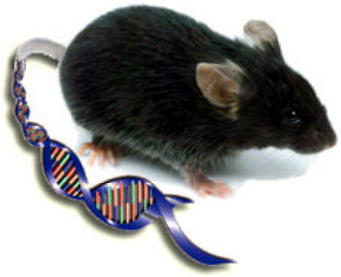
Mouse Genome Monthly



MGSC

Issue #2 December 2001

The Latest Progress From the Mouse Genome Sequencing Consortium

This is the second in a series of newsletters that are being produced on approximately a monthly schedule, with the goal of informing the scientific community about the progress on sequencing and annotating the mouse genome, and about updates on the availability of information and resources generated by this project.

Status Reports

Sequencing of the mouse genome. As of December 14, the total number of mouse whole genome shotgun traces in the Trace Archives was 31,692,551 (http://trace.ensembl.org & http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?). This corresponds to >4.4X coverage of the genome in bases of quality greater than Phred 20. Whole genome shotgun sequence production by the members of the Mouse Genome Sequencing Consortium (http://mouse.ensembl.org/) is scheduled to continue through the end of December 2001. Most of the data that will be generated during the last month of Phase 3 of the program will come from shotgun clones with either 10kb or 40kb (fosmid) inserts. The status of the assembly of the whole genome shotgun data set is discussed below.

In addition, mouse genome sequence data can also be obtained from the HTGS division of GenBank. As of December 14, there were 76,096 kb of finished non-redundant data (2.4% of the genome) and 559,376 kb of unfinished data (17.5% of the genome, but there is redundancy).

Mouse BAC map. As of December 15, the current bacterial clone map of the mouse genome consists of 539 clone contigs covering approximately 90% of the mouse genome. This map was derived from an initial set of 7,500 clone contigs assembled using the FPC program on the restriction enzyme fingerprint data obtained at the British Columbia Genome Sequence Centre. Mouse BAC end sequences obtained by The Institute for Genome Research have been used to align the mouse contigs to the human genome to accelerate the manual joining process. The synteny information has not been used, however, to create joins. 6,800 markers from radiation hybrid and genetic maps of the mouse have been integrated into the database to confirm contig order and orientation. Further refinement of the map is in progress at the Wellcome Trust Sanger Institute and Washington University Genome Sequencing Center. For more details see: http://mouse.ensembl.org/Mus_musculus/fpcview?chr=chr1&vc_start=1&vc_end=100000&x=32&y=13.

Other map information. NCBI provides several mouse maps (genetic and RH) in its Map Viewer, and is also showing a human-mouse comparative map. Human-mouse comparative information is also available on the UCSC browser and at Ensembl.

Mouse genome assemblies. Periodically, the mouse whole genome shotgun data set is assembled with both of two whole genome assemblers that are currently under development. Phusion has been developed by the Sanger Institute and Arachne has been developed by the Whitehead Institute; the on-going evaluation and development of these programs is being done by the two groups in collaboration. Several of the individual assemblies have been made available for searching and/or for download. *However, users are reminded that, at present, the sequence data are not ideally assembled and not all contaminants have been removed.* For sequence searching, go to: http://mouse.ensembl.org/Mus_musculus/ssahaview/, where the Phusion assembly of November 6 and the Arachne assembly of October 26 may be searched by the SSAHA Search Server. Searches can also be done against the database of finished and draft sequenced mouse clones.

Based on a November 9 freeze of the trace data, both assemblies give over 90% coverage of the mouse genome at the contig level, and over 95%

coverage at the ordered and oriented supercontig level. For both assemblies, the N50 contig size is about 10kb (i.e. half the nucleotides in assembled contigs are in contigs of greater than 10kb) and the N50 supercontig size is about 90kb. Over 80% of each of the assemblies can be tied to the map using the assembled BAC end reads provided by TIGR. Thus in the near future, any given sequence query will have greater than 72% chance of being located at a precise location in the mapped genome.

The next round of assembly is scheduled to be done in January 2002, after the full whole genome data set has been collected (see above). It is anticipated that this assembly will represent a substantial improvement, given that it will have added both increased depth and the paired sequences from the ends of many large insert clones.

Annotation of the mouse genome sequence. A number of groups have begun efforts to annotate the mouse genome sequence, and the groups at Ensembl, NCBI, and UCSC have made commitments to make annotated versions of the mouse genome sequence publicly available as rapidly as possible. As a start, these groups have agreed to coordinate their efforts so that they are working on a single assembled version (see above); the initial work is being done on the Phusion version of the October 15 data set. More information about the annotation of the mouse genome will be available by the time of the next Mouse Genome Monthly.

Strategy for Sequencing the Mouse Genome. A short description of the strategy being taken by the Mouse Genome Sequence Consortium was presented in issue #1 of the Mouse Genome Monthly. A more complete description has now been published in the journal *genesis; genesis* **31**(4): 137-141 (Dec.2001). The article can also be found at the Wiley Interscience web site, http://www3.interscience.wiley.com/cgi-bin/issuetoc?ID=88511897.

Full-length mouse cDNAs. Next to the genomic sequence of an organism, the resource that is most frequently requested is a complete set of full-length cDNA clones. The NIH has organized a project known as the Mammalian Gene Collection (MGC; see Strausberg et al. *Science* **286**: 455-457[1999]) to address this need. The purpose of the MGC is to provide a representative full-length (open reading frame) sequence and cDNA clone for every expressed gene of the human and the mouse. The MGC effort has been in operation since September 2000 and, as of December 17, the project has generated 8902 verified full ORF cDNA clones and sequences from the human and 3556 from the mouse. At the same time, close to 20,000 additional clones are in process for sequencing and verification and over 300,000 clones have been isolated for initial evaluation. All of those clones, as well as a 5' EST read from each, are also available. A full description of the project, including instructions for obtaining clones and sequences, can be found at http://mgc.nci.nih.gov/.

Data Access. Here again is a list of handy web sites containing information related to the MGSC, sequence data and the laboratory mouse.

http://mouse.ensembl.org -- The primary website of the Mouse Genome Sequencing Consortium
http://www.ncbi.nlm.nih.gov/genome/guide/M_musculus.html -- mouse genome resources at the NCBI
http://genome.uscc.edu -- mouse genomic sequence reads aligned against the human draft sequence in a usable browser
http://trace.ensembl.org/ and http://www.ncbi.nlm.nih.gov/Traces/trace.cgi? -- raw data underlying all of the sequence generated in the mouse genome sequencing effort and other components of the Human Genome Project
http://www.nih.gov/science/models/bacsequencing/ -- to submit requests for sequencing of individual, or small numbers of, BACs of high biological interest
http://www.nih.gov/science/models/ -- information about NIH programs for analysis of model organisms
http://mrcseq.har.mrc.ac.uk -- the MRC UK Mouse Sequencing Programme
http://www.informatics.jax.org/mgihome -- integrated access to data on the genetics, genomics, and biology of the laboratory mouse
http://mgc.nci.nih.gov -- the Mammalian Gene Collection

Published by the Mouse Genome Sequencing Consortium and NHGRI, in consultation with the Mouse Sequencing Liaison Group, which is comprised of members of the mouse research community.

Questions or Comments. Is there anything that you would like to see in future issues of the Mouse Genome Monthly? Send comments to the Mouse Sequencing Liaison Group:
(email: Mouse_Sequencing_Liaison@nhgri.nih.gov).