STC Production on Human BACs

Archive Provided for Historical Purposes

◆ Home ◆ STC Project History ◆ 1995 Meeting ◆ Articles ◆ Contacts ◆ Links ◆

◆ HGP Sequences ◆ HGP Research ◆

Several types of DNA library resources were sponsored by the DOE before and during the Human Genome Program (HGP). These included both prokaryotic and eukaryotic vector systems, and clone libraries representing single chromosomes. Bacterial Artificial Chromosomes (BACs) became the most broadly used resource for several reasons. The large size was a good match for capabilities of high throughput sequencing centers. As contrasted to some earlier resources, chimerism (having gene segments from multiple chromosome sites combined in one clone) is substantially if not completely absent. With some interesting exceptions , the BACS are stable in their bacterial hosts. In support of the functional analysis of genes, the BACs are very useful for making transgenic animals with segments of human DNAs. A brief history of BAC development is available in a preface to a 2003 issue of Methods in Molecular Biology , wherein details of BAC related protocols reside.

One particular BAC project was crucial to the timely completion of human genome sequencing. (See history .) In a 1996 initiative, the DOE Office of Biological and Environmental Research sponsored the production of sequence tag connectors (STCs) for the BACs being used in human genome sequencing. (STCs are sequence reads at the ends of cloned DNA segments; they mark the boundaries of the cloned DNA.) This publicly available resource has served both the international public collaboration and Celera Genomics Inc . in the generation of the human genome sequence.

The BACs representing a genome can together serve as a scaffold on which much shorter DNA sequence assemblies can be located. The STCs have particular roles in identifying BACs:

- represent genomic regions still not sequenced;
- facilitate extension of a sequenced genomic region;
- span regions resistant to sequencing biochemistry; and
- spans can provide quality checks on sequence assemblies.

The data resources generated under the STC initiative thus speeded human genome sequencing worldwide. STC generation is now an integral component of most genome sequencing projects.

The target of one STC for every three kilobases of the human genome was achieved during the year 2000. The primary STC production sites were The Institute for Genomic Research (TIGR) and the former Sequencing Center of the Department of Molecular Biotechnology, University of Washington (UWMB), with capabilities

now at the Institute for Systems Biology with director Leroy Hood . Much more detail is in this history , or can be provided by Marvin.Stodolsky@science.doe.gov .

Complementary mapping resources

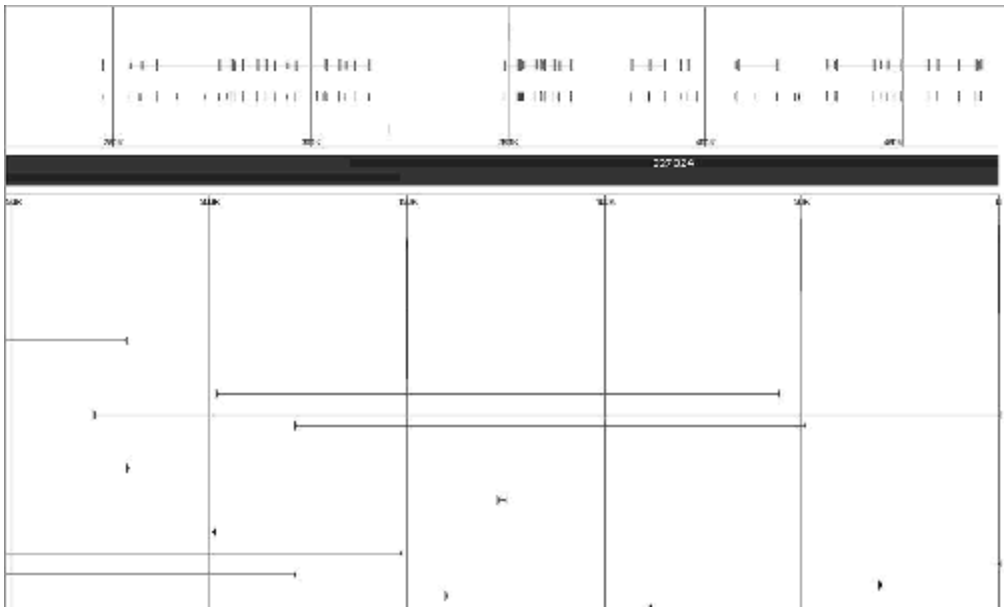
The STC data resource is complemented by several other types of mapping information. FISH (fluoresence in situ hybridization) mapping by J. Korenberg provided the first quality validation of BAC resources. The NCI extramural division, in conjunction with DOE, has awarded a consortium of three grantees to generate a Mapped Human Bacterial Artificial Chromosome (BAC) Clone Resource. Within the consortium, Barbara Trask has extending FISH analyses to newer BAC libraries. FISH mapped BACs with STCs have been added to Radiation Hybrid (RH) maps at Stanford U. RH STS markers and gene-based EST markers served to identify corresponding BACs by teams under Pieter de Jong and Ung-Jin Kim (while formerly at CalTech). Restriction fingerprints of BACs generated at UWMB and with NIH/NHGRI support at the Washington University Genome Sequencing Center aided contig construction and validation, in preparations for sequencing.

Enriching BACs for Sequencing with Sequence Tag Connectors

Several types of chromosome maps supporting biomedical research have been constructed during the Human Genome Program (HGP). In the preparation for high-throughput chromosomal sequencing, the most valuable are megabase-scale assemblies of overlapping DNA clones (contigs). Building long contigs, however, has proven a difficult task. The contig maps of chromosome 16 developed at LANL and chromosome 19 developed at LLNL were largely complete in 1995. Other chromosomes were much less ready for high-throughput sequencing.

Obtaining a substantially uniform representation of the genome using recombinant DNA clones was itself a problem, up to a few years ago. The problem was solved by the DOE-supported development of the more stable and larger recombinant BACs (bacterial artificial chromosomes) by the team of Melvin Simon at CalTech, with later process improvements by the team of Pieter de Jong . To support the contig building requirements of sequencers, **sequence tag connectors (STCs)** for the BACs are now bring generated.

STC ideas were first used by George Church and evolved in several smaller-scale , sequencing projects. Acquiring STC datasets for BACs representing a deep coverage of the whole human genome was advocated in 1995-6 (Venter, J.C., Smith, H.O., and Hood, L.E., Nature 381: 364-366) at sequencing workshops and a BAC resources meeting. A primary utility is illustrated below. The BACs whose STCs overlap an already sequenced region are candidate clones for extension of the sequence.



BAC End Sequencing Extends Contigs . Software tools are helping to position STCs. One tool, provided by the Genome Channel, allows investigators to view the contig positions of more than 15,000 BAC end sequences and their relationships to other clones and predicted genes and exons (gene-coding regions). In the figure, the black bar represents 250 kb of a much longer contig. Below the bar, the long horizontal lines denote BAC clones, of which the first, fifth, and sixth are candidates for extending the seed contig to the left. Above the bar, vertical tick marks indicate exons as predicted by GRAIL software. Exons connected by short horizontal lines represent putative gene models for the contig's forward DNA strand. (From *Human Genome News* v10n1-2)

Upon receipt of applications to a DOE 1996 HGP research solicitation, Ari Patrinos implemented a fast-track,

special review panel for those relevant to the contig problem. Overall the panel found the STC strategies meritorious but recommended pilot projects rather than an immediate genome scale implementation. Thus STC production protocols could be refined and economics clarified. Projects were initiated at six sites with a total of $5 million in September, 1996.

Several months later a workshop and review was held to assess progress. Each team had developed substantive and useful results. A major recommendation emerging from subsequent discussions was that DOE should maintain its support near the current level, about $5M/yr, but that a STC production phase should be implemented only at the sites achieving the highest-quality sequence reads. These reads would enable the more-demanding design of sequence tag sites (STSs) in addition to serving as STCs. STSs support other mapping methodologies, including BAC positioning on radiation hybrid maps, which are an important complement to contig maps.

After a transition phase, high-throughput STC production was implemented only at the The Institute for Genomic Research (TIGR) , initially under Mark D. Adams, and at the University of Washington with production managed by Gregory Mahairas of Leroy E. Hood's Department of Molecular Biotechnology (UWMB). A September 1998 site review at the newly opened High Throughput Sequencing Center of the UW Department of Molecular Biotechnology reaffirmed plans for completing the STC projects at TIGR and UWBC by fall 1999.

This timeline has now been shortened, however. In March 1999 the consortium of major sequencing centers announced a short-term objective of generating a draft sequence of the human genome within a year. Availability of the full BAC STC datasets was found crucial to achieving this goal. TIGR and the University of Washington consequently reprogrammed their ongoing projects. With additional support from DOE, STC production was expected to be substantially complete in July 1999.

The UWMB datasets already include include restriction fingerprints for BACs of the CalTech library. Extension of fingerprinting to the RPCI BACs is planned. The fingerprints will help sequencing teams validate candidate BACs overlapping their sequenced regions by distinguishing them from those that merely have limited homology within their STCs and probably represent distant chromosomal loci. With the increasing evidence of duplicated regions within the genome, great care is necessary for validating contig extension before commitment to expensive sequencing.

For the CalTech BACs there is an expanding correlation with cDNAs of the Unigene collection. This will enable the concurrent sequencing of BACs with the messenger RNAs (as represented by cDNAs) they putatively encode. This research area is complemented by a DOE-initiated series Workshops on Complete DNA Sequencing addressing international coordination of cDNA sequencing. Recognition and specification of gene-coding segments of chromosomes is greatly aided when both genome and cDNA sequences are available.

Recent reports from the STC teams were presented at the January 1999 DOE HGP Contractors and Grantees Workshop , together with many other reports relevant to genome sequencing. Detailed information and protocols are on-line at The Institute for Genomic Research (TIGR) and the Department of Molecular Biotechnology's Sequencing Center. Both teams, along with the DOE Genome Annotation Consortium , are providing online tools to aid sequencers worldwide, in the identification of BACs needed for contig extension.

For major sequencing centers, the BAC libraries are available directly from CalTech and RPCI. Sites that require fewer BACs can obtain them through commercial suppliers or regional resource centers in Europe , after identification of contig-extension candidates, by comparisons between the STC database and chromosome seed sequences.

At an October 1998 meeting of sequencing team leaders at the NIH/NHGRI it was recognized that although large

validated contigs remain the most desirable inputs, sequencing begun on single interesting BACs also has a substantial role to play in the total HGP effort. The STC datasets will be particularly useful for sequencing so initiated, as it will enable the rapid construction of contigs to guide sequence extension from the initial loci utilized. Use of STC data is integral to the whole human genome shotgun-sequencing strategy of Celera, Inc.

Genome projects for other species in which STC datasets are either in use or planned currently include *Arabidopsis thaliana* and the mouse .

Please e-mail any comments and suggestions for further related links to Marvin Stodolsky for the DOE Human Genome Task Group.

The paragraphs below correspond to some of the Hot Links in the preceding text.

Review of the STC related applications:
The applications to the HGP solicitation were received in April 1996 and reviewers for the special panel were obtained during May. The panel represented genomics efforts in seven countries and expertise in human and mouse genetics, mapping, sequencing, informatics and management. In preparation for a joint discussion, reviewers received the applications and returned initial critiques by e-mail. Some requests for clarification were forwarded to applicants and their responses returned to the anonymous reviewers. Outstanding differences in reviewer opinions were listed, in preparation for a joint conference call in July 1996. Staff of the DOE, NIH and NSF were listen-in observers, with M. Stodolsky coordinating for the DOE. The individual reviewer critiques were subsequently completed and sent to DOE . Following a final assessment by DOE Human Genome Task Group (HGTG) staff, pilot projects at six sites were initiated, with funds transmitted in September 1996.

Pilot Projects

The BAC libraries were provided by the teams under:

> Melvin Simon at the California Institute of Technology, (CalTech)
> Pieter de Jong at Children's Hospital Oakland Research Institute [formerly at the Roswell Park Cancer Inst. (RPCI)].

A basic technical problem was to purify BAC DNAs economically but with quality high enough to obtain good sequence reads. This task was addressed by:

> the CalTech team,
> the RPCI team,
> Skip Garner and Glen Evans at University of Texas SW Medical Center,
> a team under Leroy E. Hood at the University of Washington, Department of Molecular Biotechnology, (UWMB)
>
> a team under M.D. Adams at TIGR (The Institute for Genomic Research) following Adams transfer to Celera, Inc. , STC production at TIGR is now managed by William Nierman and Shaying Zhao .

An analysis of regions with chromosome segment duplications discovered in the laboratory of J. Korenberg at the Cedars Sinai Medical Center was extended, because duplications can pose troublesome ambiguities to contig

map construction. Other problematic cases are described in research by Evan Eicher and colleagues:

Eichler, EE, Lu, F, Shen, Y, Antonucci, R, Doggett, NA, Moyzis, RK, Baldini, A, Gibbs, RA, Nelson, DL. (1996) Duplication of the Xq28 CDM-CTR region to 16p11.1: A novel pericentromeric-directed mechanism for paralogous genome evolution. Hum. Molec. Genet. 5:899 912

Eichler, EE, Budarf, ML, Rocchi, M, Deaven, LD, Doggett, NK, Nelson, DL, Mohrenweiser, H. (1997) Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. Hum. Molec. Genet. 6: 991-1002.

Eichler, EE. (1998) Masquerading repeats: Paralogous pitfalls of the human genome. Genome Res. 8: 758-762.

Eichler, EE, Hoffman, SM, Gordon, LA, McCready, P, Lamerdin, JE, Mohrenweiser, HW. (1998) Complex beta-satellite repeat structures and the expansion of the zinc-finger gene cluster in 19p12. Genome Res. 8 : 791-808.

and by Barbara Trask and colleagues:

Trask, B.J., Friedman, C., martin-Gallardo, A., Rowen, L, Akinbami, C., Blankenship, J, Collins, C, Giorgi, D., Iadonato, S., Johnson, F., Kuo, W.-L., Massa, H., Morrish, T., Naylor, S., Nguyen, O.T.H., Rouquier, S., Smith, T., Wong, D.J., Youngblom, J., van den Engh, G. (1998) Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. Human Molecular Genetics 7: 13-26.

Trask, B.J., Massa, H., Brand-Arpon, V., Chan, K., Friedman C., Nguyen, O.T., Eichler, E., van den Engh, G., Rouquier, S., Shizuya H., Giorgi, D. (198) Large multi-chromsomal duplications encompass many members of the olfactory receptor gene family in the human genome. Human Molecular Genetics 7: (in press).

Pilot projects workshop and review on May 29, 1997

GRANTEES

The Institute for Genomic Research, abstract : Mark Adams, Steve Rounsley, Jenny Kelley and Hamilton O. Smith from Johns Hopkins University

Roswell Park Cancer Institute, abstract : Pieter de Jong, Joseph Catanese

University of Texas Southwestern Medical Center, abstract : Glen A. Evans and Harold R. Garner

University of Washington, Department of Molecular Biotechnology, abstract : Leroy H. Hood, Greg Mahairas, Todd Smith and Keith Zackrone

California Institute of Technology, abstract : Melvin I. Simon and Ung-Jin Kim

Cedars-Sinai Medical Center, abstract : Julie R. Korenberg

REVIEWERS:

Larry L. Deaven, Los Alamos National Laboratory
Trevor L. Hawkins, MIT Whitehead Inst.
Stanley Letovsky, Genome Data Base at Johns Hopkins University
David L. Nelson, Baylor College of Medicine
Michael Palazzolo, Lawrence Berkeley National Laboratory
Richard M. Myers, Stanford University School of Medicine
Lisa Stubbs, Oak Ridge National Laboratory

OBSERVERS/DISCUSSANTS:

Elbert W. Branscomb, DOE Joint Genome Institute
Robert W. Cottingham, Genome Data Base, Johns Hopkins University
Norman Doggert, Los Alamos National Laboratory
Sylvia Spengler, Lawrence Berkeley National Laboratory

U.S. GOVERNMENT STAFF:
DOE/OBER - Marvin Frazier, Daniel W. Drell, Arthur Katz , Marvin Stodolsky, David Thomassen
NIH/NHGRI - Mark S. Guyer, Adam Felsenfeld, Jane Peterson, Jeffery Schloss
NIH/NCI - Carol Dahl

STS versus STC requirements

For a sequence read to be useful as an STC, it need contain only a sequence segment unique to the source genome. For a read to be suitable for STS design, it must include two unique segments within the source genome. These segments must additionally lack features that would hinder priming for or read through by DNA polymerases used in the polymerase chain reaction (PCR). Thus the requirements for STS design are more stringent than those for STC usage. A higher-quality, longer sequence read is generally necessary to support the more demanding STS design requirements.

A review of progress and plans at the recently opened sequencing facility used by the University of Washington, Department of Molecular Biotechnology was held in September 1998. Leroy E. Hood and Gregory MacHarris gave presentations and production projections (see abstract).

The reviewers were:

Ellson Chen, Applied Biosystems Division of Perkin Elmer, Inc.

David Nelson, Baylor College of Medicine

Robert Robbins, Fred Hutchinson Cancer Research Center

Elbert Branscomb attended as an observer from the DOE Joint Genome Institute .

DOE was represented by Marvin Frazier, Director of the OBER Life Sciences Division

---

Validation of contigs

Clones that may overlap each other can be identified by a variety of methodologies, but none are foolproof by themselves. There may be some inadvertent representation of distant genomic loci. Also some clones may be defective due to accidents in construction or subsequent DNA rearrangements. Contig structure can be tentatively validated before sequencing by assessing whether putative overlaps contain such common representative features as restriction sites or STS markers. For members of a candidate contig passing validation tests, the one with minimal overlap of the previously sequenced region is the optimal choice for extending the seed region.

In the United States, Genome Systems Inc. and Research Genetics distribute clonal resources and provide screening services. In Europe, similar services are provided by the U.K. Human Genome Mapping Project Resource Centre and the German Resource Center .

Meetings

BAC 1995 Meeting Information

See STC Project History for later BAC developments.

Agenda and Attendees

Abstracts

- Large Human and Mouse PAC Libraries for Physical Mapping and Genome Sequencing, and More Versatile Cloning Vectors
  Eirik Frengen(1,5), Joe Catanese, Baohui Zhao, Chenyan Wu, Xiaoping Guan, Chira Chen, Eugenia Pietrzak, Julie Korenberg(3), Joel Jessee(4), Panayotis A. Ioannou(2), Hans Prydz and Pieter J. de Jong. (1)Department of Human Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263, (2)The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus, (3)Cedar Sinai Medical Center, Los Angeles, CA 90048, (4)Life Technologies, Gaithersburg, MD 20898, (5)Biotechnology Centre, Oslo, Norway.

- Evaluation of the Bacterial Artificial Chromosome Cloning System for Crop Plants
  Rod A. Wing, Texas A&M University, Soil & Crop Sciences Department, Texas A&M BAC Center, College Station, TX 77843-2123

- Towards a globally integrated, sequence-ready BAC map of the human genome
  Ung-Jin Kim, Hiroaki Shizuya, and Melvin I. Simon.
  Division of Biology, 147-75, California Institute of Technology, Pasadena, CA 91125

- Progress Towards the Construction of BAC Libraries from Flow Sorted Human Chromosomes Jonathan L. Longmire. Nancy C. Brown, Deborah L. Grady, Evelyn W. Campbell, Mary L. Campbell, John J. Fawcett, Phil Jewett, Robert K. Moyzis, and Larry L. Deaven.
  Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545

- Large Human and Mouse PAC Libraries for Physical Mapping and Genome Sequencing, and More Versatile Cloning Vectors
  Joe Catanese[1], Baohui Zhao[1], Eirik Frengen[1], Chenyan Wu[1], Xiaoping Guan[1], Chira Chen[1], Eugenia Pietrzak[1], Panayotis A. Ioannou[2], Julie Korenberg[3], Joel Jessee[4] and Pieter J. de Jong[1]. [1]Department of Human Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263, [2]The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus, [3]Cedar Sinai Medical Center, Los Angeles, CA 90048, [4]Life Technologies, Gaithersburg, MD 20898.

- BACs, PACs and the Structure of the Human Genome
  J. R. Korenberg, X-N. Chen, S. Mitchell, Z. Sun, E. Vataru, U-J. Kim, P. de Jong, M. Simon, T. J. Hudson, B. Birren, E. Lander, J. Silva, X. Wu.
  Cedars-Sinai Research Institute, Los Angeles, CA.

Articles and Resources

Articles

- "BAC End Sequencing Speeds Large and Small Projects," Human Genome News, Vol.10, No.1-2 February 1999
- "BAC End-Sequencing Projects Initiated: New Strategy Bypasses Contig Mapping," Human Genome News, July-September 1996; 8(1)
- Journal article discussing BES.
  Wu, C., Zhu, S., Simpson, S. and de Jong, P.J. (1996). DOP-vector PCR: a method for rapid isolation and sequencing of insert termini form PAC clones. Nucl. Acids Res. 24(13), 2614-5.
- "Analysis of Sequence-Tagged-Connector Strategies for DNA Sequencing" by Andrew F. Siegel, Barbara Trask, Jared C. Roach, Gregory G. Mahairas, Leroy Hood, and Ger van den Engh. Genome Research, March 1999 (v. 9, issue 3, 297-307) Abstract available

Resources

- Genome Systems Inc.
  Specializes in assembling genomic libraries, equipment, expertise, and technical support staff that are inaccessible to the typical scientist.

- Research Genetics
  Supplies custom oligo synthesis; databases on I.M.A.G.E. consortium cDNA clones, MapPairs and GenePairs; and BAC library screening and mouse BAC libraries.

Questions/Comments? contact:
Sheryl Martin (martinsa@ornl.gov)

Site produced by:
Human Genome Management Information System of Oak Ridge National Laboratory .

Disclaimer

Images adapted from:

- Resource for Molecular Cytogenetics, Lawrence Berkeley National Laboratory
- Utah Genome Center
- Children's Hospital Oakland Research Institute and
- Others

Related Web Sites

- Children's Hospital Oakland Research Institute
  Contains information on currently available BAC and PAC genomic DNA libraries; high density colony hybridization filters, BAC/PAC cloning vectors, laboratory staff, and pertinent recent publications

- University of Washington, Department of Genome Science
  This site contains software and/or documentation for RepeatMasker, phrap, phred, and other programs.

- Utah Center for Human Genome Research
  Projects at the center encompass large-scale sequencing, capillary, genotyping, and ELSI.

- The Institute for Genomic Research (now J. Craig Venter Institute)
  The Institute of Genomic Research is a not-for-profit research institute with interests in structural, functional, and comparative analysis of genomes and gene products.

- Human Artificial Episomal Chromosome (HAECS) Outlines the strategies to bridge contig gaps occurring at human genomic regions which cannot be cloned and/or maintained faithfully in bacteria using the large cloning systems.
  - Information request for contig gaps
  - Abstract

- Molecular Genetics Labs
  This site located at Cedars-Sinai Research Institute contains their integrated YAC/BAC/PAC resource for the human genome, chromosome 21 phenotypic mapping project, and gene mapping efforts.

Other Web Sites of Interest:

- Human Genome Project Information
- Genetic Meetings Home Page
- *Science* Human Genome Special Issue

- *Nature* Human Genome Special Issue

Reprinted with permission from *Bacterial Artificial Chromosomes: Methods and Protocols*, 2003 issue of Methods in Molecular Biology . Authors: Shaying Zhao and Marvin Stodolsky

---

Preface

Several developmental and historical threads are displayed and woven in this Volume. The use of large insert clone libraries is the unifying feature, with many diverse contributions. The editors have had quite distinct roles. Shaying Zhao has managed several BAC end sequencing projects. Marvin Stodolsky in 1970-80 contributed to the elucidation of the natural bacteriophage/prophage P1 vector system. Later he became a member of the Genome Task Group of the Department of Energy (DOE), through which support flowed for most clone library resources of the Human Genome Program. Some important historical contributions are not represented in this volume. This Preface in part serves to mention these contributions and also briefly surveys historical developments.

Nathan Sternberg (deceased) contributed substantially in developing a PAC library for drosophila, which utilized a PI virion based encapsidation and transfection process. This library served prominently in the Drosophila Genome Project collaboration. PACs proved easy to purify so that they substantially replaced theYACs earlier used. Much of the early automation for massive clone picking and processing was developed at the collaborating Lawrence Berkeley National Laboratory. However, the P1 virion encapsidation system itself was too fastidious, and PI virion based methods did not gain popularity in other genome projects.

Improving clone libraries was an early core constituent of the DOE genome efforts. Cosmid based libraries with progressively larger inserts were developed within the DOE National Laboratories Gene Library Program. But quality control tests by P. Youdarian indicated that perhaps 25% of human insert cosmids had some instability, possible due to the multi-copy property of the system. Both for this reason and to provide for larger inserts of cloned DNAs, DOE supported investigation of several new cloning systems. Of the eukaryotic host systems, the Epstein Bar virus based system from Jean-M. Vos (deceased) was successful indeed. But the added costs and care needed for use of eukaryotic cells precluded its wide adoption in HGP production efforts.

Among the bacterial host systems, two developed in the lab of Melvin Simon provided pivotal service. Ung-Jin Kim developed fosmids. They are maintained as single copy replicons and utilize the reliable encapsidation processes developed for cosmids. Fosmids proved to be highly stable. BACs were developed by Hiroaki Shizuya. They were introduced into E. coli by electroporation and stability was generally good, though there is an unstable BAC minority (1). This BAC resource emerged after the chimeric properties of the largeYACs was recognized. BACs were thus initially viewed with appropriate suspicion. But at the nearby Cedar-Sinai Medical Center, J. Korenberg and X.N. Chen implemented a very efficient FISH analysis. They found that chimerism of the BACs, in any, was at worst around 5% and the BACs were well distributed across all the chromosomes. Overall human genome coverage was estimated in the 98-99% range, with even centromeric and near telomeric regions represented.

Two examples of this good coverage soon emerged. Isolation of the BRAC1 breast cancer gene had failed with all other clone resources. But when Simon's group was provided with a short cDNA probe, they soon returned a BAC clone carrying an intact BRAC1 gene. Pieter de Jong had acquired the technology of cloning long DNA inserts from the Simon lab, initially using a PAC vector and electroporation. After a first successful library, DOE advised de Jong to broadly distribute this new PAC resource. Shortly thereafter, he assembled 900 kb contig for the candidate region of the BRAC2 gene. The subsequent DNA sequence generated at the Washington University indeed revealed the BRAC2 gene. These striking easy successes stimulated broad usage of the BAC and PAC resources.

The use of end sequences of clonal inserts to facilitate contig building had been used since the 1980s in small-scale mapping and sequencing projects. Glen Evans for example was piloting with DOE support a “mapping plus sequencing” strategy on chromosome 11, before the BAC resources were available. Once a covering set of cloned DNAs with sequenced ends is generated, clones to efficiently extend existing sequence contigs can be chosen (3). As the need for high throughput genome sequencing to meet HGP timelines became imminent, only a few human chromosomes had adequate contig coverage. L. Hood, H. Smith and C. Venter proposed a Sequence Tag Connector (STC) strategy to alleviate this bottleneck. With application to the entire human genome, concurrent BAC contig building and sequencing would be implemented.

The DOE instituted a fast track review of two STC applications in the spring of 1996 (2). One was from a team comprised of L. Hood, H. Smith and C. Venter, and the second from a team comprised of G. Evans, P. de Jong and J. Korenberg. A panel with broad international representation reviewed applications from two teams. Interested colleagues from the NIH and NSF were observers. While the overall STC concept was reviewed favorably, initial pilot implementations to better define the economics were recommended. A year later, progress was reviewed and a DOE commitment to a full scale implementation was made. At the request of the NIH, the DOE later increased support to accelerate a 20 fold coverage of the genome.

The STC data set has had multiple beneficial roles. Sequence Tag Sites (STSs) were defined within the STC sequences and used to enrich the Radiation Hybrid (RH) maps of the genome, thus providing for an early correspondence of the RH maps and the maturing contig maps. Validity constraints on sequence contigs were provided by the spanning BACs. Most broadly, the STC resource had an indispensable role for both the strategies of Celera Genomics Inc., and the international public sector collaboration, in the rapid generation of draft sequences of the human genome. The STC strategy is now implemented in many current genomic projects, including the NIH sponsored mouse and rat genome programs.

Herein there is provided a near comprehensive presentation of the protocols and resources developed for BACs in recent years. The book covers four topics about BACs: 1) library construction, 2) physical mapping, 3) sequencing, and 4) functional studies in the companion volume. The laboratory protocols follow the successful series format with a clear sequence of steps followed by extensive troubleshooting notes. The protocols cover simple techniques such as BAC DNA purification to complex procedures such as BAC transgenic mouse generation. Both routine and novel methodologies are presented. Besides protocols, chapter topics include scientific reviews, software tools, database resources, genome sequencing strategies and case studies. The book should be useful to those with a wide range of expertise from starting graduate students to senior investigators. We hope this book will provide useful protocols and resources to a wide variety of researchers, including genome sequencers, geneticists, molecular biologists and biochemists studying the structure and function of the genomes or specific genes.

We would like to thank all those involved in the preparation of this volume, our colleagues and friends for helpful suggestions, and Professor John Walker, the series editor, for his advice, help and encouragement.

Shaying Zhao, Marvin Stodolsky

1. http://www.ornl.gov/sci/techresources/meetings/ecr2.pdf
2. Enriching BACs for Sequencing with Sequence Tag Connectors

Updated sequence information at:

- *Science* Human Genome Special Issue
- *Nature* Human Genome Special Issue

The Draft Sequence Consortium

The consortium's goal is to produce a working draft covering at least 90 percent of human genome sequence within one year. The sequencing strategy involves determination of the sequence from mapped segments of DNA from known locations in the genome. These data are then assembled in overlapping stretches that reflect the accurate orientation of the DNA in the genome. In plans drawn up in the fall of 1998, Genome Project leaders projected completing the working draft by December 2001. The new consortium goal advances this timetable by more than a year and a half. The working draft will then serve as the scaffold for the painstaking but critical work of finishing, which involves closing gaps and correcting errors, leading to completion of the permanent high-quality, human DNA sequence by 2003 at the latest.

The five largest sequencing laboratories have joined together in a tightly knit collaboration with weekly meetings, shared materials, and shared protocols. The NHGRI funded laboratories will be responsible for producing approximately 60 percent of working draft sequence. DOE's Joint Genome Institute and the Sanger Centre will be responsible for producing approximately 10 percent and 33 percent respectively. "As one of the founders of the Human Genome Project, the Department of Energy is gratified to see the launch of the final stage of this project that promises such benefit to humanity," said Under Secretary of Energy Ernest Moniz.

Excerpted from:
http://www.nhgri.nih.gov/NEWS/pilot_project_completion.html

Earlier Projects Utilizing STC Tools

STC ideas were first used in projects led by George Church and were further evoled by other research teams.

STCs served in cosmid scale sequencing by Edwards, A. and Caskey, T.C. (1991). "Closure strategies for random DNA sequencing." In Methods: A Companion to Meth. Enzymol. 3: pp. 41-47 (Academic Press, New York).

An extension to YACs was implemented by Chen, E.Y., Schlessinger, D. and Kere, J. "Ordered shotgun sequencing, a strategy for integrated mapping and sequencing of YAC clones." Genomics 17, 651-656 (1993).

In 1995 a chromosome 11 pilot project began using cosmids with a later transition to BACs. It was supported by DOE Grant DE-F603-95ER62055 to the Salk Institute with PI G. Evans, now at the University of Texas SW Medical Center.

Whole microbial genome shotgun sequencing at TIGR was complemented by cosmid STCs to establish long range order.

A New Cooperative Strategy for Sequencing the Human and Other Genomes

J. Craig Venter , Hamilton O. Smith, and Leroy Hood

(Submitted to Nature)

Institute for Genomic Research, Johns Hopkins Medical School, Department of Molecular Biology and Genetics, University of Washington, Department of Molecular Biotechnology

One of the principal goals of the international Human Genome Project is to sequence, in a cooperative venture, the entire estimated 3 billion base pairs (bp) of DNA contained in the 24 different human chromosomes. The order of the nucleotides across each chromosome (the sequence maps) will permit the identification of the 100,000 or so human genes and provide the framework for studying how certain DNA variations among humans predispose to various diseases. This project, initiated in 1990 by the United States government (the National Institutes of Health and the Department of Energy), has been joined by the United Kingdom, France, Germany, and Japan. The cost was estimated to be $3 billion over 15 years. The first five year period focused on genetic and physical mapping (1). The genetic map contains polymorphic DNA markers scattered evenly across the genome; the physical map is generated from overlapping DNA fragments covering the 24 human chromosomes. We are now moving into the more complex sequencing phase (2, 3). We propose here a new approach to sequencing the human genome that would greatly simplify the procedure and facilitate international cooperation between large genome centers and small groups. Moreover, it will greatly facilitate the sequencing of biologically interesting chromosomal regions like gene families, such as those encoding neural and olfactory receptors, as well as smaller genomes from simpler organisms.

The most common approach to sequencing the human genome involves a three-stage divide and conquer strategy (Figure 1) employing the construction of three different clone libraries from human chromosomal DNA randomly cut, fractionated into differing size classes and then inserted into distinct vectors capable of propagating the DNA fragments in appropriate hosts (e.g. bacteria or yeast) (Table 1). (A clone is a vector with one inserted fragment of human DNA. A clone library is the entire collection of the fragments of human DNA that are integrated into a particular type of vector in one experiment.) (i) Low resolution physical maps of each chromosome are prepared by identifying shared landmarks (e.g. unique PCR [sequence tagged sites or STSs] or restriction enzyme sites) on overlapping yeast artificial chromosome (YAC) clones. (ii) High resolution or sequence-ready maps are then prepared by randomly cutting and subcloning YAC inserts into cosmid vectors. A map is constructed by identifying their landmark overlaps. (iii) A minimally overlapping path of cosmid clones is chosen and the DNA from each clone randomly fragmented into small pieces and subcloned into the M13 phage or plasmid vectors. For each cosmid, about 800 M13 clones are sequenced (average 400 base pairs) and assembled computationally into the sequence of the 40 kb cosmid insert. This random, or shotgun, approach ensures a high degree of accuracy because every nucleotide is, on average, sequenced 8 times (480 bases/clone x 800 clones = 320,000 bases of sequence). Most genome-wide or chromosome-specific physical maps generated to date are of low resolution and based on YACs (6).

This approach to genome-wide sequencing has several challenges and limitations. (i) The initial efforts for high resolution mapping of human chromosomes 16, 19, and 22 have been very expensive and they are still not finished; that is, the problem of obtaining complete maps without gaps is significant. Completing sequence-ready maps for these and the remaining human chromosomes still remains a daunting and expensive task. (ii) Approximately 50% of YAC clones exhibit rearrangements, deletions, and chimerisms (two or more DNA fragments inserted into one clone), thus rendering them often unsuitable as mapping and sequencing reagents.

The effort necessary to identify the defective YACs is significant. To a smaller degree, cosmid inserts also delete, rearrange, and form chimeras--aberrations that are also often difficult to detect. (iii) The human genome contains tandem (adjacent) arrays of very similar homology units (e.g. five tandem 21 kb arrays) or tandemly arrayed genome-wide repeats like the 7 kb LINEs that pose problems for high resolution mapping when the clone insert size is less than the tandem array length because the landmarks are very similar (e.g. 40 kb cosmid insert against 105 kb of DNA array). (iv) The conventional sequencing procedure is very complex and difficult to fully automate for the high throughput sequencing we hope to achieve in the future. (v) This approach makes cooperative collaborations among large and small groups difficult because a significant infrastructure is required for the high resolution physical mapping.

These problems, and two important scientific advances, have led us to propose a new strategy for cooperative sequencing of the human genome. The first advance rests in the ability to sequence and assemble megabase size prokaryote genomes with high accuracy and fidelity (7). The second is the development of bacterial artificial chromosome (BAC) libraries with human insert sizes up to 350 kb. BACs appear to faithfully represent human DNA far better than their YAC or cosmid counterparts (8). For example, the 1 Mb human a/d T cell receptor locus was mapped using only 17 BACs in contrast to the 75 or so cosmid clones that would have been required to achieve the same coverage. Detailed landmark analyses demonstrated that only one of 17 BACs had a defect (a small 6 kb deletion) (C. Boysen, personal communication). Moreover, BACs appear to be an excellent substrate for shotgun sequence analysis (e.g. 5/5 BACs, ranging in size from 89-210 kb, were successfully sequenced with this approach) (C. Boysen, personal communication) and other laboratories have also been successful in sequencing BACs. Accordingly, BAC clones appear to be excellent sequencing substrates that can be used to produce an accurate contiguous sequence.

Our new approach to genomic sequencing eliminates the need for any a priori physical mapping and uses BAC clones as the basic sequencing reagent (Figure 1). (i) A human BAC library with an average insert size of 150 kb and, on average, a 15-fold coverage of the human genome contain 300,000 clones. These will be arrayed into microtiter wells. (ii) Both ends (starting at the vector-insert points) of each BAC clone will be sequenced to generate 500 bases. The 600,000 BAC-end sequences will be scattered, on average, every 5 kb across the genome and will constitute 10% of the genome sequence. We will denote these end sequences as "sequence tagged connectors," or STCs, because they allow any one BAC clone to be connected on average to 30 others (e.g. a 150 kb insert divided by 5 kb will be represented in 30 BACs). The STCs would immediately be made available on the world wide web. (iii) Each BAC clone will be fingerprinted with one restriction enzyme to provide the insert size and detect artifactual clones by comparisons of the fingerprints with those of overlapping clones. (iv) A seed BAC of interest will be sequenced by any method and checked against the data base of STCs to identify the ~30 overlapping BACs. The two BACs exhibiting internal consistency among the fingerprints and minimal overlap at either end will be sequenced. The entire human genome could be so sequenced with slightly more than 20,000 BAC clones (Table 1).

This approach has several unique advantages. (i) The cost and effort to obtain complete low and high resolution maps is virtually eliminated; thus, the front end automation is greatly simplified (e.g. clone arraying, DNA purification, fingerprinting, and sequence reactions). (ii) The BAC clones can be made readily available to sequencing groups throughout the world through resource centers and/or commercial distributers. Large centers could sequence multiple BAC clones forming major contigs while small groups could contribute one or a few BAC sequences. (iii) As improved techniques for generating BAC or other yet to be developed libraries appear, reasonable numbers of these new clones could easily be added to the clone collection. (iv) It appears likely this approach will obviate the significant problem of closure for high resolution physical mapping. (v) The existing chromosomal landmarks, STS, or PCR-specific sites, and EST, or partial cDNA sequences, can be easily placed on the BAC clones, adding additional markers for BAC clones and taking significantly advantage of any associated biological information. (vi) The 10% of the genome obtained in the STCs can be searched against the

sequence data base to identify many interesting landmarks (e.g. genes, STSs, EST, etc.) that could locate the BAC clone on the preexisting chromosomal maps. (vii) Chromosomal regions of key biological interest can be sequenced first. (viii) The human genome can be sequenced earlier and for less cost (e.g. the savings on high resolution physical mapping). (ix) The STC approach will provide useful clones for biological studies even at the very early STC sequencing stages when only 3- to 4-fold coverage is achieved. (x) This would be an extremely efficient strategy for sequencing compact genomes (e.g. prokaryotes and single celled eukaryotes), as well as the model organisms the genome project has committed to sequence (e.g. E. coli, nematode, Drosophila, and mouse--the yeast genome is finished).

Arraying and DNA preparation facilities could readily make the end-sequenced BAC clones available to the world-wide genome community. BAC clones could be readily mailed and BAC-end sequences or STCs and fingerprints would be available on the world wide web, as would the identity of any clone selected for sequencing. Several research teams could thus work on the same chromosomal region without unintended duplication of effort. This would facilitate international cooperation. With our proposed strategy, participating laboratories could sequence the BAC inserts by any cost-effective method. Likewise, any DNA sequencing chemistries now in existence could be used, as well as any future chemistries.

The complete set of BAC-end sequences and fingerprints could be obtained in two years or less employing, for example, 30 Applied Biosystems 377 sequencers for a total cost of $5-10 million. This cost is a small fraction of the yet to be incurred cost of sequence-ready physical mapping.

A highly cooperative combination of large genome centers and small groups could finish the entire human genome sequence in under ten years. The current cost of DNA sequencing is around $0.30/finished bp in the most efficient laboratories and it is anticipated that it will fall to $0.10-0.25/base in the next one to three years. At these costs, the total sequencing costs for the entire genome would be less than the genome funds expended to date. The British foundation, The Wellcome Trust, recently announced funding the Sanger Center to sequence 1/6 or more of the human genome (9). Scientists in France, Germany, and Japan are discussing doing as much as 10% of the human genome each, while the United States' effort is just beginning with the establishment of six genome sequencing centers for pilot scale-up studies (10). These large centers around the world, in conjunction with many smaller groups, require an improved approach to genome coordination. The sequence tagged connector strategy proposed here offers a powerful new approach to sequencing the human and other genomes with a maximized level of international cooperation, and with all participants working on an equal basis in a self-regulating, open scientific effort.

References

1. Watson, J.D. Science 248:49 (1990).
2. National Research Council, Mapping and Sequencing the Human Genome (National Academy Press, Washington, D.C., 1988).
3. Marshall, E. Science 268:1270-1271 (1995).
4. Hudson, T. J., et al., Science 270:1945-1954 (1995); Marx, J., Science 270:1919-1920 (1995).
5. Chumakov, I. M., et al., Nature 377 (suppl.):175-298 (1995).
6. Olsen, M.V. Science 270:394-396 (1995).
7. Fleischmann, R.D., et al. Science 269:496-512 (1995); Fraser, C.M., et al. Science 270:397-403 (1995).
8. Kim, U.J., et al. Nucleic Acids Research 20:1083-1085 (1992). Shizuya, H., et al. Proceedings of the National Academy of Sciences U.S.A. 89: 8794-8797 (1992).
9. Dickson, D., Nature 378:120 (1995).
10. Marshall, E. and Pannisi, E., Science 272:188-189 (1997)

Discussion of Human Genome Mapping/Sequencing Strategy to Meet Five-Year Goals October 20, 1998

NHGRI staff met with several investigators on Tuesday, October 20, 1998 to discuss mapping strategies for identifying clones for sequencing.

Present at the meeting were Bob Waterston, David Cox, Elbert Branscomb, Eric Lander, Phil Green, and Richard Gibbs. NHGRI staff present were Francis Collins, Adam Felsenfeld, Elke Jordan, Jane Peterson, and Mark Guyer. NCBI staff present were Greg Schuler and Jim Ostell. David Bentley, Jane Rogers, and Alan Coulson of the Sanger Centre joined in the first half of the discussion by videoconference.

Until now, strategies for selecting clones for genomic DNA sequencing have largely been directed toward the generation of a sequence-ready map, or a minimal tiling path of BAC clones. Recently, however, questions have been raised about the ability of this approach to meet the throughput demands of the new timetable for completing the human genomic DNA sequence. A "sequence-driven" alternative, in which clone mapping is done subsequent to sequencing, has been offered. Prior to the meeting, the pros and cons of the map first-sequence later and the sequence first-map later alternatives were summarized by Phil Green.

The main objections to the sequence-driven approach are:

1. It does not recognize defective BACs prior to sequencing them.
2. By abandoning regionality, it blurs the lines of responsibility for finishing (particularly at the gap-closure phase); and may also diminish motivation for some sequencers, particularly smaller ones, who prefer to get credit for particular regions.
3. It defers gap-closure issues until late in the project.

The main objections to the regional map-driven (STS + fingerprint) approach are:

1. It likely will not be able to deliver clones at a rate adequate to feed sequencing capacity, at least early in the project (first 18 months or so). While some centers will be able to generate a sufficient supply of mapped, sequence-ready templates for their own purposes, others will not.
2. The quality (e.g. contig size and accuracy) of the maps that will be developed and the effects of biases in STS distribution, are difficult to predict.
3. It is difficult to adjust the strategy in the face of fluctuations in capacity among centers. Some centers will likely generate more capacity than anticipated, others will fall short, relative to the sizes of their assigned regions, and transferring responsibility and resources to compensate for this will be nontrivial.
4. The goal of identifying and sequencing gene rich-regions early in the project may not be attainable using an STS-driven approach, if the EST map proves to have inadequate resolution.

There are some additional objections that apply to both strategies (e.g. potential sensitivity to large scale repeats and to biases in clone coverage).

Prior to the meeting, a third, hybrid, strategy emerged. This would encompass both map-driven (regional) and sequence-driven (random) approaches; centers could choose to pursue one, or the other, or both. The random approach would start by end-sequencing randomly chosen BACs. If the end sequence hit a clone or contig that was mapped to a chromosomal region already being actively sequenced, no further work on that BAC would be done by the "end sequencer." The end sequence data would be submitted to a central server (see below), so that the regional center responsible for that region could use the BAC, if desired. However, if the end sequences were

informative but did NOT hit a mapped contig, that BAC would then be lightly sequenced (0.5X coverage) by the "end sequencer." That data would again be analyzed using the central server, and a decision as to whether to sequence the BAC further would be made on the basis of minimum overlap and/or gene content. If the decision were made to complete the shotgun sequencing of the BAC (9-10X coverage) it would also be RH mapped. Once whole genome fingerprint-derived contigs become available (from the effort now under way at Wash. U.), BACs would preferentially be selected from them to help reduce the incidence of defective BACs. The participants considered that such a hybrid strategy would best meet the needs and interests of the sequencing groups. The feasibility of the approach would be dependent on the development of a central server, which would maintain a list of BACs for which any useful information was available.

Important information (to be contributed by all sequencers) to have in the database, would include end sequences, chromosome location (or "unmapped"), fingerprint data, extent of overlap with other BACs, and the data from the 0. 5X sequencing. It was also agreed that this should be a publicly accessible resource.

Later in the discussion, this server was described as version 2.0 of the Human Genome Sequencing Index. It was envisioned as playing the role of the site at which "claims" for sequencing responsibility would be established. The attendees suggested that the criterion for establishing priority for sequencing a region would be real data from that region, e.g. submission of identifying information about a clone or contig to the central server. Such information could include clone name or actual sequence data. The NCBI representatives agreed to look into the feasibility of establishing this server in the near future.

Other centralized resources/services were discussed. In addition to the BAC end sequences, BAC fingerprints, and high resolution RH maps currently being produced, a resource set of chromosome-specific STS hits on the BAC library was considered to be of value, understanding that this was to be a public mapping resource and not a means of establishing sequencing claims.

As for sequencing responsibilities, the sequence-driven approach was recognized as being most useful during the early phases of genomic sequencing. As sequence data are accumulated, closing the gaps between the growing contigs will require even the non-regional groups to take on regional responsibility. At the present time, however, assuming responsibility for regions (which at this point are more likely to be blocks of 10Mb or so, rather than entire chromosomes) will be based on a group's having actually produced and submitted to the server sufficient data to establish a valid claim. Once a group (map-driven or sequence-driven) starts the "heavy" shotgun sequencing of a clone it is committing to finishing that clone, even if it falls in a region that is being sequenced by another group; such a situation could arise if a random clone selected for sequencing is in a "claimed" region, but has no overlap with anything already in the server. It was also recommended that such claims should only be valid for a reasonable period of time in order to avoid a situation where one group's failure to complete a small region delays closure/completion of a larger region. A similar approach to and degree of coordination of the sequencing effort among participants was used successfully in yeast.

Finally, the new goal for a (better than 90%) complete working draft was discussed in the context of a hybrid plan in which most of the sequence generated would either be very light (0.5X) or complete shotgun sequencing. The importance of a complete working draft sequence, as defined in the new five-year plan, by 2001 was reaffirmed. However, some participants suggested that it would not be a great deal of effort to expand coverage from 0.5X to 3X rather quickly and, consequently, concentration of sequencing effort on light shotgun and finished sequencing in the next two years could be an acceptable strategy, with determination of how much to expand the 0.5X coverage to be made toward the end of the year 2000, as sequencing capability at that time is evaluated.

Excerpted from the report:

PRIORITY SETTING MEETING FOR MOUSE GENOMICS AND GENETICS RESOURCES:

FIRST FOLLOW-UP MEETING

OCTOBER 5, 1998

SUMMARY OF MEETING: POINTS OF CONSENSUS AND ACTION ITEMS

A. STRUCTURAL GENOMICS
1. The Mouse Strain to be Sequenced
2. Sequencing Strategy

The recommended strategy for sequencing the mouse is to build the physical map as the sequence is generated. A three-phase strategy was proposed:

Phase I (Random): Researchers would sequence individual BAC clones at random until 30 percent of a 15X library was sequenced. The BAC clones would be selected at random or based on scientific interest. BACs would be sequenced and STSs from sequenced BACs would then be mapped onto a common high resolution RH mapping panel. All information would be entered into a common database.

Phase II (Directed): The sequencing strategy would change over to extending sequencing off the ends of BACs using information from the BAC fingerprinting or the BAC end sequencing projects to direct the extension. This phase would be in place until about 90 % of the genome had been sequenced.

Phase III (Closure): This might require special libraries (library made with different enzymes, smaller insert size libraries, etc) for gap filling.

The complete text of this report will be available after review by the participants. For details contact Betty Graham at NIH/NHGRI

Large-Scale BAC End Sequencing to Aid Sequence-Ready Map Construction

Mark D. Adams, Steve Rounsley, Casey Field, Jenny Kelley, Steve Bass, Brook Craven, and J. Craig Venter
The Institute for Genomic Research, Rockville, MD 20850 mdadams@tigr.org

As the genome project shifts into the large-scale sequencing phase, an overwhelming technical challenge resides in developing an efficient method for producing minimum tiling paths of sequence-ready clones across the entire genome. To illustrate this challenge, consider what a sequencing center that proposes to finish 100 megabases (Mb) of DNA sequence per year must do: on average, each day 2-3 bacterial artificial chromosomes (BACs), averaging 150 kilobases (kbp) in length each, must be sequenced.

Deep representation BAC libraries are currently being developed in Dr. Mel Simon's laboratory at CalTech and in Dr. Pieter de Jong's laboratory at the Roswell Park Cancer Institute. The DNA for preparation of these libraries was prepared from volunteer donors who had given fully informed consent to use of their DNA for the human genome project. Extensive provisions for maintaining the anonymity of the donors are in place. These libraries will form the core resource used by sequencing centers in the US and around the world for sequencing the human genome.

BAC end sequencing is used to precisely mark the position of each BAC clone relatively to completely sequenced BACs. Exhaustive end-sequencing of BAC clones from 16p is currently underway as is a larger project to end-sequence up to 300,000 BAC clones to provide markers for construction of sequence-ready maps across the genome.

We are in the midst of obtaining end-sequence data from each of 300,000 clones from these new libraries. These end sequences will be invaluable for making highly efficient use of these clones to construct sequence-ready physical maps and to select clones for sequencing. At the proposed level of redundancy (about 15 genome equivalents), the end sequences will provide a sequence marker of 300-500 base pairs (bp) on average every 5,000 bp across the genome. Currently over 13,000 BAC end sequences have been produced and the sequencing rate is continuing to increase.

A 96-well mini-prep method has been developed that permits rapid purification of BAC DNA with sufficient quantity and purity for direct sequencing. Use of new BigDye terminator chemistry from ABI has made BAC end-sequencing quite robust.

A PAC/BAC End Sequencing Data Resource for Sequencing the Human Genome

Grant No.: DE-FC03-96ER62294
Principal Investigator: Pieter J. de Jong

Progress Report: 9/15/96- 7/7/97

The pilot project "A PAC/BAC End Sequencing Data Resource for Sequencing the Human Genome" has as it immediate goal the elucidation of the DNA sequence from the ends of PAC/BAC clones that comprise the current human genomic DNA libraries. Over the past ten months, our laboratory has set out to develop a high throughput DNA sequencing system based on the use of PCR-amplified clone ends as sequencing template. The initial four months of the project was directed toward the acquisition of the necessary capital equipment (two ABI 377 DNA Sequencers, MJ Research DNA Tetrad Thermal Cycler, Hamilton Microlab 2200 pipetting robot) and the hiring of technical personnel to perform the work. By incorporating a DOP-Vector PCR approach (Wu C.,et. al., 1996, NAR, 24:2614-2615.) to generate multiple copies of the clone ends, our laboratory has "rescued" approximately 7000 PAC clone ends during the past five months. These clone ends have been generated both randomly from the RPCI6 Human PAC library (about 4600 ends) and also a specific set of clone ends (approx. 2400 ends) covering a 3Mb region of chromosome 14 were sequenced from the RPCI1, 3, 4, 5 and 6 Human PAC Libraries. We have performed automated DNA sequencing using ET-dye labeled primers (Amersham) on these ends and have analyzed the data in detail. 2100 (30.1%) of the clone ends sequenced resulted in data that did not meet our minimum acceptable requirements of no more that 3 non-base calls in a 25 base window and/or less that 10% non-base calls in the remaining sequence. In addition, all "good" sequences must pass through Phred analysis and report greater than 20 to be considered as acceptable. The acceptable clone DNA end sequences (4900 ends) have an average Phred trimmed read length of 384 bases and an average Phred non-base call percentage of 4.25. More recent work on a much smaller set of clone ends (128) incorporating new dRhodamine dye terminator chemistry (ABI) indicates potential improvement in the quality of the data generated. For this set of clone ends the failure rate drops to 21.9%, the average read length increases to 443 bases while the percentage of non-base calls drops to 2.52%. Further work is in progress using the dRhodamine dye terminators to determine whether these results remain higher once a statistically significant number of clone ends has been sequenced. DNA sequence homology analysis (BLAST) indicates that 63.7% of the ends have unique sequence (no acceptable homology to known sequences in current data banks). 23.8% of the end sequences generated have homology to known EST's and STS's, while 3.2% identify with Line repeat elements and 9.3% to ALU repeat sequences. Detailed costs analysis has been performed based on our DNA sequencing success rate relative to the amount of research funds spent to date. When total dollars spent is compared to the total number of successful DNA end sequences generated, the costs per clone end is $72.79. After subtracting out all major equipment purchases and leases that were required to get the project off the ground, the cost drops to $42.45 per end. When only reagent and supply expenditures are considered, each end costs $22.66 to generate. The above costs analysis reflects the four month period when the DNA sequencing laboratory was being set up and essentially no data was being produced. In addition, two more months passed in which we were ramping-up our small scale techniques into high throughput PCR and DNA sequencing procedures. At our current level of production, we are able to generate successful DNA sequence from PCR end- rescued clone ends at a cost of $10.01 per end. This costs reflects all personnel, reagent and supply expenditures incurred but does not include any amortization of capital equipment employed in the project. At our current average read length this cost per clone end translates to a cost of $0.0261 per base.

A PAC/BAC End-sequence Database for Human Genomic Sequencing

Glen A. Evans, Dave Burbee, Chris Davies, Trey Fondon, Tammy Oliver, Terry Franklin, Lisa Hahner, Shane Probst and Harold R. (Skip) Garner
Genome Science and Technology Center and McDermott Center for Human Growth and Development, The University of Texas Southwestern Medical Center at Dallas.

While current plans call for completing the human genome sequence in 2003, major obstacles remain in achieving the speed and efficiency necessary to complete the task of mapping and sequencing. We proposed a novel approach to large scale construction of sequence-ready physical clone maps of the human genome utilizing end-specific sequence sampling. A earlier pilot project was initially carried out to develop a GSS (genomic sequence sampled) map of human chromosome 11 by sequencing the ends of 17,952 chromosome 11 specific cosmids. This chromosome 11-specific end-sequence database allows rapid and sensitive detection of clone overlaps for chromosome 11-sequencing. In this DOE funded pilot project, we proposed a effort to evaluate the utility of PAC and BAC end-sequences representing the entire human genome as a tool for complete, high accuracy mapping and sequencing. We utilized total genomic PAC/BAC libraries constructed by P. de Jong, RPCI, followed by end-sequencing of both ends of each clone in the library and limited regional mapping of a subset of clones as sequencing nucleation points by FISH (Fluorescence in situ hybridization). To initiate regional analysis, a single clone was be sequenced by shotgun or primer directed sequencing, the entire sequence used to search the end-database for overlapping clones, and the minimal overlapping clones for extending the sequence selected. We demonstrated that this approach allows a rational and efficient simultaneous mapping and sequencing of portions of the genome, as well as expediting the coordination and exchange of information between large and small groups participating in the human genome project. This pilot project involved automated end-sequencing of approximately 5000 PAC and BAC clones representing portions of chromosome 11 as well as portions of the entire human genome. The clones and resulting end-sequence data base were evaluated for their use in 1) nucleating regions of interest for large scale sequencing concentrating on regions of chromosome 11, 2) the evaluation of mapping accuracy and integrity and 3) the evaluation of random clone end sequence libraries. A technique for high throughput DNA sequencing was developed using a Beckman/Sagian robotic system, ABI 377 automated sequencers and automated sequence data processing, annotation. FISH analysis of a sample of over 200 PAC clones was carried out and to define the chimera rate in existing PAC libraries. Continuation of this project would include PAC and BAC end-sequencing of a sufficient number of clones to represent the entire human genome. However, with the premature termination of this pilot project, current goals are directed towards completion of the work underway and preparing data for publication.

The Sequence Tagged Connector (STC) approach to genomic sequencing: Accelerating the complete sequencing of the human genome

Gregory G. Mahairas, Keith D. Zackrone, Stephanie Tipton, Sarah Schmidt, Alan Blanchard, Anne West, and Leroy Hood.
Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195.

The STC approach has been proposed as an attractive strategy to provide a sequence ready scaffolding for the efficient and directed sequencing of the complete human genome (1). This effort has been undertaken through a collaborative effort between the California Institute of technology, TIGR and the University of Washington, and funded through the U. S. Department of Energy. The approach entails the sequencing of the ends of 300,000 Bacterial Artificial Chromosomes (BACs) that constitute a 20X deep Human DNA library to construct a sequence ready scaffold of the human genome.

At the Univ. of Washington we have assembled a high throughput automated end sequencing and fingerprinting process with its associated informatics. BAC clones are robotically inoculated from 384 well plates into 4 ml 96 well culture format, grown and the BAC DNA robotically extracted using AutoGen 740 robots. BAC template DNA from the AutoGen is then robotically transferred into 96 well microtiter plates from which DNA sequencing and fingerprinting reactions are setup. DNA fingerprinting is performed using conventional agarose electrophoresis, digestion with a single restriction enzyme (EcoRV) followed by automated imaging and analysis. DNA sequencing is performed using PE-ABD High Sensitivity dye primers and ABI 377 DNA sequencers. Laboratory protocols, automated data production, data processing, quality control measures and LIMS will be described in detail. During a 50 day period the STC laboratory sequenced 23317 BAC ends (STCs). 19224 (82.4%) where greater than 100 bp non trimmed and the average nontrimmed read length was 388 bp fro a total of 7.46 Mb (.25%) of the genome. 29 % of the STCs contained repetitive DNA but less than 11% where entirely repeat. 12% of the repetitive DNA were LINE sequence, 4.6% LTR, 6.7% SINE sequence and 1.3 % of the STCs contain a microsatellite or simple sequence repeat. The total G + C content was 40 % and the average CpG content was .28, both expected numbers for human genomic DNA. 224 STCs had CpG scores of 1 representing CpG islands. 3103 STCs (16.8%) hit the EST, non-redundant nucleotide or Sixframe database. 1103 STCs hit the EST database (DB) 517 of which hit only the EST, 1087 STCs hit the nr nucleotide DB, 471 of which hit the nr nucleotide DB only and 913 STCs hit the nr protein DB 500 of which hit only the nr protein DB. 181 STCs (1%) hit all three databases, 131 hit nr nuc. and nr protein DBs, 101 hit the EST and nr prot. DBs and 304 hit EST and nr nucleotide DBs i.e. 4% hit more than one of these Dbs and probably represent genes.

1. Venter, J. C., Smith, H. O., and Hood, L. (1996) Nature 381: 364-366

BAC End Sequencing - The Caltech Contribution

Ung-Jin Kim, Hiroaki Shizuya, and Mel Simon

The following is a progress report covering the last period of work on the BAC end sequencing (BES) project. The group at Caltech has undertaken three tasks: 1) supplying arrays of BAC clones to the Hood Lab and to TIGR for high throughput end sequencing; 2) end sequencing, arraying, and checking BACs in the chromosome 16 region to evaluate the efficacy of end sequencing and the choice of minimal paths for high throughput sequencing of multi-megabase regions of the genome; and 3) developing a model system using chromosome 22 to end sequence and fingerprint BACs corresponding to chromosome 22, allowing a quantitative analysis of the utility of BAC end sequencing in providing markers and BAC substrates.

With respect to the first task, supplying BACs, we have generated and developed arrays of BACs corresponding to approximately 50,000 clones with quality control measures to identify empty wells and deficient clones in order to provide an appropriate set of substrates for large-scale BAC-end sequencing. This library was shipped to Dr. Hood. His group extracted DNA and sequenced approximately 20,000 BAC ends from this material. With regard to our second objective, we identified together with TIGR a region corresponding to 20 megabases that will be involved in high throughput sequencing at TIGR. Using a variety of probes provided by the Los Alamos National Laboratory, we selected over 2,000 BACs that map to this region of chromosome 16. We extracted DNA from these BACs using the Autogen robot and determined end sequences for most of the clones. These end sequences in general have yielded from 300-500 base pairs of usable sequence. When they were compared to BACs for which complete sequences had been obtained, we were able to position many of the end sequences and thus map the precise position of the BAC. This allowed us to close holes in the map of available BACs and to demonstrate that one could use this technique to select the BAC with minimal overlap. Furthermore, using end sequences, we generated end clones. These were used to rescreen the library and to find other adjacent BAC clones. Thus far we have screened a 12X human library (~250,000 clones) for BACs that map to the chromosome 16 region. It is clear from the end sequences, from the match ups and from the ability to saturate the region with relatively randomly distributed BACs that this approach, i.e. sequence then map, will be extremely useful in high throughput sequence determination of specific contiguous regions of the human genome. Our third goal was to use a whole chromosome, chromosome 22, as a quantitative demonstration of the utility and economy of the BAC end sequencing method. Together with TIGR, we determined the end sequences of approximately 700 BACs and we are in the process of completing the BAC coverage of chromosome 22. We expect to cover this region with 3,000-4,000 BACs and complete those end sequences. Chromosome 22 is now being sequenced by a number of groups. We can use their data to position the BAC ends and get a quantitative estimate of the utility of this approach to provide substrates with minimal overlap. In addition, we developed a method using the ABI sequencer to fingerprint clones. Our initial approach has demonstrated the feasibility of this method. We expect that we can now demonstrate its application by completing a deep, fingerprinted, overlapping BAC map of the 40 megabases for chromosome 22. This, together with the end sequences and comparisons with already sequenced regions will allow us to get numbers regarding the relative costs and savings of this approach.

We are continuing to develop new BAC libraries using DNA that has been obtained through the approval of the Caltech IRB with appropriate consent, confidentiality, and anonymity. These new libraries will be available within the next few months and all of our work will shift to the new libraries. We believe that we can demonstrate clearly the enormous utility of end sequencing BACs from the new libraries as an adjunct to, and as a preliminary approach to, high throughput sequencing.

ONE ORIGIN OF MAN: PRIMATE EVOLUTION THROUGH GENOME DUPLICATION

Julie R. Korenberg, Xiao-Ning Chen, Steve Mitchell, Rajesh Puri, Zheng-Yang Shi nd Dean Yimlamai
Medical Genetics Birth Defects Center, The CSMC Burns & Allen Research nstitute, UCLA School of Medicine, Los Angeles, CA

Chromosome duplication is a force that drives evolution. We now suggest that this may also be true of the primates and that the resulting duplications in part determine the spectrum of human chromosomal rearrangements. To investigate the existence and origin of duplications in the human genome, and their consequences, 5,000 bacterial artificial chromosomes (BACs) were mapped at 2-5 Mb resolution on human high resolution chromosomes by using fluorescence in situ hybridization. A subset of 469 of these was defined that generated two or more signals, excluding those located in regions of known repeated sequences, viz., the regions of centromeres, telomeres and ribosomal genes. Although a subset of these multiple site BACs represent the chimeric artifacts of cloning, derived from two different chromosomal regions, others reflect regions of true homology in the human genome.

Two questions were considered; first, the extent to which the multiple sites of hybridization of single BACs within single chromosomes reflected the breakpoints of naturally occurring human inversions, and second, the extent to which these same multiple hybridization points reflected the chromosomal inversion points in primate evolution. For human inversions, the results of the analyses revealed a total of 124 BACs (2.5%) mapping to two or more sites on the same chromosome, of which 81 (65%) mapped to one of 27 distinct human inversion sites, the largest share of which recognized the well-established pericentromeric inversions of chromosomes 1, 2, 9, and 18, as well as the paracentric inverted region of chromosome 7q11/q22. From this, we infer that meiotic mispairing involving the homologous regions may be responsible for the inversions.

With respect to primate evolution, a significant proportion of inversion breakpoints that characterize the chromosomal changes seen in the evolution of the great apes through man, are also reflected in the distribution of BAC multiple intrachromosomal sites. Further analyses of the 29 independent BACs recognizing the pericentromeric region of human chromosome 9 suggest at least three classes, two of which recognize only single sites in Pan troglodytes.

These data suggest that inversions occurring through primate evolution may generate small duplications that, although they can cause chromosomal imbalance in single individuals, they also provide the additional genetic material for speciation.

December 4, 1995
U.S. Department of Energy Workshop on Bacterial Artificial Chromosomes (BACs)

Agenda

8:30-10:00 a.m.
Presentations on experiences with large insert libraries. Each talk was 15 minutes.
E. Branscomb
P. de Jong
D. Page
E. Lai
B. Birren
J.F. Cheng

10:30-12:00 noon
Additional Presentations
C. Venter
C. Amemiya
R. Wing
M. Simon
M. Adams

1:00-2:00 p.m. Discussion of problems and the need for a public Large Insert Library Resource: Effort? Cost? Starting Material? Mechanisms for access?

2:00-2:30 p.m. Presentations of proposal(s) for libraries

3:30-5:00 p.m. Discussion of proposals for specific resources

Attending HGCC members and Ex Officio

Benjamin J. Barnhart
Health Effects Research Division
U.S. Department of Energy
Office of Health and Environmental Research
Barnhart@mailgw.er.doe.gov

Elbert W. Branscomb
Human Genome Center/BBRP
Lawrence Livermore National Laboratory
branscomb1@llnl.gov

Michelle Broido
Office of Health and Environmental Research
U.S. Department of Energy
michelle.broido@mailgw.er.doe.gov

Charles Cantor

Robert K. Moyzis
Human Genome Center
Los Alamos National Laboratory
moyzis@telomere.lanl.gov

Mohandas Narla
Human Genome Center
Lawrence Berkeley National Laboratory
mohandas_narla@macmail.lbl.gov

Aristides Patrinos
Health and Environmental Research
U.S. Department of Energy
Ari.Patrinos@oer.doe.gov

Melvin I. Simon
Biology Division

Center for Advanced Biotechnology
Boston University
crc@enga.bu.edu

Anthony Carrano
Human Genome Center/BBRP
Lawrence Livermore National Laboratory
carrano1@llnl.gov
avc@sts.llnl.gov

Francis S. Collins
National Center for Human Genome Research
National Institutes of Health
fc23a@nih.gov

Daniel W. Drell
Human Genome Program
Office of Health and Environmental Research
U.S. Department of Energy
daniel.drell@oer.doe.gov

Marvin Frazier
Office of Health and Environmental Research
U.S. Department of Energy

Gerald Goldstein
Medical Application and Biological Research
Office of Health and Environmental Research
U.S. Department of Energy
gerald.goldstein@oer.doe.gov

Roland F. Hirsch
U.S. Department of Energy
roland.hirsch@oer.doe.gov

David Kingsbury
Genome Database
Johns Hopkins University
dkingsbu@gdb.org

California Institute of Technology
simon@starbase1.caltech.edu

David A. Smith
Health Effects and Life Sciences
U.S. Department of Energy
David.A.Smith@oer.doe.gov

Hamilton O. Smith
Johns Hopkins University School of Medicine
ham_smith@qmail.bs.jhu.edu

Lloyd M. Smith
Department of Chemistry
Analytical Division
University of Wisconsin-Madison
lmsmith@fozzie.chem.wisc.edu

Jay Snoddy
Human Genome Task Group
U.S. Department of Energy
jay.snoddy@oer.doe.gov

Sylvia Spengler
Human Genome Program
Lawrence Berkeley National Laboratory
SJSpengler@lbl.gov

Marvin Stodolsky
Human Genome Task Group
HELSRD
U.S. Department of Energy
Marvin.Stodolsky@oer.doe.gov

David Thomassen
HELSRD
Office of Health and Environmental Research
U.S. Department of Energy
David.Thomassen@mailgw.er.doe.gov

John Wooley
Office of Health and Environmental Research
Office of Energy Research
U.S. Department of Energy
wooley@er.doe.gov

Attending Invitees

Mark D. Adams
The Institute for Genomic Research
mdadams@tigr.org

Chris T. Amemiya
Center for Human Genetics
Boston University School of Medicine
camemiya@bu.edu

Bruce Birren
Whitehead Institute/MIT Center for Genome Research

Jan-Fang Cheng
Human Genome Center
Lawrence Berkeley National Laboratory
jcheng@genome.lbl.gov

Pieter de Jong
Human Genetics Department
Roswell Park Cancer Institute
pieter@dejong.med.buffalo.edu

Eric Lai
Glaxo Wellcome Inc.
ehl21107@ussun4e.glaxo.com

David Page
Whitehead Institute
Massachusetts Institute of Technology

Lisa Stubbs
Biology Division
Oak Ridge National Laboratory
stubbs@bioax1.bio.ornl.gov
stubbslj@ornl.gov

J. Craig Venter
The Institute for Genomic Research
jcventer@tigr.org

Rod A. Wing
Department of Soil and Crop Science
Crop Biotechnology Center
Texas A&M

LARGE HUMAN AND MOUSE PAC LIBRARIES FOR PHYSICAL MAPPING AND GENOME SEQUENCING, AND MORE VERSATILE CLONING VECTORS*

Eirik Frengen *(1,5)* , Joe Catanese, Baohui Zhao, Chenyan Wu, Xiaoping Guan, Chira Chen, Eugenia Pietrzak, Julie Korenberg *(3)* , Joel Jessee *(4)* , Panayotis A. Ioannou *(2)* , Hans Prydz and Pieter J. de Jong, (1)Department of Human Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263, (2)The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus, (3)Cedar Sinai Medical Center, Los Angeles, CA 90048, (4)Life Technologies, Gaithersburg, MD 20898, (5)Biotechnology Centre, Oslo, Norway.

Recently, we have developed procedures for the cloning of large DNA fragments using a bacteriophage P1 derived vector, pCYPAC1 (Ioannou et al. (1994), Nature Genetics 6: 84-89). A slightly modified vector (pCYPAC2) has now been used to create a 15- fold redundant PAC library of the human genome, arrayed in more than 1,000 384-well dishes. DNA was obtained from blood lymphocytes from a male donor. The library was prepared in four distinct sections designated as RPCI-1, RPCI-3, RPCI-4 and RPCI-5, respectively, each having 120 kbp average inserts. The RPCI-1 segment of the library (3X; 120,000 clones, including 25% non-recombinant) has been distributed to over 40 genome centers worldwide and has been used in many physical mapping studies, positional cloning efforts and in various large-scale DNA sequencing enterprises.

Screening of the RPCI-1 library by numerous markers results in an average of 3 positive PACs per autosome-derived probe or STS marker. In situ hybridization results with 250 PAC clones indicate that chimerism is low or non-existing. Distribution of RPCI-3 (3X, 78,000 clones, less than 1% non-recombinants, 4% empty wells) is now underway and the further RPCI-4 and -5 segments (< 5% empty wells) will be distributed upon request. To facilitate screening of the PAC library, we have provided the RPCI-1 PAC library to several screening companies and non-commercial resource centers. In addition, we are now distributing high-density colony membranes at cost-recovery price, mainly to groups having a copy of the PAC library. The combined RPCI-1 and -3 segments (6X) can be represented on 11 colony filters of 22x22 cm, using duplicate colonies for each clone. We are currently generating a similar PAC library from the 129 mouse strain.

To facilitate the additional use of large-insert bacterial clones for functional studies, we have prepared new PAC & BAC vectors with a dominant selectable marker gene (the blasticidin gene under control of the beta-actin promoter), an EBV replicon and an update feature. This feature utilizes the specificity of Transposon Tn7 for the Tn7att sequence (in the new PAC and BAC vectors) to transpose marker genes, other replicons and other sequences into PACs or BACs. Hence, it facilitates retrofitting existing PAC/BAC clones (made with the new vectors) with desirable sequences without affecting the inserts. The new vector(s) are being applied to generate second generation libraries for human (female donor), mouse and rat.

*Supported in part by grants from the Office of Health and Environmental Research of the U.S. Department of Energy (#DE-FG02-94ER61883) and the National Center for Human Genome Research, National Institutes of Health (#1R01RG01165).

Evaluation of the Bacterial Artificial Chromosome Cloning System for Crop Plants

Rod A. Wing , Texas A&M University, Soil & Crop Sciences Department, Texas A&M BAC Center, College Station, TX 77843-2123.
Email: rodwing@tam2000.tamu.edu

Most plant and animal genes are known only by their phenotype (e.g. bacterial disease resistance in plants or cystic fibrosis in humans). A technology, termed "map-based gene cloning", has been developed to isolate such genes based on the position of a target on a genetic map. Since 1986, map-based cloning has been used successfully to isolate over 35 human genetic disease genes and over 20 plants genes. A crucial element of map-based cloning requires the availability of large insert DNA libraries (YACs or BACs) containing DNA inserts from 150 kb to 1000 kb.

Over the past two years our laboratory has explored the use of the BAC cloning system for plants and animals. Based on our studies, the BAC system is emerging as the system of choice for construction of plant genomic libraries with average insert sizes of 150 kb. The BAC vector, pBeloBAC11 (courtesy of M.Simon, Cal Tech), is derived from the endogenous E.coli F-factor plasmid which contains genes for strict copy number control and unidirectional origin of DNA replication. Additionally, pBeloBAC11 has three unique restriction enzyme sites (HindIII, BamH1 & SphI) located within the LacZ gene which can be used as cloning sites for megabase-size plant DNA. BAC libraries are generated by ligating size-selected restriction digested DNA with pBeloBAC11 followed by electroporation into E. coli. BAC library construction and characterization is extremely efficient when compared to YAC library construction and analysis. At present we have constructed representative BAC libraries for Sorghum bicolor (3), Oryza sativa ssp. japonica and indica (4) , Arabidopsis thaliana (2) and bovine (1). We are currently constructing libraries for cotton, sugarcane, tomato, wheat, sorghum propinquum, and corn. Our strategy has been to construct libraries from the parents of widely used mapping populations. We anticipate that by the summer of 1996 we should have the largest collection of plant and animal BAC libraries in the world.

A long term objective our BAC research is to establish a "BAC Center" at Texas A&M University (http://http.tamu.edu:8000/~creel/TAMBAC.html) which will serve as a genetic resource for the plant and animal genome communities world wide. The BAC Center will provide several functions:

- The construction and maintenance of high quality BAC libraries for plant and animal genomes.
- Robotic screening of BAC libraries with probes linked to important genes.
- BAC clone distribution.
- Integrated physical mapping - FISH (metaphase, interphase and meiotic)

For more informaion concerning BAC work please see our World Wide Web Page at http://http.tamu.edu:8000/~creel/TAMBAC.html. This page describes our Arabidopsis BAC library, The Texas A&M BAC Center, and contains a manual for constructing plant BAC libraries

References and Abstracts

Cai, L., J.F. Taylor, R.A. Wing, D.S. Gallagher, S.S. Woo, and S.K. Davis. 1995. *Construction and characterization of a bovine bacterial artificial chromosome library* (Genomics 29:413-425).

A bacterial artificial chromosome (BAC) library has been constructed for use in bovine genome mapping using the pBeloBAC11 vector. Currently, the library consists of 23,040 clones, which achieves a 70% probability (P =

0.70) of the library containing a specific unique DNA sequence. An average insert size of 146 kb was estimated from the analysis of 77 randomly selected BAC clones produced by one or two rounds of size selection. The bovine DNA inserts proved to be very stable for at least 100 cell generations. No chimeric clones were detected among 11 large, size-selected BAC clones using fluorescence in situ hybridization (FISH) on metaphase bovine chromosomes. The polymerase chain reaction (PCR) was used to screen the library for single-copy nuclear sequences. Thirty-three of 46 (72%) sequences were present in the library in at least one copy, which is consistent with the estimated 70% probability of this library containing a unique DNA sequence. A BAC clone containing the 3-beta-hydroxy-5-ene steroid dehydrogenase (HSD3B) gene was physically mapped to bovine chromosome 3 by FISH. Two new microsatellite markers were isolated from the HSD3B-positive BAC clone as sequence-tagged sites for genetic mapping. These markers cosegregated, and no recombinants were detected in 193 informative meioses. Plasmid end rescue and the inverse polymerase chain reaction methods were used to rescue both ends of this BAC clone, and chromosome walking was performed using PCR primers designed within the end region sequences. Based on our experimental results, the BAC system provides a very useful tool for complex genome analysis.

Choi, S.D., R. Creelman, J. Mullet, and R.A. Wing. 1995. *Construction and characterization of a bacterial artificial chromosome library from Arabidopsis thaliana* Weeds World. 2: 17-20.

We constructed an ordered 3,948 clone Arabidopsis BAC library. The library has a combined average insert size of 100 kb (n=54). Assuming a haploid genome size of 100,000 kb, the BAC library contains *3.95 haploid genome equivalents with a 98% probability of isolating a specific genomic region. The library was screened with five Arabidopsis cDNA probes and one tomato probe and all probes hybridized to at least one (in most cases three) BAC clones in the library.

*Note: This library is now 12X and is maintained in the Texas A&M BAC Center and Arabidopsis Biological Resource Center (ABRC), Ohio State University.

Woo S.-S., J. Jiang, B. S. Gill, A. H. Paterson, and R. A. Wing. 1994. *Construction and characterization of a bacterial artificial chromosome library for* Sorghum bicolor. Nucleic Acids. Res. 22:4922-4931.

The construction of representative large insert DNA libraries is critical for the analysis of complex genomes. The predominant vector system for such work is the yeast artificial chromosome (YAC) system. Despite the success of YACs, many problems have been described including: chimerism, tedious steps in library construction and low yields of YAC insert DNA. Recently a new E.coli based system has been developed, the bacterial artificial chromosome (BAC) system, which offers many potential advantages over YACs. We tested the BAC system in plants by constructing an ordered 13,440 clone sorghum BAC library. The library has a combined average insert size, from single and double size selections, of 157 kb. Sorghum inserts of up to 315 kb were isolated and shown to be stable when grown for over 100 generations in liquid media. No chimeric clones were detected as determined by fluorescence in situ hybridization of ten BAC clones to metaphase and interphase S.bicolor nuclei. The library was screened with six sorghum probes and three maize probes and all but one sorghum probe hybridized to at least one BAC clone in the library. To facilitate chromosome walking with the BAC system, methods were developed to isolate the proximal ends of restriction fragments inserted into the BAC vector and used to isolate both the left and right ends of six randomly selected BAC clones. These results demonstrate that the S. bicolor BAC library will be useful for several physical mapping and map-based cloning applications not only in sorghum but other related cereal genomes, such as maize. Furthermore, we conclude that the BAC system is suitable for most large genome applications, is more 'user friendly' than the YAC system, and will likely lead to rapid progress in cloning biologically significant genes from plants.

Zhang, H.B., S.D. Choi, S.S. Woo, Z.K. Li, and R.A. Wing. 1996. *Construction and characterization of two rice bacterial artificial chromosome libraries from the parents of a permanent recombinant inbred mapping*

population. (Molecular Breeding, in press).

Rice is a leading grain crop and the staple food for over half the world population. Rice is also an ideal species for genetic and biological studies for cereal crops and other monocotyledonous plants because of its small genome and well developed genetic system. To facilitate rice genome analysis leading to physical mapping, the identification of molecular markers closely linked to economic traits, and map-based cloning, we have constructed two rice bacterial artificial chromosome (BAC) libraries from the parents of a permanent recombinant inbred mapping population (Lemont and Teqing) consisting of 400 F9 recombinant inbred lines (RILs). Lemont (japonica) and Teqing (indica) represent the two major genomes of cultivated rice, both are leading commercial varieties and widely used germplasm in rice breeding programs. The Lemont library contains 7296 clones with an average insert size of 150 kb, which represents 2.6 rice haploid genome equivalents. The Teqing library contains 14,208 clones with an average insert size of 130 kb, which represents 4.4 rice haploid genome equivalents. Three single copy DNA probes were used to screen the libraries and at least two overlapping BAC clones were isolated with each probe from each library, ranging from 45 to 260 kb in insert size. Hybridization of BAC clones with chloroplast DNA probes and fluorescent in situ hybridization using BAC DNA as probes demonstrated that both libraries contain very few clones of chloroplast DNA origin and are likely free of chimeric clones. These data indicate that both BAC libraries should be suitable for map-based cloning of rice genes and physical mapping of the rice genome.

Towards a globally integrated, sequence-ready BAC map of the human genome

Ung-Jin Kim, Hiroaki Shizuya, and Melvin I. Simon

Division of Biology, 147-75, California Institute of Technology, Pasadena, CA 91125

BACs and Fosmids are stable, non-chimeric, highly representative cloning systems. The BACs maintain large fragment genomic inserts (100-300 kb)[1] and their DNA is easily prepared for most types of experiments including DNA sequencing.[2] We have been improving BAC cloning techniques and constructed > l0X human and mouse BAC libraries. As BACs are proving to be the most efficient reagents for genomic sequencing, we intend to increase the depth of the library up to 30X genomic equivalence to be able to construct optimal contig maps from which one could select minimally overlapping BAC sets for genomic sequencing. The possibility of using BACs as a generalized tool to build a global physical map was explored in our on-going chromosome 22 mapping project. Approximately 700 mapped markers including cDNAs, ESTs, STSs, cosmids, Fosmids, and other landmarks were used to screen the first 4X library. The density of the landmarks in this approach was approximately 1 per every 50-60 kb stretch of chromosome 22q. Many of these markers have been ordered on the YAC-based framework map,[3] allowing rapid and precise localization of BAC contigs along the long arm of chromosome 22. Over 80% of the chromosome has been covered by BACs that have been identified and mapped to corresponding loci by markers. We currently have more than 1,000 chromosome 22-specific BACs, or on the average 3X coverage of chromosome 22q, which are now being characterized by restriction fingerprint analysis and the extent of overlaps between the clones in the contigs determined. Closure of gaps is being sought by screening deeper BAC library with markers and BAC end probes.

Currently large numbers of human genes that have been discovered and exist in the form of sequence-tagged cDNAs or ESTs are being assigned to genomic subregions via YACs and radiation hybrids. Because the landmarks from the YAC framework map have allowed rapid assembly of BAC maps on the chromosome 22q arm, it is feasible to employ the ESTs from the radiation hybrid/YAC frameworks as landmarks and rapidly assemble BACs to generate genome-wide BAC contig maps. Approximately 30,000 such landmarks will correspond to a density of 1 landmark in less than 100 kb of euchromatin. We are planning to utilize initially 30,000 mapped ESTs or cDNAs to construct BAC contigs on the entire genome. The resulting BAC-EST maps, even before its completion, will provide high resolution EST (or gene) maps, and more importantly, entry points for gene finding and large scale genetic sequencing.

* Supported by a Department of Energy grant # FG0389ER60891.

[1] Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y., and Simon, M.I. (1992) *Proc.Natl. Acad. Sci. USA* **89** , 8794-8797.
[2] Kim, U.-J., Birren, B.W., Yu-Ling Sheng, Tatiana Slepak, Valena Mancino, Cecilie Boysen, Hyung -Lyun Kang, Melvin I. Simon, and Hiroaki Shizuya, submitted.
[3] Collins, J.E. et al. (1995) *Nature* , in press.

Abstract scanned from text submitted for January 1996 DOE Human Genome Program Contractor-Grantee Workshop.

Progress Towards the Construction of BAC Libraries from Flow Sorted Human Chromosomes*

Jonathan L. Longmire. Nancy C. Brown, Deborah L. Grady, Evelyn W. Campbell, Mary L. Campbell, John J. Fawcett, Phil Jewett, Robert K. Moyzis, and Larry L. Deaven

Life Sciences Division and Center for Human Genome Studies, Los Alamos National Laboratory, Los Alamos, NM 87545

Over the course of the National Laboratory Gene Library Project (NLGLP) we have constructed a series of DNA libraries from flow sorted human chromosomes. Small insert, complete digest libraries cloned into the EcoRI site of Charon 21 A are available from the American Type Culture Collection, Rockville, MD. Partial digest libraries cloned into cosmid (sCos1) or phage (Charon 40) vectors have been constructed for chromosomes 4, 5, 6, 8, 9,10,11,12, 13,14,15,16,17, 20, X and Y. Purity estimates by *in situ* analysis of sorted chromosomes, flow karyotype analysis, and plaque or colony hybridization indicate that most of these libraries are 90-95% pure. Additional cosmid library constructions, 5-10X arrays of libraries into microtiter plates, and high density membrane arrays of libraries are in progress. We have also constructed a limited number of human chromosome-specific YAC libraries. In addition, we have constructed chromosome-specific M13 or pBluescript libraries for generating STS markers and for selection of chromosome-specific inserts from total genomic YAC libraries.

Because of the advantages of large insert size and stability associated with BAC cloning systems, we are currently attempting to adapt the pBelloBAC vector for use with flow sorted human chromosomes. The technical challenges involved in accomplishing this goal include developing methodologies that will allow predictable partial digestion of very small masses of DNA embedded in agarose plugs and improving BAC cloning efficiencies to allow construction of libraries from microgram quantities of chromosomal DNA. Currently, we are making modifications to pBelloBAC that include adding *Sac* ll and *Cla* l restriction sites into the cloning region of the vector. In addition, we have modified and significantly increased the efficiency of methods that are used to recover flow sorted chromosomes into agarose plugs prior to DNA isolation. These improvements together with new methods enabling partial digestion of chromosomal DNA samples (in progress) could allow the construction of BAC libraries from flow sorted human chromosomes.

*This work was supported by the USDOE under contract W-7405-ENG-36.

Abstract scanned from text submitted for January 1996 DOE Human Genome Program Contractor-Grantee Workshop.

Large Human and Mouse PAC Libraries for Physical Mapping and Genome Sequencing, and More Versatile Cloning Vectors

Joe Catanese[1], Baohui Zhao[1], Eirik Frengen[1], Chenyan Wu[1], Xiaoping Guan[1], Chira Chen[1], Eugenia Pietrzak[1], Panayotis A. Ioannou[2], Julie Korenberg[3], Joel Jessee[4] and Pieter J. de Jong[1]

[1]Department of Human Genetics, Roswell Park Cancer Institute, Buffalo, NY 14263, [2]The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus, [3]Cedar Sinai Medical Center, Los Angeles, CA 90048, [4]Life Technologies, Gaithersburg, MD 20898.

Recently, we have developed procedures for the cloning of large DNA fragments using a bacteriophage P1 derived vector, pCYPAC1 (loannou et al. (1994), Nature Genetics 6: 84-89). A slightly modified vector (pCYPAC2) has now been used to create a 15-fold redundant PAC library of the human genome, arrayed in more than 1,000 384-well dishes. DNA was obtained from blood lymphocytes from a male donor. The library was prepared in four distinct sections designated as RPCI-1, RPCI-3, RPCI-4 and RPCI-5, respectively, each having 120 kbp average inserts. The RPCI-1 segment of the library (3X; 120,000 clones, including 25% non-recombinant) has been distributed to over 40 genome centers worldwide and has been used in many physical mapping studies, positional cloning efforts and in various large-scale DNA sequencing enterprises. Screening of the RPCI-1 library by numerous markers results in an average of 3 positive PACs per autosome-derived probe or STS marker. In situ hybridization results with 250 PAC clones indicate that chimerism is low or non-existing. Distribution of RPCI-3 (3X, 78,000 clones, less than 1% non-recombinants, 4% empty wells) is now underway and the further RPCI-4 and -5 segments (< 5% empty wells) will be distributed upon request. To facilitate screening of the PAC library, we have provided the RPCI-1 PAC library to several screening companies and non-commercial resource centers. In addition, we are now distributing high-density colony membranes at cost-recovery price, mainly to groups having a copy of the PAC library. The combined RPCI-1 and -3 segments (6X) can be represented on 11 colony filters of 22x22 cm, using duplicate colonies for each clone. We are currently generating a similar PAC library from the 129 mouse strain.

To facilitate the additional use of large-insert bacterial clones for functional studies, we have prepared new PAC & BAC vectors with a dominant selectable marker gene (the blasticidin gene under control of the beta-actin promoter), an EBV replicon and an "update feature". This feature utilizes the specificity of Transposon Tn7 for the Tn7att sequence (in the new PAC and BAC vectors) to transpose marker genes, other replicons and other sequences into PACs or BACs. Hence, it facilitates retrofitting existing PAC/BAC clones (made with the new vectors) with desirable sequences without affecting the inserts. The new vector(s) are being applied to generate second generation libraries for human (female donor), mouse and rat.

Supported in part by grants from the Offfice of Health and Environmental Research of the U.S. Department of Energy (#DE-FG02-94ER61883) and the National Center for Human Genome Research, National Institutes of Health (#1R01RG01165).

Abstract scanned from text submitted for January 1996 DOE Human Genome Program Contractor-Grantee Workshop.

BACs, PACs and the Structure of the Human Genome

J. R. Korenberg, X-N. Chen, S. Mitchell, Z. Sun, E. Vataru, U-J. Kim[1], P. de Jong[3], M. Simon[1], T. J. Hudson[2], B. Birren[2], E. Lander[2], J. Silva[2], X. Wu[2].

Cedars-Sinai Research Institute, Los Angeles, CA.

Not all that glitters is single copy sequence. In order to study genome organization and to provide an integrated, genomic framework for gene isolation, sequencing and mapping, we have established a Mapped BAC/PAC Resource. The goal is to represent unequivocally, 0.8-1.2X of the human genome in a stable framework resource, integrated at 1-5,000 loci with the RH, genetic and STS maps.

Current Resource: Human
The current Mapped BAC/PAC Resource now defines 4,300 sites, and represents about 18% of the human genome. We have assigned 4,000 of 17,000 BAC/PAC clones, including 3750 BACs and 250 PACs, to regions of 2-6 Mb by using fluorescence in situ hybridization, and have integrated 91 BACs with the genetic, YAC/STS, and radiation hybrid (RH) maps by using PCR of 1,000 markers to screen the 17,000 BACs. More than 250 sites are non-tandem repetitive sequence sites; 264 BAC/PACs recognize alpha satellite sites, of which 143 were selected with a consensus alpha oligonucleotide, 102 are specific to a single chromosome and the totality of all alpha-BACs now recognize all chromosomes except 10 and the Y. Finally, BACs selected by a TTAGGG consensus oligonucleotide recognize 18 telomeres, 5 of which are specific to a single chromosome.

Information on the Resource, is available on the WWW site *http://www.csmc.edu/genetics/korenberg/korenberg.html* that includes request forms and agreements. To facilitate distribution, screening, and aneuploidy applications, 2,902 BACs were rearrayed to reflect the true chromosomal organization, from chromosome 1p through 22q.

Mouse
Using high resolution techniques, 100 BACs have been mapped to single bands in the mouse genome.

Human Disease and Genome Organization
Analysis of the 227 BACs on chromosome 7 suggests a novel genomic structure involving clustered low-copy repetitive sequences whose arrangement likely predisposes to the deletions responsible for Williams syndrome. Similar clustering of other subsets of BACs suggests that this structure may be a model for the existence of additional subsets of low copy interspersed repeated sequences that account not only for deletions responsible for human disease syndromes but also for a subset of somatic deletions and rearrangements responsible for cancers.

The Mapped BAC /PAC Resource now provides rapid approaches to genome organization and a rapidly integrated and flexible framework for mapping and sequencing the human genome.

[1] Caltech, Pasadena, CA;

[2] Whitehead Institute/MIT, Boston, MA;

[3] Roswell Park Cancer Institute. Buffalo. NY.

Abstract scanned from text submitted for January 1996 DOE Human Genome Program Contractor-Grantee Workshop.

Dear Human Genome Researchers,

Our laboratory has developed strategies to bridge contig gaps occurring at human genomic regions which cannot be cloned and/or maintained faithfully in bacteria using the large cloning systems (PI/BAC/PAC,etc). Such strategies derive from the original HAEC technology which allows direct cloning of DNA in human cells as Human Artificial Episomal Chromosomes (Nature Genetics 8:33-41, 1994; Methods in Molecular Genetics 8:167-188, 1996).

We have now entered the phase of testing such technology for proof-of-concept. We will provide the human genome community with a resource to isolate and sequence refractory contig gaps. Hence, we are interested in hearing from the human genome community about persistent and refractory genomic gaps in order to identify potential candidates for the HAEC bridging technology. Suitable candidates are contig gaps which have been thoroughly tested on highly redundant large bacterial and yeast libraries and also shown to be missing in libraries prepared with different restriction enzymes.

Interested human genome laboratories should send their email reply to the attention of Jean-Michel Vos at the following address: angemari@gibbs.oit.unc.edu.

Dr. Jean-Michel H. Vos 349 Lineberger Comprehensive Cancer Center School of Medicine CB#7295 University of North Carolina at Chapel Hill Chapel Hill, NC 27599-7295 Fax: 919-966-3015
Vos lab web site : http://www.med.unc.edu/wrkunits/3ctrpgm/lccc/voslab

---

This HAEC research is sponsored by the Department of Energy Human Genome Program.

Human Artificial Episomal Chromosome (HAECS) for Cloning, Shuttling and Functional Assay of Large Genetic Units in Human and Rodent Cells*

Min Wang, Panayotis A. Ioannou, Michael Grosz, Subrata Banerjee, Evy Bashiardes**, Michelle Rider, Tian-Qiang Sun*** and Jean-Michel H. Vos*****

Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC [Vos: 919 966-3036 (Phone); 919 966-3015 (Fax) and *vos@tmed.unc.edu* (E-mail)].

Of some 100,000 human genes, only a few thousand have been cloned, mapped or sequenced so far. Much less is known about other chromosomal regions such as those involved in DNA replication, chromatin packaging, and chromosome segregation. Construction of detailed physical maps is only the first step in localizing, identifying and determining the function of genetic units in human cells. Studying human gene function and regulation of other critical genomic regions that span hundreds of kilobase pairs of DNA requires the ability to clone an entire functional unit as a single DNA fragment and transfer it stably into human cells.

We have developed a human artificial episomal chromosome (HAEC) system based on latent replication origin of the large herpes Epstein-Barr virus (EBV) for the propagation and stable maintenance of DNA as circular minichromosomes in human cells.[1,2] Individual HAECS carried human genomic inserts ranging from 60 to 330 kb and appeared genetically stable. An HAEC library of 1500 independent clones carrying random human genomic fragments with average sizes of 150 to 200 kb was established and allowed recovery of the HAEC DNA. This autologous HAEC system with human DNA segments directly cloned in human cells provides an important tool for functional study of large mammalian DNA regions and gene therapy.[3,4]

Current efforts are focused on (a) shuttling large BAC/PAC genomic inserts in human and rodent cells and (b) packaging BAC/PAC/HAEC clones as large infectious Herpes Viruses for shuttling genomic inserts between mammalian cells and (c) constructing bacterial-based human and rodent HAEC libraries. (a) We have designed a "pop-in" vector, which can be inserted into current BAC-or PAC-based clone via site-specific integration. This "CRE-LOXP"-mediated system has been used to establish BAC/PAC up to 250 kb in size in human cells as HAECS. (b) We have obtained packaging of 160-180 kb exogenous DNA into infectious virions using the human lymphotropic Epstein-Barr virus. After delivery into human beta-lymphoblasts cells the HAEC DNA was stabley established as 160-180 kb functional autonomously replicating episomes.[5,7] We have also generated a hybrid BAC/HAEC vector, which can shuttle large DNA inserts, i.e., at least up to 260 kb, between bacteria and human cells. Such a system is being used to develop large insert libraries, whose clones can be directly transferred into human or rodent cells for functional analysis. These HAEC-derived systems will provide useful molecular tools to study large genetic units in humans and rodents, and complement the functional interpretation of current sequencing efforts.

* Supported by the Office of Health and Environmental Research, Human Genome Program, Department of Energy, under Contract No. DE-FG05-91ER61135

** The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus.

*** Department of Biochemistry and Biophysics; University of North Carolina, Chapel Hill, North Carolina 27599.

[1] Sun, T.-Q., Fenstermacher, D. & Vos, J.-M.H. Human artificial episomal chromosomes for cloning large DNA in human cells *Nature Genet* **8** , 33-41 (1994).

[2] Sun, T.-Q. & Vos, J.-M.H. Engineering of 100-300 kb of DNA as persisting extrachromosomal elements in human cells using the HAEC system in *Methods molec. Genet.* (ed. Adolph, K.W.) (Academic Press, San Diego, CA, 1995).

[3] Vos, J.-M.H. Herpesviruses as Genetic Vectors in *Viruses in Human Gene Therapy* (ed. Vos, J.-M.H.) 109-140 (Carolina Academic Press & Chapman & Hall, Durham N.C., USA & London, UK, 1995).

[4] Kelleher, Z. & Vos, J.-M. Long-Term Episomal Gene Delivery in Human Lymphoid Cells using Human and Avian Adenoviral-assisted Transfection. *Biotechniques* **17** , 1110-1117 (1994).

[5] Banerjee, S., Livanos, E. & Vos, J.-M.H. Therapeutic Gene Delivery in Human beta-lymphocytes with Engineered Epstein-Barr Virus. *Nature Medicine, Accepted*

[6] Sun, T.-Q., Livanos, E., & Vos, J.-M.H. Infectious HAECS for Disease Correction. *Nature Medicine* , Submitted.

[7] Wang, S. & Vos, J.-M.H. An HSV/EBV based vector for High Efficient Gene Transfer to Human Cells in vitro/in vivo. *Submitted*

.

Abstracts scanned from text submitted for January 1996 DOE Human Genome Program Contractor-Grantee Workshop.