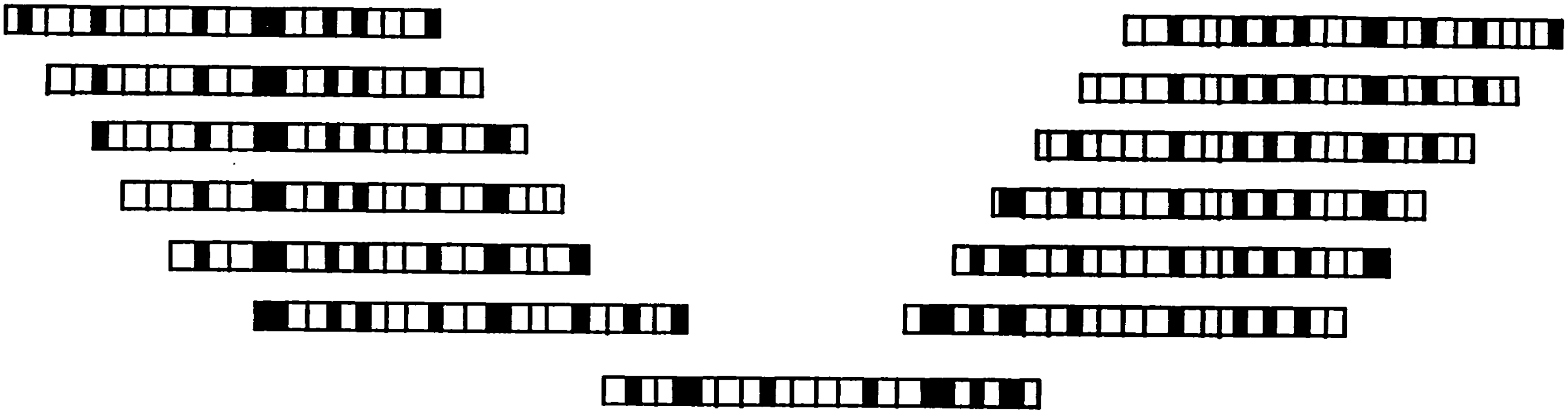
Web access to sequencing status



DerBrowser:

http://www.mpimg-berlin-dahlem.mpg.de/~andy

Preselection of shotgun clones



Projects completed

Xp22

+ Region: DXS8254 - DXS1683, containing the PEX gene
+ Size: 243 kb, contiguous
+ Status: complete
+ Accession number: Y10196

Projects in progress

21q22.3

+ Region: D21S349 - MX1
+ Size: 500 kb
+ Status: 3 cosmids and 2 PACs at different sequencing stages, other shotgun libraries in preparation

Xq28

+ Region: DXS304 - DXS1345, proximal to MTM gene
+ Size: 320 kb (one cosmid in region sequenced previously)
+ Status: finishing stage

Xq13

+ Region: GJB1 - DXS559
+ Size: 500 kb
+ Status: shotgun libraries in preparation

Xq13

+ Region: DXS227
+ Size: 150 kb
+ Status: shotgun libraries in preparation

Xq12

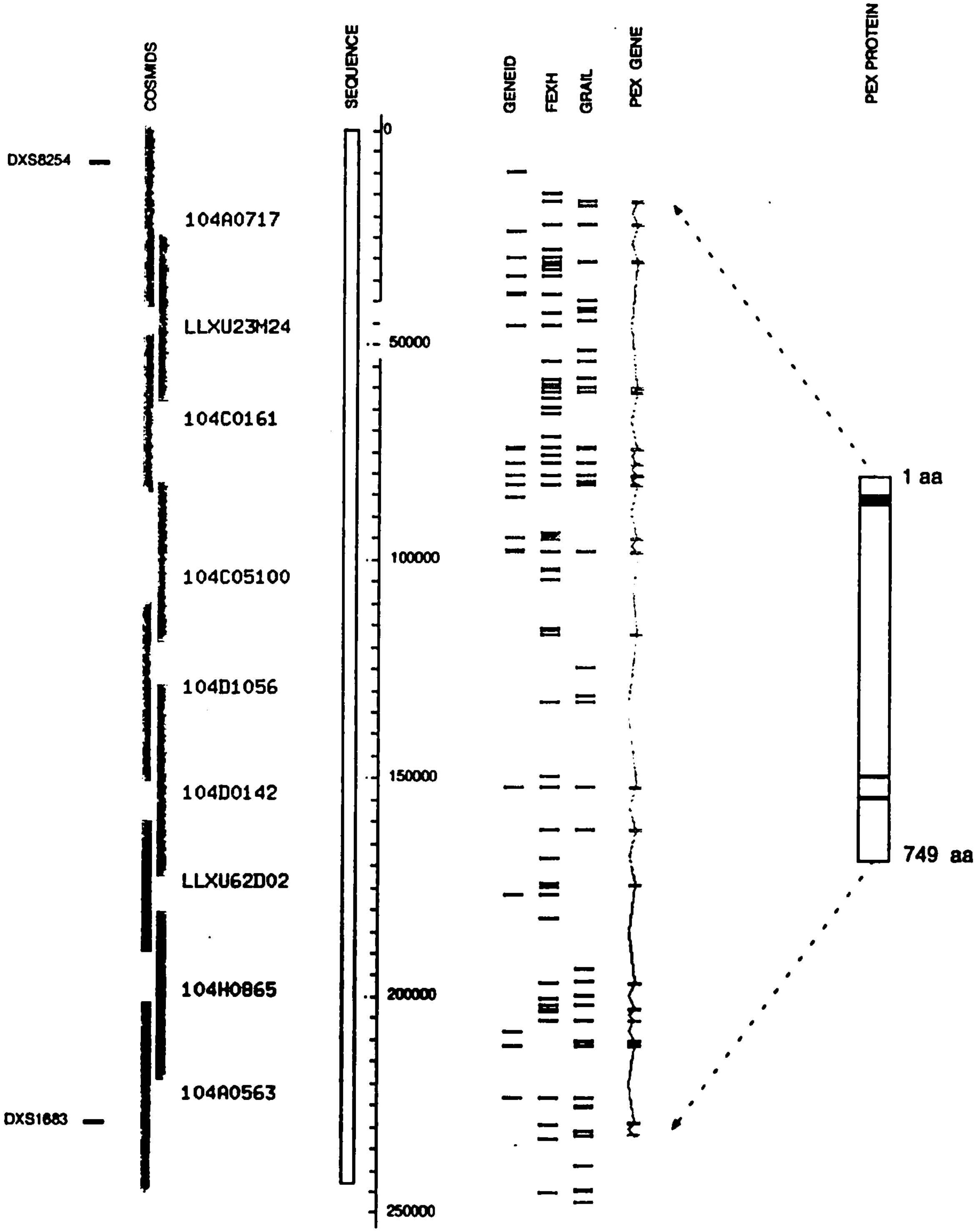
+ Region: DXS908
+ Size: 150 kb
+ Status: shotgun libraries in preparation

17p11

+ Region: D17S71 - D17S58
+ Size: 1000 kb
+ Status: shotgun libraries in preparation

Data quality

- Attempt to close all gaps
- Double stranding/alternative chemistry
- Cover all regions by sequence from more than one shotgun clone
- Attempt to resolve all problematic regions
- Confirm sequence by comparison to restriction digests



Accession number: Y10196

Assembly and analysis of sequence

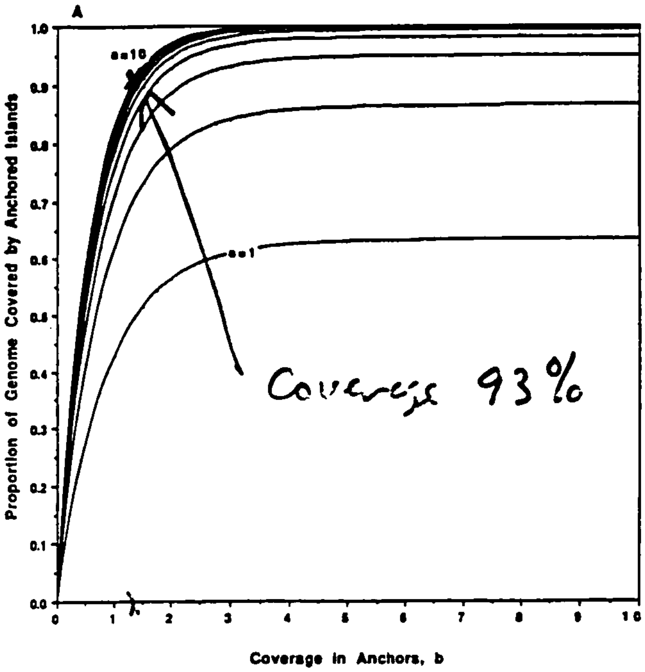
- Staden package: pregap programs, xgap and gap4
- Phred/Phrap (P. Green) and Phrap2Gap (Sanger Centre)
- Gene prediction: Grail, Genefinder (V. Solovyev), Xpound
- Masking of repeats: Repeat Masker/Repbase (A. Schmidt) and Blastn/Simple.db with XBLAST (J-M Claverie)
- Database searches: Blastn and Blastx/ nr and dbEST
- Search and analysis tools: Seqsplit/Blastunsplit and MSPcrunch/Blixem (E. Sonnhammer and R. Durbin)
- Data storage and visualisation: Acedb

Shotgun cloning and sequencing

- Starting DNA: CsCl purified cosmid/PAC
- Standard shotgun cloning: insert sizes 1.2-1.8 kb, sequencing vector: pUC18
- Clones picked in microtitre dishes, inserts PCR amplified
- Cycle sequencing performed using ABI Catalyst, reactions run on ABI 377s
- Data collection, transfer to Unix environment
- Gap closure/finishing after assembly: reverse reads, directed primer walking, PCR

Chr. 21 - construction of sequence-ready maps

- Libraries: Chr. 21 cosmid and whole genome PAC (and BAC)
- Hybridisation screening using STS probes and riboprobes (extension of existing contigs, and anchoring of new ones)
- FISH mapping of selected clones
- Contigs also contributed by collaborating groups
- Restriction digests to aid selection of a minimal tiling path
- Higher resolution fingerprinting performed in selected regions
- End sequencing of clones to aid gap closure

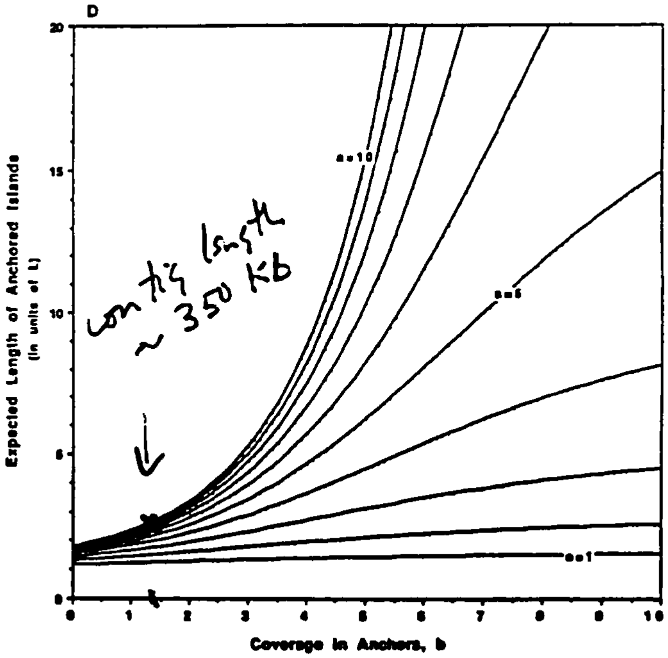


Coverage 93%

ASSUME - 1STS/100 KB.

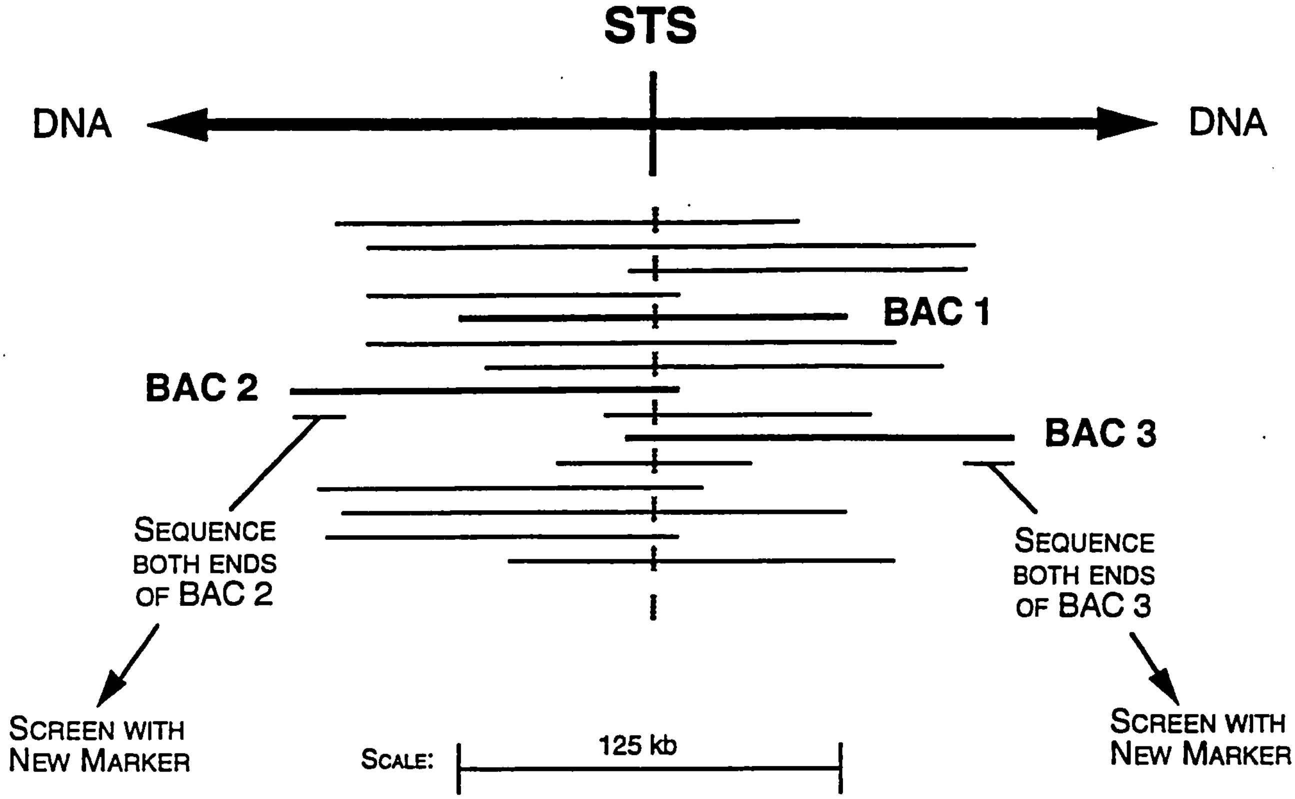
ASSUME - 10X LIBRARY

ASSUME - CLONE ~ 140 KB

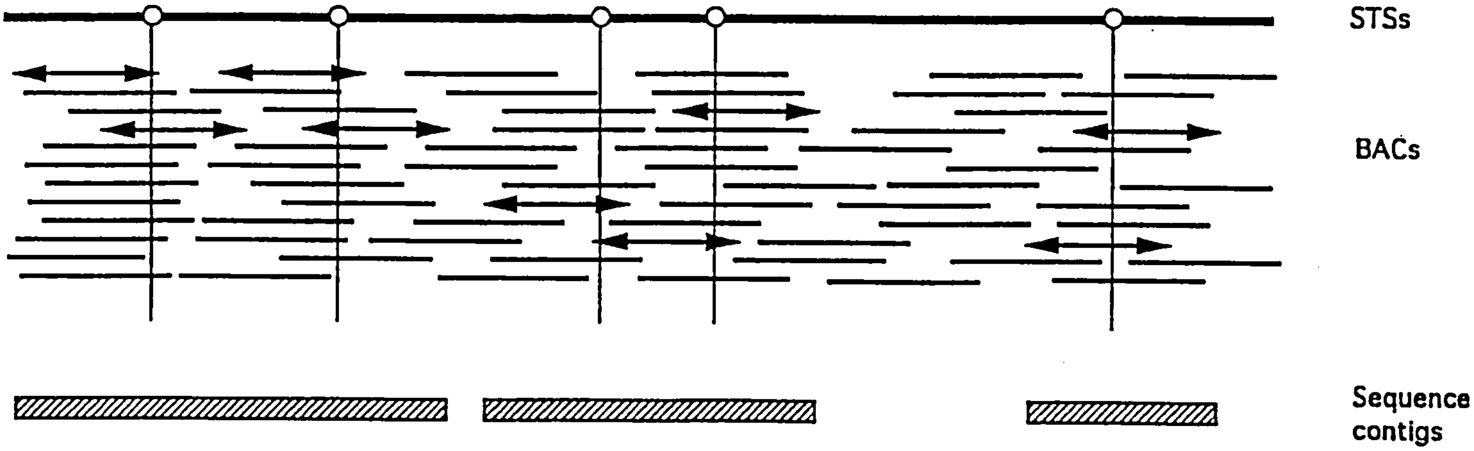


contig length
~ 350 Kb

HUDSON


STS
DNA
DNA
BAC 1
BAC 2
BAC 3
SEQUENCE BOTH ENDS OF BAC 2
SEQUENCE BOTH ENDS OF BAC 3
SCREEN WITH NEW MARKER
SCREEN WITH NEW MARKER
SCALE:
125 kb

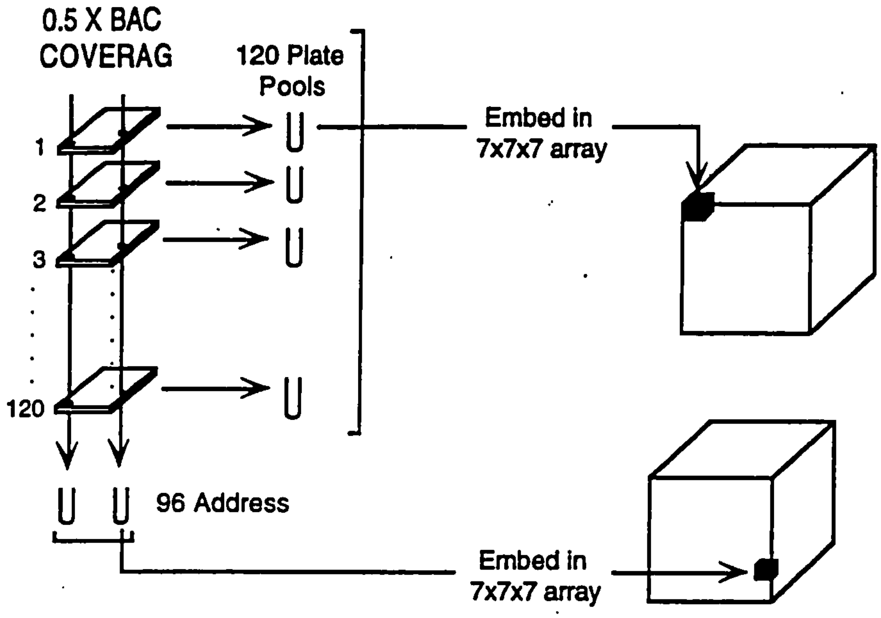
Hudson



Hv0520

1. BAC POOLING SCHEME

PRIMARY **SECONDARY**



0.5 X BAC COVERAG

120 Plate Pools

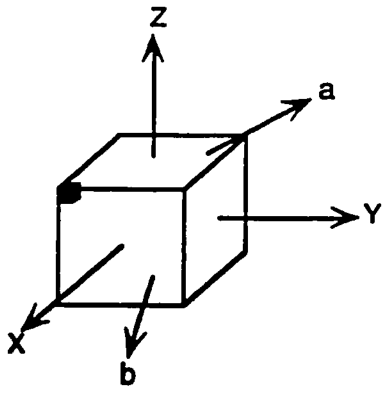
1
2
3
120

96 Address

Embed in 7x7x7 array

Embed in 7x7x7 array

2. SCREENING BAC POOLS



70 PCR Assays For a 0.5X Library

3. SCREENING 20X BAC LIBRARIES

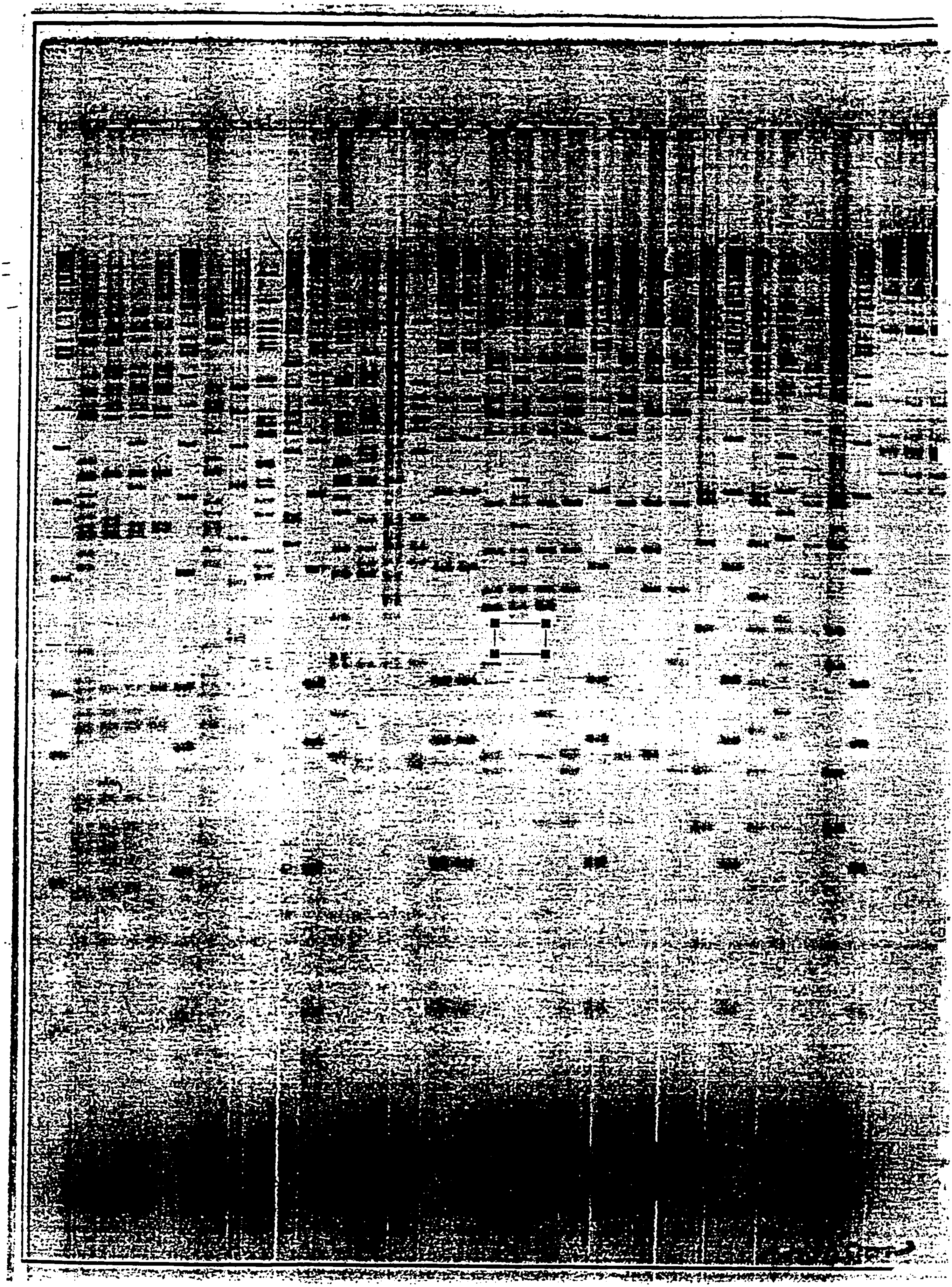
2800 PCR ASSAYS for a 20X Library

GENOMATRON: **300,000 PCRs/day**

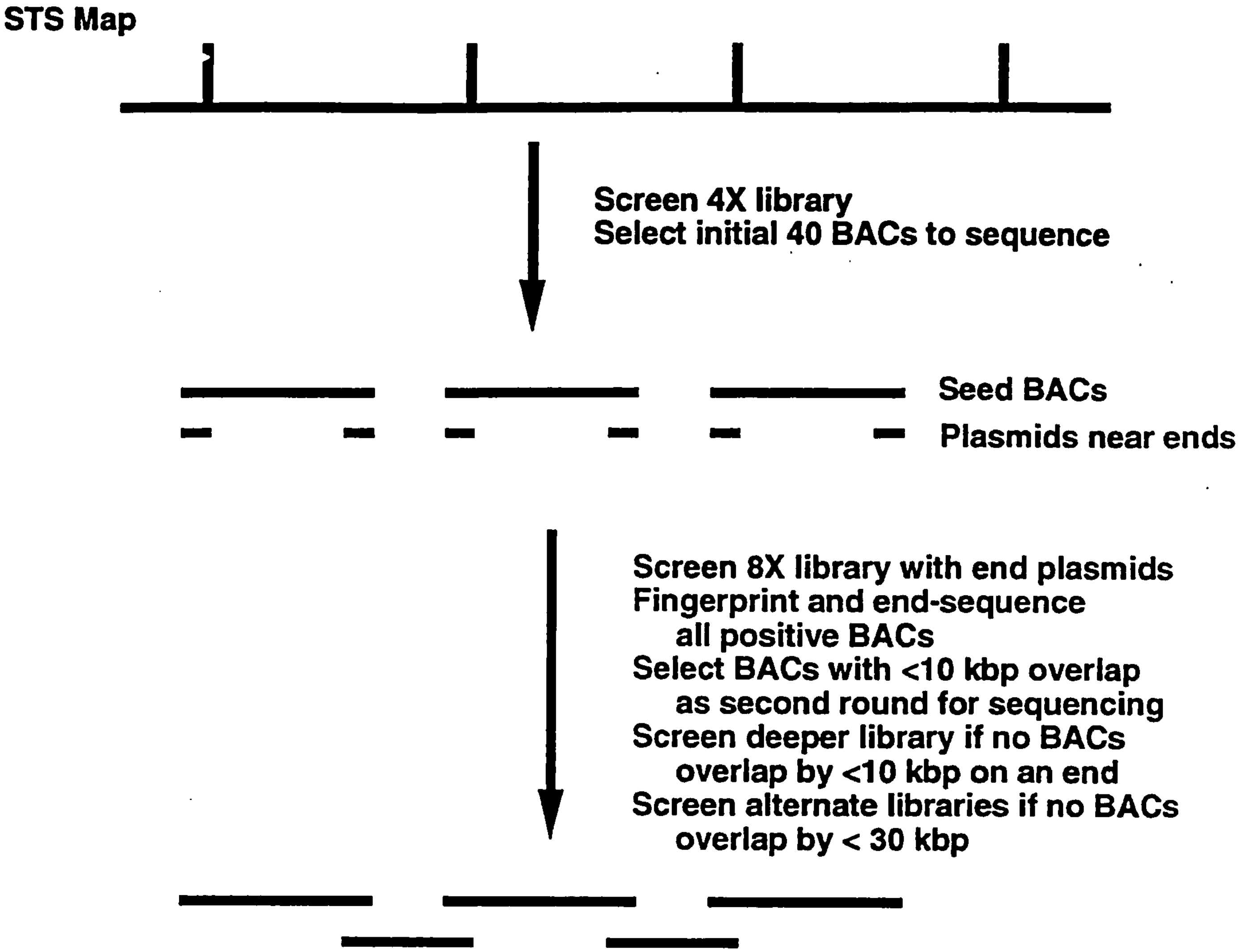
CAPACITY: **100 STSs screened/day**

Hudson

C:\I ATA\adedig10.ds , Range = 0.000-100004 rfu, 106%



TIGR/CalTech Mapping Strategy



STS Map

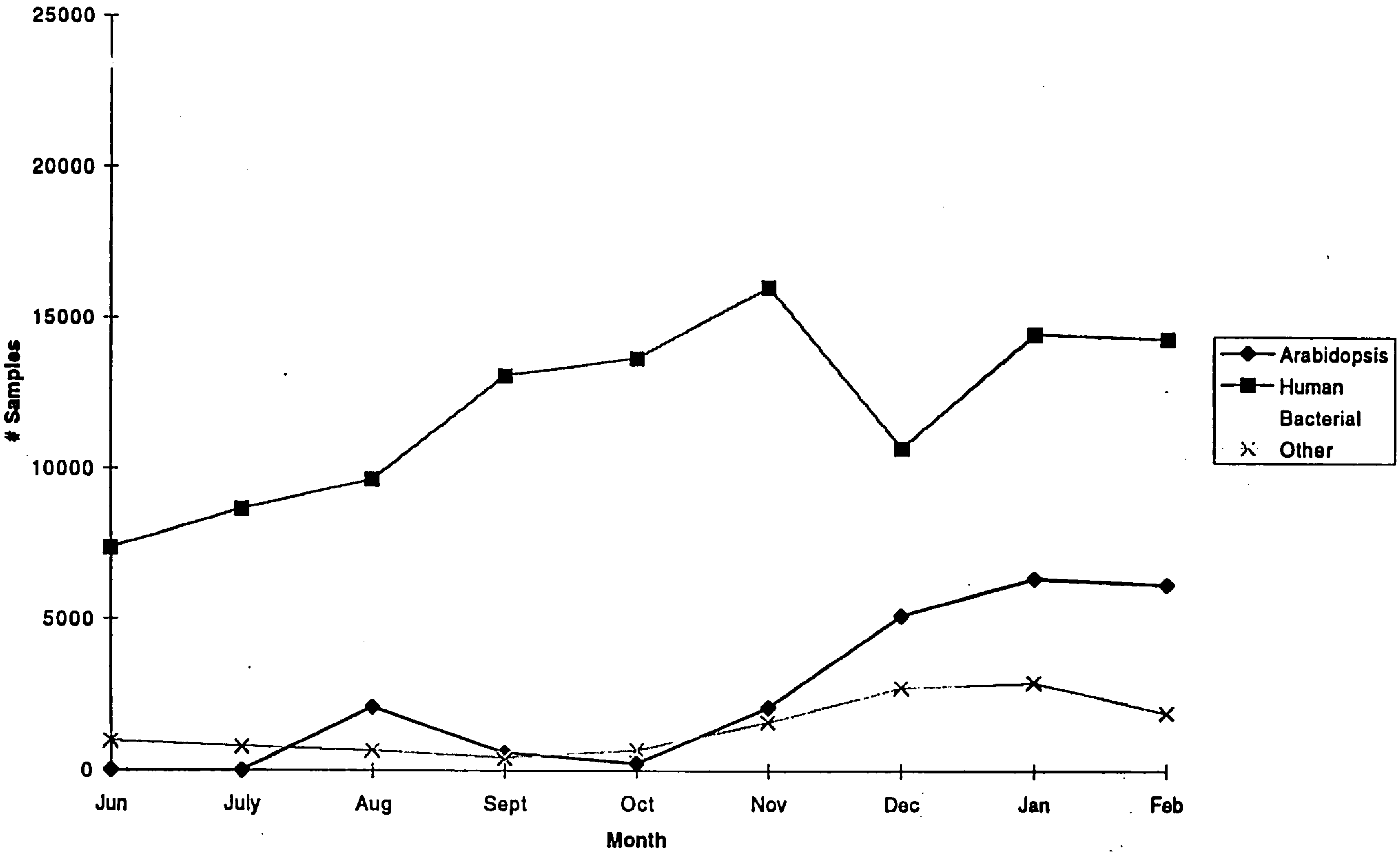
Screen 4X library
Select initial 40 BACs to sequence

Seed BACs

Plasmids near ends

Screen 8X library with end plasmids
Fingerprint and end-sequence
all positive BACs
Select BACs with <10 kbp overlap
as second round for sequencing
Screen deeper library if no BACs
overlap by <10 kbp on an end
Screen alternate libraries if no BACs
overlap by < 30 kbp

Sequencing by Project 6/96-2/97



Summary

Category	# of BACs	Size
Submitted to GenBank	18	2,643,073
Closure	5	735,000
Random	2	360,000
Ready for random	12	1,875,000

Library Team

Cheryl Phillips, Kun Shen, Marie LaBombard

Random Team

10 377xl, 9 373, 1 373xl, 5 Catalyst
Joyce Fuhrmann, Tanya Mason, Steve Bass, Paul Sadow, Jen Tench, Lisa Jiang, Roy Sittig

Closure Team

Rhonda Brandon, Kurt d'Andrea, Sean Sykes, Tracy Spriggs, Tammy Lockwood

Gene List

G1 to S Phase transition protein 1, GST1
B cell maturation protein
hypothetical protein CIT987SK_2A8_1
extoses like gene (partial)
hypothetical protein CIT987SK_362G6_1
hypothetical protein CIT987SK_362G6_2
T-complex protein 1, Beta subunit (TCP-1-BETA), partial
Human gene for Myosin heavy chain (partial)
Multidrug resistance-associated protein isolog
Multidrug resistance-associated protein
pM5
eIF-3 p110 subunit

12 genes 2,363,073 bp

OR

1 gene per 196 kbp (!)

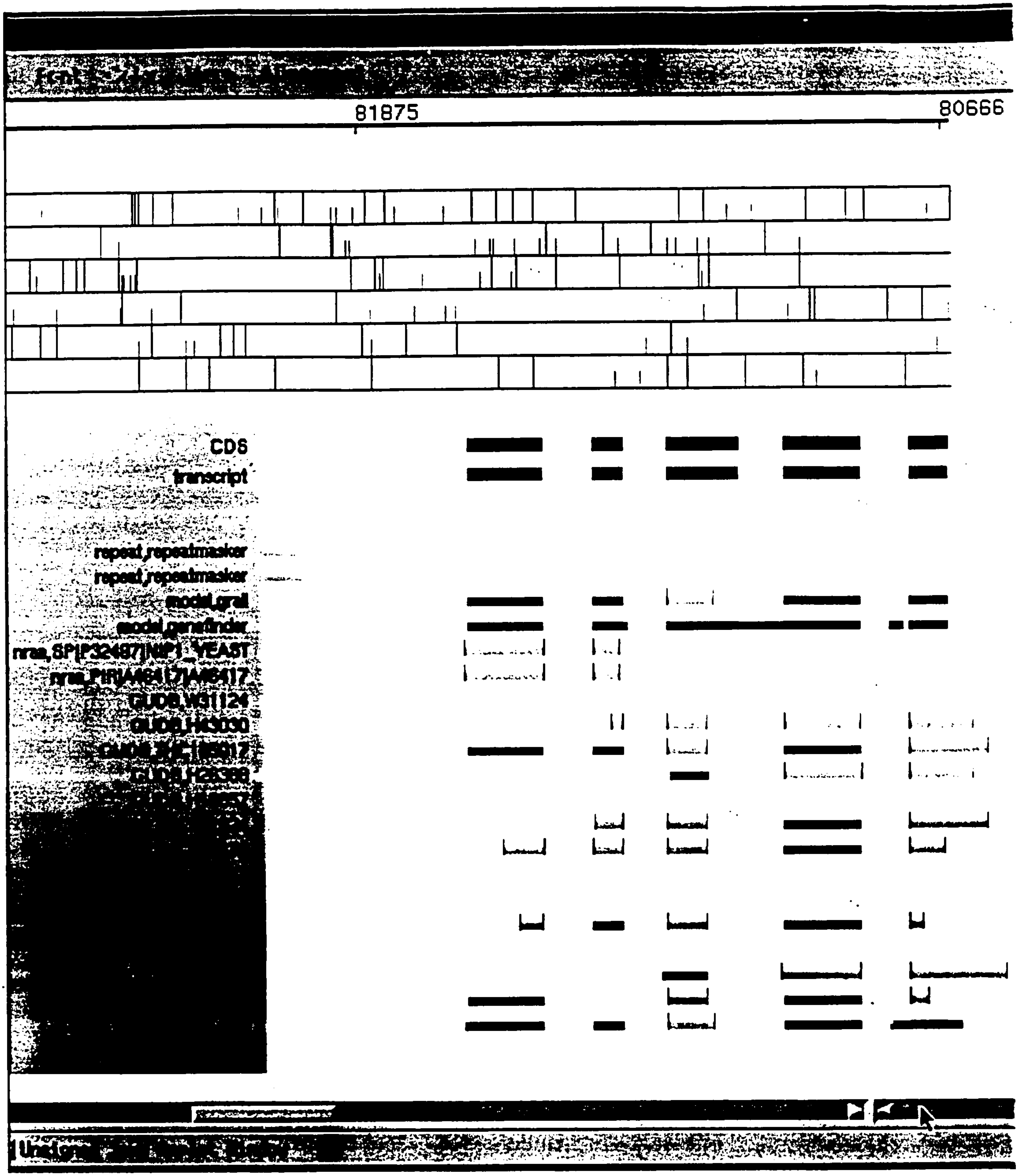

81875
80666
CDS
transcript
repeat,repeatmasker
repeat,repeatmasker
model,grail
model,genefinder
nraa,SP|P32497|NIP1_YEAST
nraa,PIR|A46417|A46417
GUDB,W31124
GUDB,H43030
GUDB,H25368

Table of Double Chemistry Effort and Results

BAC name	Total Len	Single-strand	Terminators	Terms in area	Bases Changed
C16Q	227,403	25,110	209	74	3
CPBA	136,182	14,991	202	73	2

French Sequencing Center

Centre National de Séquençage

Budget	14 M $
Staff	110-120
Location	Evry (near Généthon)
Starting	Summer 1997
Projects	To be submitted

Project Evaluation

Scientific Committee

- scientific quality
- feasibility, opportunity
- scientific interest
- scientific priority

Steering Committee

- political recommendations about projects
- priority decisions
- recommendations about policies on data release and intellectual property.

Stanford Human Genome Center

Chromo 21

EPM1 1.2 Mb

DS .4 Mb

Chromo 4

4q25 5 Mb

Finished 100 kb

In GenBank 1.2 mb >3 kb

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

A T G C

~~1 2 3 4 5 6 7 8 9 10 [illegible] 20~~

Summary of targets

Main projects

Work is in progress on the following five chromosomes. Selected regions are the subject of early effort as listed, but further mapping and clone isolation is under way for the majority of each chromosome. See individual project pages for further information.

These regions of Chromosomes 22 and X are being sequenced jointly with GSC, St Louis.

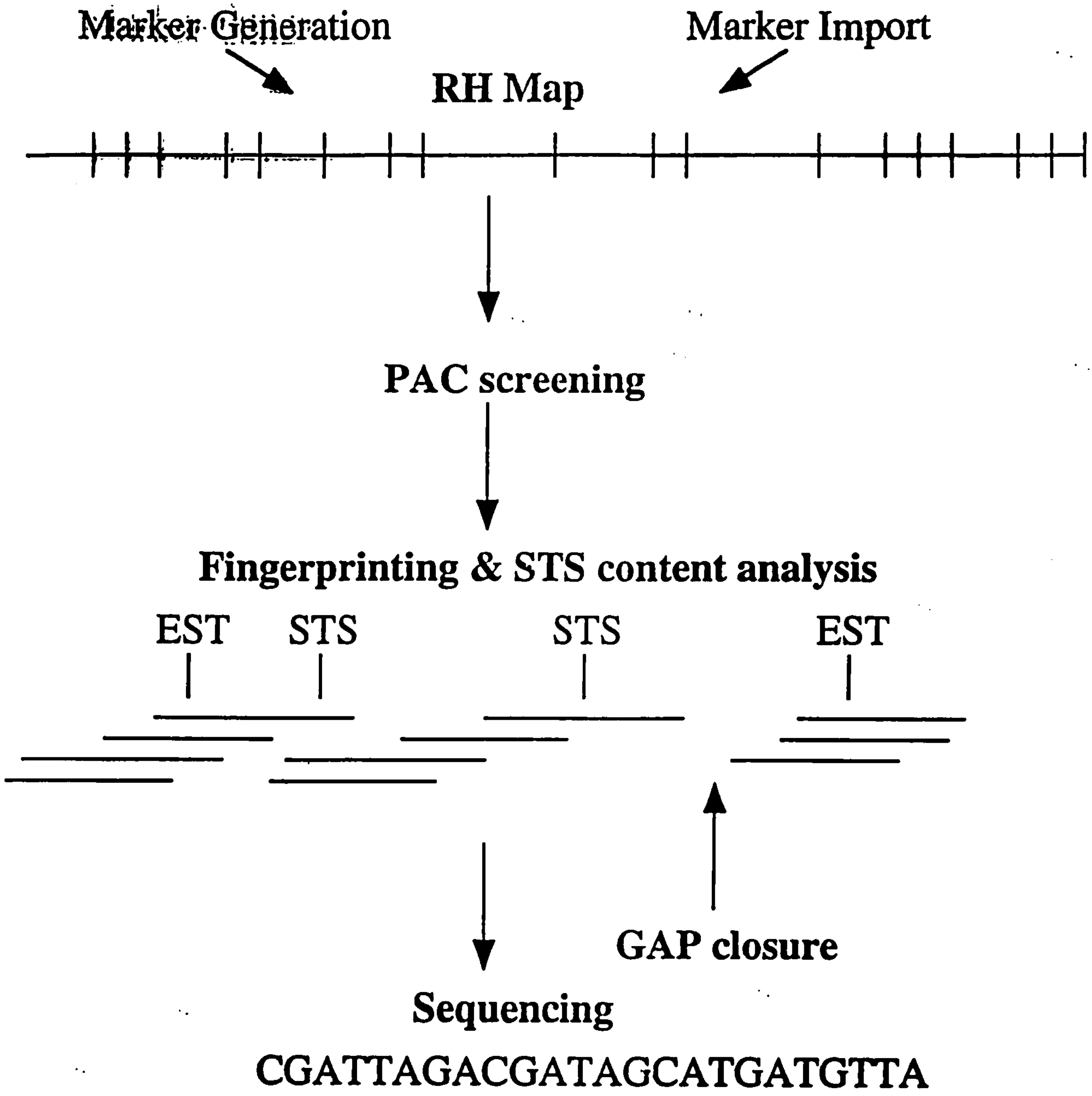
Chromosome 1	300 Mb	1p35-1pter 1pcen-1p13 1q22
Chromosome 6	160 Mb	6p21.3 6p23 6q21 6q27
Chromosome 20	80 Mb	20q11.2-13.1
Chromosome 22	25 of 45 Mb	22q12-13
X chromosome	90 of 150 Mb	Xp Xq22 Xq23-26

Sequencing collaborations

We undertake collaborations to sequence limited regions of specific interest, as listed below:

Chromosome	Size	Region
3p21.3	0.3 Mb	The LUCA6 region
4p	1.6 Mb	The HD region
11p13	0.2 Mb	The PAX6 region
11p15.5	80 kb	
12		The MODY3 region
13q12	0.9 Mb	The BRCA2 region
16p	0.3 Mb	The globin region

See also:


Marker Generation
Marker Import
RH Map
PAC screening
Fingerprinting & STS content analysis
EST
STS
STS
EST
GAP closure
Sequencing
CGATTAGACGATAGCATGATGTTA

Sanger Centre Summary of Human Progress
(all figures are Mb except markers)

Chromosome	1	6	20	22	X	Other	Total
S.C. region	300	160	80	25	90		655
Markers working	3029	2720	1268	951 (+166)	866 (+127)		
[Markers/Mb]	[10.1]	[17.0]	[15.9]	[21.1]	[9.6]		
Coverage in bacterial clones	20	23.9	5.2	19.3	29.0		97.4
Ready for seq	0.8	4.4	1.0	5.0	10.2		17.5
Unfinished seq	0	1.9	0	5.4	4.1	0.5	11.9
Finished seq	0	0.6	0	3.1	7.5	3.4	14.6
Total seq on ftp	0	2.5	0	8.5	11.6	3.9	26.5

Sanger Centre Total Sequence Output (Mb)
February 1997

	Unfinished	Finished	**Total in Public Domain**
Nematode	9.7	29.7	**39.4**
Human	13.8	14.2	**28.0**
Yeasts	0.0	6.2	**6.2**
TB	1.1	2.1	**3.2**
TOTAL	24.6	52.2	**76.8**

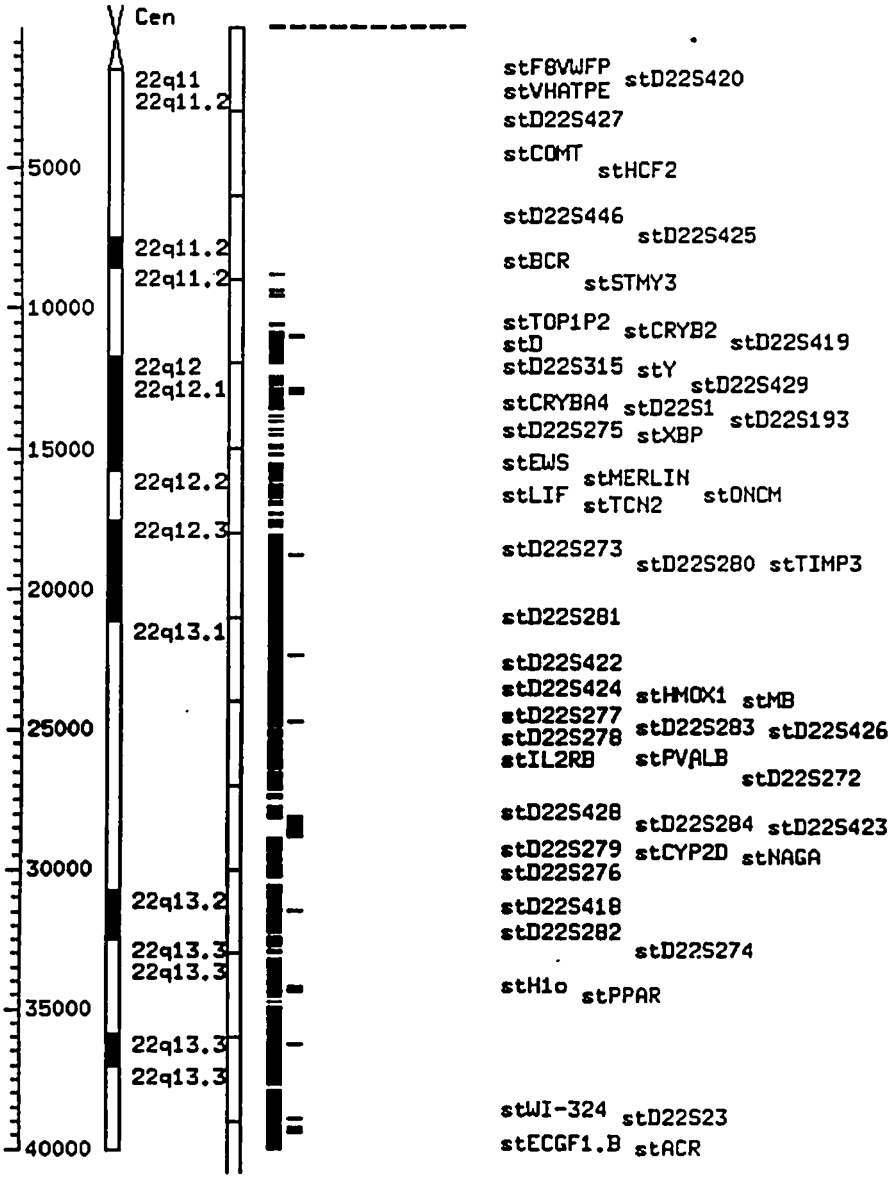
TOTAL FINISHED LAST YEAR	34 Mb

Map Status for Chromosome 22

The picture on the left shows the current status of sequencing for **Chromosome 22.**

Click on the **column of white boxes** to zoom in on an interval of the chromosome. Click on the **red boxes** to see the clones being sequenced. Click on the marker names to see a report for that marker (this is still under development).

All sequence data for this region is available from the human sequence directory of our FTP site.



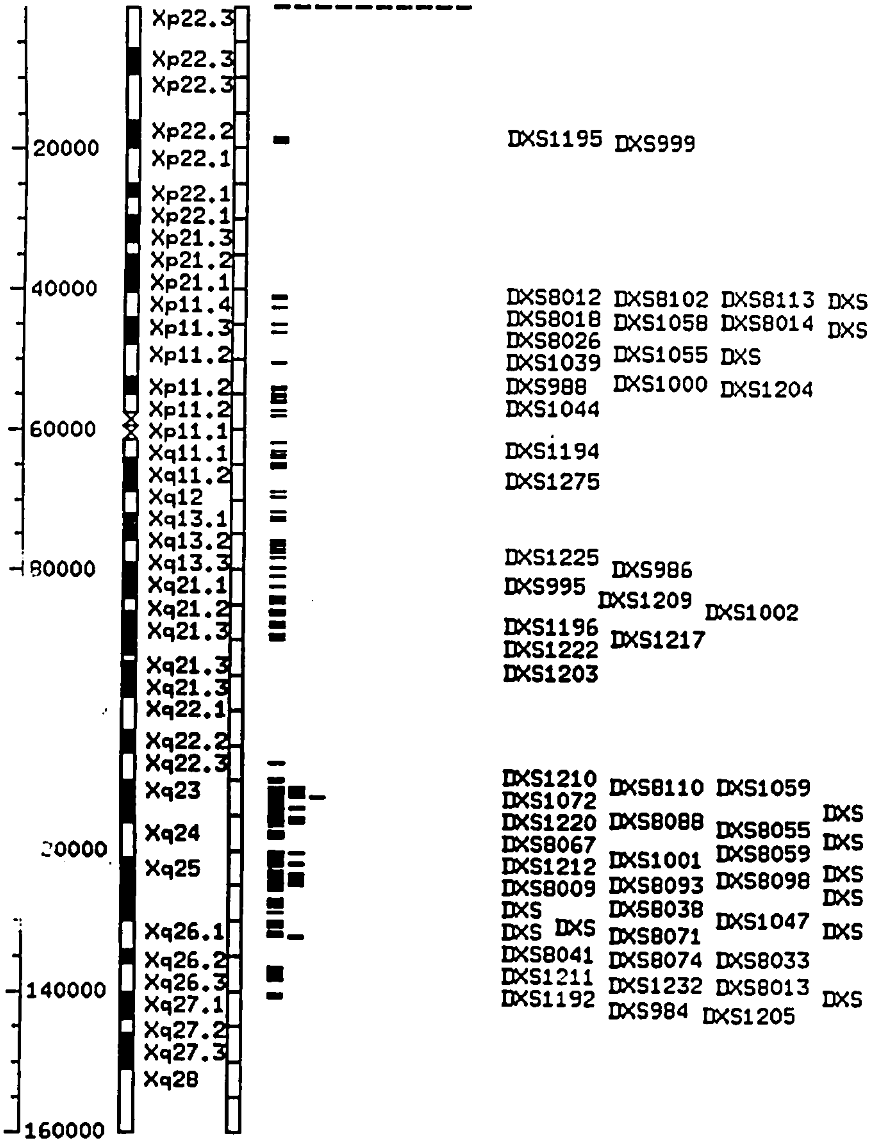
The graphical display was made using acedb

Map Status for Chromosome X

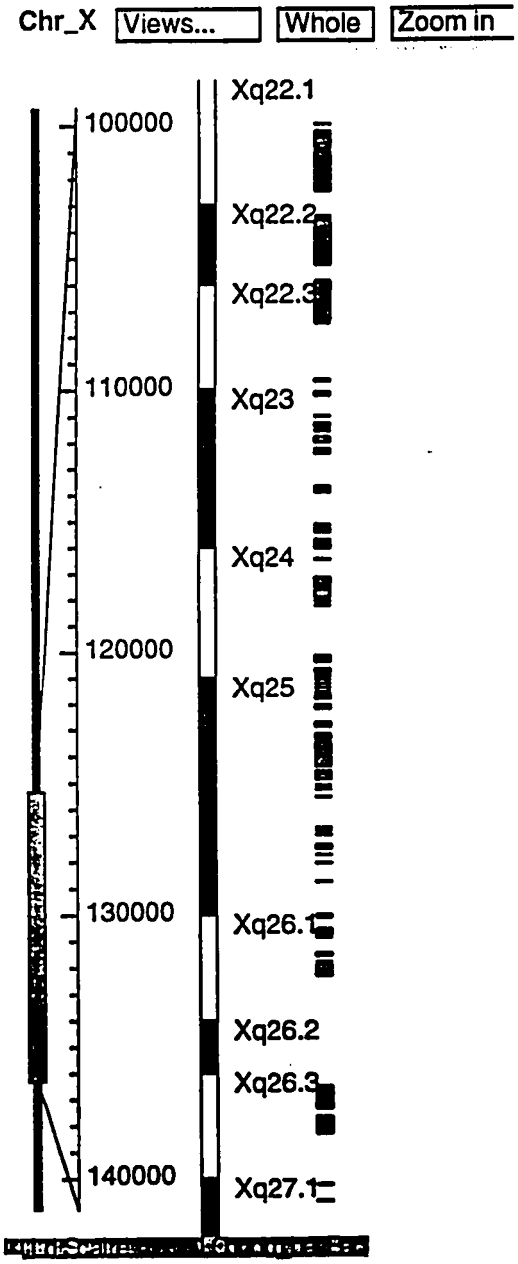
The picture on the left shows the current status of sequencing for **Chromosome X**.

Click on the **column of white boxes** to zoom in on an interval of the chromosome. Click on the **red boxes** to see the clones being sequenced. Click on the **marker names** to see a report for that marker (this is still under development).

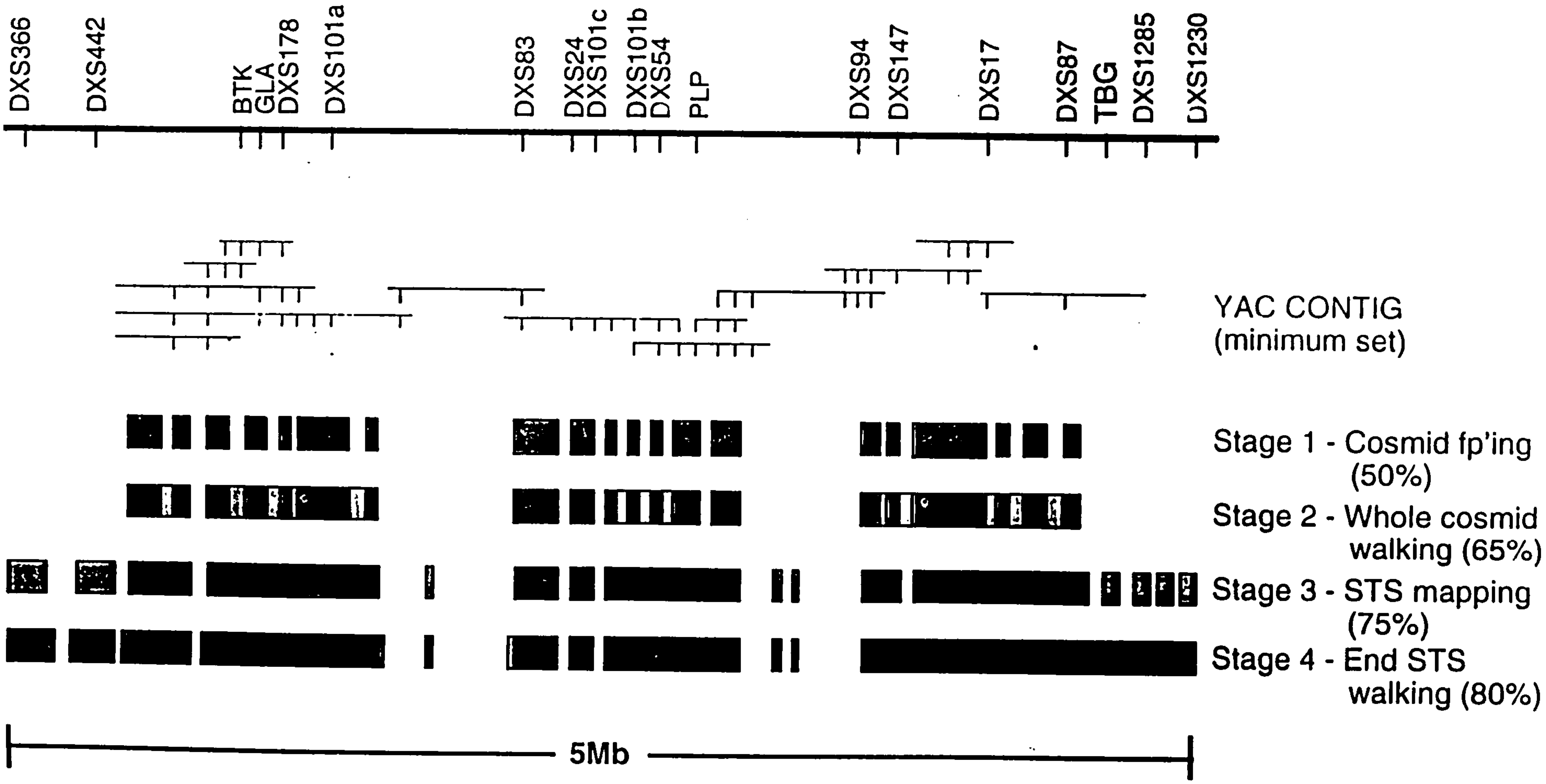
All sequence data for this region is available from the human sequence directory of our FTP site.



The graphical display was made using acedb


Chr_X
Views...
Whole
Zoom in
100000
110000
120000
130000
140000
Xq22.1
Xq22.2
Xq22.3
Xq23
Xq24
Xq25
Xq26.1
Xq26.2
Xq26.3
Xq27.1

Cosmid Coverage In Xq22 From DXS366 To DXS1230



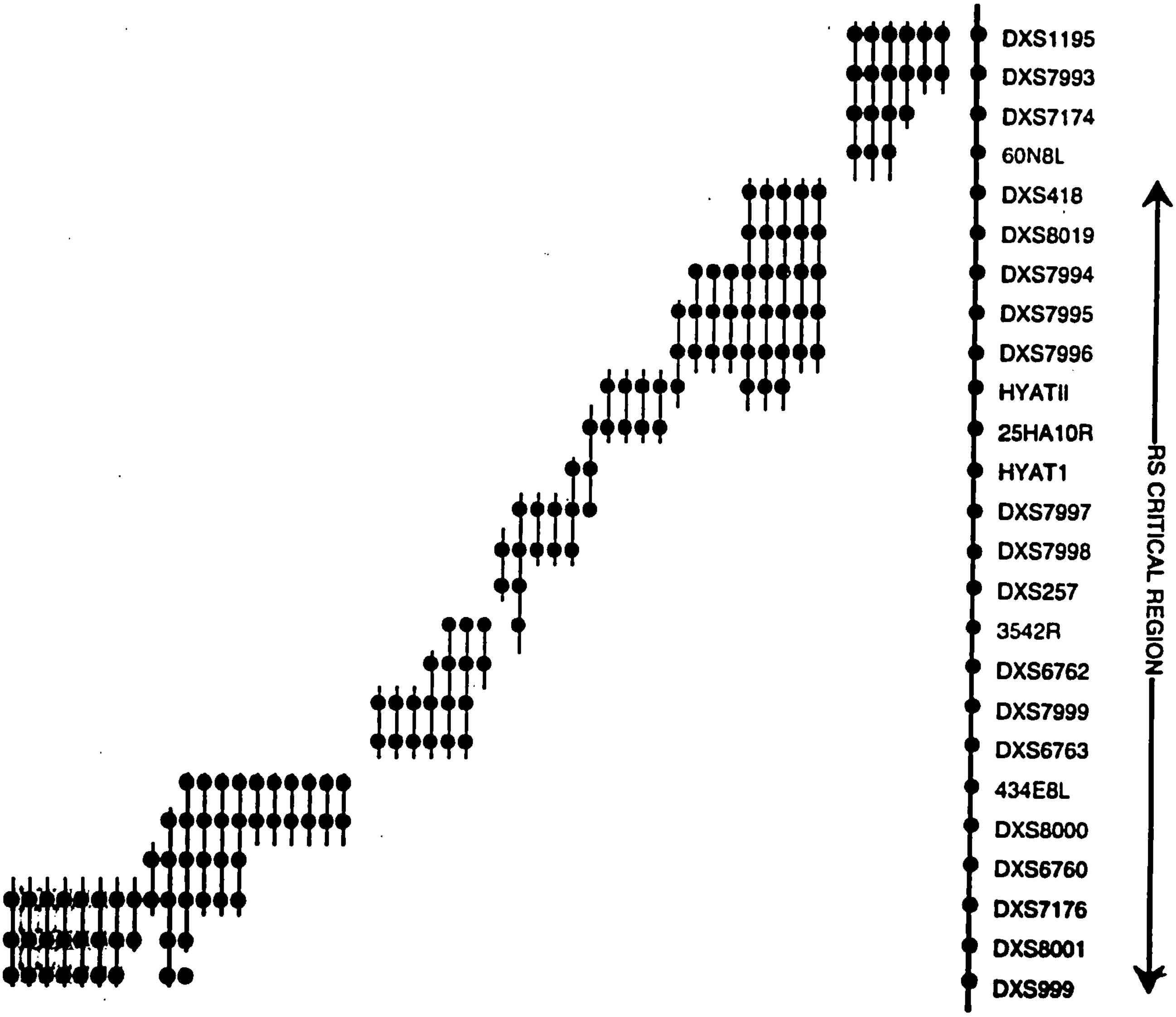
YAC CONTIG (minimum set)

Stage 1 - Cosmid fp'ing (50%)

Stage 2 - Whole cosmid walking (65%)

Stage 3 - STS mapping (75%)

Stage 4 - End STS walking (80%)


DXS1195
DXS7993
DXS7174
60N8L
DXS418
DXS8019
DXS7994
DXS7995
DXS7996
HYATII
25HA10R
HYAT1
DXS7997
DXS7998
DXS257
3542R
DXS6762
DXS7999
DXS6763
434E8L
DXS8000
DXS6760
DXS7176
DXS8001
DXS999
RS CRITICAL REGION

GENOME SEQUENCE QUALITY CRITERIA

- Fidelity:
  - *2X validation* of all sequence-ready clones, using method adequate to detect small ($< 1kb$) coligations, deletions, transposon insertions
- Accuracy:
  - Error rate $< 1/10kb$
  - Base-specific error probabilities submitted with sequence
  - Independent test of assembly accuracy
- Contiguity:
  - All gap sizes estimated
  - All sequence contigs *oriented* and *ordered* within the chromosome

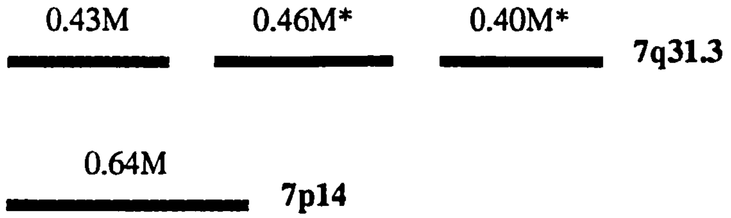
UWGC SEQUENCING STRATEGY: KEY FEATURES

- MCD mapping
  - Clone validation
  - Better tiling paths
  - More efficient finishing (gap closure)
  - Assembly verification
  - Current cost: $.05 to $.12 per bp
- Long reads
  - Reduce finishing and assembly problems
  - Raise machine costs, lower all other costs
- Objective finishing criteria based on error probabilities

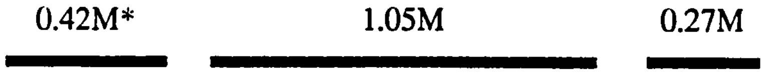
UNIV. OF WASHINGTON GENOME CENTER
MAMMALIAN SEQUENCING PROJECTS

- Human chromosome 7
- Human HLA Class I
- Mouse T-cell receptor alpha

Human chromosome 7



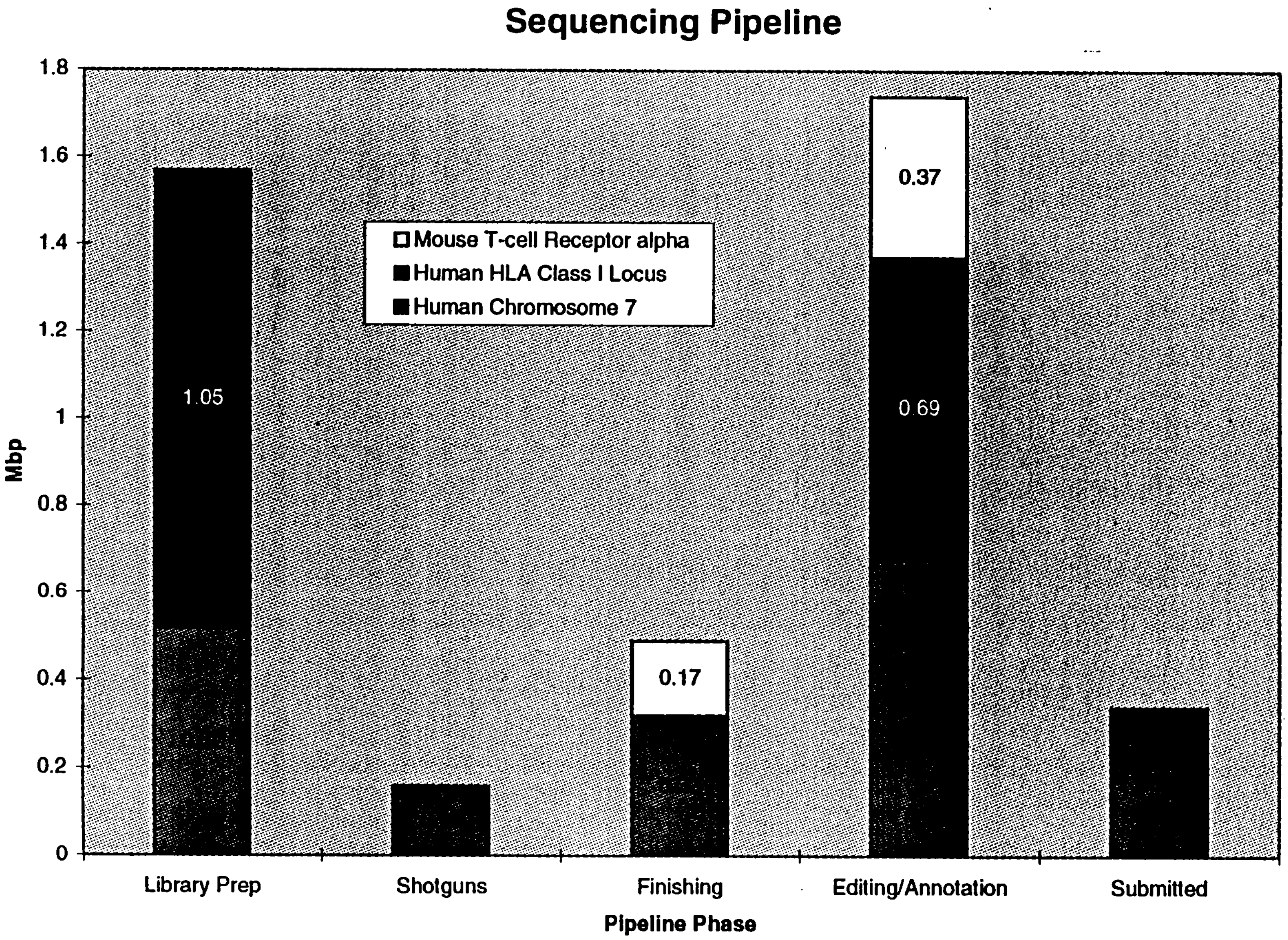
Human HLA class-I locus



Mouse T-cell receptor α

This 1M region is covered by 75 BACs and is being MCD mapped by a combination of BAC-to-cosmid subcloning and direct BAC fingerprints. The details are in another figure.

Color code: ■ mostly sequenced, ■ being sequenced


Sequencing Pipeline
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Mbp
□ Mouse T-cell Receptor alpha
■ Human HLA Class I Locus
■ Human Chromosome 7
1.05
0.17
0.37
0.69
Library Prep
Shotguns
Finishing
Editing/Annotation
Submitted
Pipeline Phase

INTERNAL ACCURACY ASSESSMENT

- MCD mapping.

  Test: MCD maps are compared to sequence-predicted maps.

  Results:

  - No mapping errors thus far in HLA and Chr. 7 regions (1.2Mb finished sequence).

- Sequencing.

  Test: All cosmids are independently finished, and sequences of overlapping same-haplotype cosmids are compared.

  Results:

  - Chr. 7:
    0 discrepancies in 2 X 38802 bp
  - HLA:
    2 discrepancies in 2 X 43084 bp
    - 1 mismatch (phrap error – incorrect read selected)
    - 1 apparent cosmid mutation (12 bp insertion/deletio in repeat region)

MCD MAPPING

- Start with large clones (YACs or BACs) from region of interest; 2X depth
- Subclone into cosmids; 20-30X depth
- Restriction digests with three enzymes
- Construct map of restriction sites & clone ends

sWSS106
Segmap V. 3.40 Data File Date: Wed Oct 30 19:54:22 1996
Chromosome 7 q31.1 q34

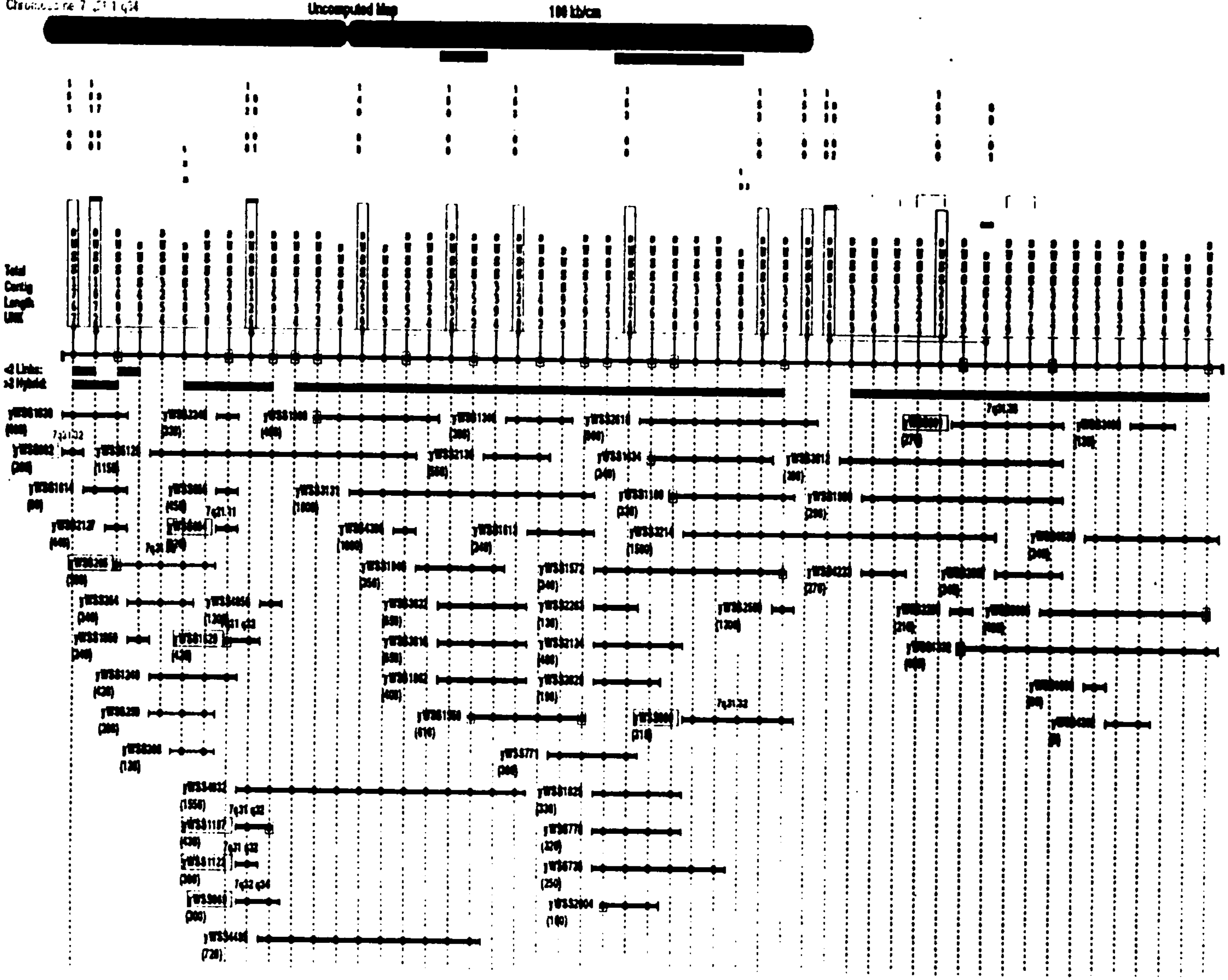


Table 1. Summary of YAC→cosmid MCD maps for portions of human chromosome 7.

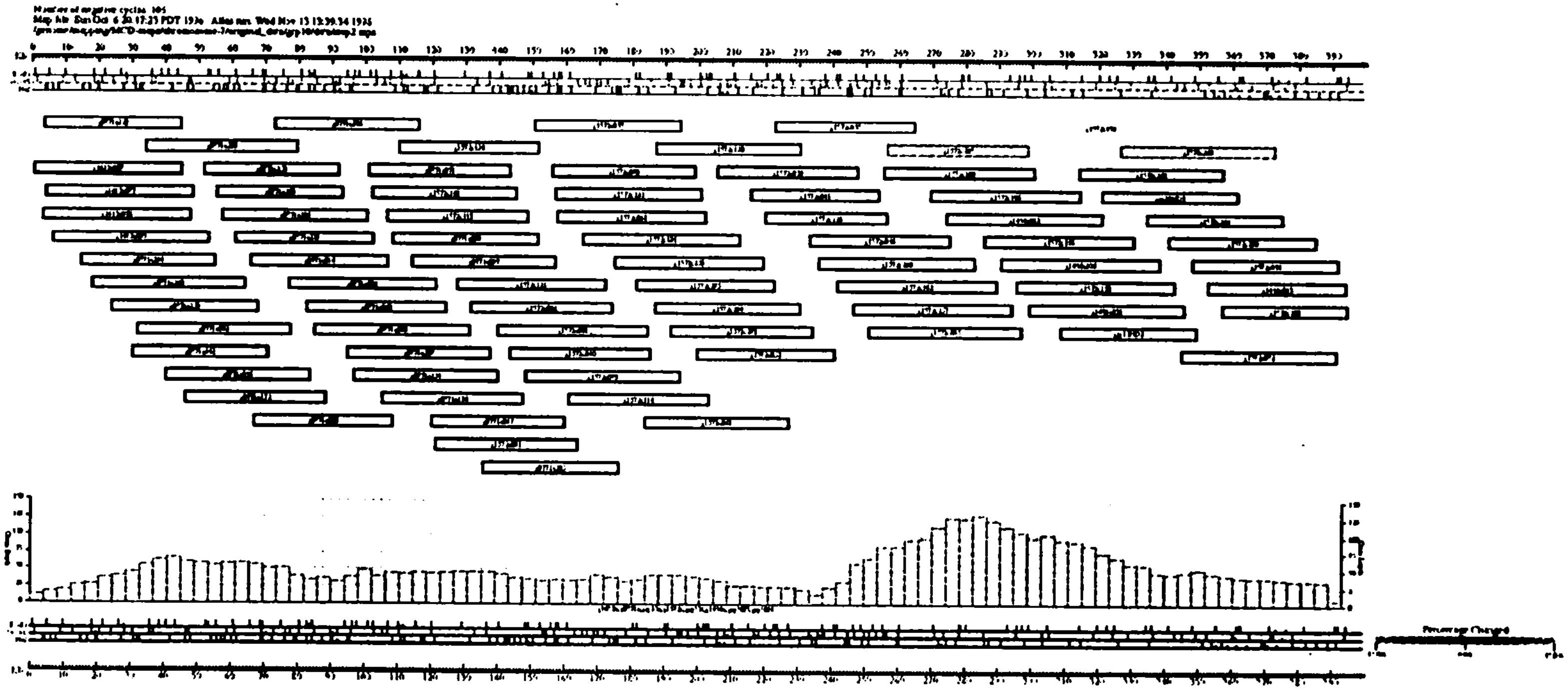
Chr-7 YACs	Coverage[a]	N_f[b] (EcoRI)	N_f[b] (HindIII)	N_f[b] (NsiI)	Co-ligations[c]	Map Size[d] (Kbp)
yWSS771	30.3	9.8 / 1.2	8.4 / 1.2	11.4 / 1.2	2.8%	44+170
yWSS1346	29.2	10.5 / 1.2	12.4 / 1.3	10.0 / 1.3	3.0%	281
yWSS1434	20.5	7.4 / 1.3	6.8 / 1.4	7.4 / 1.6	7.8%	156
yWSS1564	16.7	9.2 / 1.3	10.4 / 1.5	9.8 / 1.3	7.9%	640
yWSS1572	31.5	8.0 / 1.2	9.1 / 1.2	9.0 / 1.3	4.5%	292
yWSS1613	26.3	10.6 / 1.2	10.6 / 1.1	11.5 / 1.3	3.5%	136+56
yWSS1862	23.4	8.4 / 1.2	11.0 / 1.2	11.6 / 1.3	3.4%	261
yWSS1980	20.7	8.3 / 1.1	8.5 / 1.1	10.8 / 1.1	5.7%	278

[a] Coverage is calculated assuming a 40 Kbp insert size. Clones left out of the map because they could not be uniquely placed are included in the coverage calculation, while co-ligations and yeast impurities are not. When there are two contigs, we simply add their sizes to compute the coverage.

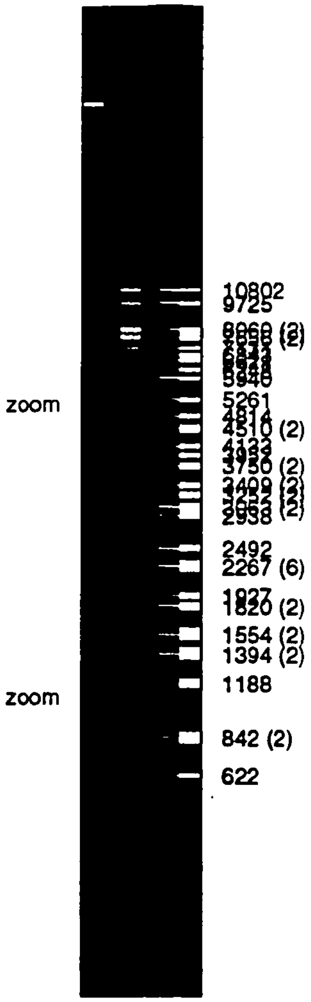
[b] N_f refers to the average number of fragments observed in a clone, which is the first number given in each row. The second number indicates the average number of fragments per fragment group, an indication of how well ordered the restriction fragments are in the maps. Contigs smaller than 100 Kbp are not included when summarizing fragments per fragment group.

[c] Co-ligations are cosmids that contain both a human insert from the targeted region and an unrelated piece of DNA that is inserted between the end of the human insert and the cosmid vector.

[d] Map sizes are based on the sum of the restriction-fragment sizes. The gap in the overlap region between YACs yWSS771 and yWSS1613 has not yet been closed. These maps agree perfectly with each other on either side of the gap, and both maps stop at exactly the same places.

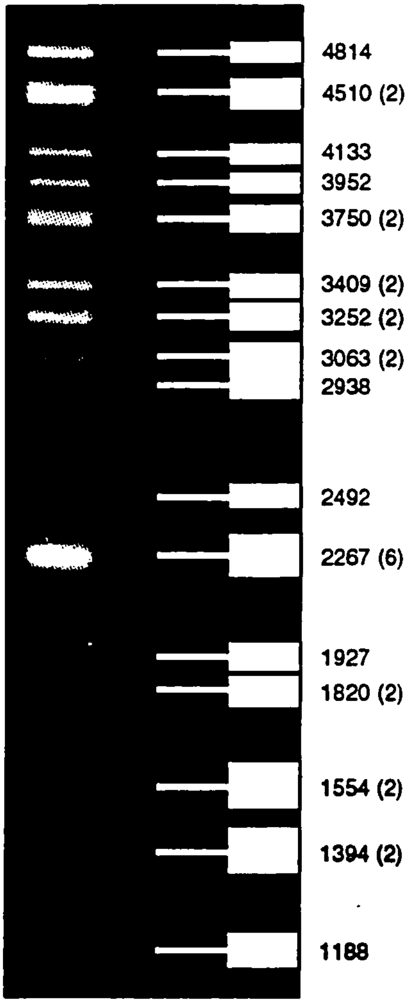


■ Library Prep ■ Shotguns □ Finishing ▨ Editing/Annotation ■ Submitted lane number 28



ma24 (HindIII)

on file "ma13"



zoom subregion

```
-------- EcoRI --------     ------- HindIII -------     -------- NsiI ---------
  MCD map         Clone       MCD map         Clone       MCD map         Clone

    :                             :                             :
----------                      691.28                    ----------
   2084.47                    ----------                     4799.00
----------                       4268.57                  ----------
   1122.77                    ----------                     1561.94
----------                       1104.83                     2559.12
   5079.10                    ----------                  ----------
----------                       1800.14                     9148.59
   1123.18                    ----------                  ----------    ----------
----------                       1973.64                     5048.52       5052.00
   1273.74                    ----------                      ???        * 5709.00
----------                       1858.81                  ----------
   9915.76                    ----------                     1575.81       1586.00
----------    ----------         3876.31                  ----------
   3465.80       3462.00      ----------                     3378.90       3378.00
    ???        *13976.00          974.84                  ----------
----------                    ----------                   * 4335.94        ???
   1673.51       1676.00         2944.00                  ----------
----------                    ----------                     6350.05       6343.00
   3330.49       3327.00         5435.75                  ----------
----------                    ----------    ----------       2141.53       2146.00
   1221.67       1223.00         4864.69       4860.00    ----------
----------                        ???        * 8374.00       6769.09       6762.00
 *12709.65        ???         ----------                  ----------
----------                       1550.16       1550.00        630.39        629.00
   9836.67       9778.00      ----------                  ----------    ----------
----------                        768.44        768.00      10373.77
   1049.14       1052.00      ----------                  ----------
----------    ----------         1111.40       1111.00       1582.33
   4244.29                    ----------                  ----------
----------                     * 6975.33        ???         14942.58
   3008.12                    ----------                  ----------
----------                       2127.50       2130.00       1222.25
   7014.18                    ----------                  ----------
----------                       2769.91       2770.00        970.00
   3112.29                    ----------                     4153.25
----------                       1789.55       1791.00    ----------
   5941.43                       1355.84       1353.00       2833.33
----------                    ----------                  ----------
   2019.67                       1553.21       1550.00       6344.00
   8330.00                    ----------                  ----------
----------                       2300.24       2301.00        961.00
   2650.00                    ----------                     3832.00
----------                       7324.61       7304.00    ----------
   3514.00                    ----------    ----------       1364.67
----------                       7077.58                     1755.33
   2361.40                    ----------                     4019.83
----------                       8837.62                     5315.67
    842.83                    ----------                  ----------
   1113.00                       1695.92                     7826.00
   4335.00                       3706.42                      797.40
----------                    ----------                     1693.60
    :                             :                             :
```

A transposon-insertion detected on chromosome-7 yWSS1346. Every enzyme domain in the aberrant clone has one extraneous fragment that cannot be matched to the MCD map. However, if something like 1400-bp is subtracted from each of these 3 extraneous fragments, the clone can be mapped in.

OBJECTIVE PROCEDURE TO ACHIEVE DEFINED ERROR RATE

- Following shotgun assembly, estimate error probability at each consensus base position; compute expected number of errors for entire cosmid or BAC.
- Finishing: collect enough additional data, or edit, in regions of highest error probability ("gaps") to force expected number of errors below 1 per 10 kb.
- Periodically, for selected cosmids, test agreement between expected number of errors and actual number of errors (relative to "gold standard").

- Monitor raw data quality using per read distribution of error probabilities.
- Explore optimal (least expensive) shotgun / finishing tradeoff yielding target error rate.

CURRENT TECHNOLOGY DEVELOPMENT

- MCD mapping:
  - BAC restriction digests
  - Automated clone anomaly detection
- Sequence assembly and editing:
  - Phrap: Improved error probabilities, resolution of large exact repeats, use of map information, reassembly directives
  - Phred: Lane processing, compression resolution
  - Consed: Tags, custom navigation

UNIVERSITY OF WASHINGTON GENOME CENTER

Maynard V. Olson, Director

MCD Mapping

Jun Yu, Leader	(ft)
Ying Ge	(ft)
Zahra Magness	(ft)
Ruolan Qiu	(ft)
Channakhone Saenphimmachak	(ft)

Sequencing

Shawn Iadonato, Leader	(ft)
Cindy Desmarais	(ft)
Thomas Gilbert	(ft)
Kim Harris	(ft)
Lloyd Lytle	(ft)
Oanh Nguyen	(pt)
Quynh Pham	(ft)
Karen Phelps	(pt)
Steven Swartzell	(ft)

Software Development

Brent Ewing	(ft)
David Gordon	(ft)
Arian Smit	(ft)
Ed Thayer	(ft)
Colin Wilson	(ft)

Production Informatics/Map Finishing

Gane Wong (ft) and Charles Magness (pt)

Kerry Bubb	(ft)
Jina Chang	(pt)

UNIVERSITY OF WASHINGTON GENOME CENTER

Maynard V. Olson, Director

Collaborators:

NHGRI:	Eric Green
Fred Hutch Cancer Research Center:	Dan Geraghty, Thierry Guillaudeux, Marta Janer
University of Washington - Molecular Biotechnology:	Leroy Hood, Inyoul Lee, Lee Rowen
University of Washington - Computer Science:	Richard Karp
Washington University - Computer Science:	Will Gillett, Liz Hanks

* EVERYONE – PLEASE EDIT THIS TABLE FOR YOUR CENTER, AND RETURN TO FRANCIS COLLINS BY FRI. AFTERNOON COFFEE/TEA.

Human Sequence Production

Investigator	Cumulative Finished Sequence	Predicted 3/1/97 – 2/28/98	 3/1/98 – 2/28/99
Sulston	14.6	35	80
Waterston	1.9	12	24
Lander/ Hudson/Hawkins	2.1	20	80
Adams	2.6	11	
Gibbs	3	15	100
Cox	0.1	5	
Lehrach	0.24	1	2
Weissenbach			
Mattick			
Rosenthal	1.5	6	12
Bloecker*		1	2
Green/Olson	0.34		
Chen	2.4	3	
Sakaki*	2.7	3.4	}30
Other Japanese efforts	10^{-4} acc	12	
Evans	1.6 (in Genbank)	5	50 ←
Palazzolo	4		
Roe	3.8		

[PLEASE FILL IN ANY GAPS! — AND DON'T BE OFFENDED AT ERRORS!]

* Not present, but reported on by others

* EVERYONE – PLEASE EDIT THIS TABLE FOR YOUR CENTER, AND RETURN TO FRANCIS COLLINS BY FRI. AFTERNOON COFFEE/TEA.

Human Sequence Production

Investigator	Cumulative Finished Sequence	Predicted	
		3/1/97 – 2/28/98	3/1/98 – 2/28/99
Sulston	14.6	35	80 ← O.K.✓
Waterston	1.9	12	24
Lander/ Hudson/Hawkins	2.1	10	80
Adams	2.6	11	
Gibbs	3	15	100
Cox	0.1	5	
Lehrach	0.24	1	2
Weissenbach			
Mattick			
Rosenthal	1.5	6	12
Blocker*		1	2
Green/Olson	0.34		
Chen	2.4	3	
Sakaki*	2.7	3.4	} 30
Other Japanese efforts		12	
Evans	1.6		
Palazzolo	4		
Roe	3.8		

[PLEASE FILL IN ANY GAPS! – AND DON'T BE OFFENDED AT ERRORS!]

* Not present, but reported on by others

✗ EVERYONE – PLEASE EDIT THIS TABLE FOR YOUR CENTER, AND RETURN TO FRANCIS COLLINS BY FRI. AFTERNOON COFFEE/TEA.

Human Sequence Production

	Cumulative	Predicted	
Investigator	Finished Sequence	3/1/97 – 2/28/98	3/1/98 – 2/28/99
Sulston	14.6	35	80
Waterston	1.9 + 2.9 ~~submitted~~ in GenBANK	~~12~~ 24	24+ ←
Lander/Hudson/Hawkins	2.1	10	80
Adams	2.6	11	
Gibbs	3	15	100
Cox	0.1	5	-
Lehrach	0.24	1	2
Weissenbach			
Mattick			
Rosenthal	1.5	6	12
Blocker*		1	2
Green/Olson	0.34		
Chen	2.4	3	
Sakaki*	2.7	3.4	} 30
Other Japanese efforts		12	
Evans	1.6		
Palazzolo	4		
Roe	3.8		

[PLEASE FILL IN ANY GAPS! – AND DON'T BE OFFENDED AT ERRORS!]

* Not present, but reported on by others

EVERYONE – PLEASE EDIT THIS TABLE FOR YOUR CENTER, AND RETURN TO FRANCIS COLLINS BY FRI. AFTERNOON COFFEE/TEA.

Human Sequence Production

	Cumulative	Predicted	
Investigator	Finished Sequence	3/1/97 – 2/28/98	3/1/98 – 2/28/99
Sulston	14.6	35	80
Waterston	1.9	12	24
Lander/Hudson/Hawkins	2.1	20	80
Adams	2.6	11	
Gibbs	3	15	100
Cox	0.1	5	
Lehrach	0.24	1	2
Weissenbach			
Mattick			
Rosenthal	1.5	6	12
Bloecker*		1	2
Green/Olson	0.34		
Chen	2.4	3	
Sakaki*	2.7	3.4	} 30
Other Japanese efforts		12	
Evans	1.6		
Palazzolo	4		
Roe	3.8		

[PLEASE FILL IN ANY GAPS! – AND DON'T BE OFFENDED AT ERRORS!]

* Not present, but reported on by others

* EVERYONE - PLEASE EDIT THIS TABLE FOR YOUR CENTER, AND RETURN TO FRANCIS COLLINS BY FRI. AFTERNOON COFFEE/TEA.

Human Sequence Production

Investigator	Cumulative Finished Sequence	Predicted 3/1/97 - 2/28/98	3/1/98 - 2/28/99	
Sulston	14.6	35	80	
Waterston	1.9	12	24	
Lander/ Hudson/Hawkins	2.1	20	80	
Adams	2.6	11		
Gibbs	3	15	100	
Cox	0.1	5		
Lehrach	✓ 0.24	1	2	CORRECT
Weissenbach				
Mattick				
Rosenthal	1.5	6	12	
Blöcker*		1	2	
Green/Olson	0.34			
Chen	2.4	3		
Sakaki*	2.7	3.4	} 30	
Other Japanese efforts		12		
Evans	1.6			
Palazzolo	4			
Roe	3.8			

[PLEASE FILL IN ANY GAPS! - AND DON'T BE OFFENDED AT ERRORS!]

* Not present, but reported on by others

EVERYONE – PLEASE EDIT THIS TABLE FOR YOUR CENTER, AND RETURN TO FRANCIS COLLINS BY FRI. AFTERNOON COFFEE/TEA.

Human Sequence Production

Investigator	Cumulative Finished Sequence	Predicted 3/1/97 – 2/25/98	3/1/98 – 2/28/99
Sulston	14.6	35	80
Waterston	1.9	12	24
Lander/Hudson/Hawkins	2.1	20	80
Adams	2.6	11	
Gibbs	3	15	100
Cox	0.1	5	
Lehrach	0.24	1	2
Weissenbach			
Mattick			
Rosenthal ✓	1.5	6	12
Bloecker*		1	2
Green/Olson	0.34		
Chen	2.4	3	
Sakaki*	2.7	3.4	} 30
Other Japanese efforts		12	
Evans	1.6		
Palazzolo	4		
Roe	3.8		

[PLEASE FILL IN ANY GAPS! — AND DON'T BE OFFENDED AT ERRORS!]

* Not present, but reported on by others

EVERYONE – PLEASE EDIT THIS TABLE FOR YOUR CENTER, AND RETURN TO FRANCIS COLLINS BY FRI. AFTERNOON COFFEE/TEA.

Human Sequence Production

Investigator	Cumulative Finished Sequence	Predicted 3/1/97 – 2/28/98	3/1/98 – 2/28/99
Sulston	14.6	35	80
Waterston	1.9	12	24
Lander/Hudson/Hawkins	2.1	20	80
Adams	2.6	11	
Gibbs	3	15	100
✓ Cox	0.1 = .3	5	
Lehrach	0.24	1	2
Weissenbach			
✓ Mattick	0	0	? (depends on AUST'n funding)
Rosenthal	1.5	6	12
Bloecker*		1	2
Green/Olson	0.34		
Chen	2.4	3	–
Sakaki*	2.7	3.4	} 30
Other Japanese efforts		12	
Evans	1.6		–
Palazzolo	4		
Roe	3.8		

[PLEASE FILL IN ANY GAPS! – AND DON'T BE OFFENDED AT ERRORS!]

* Not present, but reported on by others

EVERYONE - PLEASE EDIT THIS TABLE FOR YOUR CENTER, AND RETURN TO FRANCIS COLLINS BY FRI. AFTERNOON COFFEE/TEA.

Human Sequence Production

	Cumulative	Predicted	
Investigator	Finished Sequence	3/1/97 - 2/28/98	3/1/98 - 2/28/99
Sulston	14.6	35	80
Waterston	1.9	12	24
Lander/Hudson/Hawkins	2.1	10	80
Adams	2.6	11	
Gibbs	3	15	100
Cox	0.1	5	
Lehrach	0.24	1	2
Weissenbach			
Mattick			
Rosenthal	1.5	6	12
Blocker*		1	2
✓ Green/Olson	.59Mb ~~0.34~~	6 Mb (?)	
✓ Chen	2.4	3.5	6.0 /
Sakaki*	2.7	3.4	} 30
Other Japanese efforts		12	
Evans	1.6		
Palazzolo	4		
✓ Roe	3.8	5to6 MB ↑	(12 mB)

[PLEASE FILL IN ANY GAPS! — AND DON'T BE OFFENDED AT ERRORS!]

* Not present, but reported on by others

* EVERYONE – PLEASE EDIT THIS TABLE FOR YOUR CENTER, AND RETURN TO FRANCIS COLLINS BY FRI. AFTERNOON COFFEE/TEA.

Human Sequence Production

Investigator		Cumulative Finished Sequence	Predicted 3/1/97 – 2/28/98	Predicted 3/1/98 – 2/28/99	
Sulston		14.6	35	80	
Waterston		1.9	12	24	
Lander/ Hudson/Hawkins		2.1	20	80	
Adams	TIGR	2.7	11	14x	50/yr
Gibbs		3	15	~~100~~	
Cox		0.1	5		
Lehrach		0.24	1	2	
Weissenbach					
Mattick					
Rosenthal		1.5	6	12	
Bloecker*			1	2	
Green/Olson		0.34			
Chen		2.4	3.5	6	
Sakaki*		2.7	3.4	}30	
Other Japanese efforts			12		
Evans		1.6			
Palazzolo		4			
Roe		3.8			

Current funds NIH

proposed Future

[PLEASE FILL IN ANY GAPS! – AND DON'T BE OFFENDED AT ERRORS!]

* Not present, but reported on by others

* EVERYONE — PLEASE EDIT THIS TABLE FOR YOUR CENTER, AND RETURN TO FRANCIS COLLINS BY FRI. AFTERNOON COFFEE/TEA.

Human Sequence Production

	Cumulative	Predicted	
Investigator	Finished Sequence	3/1/97 – 2/28/98	3/1/98 – 2/28/99
Sulston	14.6	35	80
Waterston	1.9	12	24
Lander/ Hudson/Hawkins	2.1	20	80
Adams	2.6	11	
Gibbs	3 < 3.0 MB in GenBank 'Finished' + 1.0 if all data are considered = 4.0 MB	~~15~~ 12 (our cycle is to April)	100
Cox	0.1	5	
Lehrach	0.24	1	2
Weissenbach			
Mattick			
Rosenthal	1.5	6	12
Bloecker*		1	2
Green/Olson	0.34		
Chen	2.4	3	
Sakaki*	2.7	3.4	} 30
Other Japanese efforts	–	12	
Evans	1.6		
Palazzolo	4		
Roe	3.8		

[PLEASE FILL IN ANY GAPS! — AND DON'T BE OFFENDED AT ERRORS!]

* Not present, but reported on by others

Richard Gibbs

P.T.O.

Current 3.0 MB in GenBank
total 4.0 MB if other projects
on the verge of completion are
included

Next year (April 1,'97 - April 1,'98) 15MB

Following Year (April 1, 98 — April 1,'99) 20 MB

If resources available — during 98-99 period
the scale up will be to 100 MB/year.

To: [redacted]
cc: [redacted]
From: [redacted]
Date: 02/24/97 07:46:26 AM
Subject: QC

Hi Eric,

Thanks for your very thoughtful note about quality control measures for genome sequencing. I agree that this is a critical issue, and that the round-robin exercise, while useful, is not the last word. I found your proposals very useful, and we had a good discussion at Council on this topic, though we did not reach any conclusions. I particularly like your idea about a method to determine nucleotide-level accuracy.

I am sure that this topic will get considerable air-time in Bermuda, and I am really sorry that you cannot be there. Depending on the outcome of that gathering, NHGRI may well wish to convene a working group on QC as a high priority, perhaps even in time to bring a proposal to May Council.

Thanks for writing -- as usual, your comments were thoughtful and right on the mark.

Have a great time in Indonesia with your family.

Best regards,
Francis

WHITEHEAD INSTITUTE

February 8, 1997

Dr. Francis S. Collins, Director,
National Center for Human Genome Research
National Institutes of Health
Building 38A, Room 605
9000 Rockville Pike
Bethesda, MD 20892

[redacted]

I am writing with the hope of stimulating renewed discussion by NHGRI staff and council about quality control (QC) in the human genome project.

It is well appreciated that the human genome project must attend to both quantity and quality.

In the short term, quantity is probably the greatest challenge—in that it is necessary to achieve an unprecedented scale-up in worldwide annual sequence output of mammalian genomic sequence from ~10 Mb in 1996 to the ~500 Mb by 2000 required to complete the 3 Gb human sequence by the stated goal of 2005.

Still, it is also critical to ensure quality. It is already time to begin developing QC programs appropriate for a high-throughput sequencing enviroment.

NHGRI has taken a first step by oganizing a round-robin cross-validation program, in which various centers will re-sequence one another's clones. This will certainly be a useful exercise and we all support it. However, it must be recognized that such a once-a-year round-robin does not constitute a QC program for the long term. Among other things:

- The testing is expensive and time-consuming. If done as a proper blind test, it will cost as least as much as the initial sequencing—perhaps $40,000 to re-sequence a 100,000 bp BAC.

- The testing is too sporadic to provide useful feedback, to ensure that the process has not drifted. Ideally, QC procedures should provide regular and rapid feedback.

- The testing does not even address the critical issue of whether the reported sequence actually matches the human genome, as opposed to the clone sequenced.

I think it would be a good time to launch a discussion about better, more efficient and more regular QC procedures. A good solution might be to constitute an NHGRI working group on QC to discuss the issue.

QC Issues

To illustrate the issues that might be considered by a working group, I list below some preliminary thoughts about QC. They are intended simply as examples. A working group would surely develop a more complete analysis.

First, it is useful to define precisely what needs to be quality-controlled. There would seem to be three critical issues:

(1) Nucleotide-level accuracy of reported sequence vs. clone. In a given stretch of reported nucleotides, what is the probability that a given base is incorrect? The target standard has been declared to be 10^{-4}, but we currently lack any good way to measure this rate.

Sequence assembly programs that report "quality scores" do not provide a meaningful solution to this problem. Although the quality scores have been correlated with accuracy in a few instances, there is no assurance that they correctly apply to later projects. Production processes typically change and drift—especially when they involve much human editting by new employees, new sequencer configurations, or new dye-terminators.

Rather, frequent experimental measurement of sequence accuracy is far preferable.

(2) Assembly-level accuracy of reported sequence vs. clone. Even if the local nucleotide-level accuracy is high, there may be problems of misassembly–especially owing to repeated sequences. It is important to ask whether the gross assembly of the sequence correctly reflects the clone.

(3) Fidelity of clone vs. human genome. Does the clone faithfully represent the human genome? Or, has it undergone deletions or rearrangements?

It is important to develop well-defined, efficient protocols for addressing each question--with the goal that these procedures be used by the groups producing the sequence (QC should primarily be a responsibility of each center) as well as by any independent assessor. Here are some possible examples:

(1) Nucleotide-level accuracy of reported sequence vs. clone. Measuring nucleotide accuracy is fundamentally (i) a local issue and (ii) a statistical issue.

Assumes no bias in M13 cloning

It can be approached by classical sampling. Specifically, one needs to "inspect" 30,000 bases to infer that the error rate is $<10^{-4}$. (Statistically, if no errors are found in 3N nucleotides, one can infer at the 95% confidence level that the error rate is ≤1/N. This follows from the Poisson distribution, since $e^{-3} = 0.05$.)

To test the accuracy of a BAC sequence, one could:

- prepare an M13 library consisting of short (e.g., 400 bp) inserts; and
- re-sequence random M13 clones by performing forward and reverse reads, using both dye-primers and dye-terminators. (In this fashion, one automatically has double-stranded, double-chemistry sequence for all 400 bp of the clone--without having to worry about assembly. Each clone is an independent test.)

By sequencing 90 random M13 clones from a 100 kb BAC, one would inspect 36,000 total nucleotides and ~30,000 *independent* nucleotides.

The test would be quite inexpensive: Such four-fold re-sequencing of 90 M13 clones should cost <$500 in reagents.

It might be appropriate for a large sequencing center to perform this test bi-weekly on a recently finished BAC, to test whether accuracy is being maintained. It would also be practical for an independent assessor to perform such a test monthly.

(2) Assembly-level accuracy of reported sequence vs. clone. It is broadly agreed that assembly accuracy is likely to be best verified by restriction digestion, which provides a 'global' view of a clone. However, many questions remain unanswered. Supposing that a reported sequence correctly predicts the restriction fragments (for a given set of enzymes and given fragment measurement system, with its inherent accuracy), what conclusion about accuracy can be made? Are there potential alternative assemblies that would also be consistent with the data? (The answer could involve using a computer algorithm to examine alternatives.) Accordingly, which enzymes and what measurement systems would suffice as a routine QC system?

(3) Fidelity of clone vs. human genome.

Testing genomic fidelity is a thorny problem. One approach is to check whether a clone's fingerprint agrees with the fingerprint of a few neighboring clones in a local map. However, this test is not typically done in a rigorous fashion: there is no clear requirement for the degree of coverage or for the criteria for agreement.

A more rigorous approach could involve using end-STSs (perhaps chosen after sequencing is complete) to select a dense collection of overlapping clones at both ends to build an appropriately dense map and then check the resulting fingerprints in a more systematic fashion. Would this be worth the trouble? Would this test be appropriately applied to every clone or only to a sample of clones (e.g., one in ten)? Are there more efficient methods for checking genomic fidelity?

In any case, these are merely initial thoughts. My main point is that it would be a good time to consider constituting a working group on QC, or otherwise revive discussion about this issue. I'd be glad to discuss this issue further with you, NHGRI staff or council.

With best regards,

Sincerely,

Eric S. Lander
Member, Whitehead Institute for Biomedical Research
Professor of Biology, MIT
Director, Whitehead/MIT Center for Genome Research

WHITEHEAD INSTITUTE

February 8, 1997

Dr. Francis S. Collins, Director,
National Center for Human Genome Research
National Institutes of Health
Building 38A, Room 605
9000 Rockville Pike
Bethesda, MD 20892

[redacted]

Dear Francis:

I am writing with the hope of stimulating renewed discussion by NHGRI staff and council about quality control (QC) in the human genome project.

It is well appreciated that the human genome project must attend to both quantity and quality.

In the short term, quantity is probably the greatest challenge—in that it is necessary to achieve an unprecedented scale-up in worldwide annual sequence output of mammalian genomic sequence from ~10 Mb in 1996 to the ~500 Mb by 2000 required to complete the 3 Gb human sequence by the stated goal of 2005.

Still, it is also critical to ensure quality. It is already time to begin developing QC programs appropriate for a high-throughput sequencing enviroment.

NHGRI has taken a first step by oganizing a round-robin cross-validation program, in which various centers will re-sequence one another's clones. This will certainly be a useful exercise and we all support it. However, it must be recognized that such a once-a-year round-robin does not constitute a QC program for the long term. Among other things:

- The testing is expensive and time-consuming. If done as a proper blind test, it will cost as least as much as the initial sequencing—perhaps $40,000 to re-sequence a 100,000 bp BAC.

- The testing is too sporadic to provide useful feedback, to ensure that the process has not drifted. Ideally, QC procedures should provide regular and rapid feedback.

- The testing does not even address the critical issue of whether the reported sequence actually matches the human genome, as opposed to the clone sequenced.

I think it would be a good time to launch a discussion about better, more efficient and more regular QC procedures. A good solution might be to constitute an NHGRI working group on QC to discuss the issue.

QC Issues

To illustrate the issues that might be considered by a working group, I list below some preliminary thoughts about QC. They are intended simply as examples. A working group would surely develop a more complete analysis.

First, it is useful to define precisely what needs to be quality-controlled. There would seem to be three critical issues:

(1) Nucleotide-level accuracy of reported sequence vs. clone. In a given stretch of reported nucleotides, what is the probability that a given base is incorrect? The target standard has been declared to be 10^{-4}, but we currently lack any good way to measure this rate.

Sequence assembly programs that report "quality scores" do not provide a meaningful solution to this problem. Although the quality scores have been correlated with accuracy in a few instances, there is no assurance that they correctly apply to later projects. Production processes typically change and drift—especially when they involve much human editting by new employees, new sequencer configurations, or new dye-terminators.

Rather, frequent experimental measurement of sequence accuracy is far preferable.

(2) Assembly-level accuracy of reported sequence vs. clone. Even if the local nucleotide-level accuracy is high, there may be problems of misassembly– especially owing to repeated sequences. It is important to ask whether the gross assembly of the sequence correctly reflects the clone.

(3) Fidelity of clone vs. human genome. Does the clone faithfully represent the human genome? Or, has it undergone deletions or rearrangements?

It is important to develop well-defined, efficient protocols for addressing each question--with the goal that these procedures be used by the groups producing the sequence (QC should primarily be a responsibility of each center) as well as by any independent assessor. Here are some possible examples:

(1) Nucleotide-level accuracy of reported sequence vs. clone. Measuring nucleotide accuracy is fundamentally (i) a local issue and (ii) a statistical issue.

It can be approached by classical sampling. Specifically, one needs to "inspect" 30,000 bases to infer that the error rate is $<10^{-4}$. (Statistically, if no errors are found in 3N nucleotides, one can infer at the 95% confidence level that the error rate is $\leq 1/N$. This follows from the Poisson distribution, since $e^{-3} = 0.05$.)

To test the accuracy of a BAC sequence, one could:

- prepare an M13 library consisting of short (e.g., 400 bp) inserts; and
- re-sequence random M13 clones by performing forward and reverse reads, using both dye-primers and dye-terminators. (In this fashion, one automatically has double-stranded, double-chemistry sequence for all 400 bp of the clone--without having to worry about assembly. Each clone is an independent test.)

By sequencing 90 random M13 clones from a 100 kb BAC, one would inspect 36,000 total nucleotides and ~30,000 *independent* nucleotides.

The test would be quite inexpensive: Such four-fold re-sequencing of 90 M13 clones should cost <$500 in reagents.

It might be appropriate for a large sequencing center to perform this test bi-weekly on a recently finished BAC, to test whether accuracy is being maintained. It would also be practical for an independent assessor to perform such a test monthly.

(2) Assembly-level accuracy of reported sequence vs. clone. It is broadly agreed that assembly accuracy is likely to be best verified by restriction digestion, which provides a 'global' view of a clone. However, many questions remain unanswered. Supposing that a reported sequence correctly predicts the restriction fragments (for a given set of enzymes and given fragment measurement system, with its inherent accuracy), what conclusion about accuracy can be made? Are there potential alternative assemblies that would also be consistent with the data? (The answer could involve using a computer algorithm to examine alternatives.) Accordingly, which enzymes and what measurement systems would suffice as a routine QC system?

(3) Fidelity of clone vs. human genome.

Testing genomic fidelity is a thorny problem. One approach is to check whether a clone's fingerprint agrees with the fingerprint of a few neighboring clones in a local map. However, this test is not typically done in a rigorous fashion: there is no clear requirement for the degree of coverage or for the criteria for agreement.

A more rigorous approach could involve using end-STSs (perhaps chosen after sequencing is complete) to select a dense collection of overlapping clones at both ends to build an appropriately dense map and then check the resulting fingerprints in a more systematic fashion. Would this be worth the trouble? Would this test be appropriately applied to every clone or only to a sample of clones (e.g., one in ten)? Are there more efficient methods for checking genomic fidelity?

In any case, these are merely initial thoughts. My main point is that it would be a good time to consider constituting a working group on QC, or otherwise revive discussion about this issue. I'd be glad to discuss this issue further with you, NHGRI staff or council.

With best regards,

Sincerely,

Eric S. Lander
Member, Whitehead Institute for Biomedical Research
Professor of Biology, MIT
Director, Whitehead/MIT Center for Genome Research

THE WELLCOME TRUST
MEMORANDUM

TO: All delegates FROM: Jilly Steward

DATE: 28th February 1997

Changes to delegate list at rear of programme and revised delegate list

Since the programme was printed and distributed there have been a few changes to the delegate list and a revised one is attached. However, there have been additional changes to that and we are delighted that the following delegates have been able to join us:

Dr Catherine Moody, Medical Research Council, London
Dr Graham Cameron, European Molecular Biology Laboratory, Cambridge
Dr Wilhelm Ansorge, European Molecular Biology Laboratory, Heidelberg

A final list will be distributed with the report from the meeting with the full addresses and contact numbers of the above, but it would also be helpful if you could let me know if you need any amendments made to your entry.

Jilly

THE WELLCOME TRUST

210 Euston Road
London NW1 2BE

[redacted]

ref:js/so'd/invite

16th January 1997

Dr Francis Collins
National Institutes of Health
National Centre for Human Genome Research
~~21~~Center Drive MSC 2152
Bethesda
MD 20892-2152
USA

Dear Dr Collins

The Wellcome Trust along with the National Institutes of Health and the Department of Energy are convening a second International Strategy meeting on Human Genome Sequencing to be held at the Hamilton Princess Hotel, Bermuda from 27th February - 2nd March 1997. The aim of the meeting as before is to facilitate the co-ordination of research groups funded for large-scale human genome sequencing.

I am writing on behalf of Dr Michael Morgan to invite you to attend this meeting, which will be limited to a maximum of 45 participants. The cost of your accommodation and subsistence for the duration of the meeting will be met by the meeting sponsors. However, it is anticipated that participants will be able to meet their own travel expenses.

Participants should aim to arrive on the Thursday evening and arrange their return flights on the Sunday morning.

I hope that you will be able to join us for this meeting which, I am sure, will continue the traditions established last year and provide international collaboration and co-ordination in this vital field of research.

Further details will be sent to you in the near future but I should be grateful if you could email me your reply by Monday 20th January 1997 and let me know at that stage whether your departure will be on Saturday evening or on Sunday so we can confirm accommodation requirements with the hotel on Tuesday 21st January. Please fax your completed registration form to me by no later than Monday 27th January 1997.

I look forward to hearing from you and to your participation in this meeting.

Yours sincerely

Jilly Steward
Meeting & Travel Manager

International Strategy Meeting on Human Genome Sequencing
The Hamilton Princess Hotel, Bermuda
27th February - 2nd March 1997

Title ______________________ ______________________

Surname ______________________ *Forename* ______________________

Academic Address ____

____

____

Tel No ______________________ *Fax No* ______________________

email address ____

I will be able to attend the *second International Strategy Meeting on Human Genome* **from 27th February - 2nd March 1997** **YES / NO**

Any dietary requirements ____

Accommodation dates required ______________________________ *inclusive*

AV Requirements:

Single Projection □ ***OHP*** □ ***Other*** ______________

I shall undertake my own travel arrangements and my itinerary is as follows:

Arrival Date

From ______________ Time ________ Flight No ______________

Departure Date

From ______________ Time ________ Flight No ______________

Any other comments ____

____

____

Please return this form immediately by fax to:
Mrs Jilly Steward, The Wellcome Trust,
183 Euston Road, London, NW1 2BE
Fax: 0171 611 8237

To: [redacted] @ INTERNET,

cc: [redacted]

From: [redacted]gov ("Guyer, Mark") @ INTERNET

Date: 01/31/97 08:55:00 AM

Subject: Agenda for Bermuda

→ Mtg file
Bermuda
2/27–3/1

Michael Morgan gave me John Sulston's latest draft agenda for the Bermuda meeting, which I am including in this message, along with our reactions and proposed revision for your information/comments.

Sulston draft:

I. Progress, strategies, developments

a. Reports from each sequencing group, as last year -- speakers should be asked to address the effectiveness of their strategies for constructing sequence-ready maps and cost estimates.

b. New libraries -- how will they be incorporated into production lines?

c. Brief look at technology

II. Allocation of regions/etiquette for sharing

Territorial claims -- how much sequence is appropriate to stake out? What will happen when more than one group is interested in sequencing a particular region?

III. Release criteria and timing

a. Sequence quality standards

b. Data release -- how have different groups implemented the conclusions from last year's meeting? should these conclusions be revisited? how can the usefulness or lack of usefulness of very rapid release be assessed?

IV. Interpretation

a. Annotation standards -- what level of annotation is appropriate for large-scale genomic sequencing laboratories?

b. EST sequencing/full-length cDNA sequencing -- the role of such sequences in assembling genomic sequence?

c. Mouse sequencing -- who is interested? how will the mouse get done?

****

Our (MG, JP, JS) proposed comments and revised agenda

John: Michael gave me a copy of what I assume is the most recent draft of an agenda for the Bermuda meeting. I've discussed this with some of the people around here and have the following comments:

A. In general, we like the agenda and think that it should be a pretty meaty, probably rather intense discussion. There are a lot of issues that are much more well-defined than they were last year, and this is a good opportunity to address them seriously.

B. We like the format of including specific questions/issues to be answered/addressed, and hope that these will be included in the actual agenda that is distributed. If so, however, we believe that the questions should be more complete than they are at the moment, to avoid any possibility of misunderstanding. For example, the explication of the Reports from each sequencing group currently mentions construction of sequence-ready maps and cost estimates, but doesn't specifically mention finished sequence (admittedly pretty obvious), or progress on the finishing problem or other bottlenecks.

B. As is probably becoming clear, we are very focused right now on issues of data standards, so we have suggested that the title of section III be modified to emphasize that and that this issue be addressed before the allocation question, to emphasize its importance. And we have offered some suggestions for questions to be addressed.

C. We're not sure that it will be particularly useful for each group to tell us what its sequencing costs are because it's not likely that we will hear much about how those costs were arrived at, it's still reasonably early in the scale-up game and current costs are not that related to potential final costs, and there will be a lot of posturing. However, it might well be useful to have a general discussion about establishing a uniform way of determining the cost of production sequencing.

D. When Michael was here earlier this week, we discussed having the regularization of this meeting be explicitly addressed. If it is going to be useful to have a meeting like this, at least for the next couple of years, I think it would be very helpful for people to be able to plan on it.

E. We (at NHGRI) have a certain degree of concern about the concepts of level 1, level 2, level 3 sequence that have become so common in the community of late. We think that this is not a good development because it distracts focus from the real problem of finishing sequence to a certain standard. We have no intention of giving "credit" for (or even looking at) anything less than finished sequence as we make our decisions over the next couple of years. Is this an issue that should be explicitly taken on (we have no doubt that as soon as reports start being presented, we will hear about level 1, 2, etc, which is why we propose to remind presenters to only report on finished sequence).

So, with those comments, here are our suggestions for the next draft of theagenda:

I. Progress, strategies and developments

a. Reports from each sequencing group: Speakers should be asked to address the effectiveness of their strategies, being sure to address the construction of sequence-ready maps, output of finished sequence (finished meaning of the quality that the group is willing to submit to a database as finished), current bottlenecks and plans for addressing them, in particular what progress has been made in addressing the finishing problem.

b. New libraries -- how will they be incorporated into production lines?

c. Brief look at technology

II. Sequencing quality and release

a. Sequence quality standards: Discussion of the NHGRI sequecing standard, which addresses base accuracy and coverage. What is an acceptable standard for number of gaps per Mb? Are current approaches for measuring accuracy adequate? What data would be required in considering revision of the standards?

b. Data release: How have different groups implemented the conclusions from last year's meeting? Should these conclusions be revisited? How can the usefulness or lack of usefulness of very rapid release be assessed?

III. Allocation of regions/etiquette for sharing

Territorial claims -- How much sequence is appropriate to stake out? What will happen when more than one group is interested in sequencing a particular region? What will happen when a group does not meet its commitment to complete a particular region?

IV. Cost of sequencing

Can a standard/uniform way of measuring the cost of producing sequence be agreed upon?

V. Interpretation

a. Annotation standards: What level of annotation is appropriate for large-scale genomic sequencing laboratories?

b. EST sequencing/full-length cDNA sequencing: What role can such sequences play in assembling genomic sequence?

c. Mouse sequencing: How much mouse sequence is needed to assess the usefulness of such data in (a) assembling human sequence, (b) interpreting human sequence? Who is interested? How will the mouse sequencegetdone?

VI. Future meetings. Should this meeting be held next year? Beyond next year?

THE WELLCOME TRUST

183 Euston Road
London NW1 2BE

[redacted]

Mtg file 2/28-3/1

FACSIMILIE
TRANSMISSION

TO: Francis Collins

FROM: Jilly Steward

FAX No [redacted]

DATE: 12th February 1997

No of Pages including front sheet 2
Message:

Dear Francis,

Michael and I are in the process of finalising the programme and I am writing to ask whether you would be happy to Chair Session II (attached). I should be most grateful if you could email or fax me by return as the programme is due to be going to publishing tomorrow afternoon

Final letters and programme will be couriered to all delegates on Monday.

I look forward to meeting you again on the 27th

With kind regards

Yours sincerely

Jilly Steward

Dear Jilly ✓ faxed 2-14-97

I am happy to do this.

Regards

Fran

11.00 **Morning Coffee** -***Lobby area***

1130 Yoshiyuki Sakaki
1140 Asao Fujiyama
1150 Pieter de Jong
1200 Glen Evans
1210 Michael Palazzolo
1220 Bruce Roe

12.30 **Luncheon** - ***Tiara Room, Mezzanine Floor***

1400 ***Session II SEQUENCING QUALITY AND COSTS***

CHAIRMAN: FRANCIS COLLINS

Round Table Discussion

Aims of this session are to discuss:
a) Sequence quality standards:

Should a universal standard addressing base accuracy, coverage and number of gaps per Mb be adopted?

Can a standard/uniform way of measuring the cost of producing sequence be agreed upon?

1600 ***SESSION II continues: DATA RELEASE***

CHAIRMAN: FRANCIS COLLINS

Round Table Discussion

How have different groups implemented the conclusions from last's years meeting?
Should these conclusions be revisited?
How can the usefulness of very rapid release be assessed?

1800 Close of Session

1930 **Pre Dinner Drinks** - ***Harbourfront Restaurant, Front Street***

2000 **Conference Dinner** - ***Harbourfront Restaurant, Front Street***

```
                    *********************
                    ***   TX REPORT   ***
                    *********************

TRANSMISSION OK

TX/RX NO                 4331
CONNECTION TEL               9011441716118545
SUBADDRESS
CONNECTION ID            WELLCOME TRUST
ST. TIME                 02/13 09:26
USAGE T                  01'24
PGS.                       2
RESULT                   OK
```

FAX TRANSMISSION
National Human Genome Research Institute
31 Center Drive, Bldg 31, Room 4B09
Bethesda, Maryland 20892-2152
Phone: 301-496-0844
FAX: 301-402-0837

From: Susan Saylor
Secretary To Francis S. Collins, M.D., Ph.D.
Director, NHGRI

TO: Jilly Steward

FAX NUMB[redacted]

DATE: 2/13/97

Pages including cover sheet: 2

COMMENTS: Please see attached.

```
                    *********************
                    ***   TX REPORT   ***
                    *********************

TRANSMISSION OK

TX/RX NO                  4364
CONNECTION TEL                9011441716118545
SUBADDRESS
CONNECTION ID             WELLCOME TRUST
ST. TIME                  02/14 10:43
USAGE T                   00'57
PGS.                        1
RESULT                    OK
```

THE WELLCOME TRUST

Mtg file

183 Euston Road
London NW1 2BE
[redacted]

JS/SO'D/LET/898

17th February 1997

Dr Francis Collins
National Institutes of Health
National Human Genome Research Institute
31 Center Drive MSC 2152
Building 31, Room 4B09
Bethesda MD 20892-2151
USA

Dear Dr Collins

On behalf of the co-sponsors of this meeting, I am writing to you with final instructions and arrangements in respect of your attendance at the International Strategy Meeting on Human Genome Sequencing to be held at The Hamilton Princess Hotel, Bermuda from the 27th February - 2nd March 1997. Please find enclosed a copy of the final programme.

A brochure of the hotel has already been sent to you, but contact details at the hotel are as follows:

The Hamilton Princess Hotel
P.O.Box HM 837
Hamilton HM CX
Bermuda

[redacted]

Ground transportation to and from the airport has been arranged with Bee-Line Transportation who have been sent details of all flight arrivals and departures. Please let me know **immediately** if your flight schedules have changed from those stated on your registration form.

Accommodation has been arranged for you at The Hamilton Princess Hotel for the nights of 27th, 28th February and 1st March. The co-sponsors of the meeting have arranged for the programme to commence with a cocktail reception on Thursday evening at 20.30. Because of the US governmental restrictions, arrangements have been made with the hotel for you to settle your own full account prior to departure. Could we also ask that, for ease of handling at the hotel, **all your luggage is clearly marked with your name**.

Please note that it is a requirements of The Hamilton Princess that for all evening functions dress for delegates should be smart casual with gentlemen wearing a jacket. This dress code also applies to the restaurant for the conference dinner. The dinners, including the conference dinner, are an essential part of the programme and delegates are therefore expected to attend all of these events unless previously agreed with the organisers. If you have made alternative arrangements please let me know immediately so that numbers can be amended.

It is the policy of the Trust's funding that all delegates are expected to stay for the entirety of the meeting unless personally agreed with the Trust prior to the start of the meeting.

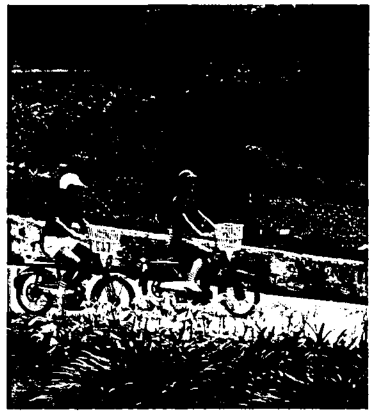
In the event of severe delays on your way to the meeting or any last minute changes to itinerary please contact me as soon as possible. I may be contacted at The Hamilton Princess from Tuesday evening, 25th February 1997 on telephone number 441-295-3000 or facsimile 441-295-1914.

I look forward to seeing you at The Hamilton Princess, and to an interesting and successful conference. In the meantime, should you require any further assistance, please do not hesitate to contact me.

With kind regards.

Yours sincerely

Jilly Steward
Meetings and Travel Manager



At The Princess you're within walking distance of Bermuda's most popular activities. Stroll along the streets of Hamilton, past Victorian buildings dipped in pastels and trimmed in wrought iron, or take a horse and buggy ride along the waterfront.

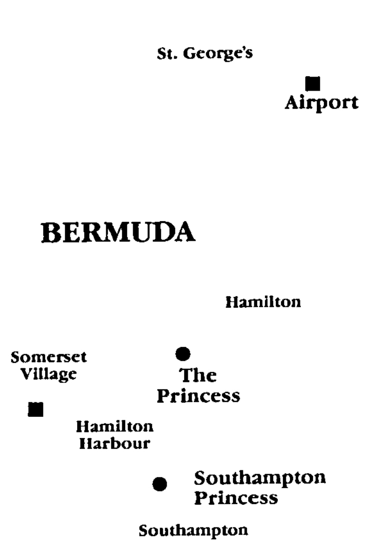
Shop in the boutiques along Front Street where you'll be tempted by sparkling crystal, British woolens and fine antiques.

For sunning and swimming, you'll find two beautiful pools at the hotel overlooking Hamilton Harbor. Or take the scenic ferry ride to the Southampton Princess Beach Club. Here you can enjoy a glorious private beach nestled in a sandy pink cove, enjoy the challenge of an 18-hole, executive golf course, and its breathtaking ocean views or play tennis on the 11 tournament-calibre courts, many lit for night play. And you can make arrangements for a wide variety of water sports from concessionaires at the Beach Club.

Whatever your pleasure, you can find it at The Princess. Where all the charm, warmth and unique hospitality of Bermuda are yours to enjoy.

Note: Some of the facilities may not be available during certain times of the year.

THE Princess



The Princess is in Pembroke Parish, overlooking beautiful Hamilton Harbor. Guests have the use of the restaurants, bars and sporting facilities of its nearby sister hotel, the Southampton Princess.

Princess® Hotels

Acapulco
ACAPULCO PRINCESS
PIERRE MARQUES
Acapulco, Mexico

Arizona
SCOTTSDALE PRINCESS
Scottsdale, Arizona

Bahamas
BAHAMAS PRINCESS RESORT AND CASINO
Freeport, Grand Bahama

Bermuda
SOUTHAMPTON PRINCESS
Southampton, Bermuda

THE PRINCESS
Hamilton, Bermuda

California
PALM SPRINGS MARQUIS HOTEL AND VILLAS
Palm Springs, California

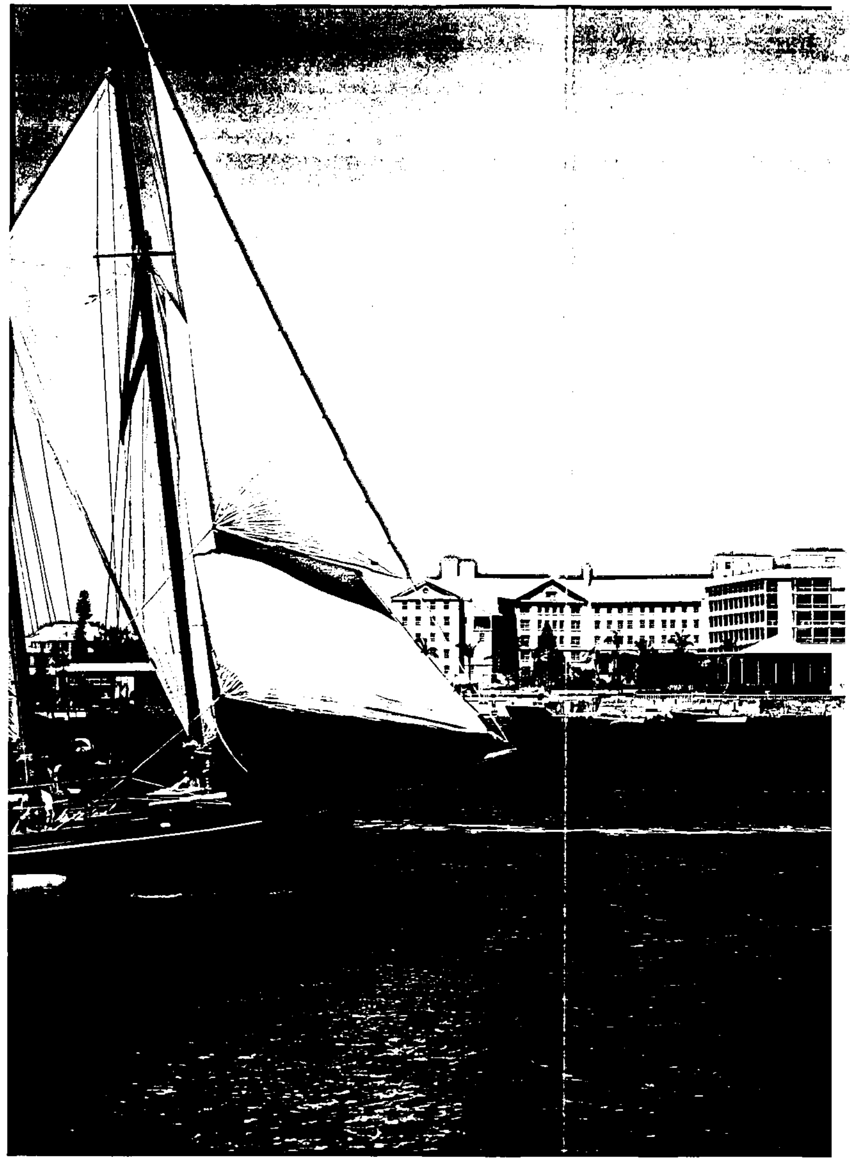
For information and reservations consult your Travel Agent.

® Represented by Princess Hotels International, Inc.

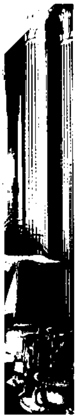
10/91R
Printed in the U.S.A.

THE Princess HAMILTON, BERMUDA

WHERE CHARM
IS A TRADITION



Picture the Bermuda of your dreams. Pastel pinks and clear blu dotted by yachts and sailboats. A place where traditions are still The Princess. Let her charm you with attentive service, magnificei guest rooms and suites. . . and with all that is Bermuda.



Bermuda at its

1884, through
sts, The Princess
hat Bermudians
lity and quality.
an unequalled

eerful, luxurious
a picturesque

e both excellent
beautiful setting
e dazzled by fine
d a spectacular
od Grill offers
den setting over-
or dinner. And
and sightseeing,
and Restaurant
ent for a casual
s simply a place

ent, take in the
where you'll be
sparkling dance
iew of Hamilton