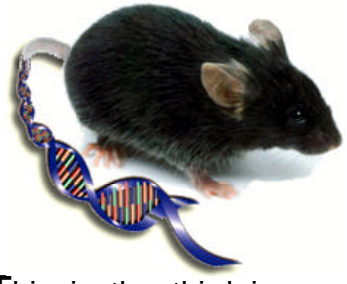
Mouse Genome Monthly



MGSC

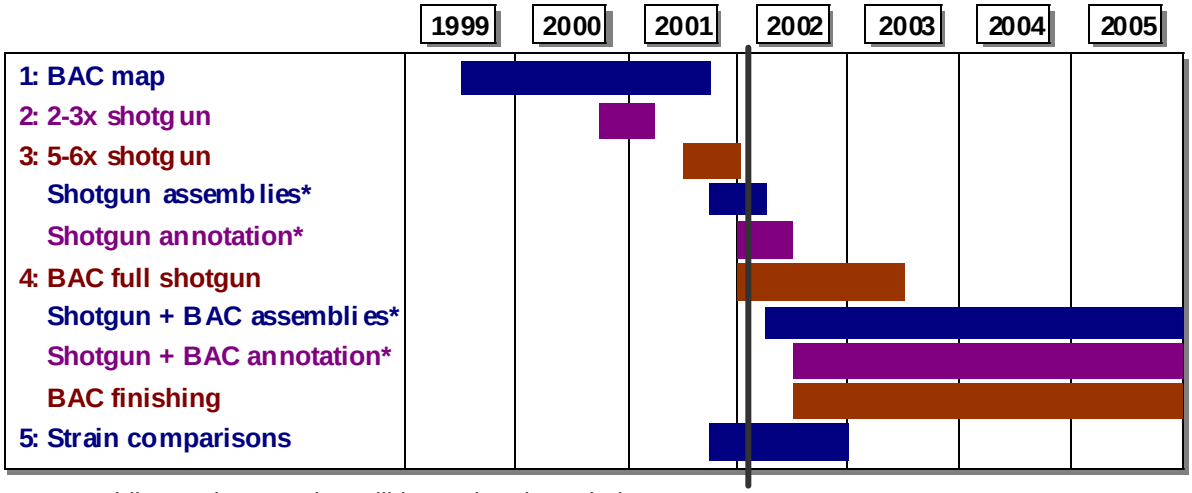
Issue #3 January 2002

The Latest Progress From the Mouse Genome Sequencing Consortium

This is the third 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 report on the sequencing of the mouse genome

- The whole genome shotgun phase of the C57BL/6J sequencing effort has been completed. 40,793,320 reads have been generated and submitted to the Trace Archive (see below for URL). This is 6X to 7X coverage of the genome.
- Assemblies will be done both at the Sanger Institute (using PHUSION) and the Whitehead Institute (using ARACHNE). The two assemblies will be compared and one will be chosen for public distribution on the MGSC website (see below) by the end of March.
- A significant improvement in sequence coverage of the genome is expected, with a reduction in the mis-assembly rate compared with the October assembly now posted. In a test assembly carried out in January using ARACHNE, 34 million reads were assembled into contigs having an N50 size of 16kb and linked into supercontigs having an N50 size of 1.2Mb.
- The mouse sequence will be finished through sequencing of individual BAC clones, incorporating the whole genome shotgun data. The tiling path of BAC clones across the genome has been selected, incorporating existing C57BL/6J BAC sequences where possible. Subclone library construction and BAC sequencing is currently under way.
- The current BAC clone map of the mouse genome consists of 325 clone contigs. As noted in the December issue, 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. 16,997 markers from radiation hybrid and genetic maps of the mouse have been integrated into the database to confirm contig order and orientation. For more details see: http://genome.wustl.edu/projects/mouse/index.php



*Assemblies and annotation will be updated regularly.

New Web sites

A. Single site for access to current assemblies of the mouse genome sequence and mouse-human alignments. In response to a number of requests, the MGSC has set up a web page, http://mouse.ensembl.org/Mus_musculus/resources.html as an interim measure to provide access to the mouse sequence data while a final assembly of the full whole genome shotgun data set is being produced. This web site contains links to a number of experimental resources that are not yet ready for submission to public databases but which nonetheless may be useful to some investigators. These include a number of the trial assemblies of the whole genome shotgun data that have been done periodically during the past few months, as well as several trial alignments between the mouse reads and the human sequence assembly that have been done with new programs under development. Investigators who use this information should understand that the information provided in these links is very recent, experimental, likely to contain errors, and subject to change at any time. The MGSC has provided these links in the spirit of openness but cannot yet assume any responsibility for the accuracy of the information beyond the collaborators' best-faith efforts to develop useful resources for the genome research community as rapidly as possible.

B. First Ensembl annotation of the mouse genome. The first view of the annotated mouse genome is sequence is available: http://www.ensembl.org/Mus_musculus. This is based on the first MGSC mouse genome assembly (data freeze of October 2001, which contained about 4X coverage). A few relevant statistics are (Last update January 28, 2002):

Ensembl gene predictions:	16,801
Ensembl gene exons:	123,672
Ensembl gene transcripts:	20,652
Contigs:	677
Clones:	677
Base Pairs:	3,320,754,623

The whole genome shotgun assembly was aligned to the joint Sanger Institute/WUGSC BAC map (frozen in Sept 2001) to provide this assembly. No clone-based sequencing has been incorporated. The assembly was analyzed via the Ensembl pipeline to predict genes and other features of interest.

With this release, http://www.ensembl.org now supports both human and mouse data sets simultaneously; as noted, mouse is located at http://www.ensembl.org/Mus_musculus/ and human is at http://www.ensembl.org/Homo_sapiens. Because of this, the new mouse site supports many of the same features of the new human release. In addition, there are "match" tracks on the ContigView page (http://www.ensembl.org/Mus_musculus/contigview) that show sequence similarity matches between the mouse and human genomes. These tracks can be used to jump between similar regions of the two genomes.

Deeper shotgun coverage, improved assembly software, refined BAC maps, more sensitive gene prediction and incorporation of finished clone sequences will improve these data over time. Updates are planned at approximately three-month intervals.

Integrating annotated sequence with biological data at JAX-MGI. The role of the Mouse Genome Informatics (MGI) efforts at The Jackson Laboratory is to support the incorporation of genome sequence annotation into the framework of biological knowledge and genetic experimental data that already exist and continue to accumulate about the laboratory mouse. The initial automated computational analysis and assembly of the mouse genome sequence is being accomplished by the MGSC.

The resources at the MGI website (http://www.informatics.jax.org) serve as an integration nexus for mouse genomic, genetic, and biological data. Among the data types currently represented at MGI are gene identification, allelic variants, maps and mapping data, mutant and strain phenotype descriptions, expression data, mammalian orthologous gene relationships, and functional classifications. Curated associations between genes and nucleotide and protein sequences are maintained, as well as data attributes of molecular segments (*e.g.,* clones, primers, etc.) including strain, tissue and sequence type (*e.g.*, mRNA, DNA, EST, etc.). The MouseBLAST server can be used to obtain both the expected BLAST output and the mouse gene represented by each hit, with links to the gene details.

The incorporation of the genomic sequence and its continued curation in the context of other biological data is critical to advancing our understanding of human biology and disease. Watch for new features and tools at the Mouse Genome Informatics site as sequence continues to play a more prominent role in the constellation of available genomic information.

Process to choose additional organisms for construction of BAC libraries and sequencing. The first meeting of the NHGRI's review panel, recently designated as the Genome Resources and Sequencing Priority Panel (GRASPP) was held in early January. The panel approved, with high priority, a request to prepare BAC libraries from several additional mouse strains, in addition to those available already from C57BL/6J male and female and 129/Sv female. These new libraries would be largely used to aid investigators in molecular identification of genetic modifiers and quantitative trait loci. After discussions with the Mouse Sequencing Liaison Group and community input (organized by Dr. David Beier), Dr. Pieter de Jong has begun to construct a 10X male library from CAST/Ei, a very popular inbred strain derived from the *M. musculus castaneus* subspecies. Several additional strains were recommended for library construction in the near future (A/J, DBA/2J and BALB/cByJ), and depending upon the demonstrated utility of the resource, additional strains (AKR/J, C3H/HeJ, and SPRET/Ei) were suggested for future libraries. The SJL/J strain was considered but it was learned that a library from SJL/J males is presently in construction.

Report from the Mouse SNP Workshop (based on a summary by David R. Beier, Brigham & Women's Hospital).

On January 30, a number of investigators attended a meeting organized by NHGRI to discuss the utilization of SNPs in mouse genetic investigation. Kerstin Lindblad-Toh (Whitehead Institute) discussed the present efforts of the MGSC to identify 50,000 SNPs by random shotgun sequencing of DNA from several inbred strains and alignment of these reads to the B/6 genome to identify SNPs. The initial three strains being analyzed are 129S1/SvImJ, C3H/HeJ and BALB/cByJ. Further candidates for SNP discovery are SJL/J, FVB/NJ, CAST/Ei and SPRET/Ei. The exact strains and number of SNPs from these strains will be determined after the analysis of the SNPs discovered in the first three strains. Dr. Lindblad-Toh also noted the intent to characterize the allele distribution of approximately a hundred SNPs in a larger set of inbred strains (Phenome list). Thus, it is expected that the final set of 50,000 SNPs will shed preliminary light on the haplotype structure across mouse strains and that for any lab strain combination roughly 10,000 SNPs should be available by inference to related lab strains. The consensus of the meeting participants was as follows:

1. The goal of obtaining a reference set of 50,000 SNPs from a small number of strains is realistic and underway at the Whitehead Institute. This master set would be valuable for investigators since it provides identification of possible SNPs that can be investigated specifically for polymorphism between strains of interest.
2. It would be desirable to have comprehensive allele distribution information for a substantial subset of these SNPs, especially as the cost of additional strain genotyping is quite small compared to that for SNP discovery and assay development. The utility of the CIDR database of MIT microsatellite marker allele distribution (http://www.cidr.jhmi.edu/mouse/mouse.html) was also noted. Given this, it was recommended that the 48 mouse strains that are being characterized as part of the mouse Phenome Project (http://www.jax.org/phenome) be analyzed for SNP alleles.
3. There was consensus that it would be useful to have identifying SNPs for every BAC in the near term (i.e. perhaps two years). This would about the same density that will be available soon, but distributed so that all BACs are marked.
4. No consensus was reached on the question of the utility of high density or complete SNP information. Researchers trying to identify QTLs believe that complete SNP information for a set of phenotyped strains would be extremely valuable. For those using mutagenesis approaches, knowing complete SNP distributions would not be particularly useful. There was more agreement that having such information for a small number of strains could be helpful for understanding genome architecture and strain relatedness, for identification of informative SNPs between strains based on inferences from haplotype structure and for allowing new questions to be addressed. Further discussion is planned.
5. Human SNP analysis is frequently done in dedicated centers using expensive technology. Because high throughput is crucial, per-reaction costs need to be minimized. Any large-scale analysis of mouse SNPs will also be most efficiently done in such centers. Still, much current mouse genetic analysis, including analysis of mutations, tends to be done in individual laboratories and rarely requires the same throughput (al though it was noted that the availability of efficient technologies tends to alter the nature of the studies that are done). For these laboratories the initial cost of instrumentation can be a significant problem. It was therefore recommended that some of the instrumentation and assays for mouse SNP detection be compatible with utilization in small laboratories.

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

http://mouse.ensembl.org -- output of the Mouse Genome Sequencing Consortium
http://mouse.ensembl.org/Mus_musculus/resources.html -- access to experimental assemblies of mouse genome
http://www.ensembl.org/Mus_musculus -- v1 annotated view of the mouse genome
http://www.ncbi.nlm.nih.gov/genome/guide/M_musculus.html -- mouse genome resources at the NCBI
http://genome.ucsc.edu -- mouse genomic sequence reads aligned against the human draft sequence in a usable browser
http://www.ncbi.nlm.nih.gov/Traces/trace.cgi? and http://trace.ensembl.org/ -- 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

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).