AGI: Data Release Policies

This page reproduces the text of a message sent on July 17, 1997 to arab-gen, the Arabidopsis bulletin board (newsgroup) by David Meinke, Chair of the Science Steering Committee of the Arabidopsis Genome Project. Includes hot-links to Web sites and e-mails addresses.

Dear Colleagues:

Attached please find **the data release policies for current participants in the Arabidopsis Genome Initiative (AGI)**. Each participant was requested to answer basic questions (shown below) concerning release of genomic sequence data. The Science Steering Committee for the Multinational Arabidopsis Genome Project requested this information to help members of the community become better informed and to encourage the most rapid release of data possible.

If you have any comments on the release policies outlined here, please direct them to members of the steering committee or post them to this Arabidopsis newsgroup.

Sincerely,

David Meinke, Chair
Science Steering Committee
Arabidopsis Genome Project

QUESTIONS ADDRESSED TO AGI PARTICIPANTS:

A. What role are you presently playing in the Arabidopsis Genome Initiative?

B. What is your policy concerning release of AGI sequence data?

C. What are the primary reasons (scientific and/or political) behind this policy?

D. Do you anticipate any changes in your policy in the future?

E. If you have a substantial amount of data not yet released, when are you committed to making this information available?

F. Will there be a similar delay in the release of future data?

G. What is the Internet address where your data can be accessed?

H. Any additional comments you wish to make with respect to data release policies for the AGI are welcome.

ANSWERS TO QUESTIONS ARE LISTED IN THE FOLLOWING ORDER:

Japanese Sequencing Program (Satoshi Tabata)
European Sequencing Program (Michael Bevan)
French Sequencing Program (Marcel Salanoubat)
SPP Consortium, Univ. Pennyslvania Component (Joe Ecker)
SPP Consortium, PGEC Component (Sakis Theologis)
SPP Consortium, Stanford Univ. Component (Nancy Federspiel)
TIGR Sequencing Program (Steve Rounsley)
Cold Spring Harbor, Washington Univ. Consortium (Dick McCombie)

I. JAPANESE SEQUENCING PROGRAM

Response from Satoshi Tabata (tabata@kazusa.or.jp)

A. We are currently sequencing portions of chomosome 5.

B. We will be releasing the finished sequences with full annotation within six months after sequencing began.

C. 1) We believe that the finished accurate sequence is most useful to users after all, though we recognize some usefulness of unassembled random sequence data for marker construction and gene finding.

2) We do not like the sequencing process to be complicated by putting an additional task in it, because we need to control the whole process of large-scale sequencing very carefully to achieve maximum performance.

3) Consistent and informative annotation would undoubtedly be very useful to users, and this process delays the data release for only two to three weeks.

D. No changes anticipated. By improving the sequencing and annotation system, the data release will be within four to six months after the start of sequencing.

E. The release will be in the middle of July, as I mentioned at the Arabidopsis meeting.

F. No. We just began preparaion for the second release that will be on September 1st.

G. KAOS (Kazusa Arabidopsis data Opening Site) at:

http://www.kazusa.or.jp/arabi/

H. Genome sequencing projects, especially of higher eukaryotes, are long-term projects, and one needs to have enough time for preparation to reach the final goal most quickly and efficiently. In our case, pilot sequencing started in February 1996 and is still under improvement. We started development of automatic annotation system in September 1996, and it is not completed yet. I am expecting that we will be able to produce up to 800 kb per month by the end of this year, but both sequencing and annotation systems have to be completed for this preformance. We are very happy feeling that people in the community are placing great expectation on us, and we would sincerely like to come up to it. So, please give us time and be tolerant a bit. Please !!

II. EUROPEAN SEQUENCING PROGRAM

Response from Michael Bevan (michael.bevan@bbsrc.ac.uk)

A. The EU network sequences on the bottom arm of chromosome 4 and provides contig information to the CSHSQ on the top arms of chromosomes 4 and 5.

B. Our policy conforms to that agreed in the Memorandum of Understanding. We make available to the public sequence that aims to be of the following standard: It is sequenced on both strands with a mixture of chemistries wherever possible. Regions in which it is difficult to obtain sequence on both strands MAY, after some additional effort, be publically released as an initial version with a "health warning". The predicted restriction map must agree, within the limits of accuracy, to the restriction fingerprint of the clone (which is anchored to the chromosome), and overlaps with adjacent clones must be checked and sequence differences identified and resolved. We are also using BAC end sequencing for overlap verifications. Initially the sequence may not be annotated using our standard protocols, but we aim to accomplish this within one month of public release. The regions sequenced as part of the first stage of the EU sequencing project were being carried out before the advent of the AGI Agreement. It was decided to maintain our focus on releasing a single contig of sequence for the main region, rather than to release the sequence of unanchored cosmid clones which were all in different states of preparedness and of unverified accuracy. This region was sent as ten 200 kb units to EMBL in early June, and these have the following accession numbers: Z97335 - Z97344 inclusive.

We have deviated from the AGI Agreement in one important respect, which requires clarification. The EU network originally agreed to number the predicted genes in order from an (arbitary) end of the clone. For the 1.9 Mb contig this proved to be a poor solution for a variety of reasons, so we have numbered the predicted genes in their known order within the contig, and in their predicted order from the centromere. It is 2.3 Mb from the distal copy of the pericentromeric repeats to the proximal extent of the published contig. We calculated this region contains about 500 genes, and our coordinates allow for a slop of 5 genes per unit (for different versions) plus another 100 genes, therefore our numbering within the contig starts at 3000, with each gene having coordinates of 3000, 3005, 3010 etc, and different versions being eg 3006, 3007 for different versions of 3005. The overall aim of this is to try to develop a chromosome-based system from an early stage. Useful comments are welcome. The AP2 region, a 450 kb contig around that gene, will soon be released when annotation is completed, hopefully within 4 weeks.

As has always been the case, anyone with a particular interest in these regions has always been helped (by providing sequence for example) to the best of our ability. The BACs sequenced in the second stage of EU sequencing that are complete, but not yet annotated, will be available from the MIPS website (see below) as soon as they are completed. There are 8 of these presently being annotated. Due to the unexpected anxiety caused by the public not being able to see these sequences before annotation, we have brought forward our release schedule for these clones and will maintain this release policy in the future. Next year we aim to release daily production sequence from some of the larger labs in the EU network.

C. The EU release policy, seen above, is an amalgamation of often conflicting requirements. We have to balance the requirements of the EC, who wish the sequence to benefit relevant sectors of industry (worldwide), with the normal scientific practice of releasing data of the highest possible standards (with available technology) and of the greatest utility, with the requirement of the public for immediate access, for a wide variety of reasons. We have worked hard to deal with these requirements and have what we hope is a satisfactory solution, at least for the present.

D. This policy will be revised from time to time, according to new developments.

E. The EU network is presently sequencing and analysing 46 BACs and P1s. 8 are completed according to the criteria described above. About 16 clones are in a "prefinal" state, ie they have between 1-10 physical gaps and are not completely double-stranded. A further 4 clones, distributed early in the project, present difficulties with respect to repeat regions that require individual attention. These will probably not be completed and available for some time.

F. It is difficult to predict the time of release of individual clones. Some clones are complete except for difficult-to-sequence regions, and we cannot predict when (if ever) these will be completed to the standards described above. We will do our best to complete these areas to high standards. These regions are, of course, often the most interesting, and subsequent analysis requires highly accurate sequence.

G. The internet addresses for EU sequence and analysis is:

http://mips.gsf.de/proj/thal/db/index.html

Here one can see the sequence of the FCA contig presented as a graphical output, with predicted genes as clickable objects leading to a database of homologies. In addition, an Arabidopsis protein sequence and structure database, also with similarity searches using Smith-Waterman methods, is available at:

http://pedant.gsf.de/

Annotated sequence, in its chromosomal context, will soon be available from:

http://synteny.nott.ac.uk/arabidopsis.html

in a more familiar ACeDB format, in addition to the usual embl format. These databases will undergo continual improvements both in the amount and type of data collected, as well as in the query tools to use, and detail of analysis. The ultimate goal of the EU plant database community is to link Arabidopsis genome sequence with that of crop plant species and with function search activities in the EU.

H. Scientists might like to begin to consider the type and degree of annotation they expect from the sequencing community, and how soon they expect to see reasonably sized contigs, precisely anchored in a chromosomal context. Finally, they might like to think about the type of databases that best suit their work, and ensure that these are supported.

III. FRENCH SEQUENCING PROGRAM

Response from Marcel Salanoubat (salanou@genethon.fr)

A. End sequence of BAC clones from the IGF and TAMU library in collaboration with TIGR.

B. Release of the data as soon as possible. The delay is caused only by the time needed to process the data.

C. Not concerned.

D. Not about the end sequence project.

E. Since we are in the begining of the project it took some time for the first release of the data. The first release will be done at the begining of this week.

F. We hope to be able to decrease the time needed for the release of the data in the future.

G. http://www.infobiogen.fr/CNS/Arabidopsis.html

The address of this site is provisional. We will have a new address in September.

H. No additional comments.

IV. SPP CONSORTIUM (U PENN COMPONENT)

Response from Joe Ecker (jecker@atgenome.bio.upenn.edu)

A. The responsibilities of the PENN group of the SPP consortium are: (1) to prepare a tiling path of overlapping BAC clones for 5.2 megabases of chromosome 1; (2) to determine the end sequences of chromosome 1 BAC clones; (3) to provide the complete finished and annotated sequences for a minimum of 5.2 Mb of chromosome 1 by 1999 (with the Stanford and PGEC groups).

B. The PENN group policy for release is immediate release of the shotgun phase and finished sequences into to Genbank. For more details see the SPP DATA RELEASE POLICY at:

http://pgec-genome.pw.usda.gov/scope.html

C. We have always felt that this policy best meets the desires of the community of Arabidopsis researchers.

As indicated by a show of hands at the 1997 Arabidopsis meeting in Madison WI, a clear majority of researchers agree with this policy. By my count, there were only two individuals out of 600 or so in the audience that felt immediate release of data was not a good policy.

D. No changes anticipated.

E. Not applicable.

F. Not applicable.

G. The web site for the PENN group is:

http://cbil.humgen.upenn.edu/~atgc/ATGCUP.html

More specifically, we present the data or links to our data for all of the following categories:

1) BAC end sequence data can be found in the Genbank dbGSS division:

http://www.ncbi.nlm.nih.gov/dbGSS/index.html

and the BAC end-sequence trace data and "blast hits" list can be viewed on our web site:

http://cbil.humgen.upenn.edu/~atgc/SPP.html

2) shotgun sequencing phase data (htgs-database) and finished BAC sequences (nr-database) can be found in Genbank:

http://www.ncbi.nlm.nih.gov/BLAST/

A graphical presentation of our finished and annotation BACs can be found at:

http://cbil.humgen.upenn.edu/~atgc/SPP.html

3) BAC mapping data for both the IGF and TAMU libraries is available on our web site:

http://cbil.humgen.upenn.edu/~atgc/physical-mapping/physmaps.html

H. No additional comments. We are too busy sequencing!

V. SPP CONSORTIUM (PGEC COMPONENT)

Response from Athanasios Theologis (theo@mendel.berkeley.edu)

A. The PGEC sequencing group as part of the SPP Consortium is responsible for the following:

1) To construct all the M13 shotgun libraries for sequencing 5.2 Mb of Chromosome 1.

2) To provide 1.7 Mb of finished and annotated sequence (1/3 of the SPP Consortium goal).

B. The PGEC policy since the beginning of the project has been immediate release of the shotgun phase and finished sequences to GenBank. For more details see our web site at:

http://pgec-genome.pw.usda.gov

C. We have always felt that the immediate release policy is the most beneficial for the advancement of Plant Biology and Agriculture worldwide.

D. No changes anticipated.

E. Not applicable.

F. Not applicable.

G. The web site for the PGEC group is:

http://pgec-genome.pw.usda.gov

More specifically, we present the data or links to our data as follows:

1) Libraries constructed and planned to be constructed can be seen on our web site:

http://pgec-genome.pw.usda.gov

2) Shotgun sequencing phase data (htgs database) and finished BAC sequences (nn database) can be found in GenBank:

http://www.ncbi.nlm.nih.gov/BLAST

3) A graphical presentation of our finished and annotated BACs can also be found at:

http://pgec-genome.pw.usda.gov

H. No additional comments.

VI. SPP CONSORTIUM (STANFORD COMPONENT)

Response from Nancy Federspiel (nfeder@sequence.stanford.edu)

A. The responsibilities of the Stanford group of the SPP Consortium are:

1) To plate libraries, pick clones, and prepare template DNA for all members of the SPP Consortium.

2) To provide the complete finished and annotated sequences for a minimum of 5.2 Mb of chromosome 1 by 1999 (with the Penn and PGEC groups).

B. The Stanford policy is immediate release of shotgun phase and finished sequences to GenBank.

C. We feel that this policy best meets the needs of the scientific community.

D. No changes anticipated.

E. Not applicable.

F. Not applicable.

G. Internet address:

http://sequence-www.stanford.edu/ara/ArabidopsisSeqStanford.html

H. No additional comments.

VII. TIGR SEQUENCING PROGRAM

Response from Steve Rounsley (rounsley@tigr.org)

A. TIGR is participating in the AGI by sequencing (via the shotgun sequencing method) and annotating BAC clones from chromosome II. In addition to this, we are generating BAC end sequences from the entire TAMU and IGF libraries in collaboration with the Centre National de Sequencage in France.

B. After generating shotgun sequences for a given BAC clone, an assembly is performed and contigs longer than 2kb generated at this stage are made publicly available. Currently this is via our ftp site, but we are working with NCBI to make them available in the HTGS division of GenBank. This will make them available for searching via the BLAST server at NCBI. After this initial release, the process of closing gaps begins, followed by annotation. The annotated BAC sequence is submitted to NCBI and appears in the Plant division of GenBank.

BAC end sequences are made available immediately after processing. They are available on our ftp site, and on our web site. The web site offers a search capability with a graphical view of the results.

C. We believe the release of assembled contigs after the shotgun sequencing is complete provides useful sequence data to the community but also provides enough time to provide some quality control over the data. However preliminary data is still just that and should be treated with caution, as contig order and structure can change during the finishing process. We are still committed to providing high quality annotation for the sequence when it is complete.

D. The only change that will occur in the foreseeable future is to have the preliminary sequence available at NCBI. This should happen within the next month. Until then, they are available on the ftp site.

E. The only sequence data that are not available are the shotgun sequences themselves that accumulate during the sequencing phase. Typically it takes less than a week to generate these sequences and then the initial assembly is performed. The contigs that result from this assembly are made available.

F. Future BAC sequences will be made available as outlined above.

G. Annotated Sequence data is available on the web:

http://www.tigr.org/tdb/at/atgenome/atgenome.html

Current status of any given clone can be found at:

http://www.tigr.org/tdb/at/atgenome/at_bacs.html

BAC end sequences can be searched at:

http://www.tigr.org/tdb/at/atgenome/bac_end_search/bac_end_search.html

Preliminary sequences and BAC end sequences can be found on our ftp site:

ftp://ftp.tigr.org/pub/data/a_thaliana/

While these can be accessed through a web browser, a more reliable and more robust method to obtain them is to use anonymous ftp.

1. Connect to ftp.tigr.org
2. login as anonymous
3. password is your email address
4. type cd /pub/data/a_thaliana/

H. No additional comments.

VIII. COLD SPRING HARBOR/WASHINGTON UNIVERSITY CONSORTIUM

Response From Dick McCombie (mccombie@cshl.org)

A. We are a consortium that is targetting the sequencing of the short arm of chromosome IV and parts of chromosome V during the period of fall 1996- fall 1999. We proposed sequencing at least 6.5-7.0 megabases during this three year period.

B. Our AGI data release policy is as follows:

1) All contigs over 1.5 kilobases will be released as they are assembled from shotgun sequence data. These contigs will be available on ftp sites and BLAST servers accessible from our web sites. The sequence assemblies in our labs will be automatically analysed and this new and updated contig data transferred to the ftp and BLAST servers on a daily basis.

2) Finished, annotated sequence will be submitted to GenBank as it is completed, with no intrinsic delay or data hold.

C. The sequencing of the Arabidopsis genome is in large part a public service to the community. As such we believe that it is important to make the data available as rapidly as possible to that community. By putting the sequences on our own ftp sites and providing BLAST search capabilities at those sites we can, and do, provide daily updates of the projects in progress in our labs.

We will continue to use GenBank as a repository for finished, high quality, annotated sequence. This is in fitting with the role GenBank was created for and has historically served.

D. No changes anticipated.

E. We do not have substantial unreleased data.

F. We will continue daily updates.

G. For Cold Spring Harbor:

http://www.cshl.org/genseq

for Washington University:

http://genome.wustl.edu/gsc/gschmpg.html

H. We have been and continue to be committed to immediate data release. We welcome the acceptance of this principle by any other AGI group that has not already done so.

David W. Meinke
Department of Botany
Oklahoma State University
Stillwater, OK 74078
Phone: 405-744-6549
FAX: 405-744-7673
Email: meinke@osuunx.ucc.okstate.edu
WWW: http://mutant.lse.okstate.edu/