Report from the Second IMAGE/Full-Length cDNA Consortium Meeting

John Quackenbush
15 September 1997
Hilton Head, SC, USA

The purpose of this second Consortium meeting was to continue to coordinate effort among groups committed to generating a publicly available Full-Length cDNA sequence resource. This would build on the success of the I.M.A.G.E. Consortium [http://image.llnl.gov/] EST Sequencing Program in avoiding duplication of effort.

This meeting followed a successful meeting held in May following the Cold Spring Harbor Genome Sequencing and Mapping Meeting. [**Note:** For additional relevant links and background material on complete cDNA sequencing, see the report of the first Workshop on Complete cDNA Sequencing (May 19, 1997).] At the May consortium meeting, the various participants outlined their efforts, which largely consisted of pilot projects. There was universal recognition of the value of the continuing full-insert and full-length cDNA sequencing efforts, but most groups had little, if any funding support for their projects.

Attending this meeting were representatives of both academic and commercial cDNA sequencing groups. Group representatives presented summaries of their work, but these presentations were interspersed with general discussion of a variety of issues, including the distinction between full-length and full-insert sequencing projects, the need for coordination to avoid duplication, and the best manner in which to implement a coordination/registration database.

Participant Reports:

Joakim Lundeberg, Royal Institute of Technology, Stockholm
Dept. of Biochemistry
joakim.lundeberg@biochem.uth.se
Funding for the European cDNA sequencing project began September 1, 1997. A total of 400,000 ECUs has been provided to each of 8 institutions. Participating institutions (see EURO-IMAGE below) will generate 8Mb of cDNA sequence spanning the full inserts of cDNA clones. There are no plans at present to create full-length cDNA libraries or to try to identify full-length clones. There are also no universal plans to coordinate sequencing at present, although Charles Auffrey is working with a number of groups to coordinate efforts. (After reading a draft of this report, Auffrey sent a description of the EURO-IMAGE project.)

Michael Rhodes, UK-HGMP Resource Center
mrhodes@hgmp.mrc.ac.uk
Using Washington University's icatools, the UK-HGMP clustered all the IMAGE clone cDNA sequences and 60,000 clones believed to be unique were identified. These are currently being rearrayed as a unique collection that will be made publicly available. However, the HGMP does not presently have funds to sequence these clones. Rather, they will attempt to coordinate with the European Consortium to have these clones sequenced.

Karl Guegler, Incyte Pharmaceuticals
karl@qmgate.incyte.com
Incyte is continuing to sequence ESTs as well as full-length clones. They are currently considering a plan to release 3' and 5' sequence data and issuing public notifications of clones that are about to be fully sequenced. Most of this will occur through Genome Systems, Inc., which was recently purchased by Incyte.

David Smoller, Genome Systems, Inc.
dave@genomesystems.com
Genome Systems has clustered the IMAGE sequences, producing 23,000 clusters and 33,000 singletons. They are now

identifying the longest 5' clones from each of the clusters and plan to rearray approximately 18,000 of these long clones and provide both clones and filters at a relatively low cost. In addition to the human clones, they are also rearraying *Drosophila* clones.

Richard Gibbs, Baylor College of Medicine
agibbs@bcm.tmc.edu
Baylor is continuing to sequence full-insert cDNA clones using the concatenation strategy that they developed. The goal is produce 1000 full-insert cDNA sequences; 150 have already been completed and submitted to GenBank and another 700 are in the pipeline from IMAGE plates 75 and 76.

LaDeana Hillier and Jeff Woessner, Washington University
lhillier@watson.wustl.edu
jwoessne@watson.wustl.edu
Wash U is attempting to sequence the full inserts from clones representing the UniGene clusters, but avoiding those clusters for which existing complete mRNA sequence exists. They are choosing clones with inserts smaller than 800 bp in size so that the clones can be sequenced completely sequenced using long reads from the clone 3' and 5' ends. At present, they have identified 8060 and have 400 completed that they are preparing for submission to GenBank. They will post the IMAGE IDs of the clones they plan to sequence on the Washington University web site. Wash U has also secured Howard Hughes funding to use the same strategy to sequence one representative full-insert clone from each unique Mouse cDNA cluster.

Stefan Weimann, DKFZ
s.weimann@dkfz-heidelberg.de
As part of the European Consortium project, DKFZ has begun sequencing of 384 clones larger than 1.5 kb (and an average insert size of approximately 2kb). They have also constructed 4 new libraries, 3 using the Clonetech CAPFINDER technique and 1 using "standard" cDNA library construction protocols.

Horst Domedy, Genzentrum, Muenchen
domedey@emb.uni-muenchen.de
As part of the European cDNA consortium, Genzentrum will generate 15,000 ESTs as well as 1Mb of additional cDNA sequence, focusing on Chromosome 21-specific cDNAs. These will be make publicly available through the German Genome Resource Center.

Babru Samal, Amgen, Inc.
bsamal@amgen.com
Amgen is looking at signal peptide-producing genes in mouse, identifying them through EST sequencing and converting them to full-length sequences using RACE/Clonetech Marathon protocols. However, there are at present no plans to make either the clones or their sequence public.

Takao Isogai and Toshio Ota, Helix Research Institute, Japan
isogai@hri.co.jp
ota@hri.co.jp
Helix and Dr. Sugano of the University of Tokyo, have constructed libraries with enhanced representation of full-length clones using an oligo-capping method and full-length sequencing has begun using these libraries. Neither the libraries nor their sequences are publicly available at this time, but they are constructing a plan to release both. Dr. Nomura's group at the Kazusa Research Institute, Japan, is sequencing approximately 2Mb of cDNA sequence per year representing about 300 cDNA clones. Dr. Sugano of Tokyo is also constructing mouse full-length cDNA libraries and those will be publicly available.

Giorgio Valle, University of Padua, Italy
valle@eos.bio.unipd.it
Padua has done their own sequence clustering. From the unique clusters, they have mapped approximately 500 clones on the GenBridge 4 Radiation Hybrid Mapping panel, and are attempting to obtain expression data for these transcripts. They are collaborating with Charles Auffrey to identify candidate full-length clones using a PCR strategy to screen 20,000 pooled clones in to identify the longest representative of these clusters.

Robert Cottingham, Genome Database
bc@gdb.org
As IMAGE clone libraries are registered, Greg Lennon submits these to GDB, so all IMAGE clones have GDB accession numbers (GDBids). In order to cross-reference mapping and sequence data, GDB is establishing links between dbEST and cDNAs in GDB. This effort would be greatly simplified if EST sequence submissions to dbEST included the GDBid for the corresponding cDNA clone in GDB. A suggested submission protocol was provided.

Levy Ulanovsky, Argonne National Laboratory
Argonne National Laboratory
levy@anl.gov
Ulanovsky and collaborators at ANL have been applying their DENS Oligowalking approach to the sequencing of cDNA clones. This technique has promise to reduce the cost of primer walking sequencing approaches by eliminating the need for custom primers.

John Quackenbush, The Institute for Genomic Research (TIGR)
johnq@tigr.org
TIGR has been working on a number of approaches to creating full-length libraries. The 5' CAP trapping protocol described by Carninci et al. (Genomics 37, 1996) has proven difficult to reproduce so a number of other approaches are being attempted. In order to begin generating full-length cDNA sequence, the TIGR Tentative Human Consensus (THC) assemblies are being used to identify candidate full-length clones. At present, 50 have been identified and completely sequenced, producing 37 full-length sequences representing 36 unique transcripts. An additional 13 full-insert sequences have also been completed but represent only partial transcripts.

Additional Participants:

Marvin Stodolsky
US Department of Energy
marvin.stodolsky@oer.doe.gov

Andreas Duesterhoeft
Qiagen GmbH
a.duesterhoeft@qiagen.de

Funding

In Europe, the European Consortium has been provided funds to continue and expand cDNA sequencing. In the US and Japan there is far less public funding for cDNA projects. However, the US National Cancer Institute and the Merck Genome Research Institute both have plans to fund projects for full-length cDNA library protocol development.

General Discussion

Following the progress reports, there was a general discussion of some of the important issues facing groups involved in full-length cDNA sequencing. The discussion identified the following important areas:

1. The need for full-length libraries.
2. The need for additional funding for full-length and full-insert sequencing.
3. An understanding of the various clustering methods and their relationships.
4. Coordination of sequencing efforts in order to avoid duplication.

While all present agreed that we should continue to work on the first three items, it was clear that it was appropriate to create an allocation database similar to the RHalloc database at EBI used to coordinate Radiation Hybrid Mapping. The following data were suggested as necessary for each clone registered in such a cDNA allocation database: clone IMAGE ID, 3' and 5' sequence Accession Numbers (extracted from dBEST), the estimated clone length, whether the clone is full-length or simply sequenced for its full insert, the date the clone was registered, and a status. Allocations would continue until the full-insert sequence was submitted to dBEST, at which time the status would change from "allocated" to

"finished"; if an allocated clone is not sequenced in a timely manner, the status should change to "expired". There was some discussion about how long a registration should be allowed to remain without the clone being finished before it expires. The consensus was that six months should be sufficient to finish sequencing. The other issue that was not resolved was the number of clones that groups should be allowed to register at any one time. Most felt that 384 clones was a good initial allocation size. Bob Cottingham agreed to contact Particia Rodriguez-Tomé of the EBI to discuss the possibility of establishing such a database.

While still at Hilton Head, Dr. Rodriguez-Tomé contacted me to ask that I let the IMAGE consortium know that she would be happy to establish such a database provided that funding be made available to support it. Dr. Rodriguez-Tomé was preparing a grant application to support a similar genomic sequencing registration database, and she felt that a full-length cDNA sequencing database would be a natural extension of her application. As the application was due immediately following her return to the EBI from Hilton Head, she asked me to provide a letter of support on behalf of the IMAGE Consortium.

Next Meeting

It was agreed that the next logical time to meet would be in May following the Cold Spring Harbor Meeting; LaDeana Hillier of Washington University agreed to coordinate that meeting.

Return to cDNA Sequencing

WCCS I

Return to Genome-Related Meetings

cDNA Sequence Submission Guidelines from the GDB

The purpose of this submission procedure is to ensure that the relevant information appears in dbEST and GDB without requiring submission to both databases. In summary the submission procedure consists of a usual submission to dbEST where the "GDB#" field in the EST File is included. The steps are:

1. Get the GDB accession id (GDBid) for the cDNA clone.

2. Create a dbEST submission including the GDBid in the GDB# field of the EST File.

3. Submit to dbEST.

By including the GDB# field in the dbEST submission, the two databases can then be linked and the information automatically integrated without having to submit separately to each database.

Submission procedure details

Each of the steps above is further explained below:

1. GDBids can be obtained for cDNA clones by using the GDB Web interface. Go to *http://www.gdb.org/*. Enter the cDNA clone name, for example "IMAGE:338479", into the Simple Search form and press Submit. This returns detail information about the clone including the Accession ID which in this case is GDB:1263853.

Alternatively, there are several programmatic methods for querying the database (contact help@gdb.org for more information). Also there is a flat file at *ftp://ftp.gdb.org/outgoing/cDNA.Z* which contains a list of all cDNA clones by GDBid and name. This file can be downloaded and used locally.

Please contact data@gdb.org if the GDBid for the cDNA clone can not be found.

2. Detailed instruction about creating a dbEST submission file can be found at
http://www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html

In order to link the two databases, the "GDB#" field in the EST File must contain the GDB accession id (GDBid) for the cDNA clone.

The correct syntax for the GDB# field in the example given above of IMAGE:338479 is: GDB#: GDB:1263853

3. Once the submission file is prepared, send it to batch-sub@ncbi.nlm.nih.gov

Please contact data@gdb.org if there are any questions. We will be happy to assist with your submission.

Letter in Support of the EBI Application

Particia Rodriguez-Tomé
EMBL Outstation
European Bioinformatics Institute
The Wellcome Trust Genome Campus
Winxton, Cambridge CB10 1SD
United Kingdom

Dear Particia Rodriguez-Tomé,

At the recent Hilton Head Genome Sequencing and Analysis Conference, members of the IMAGE Consortium and other interested individuals met to discuss plans to sequence full-length and full-insert cDNA clones. Participants represented groups from the European Community, Japan, and the United States that have begun projects to generate sequence data spanning complete cDNA clones. These projects will not only provide a much needed resource for the annotation of the emerging human genomic sequence, but will also be extremely valuable for the identification and functional classification of the complement of human genes.

The motivation for our meeting was to build on the success of the IMAGE Consortium in freely providing both EST sequence data and physical clones to the general biological. As we discussed our projects, one common theme that emerged was the need for coordination of efforts to avoid sequencing of multiple clones likely to represent the same human transcript. The participants agreed that an allocation database, similar to the RHalloc database that you and Kate Rice developed and maintained at the EBI, would be the most efficient way to avoid unnecessary and costly duplication of effort.

An allocation database would allow participants to announce their intentions to sequence particular human cDNA clones and should include, at a minimum, the clone IMAGE ID, the 5' and 3' dBEST Accessions, the target species, the clone length (if available), whether this project was part of a full-length or a full-insert sequencing effort, and contact information for the group conducting the sequencing. While registration of a clone in such a database would not preclude another group from sequencing that or a related clone, it would allow both large and small cDNA sequencing groups to make intelligent decisions about the clones they plan to sequence. We would expect groups that had allocated any particular clone to complete sequencing and submit the sequence data to dBEST, at which time the allocation status should change from "reserved" to "completed"; if a group failed to sequence any reserved clone in a timely manner, such as six months after registration, its status should change to "expired" to allow other interested parties to allocate and sequence it.

We realize that building and maintaining such a database is not a trivial task, but without it, valuable funds and resources are likely to be lost to duplicated effort. We hope that you will be able to establish such a database at the EBI and we would be strongly support your efforts to secure funding to both build and maintain it..

Yours sincerely,

John Quackenbush, Ph.D.
The Institute for Genomic Research