Mouse Genome Monthly

M
G
S
C

Issue #1 November 2001

The Latest Progress From the Mouse Genome Sequencing Consortium

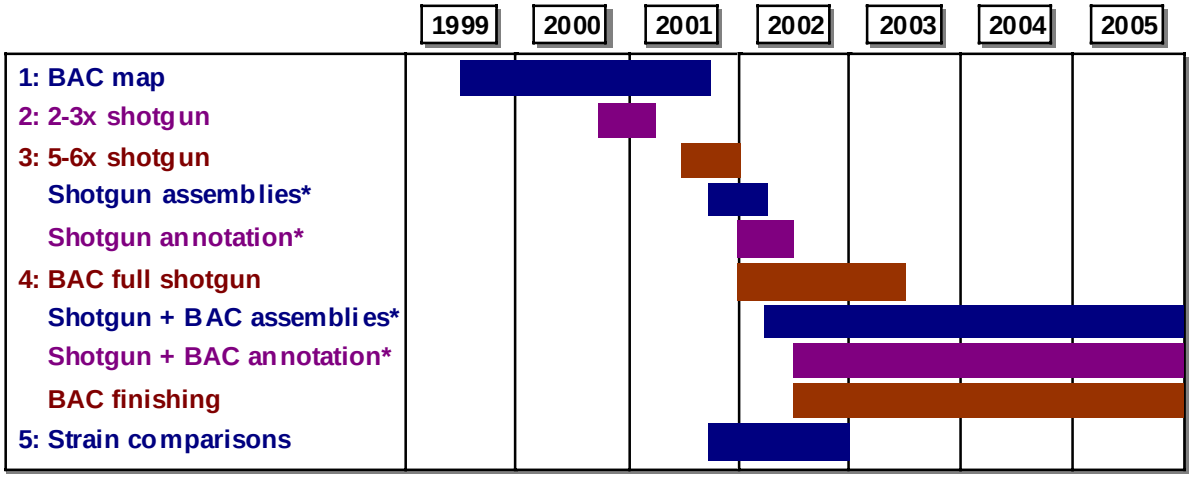
The production phase of the public mouse (C57BL/6J) genome sequencing effort has been underway for a year and is advancing rapidly. Francis Collins, Director of the NHGRI, distributed an open letter to the mouse community last summer to describe progress to that point. This newsletter, which will be produced monthly for the next several months, is among a number of additional means that have been developed to continue to keep the community of mouse researchers abreast of the progress of the sequencing of the mouse genome.

Plan for sequencing the mouse genome

The objectives of the current NHGRI/Wellcome Trust-supported mouse genome activities are to produce a robust physical map and a high quality, finished genomic sequence of the C57BL/6J strain, and to make that information publicly available with no restrictions. Further, all information generated in the interim will also be released rapidly to the scientific community. A comprehensive view of the project's data, including the results summarized below, will be provided at a central MGSC server, hosted by the Ensembl database and managed by the four collaborating groups. In addition, MGSC data will be incorporated into other key genome servers – including those at the National Center for Biotechnology Information and the University of California at Santa Cruz (URL's are given below).

The sequencing of the mouse genome currently involves two types of efforts, a genome-wide program and a targeted sequencing program:

- The genome-wide sequencing program is being undertaken by the Mouse Genome Sequencing Consortium (MGSC), a collaboration consisting of three large sequencing centers (Washington University Genome Sequencing Center, Whitehead/MIT Center for Genome Research and the Sanger Centre) and an international database (Ensembl, a joint project between the European Bioinformatics Institute and the Sanger Centre).
- Targeted sequencing programs are focused on obtaining sequence information from specific regions of high biomedical or biological significance. NHGRI-supported centers involved in this effort include Harvard Partners Genome Center, Cold Spring Harbor Laboratory Lita Annenberg Hazen Genome Center, and University of Oklahoma's Advanced Center for Genome Technology. These three groups participate in an NIH program that accepts requests for sequencing individual BAC clones containing regions of high biological interest. In the UK four centers are participating in the MRC UK Mouse Sequencing Programme, focused on 4 regions totaling 50Mb on Chromosomes 4, 13, 2 and X and these centers also accept requests for targeted sequencing of individual BACs or small contigs from research groups.



*Assemblies and annotation will be updated monthly

Project Timeline *(see figure)*

Stage 1. *BAC map (complete)*
- ~305,000 BACS fingerprinted; currently assembled into approximately 600 contigs;
- the clone contigs have been assigned to mouse chromosomes via ~14,000 RH markers associated with BACs;
- they have also been assigned to human conserved segments via BAC end sequences.

Stage 2. *2-3x whole genome shotgun coverage (complete)*
- ~17M individual reads deposited in public databases); most useful for comparative studies with the human sequence.

Stage 3. *5-6x shotgun coverage*
- 21-23M more traces will provide longer initial sequence contigs, greater continuity, anchored assembly, improved analysis, and SNP discovery/mapping;
- ongoing assembly and annotation, will be updated monthly as data improve.

Stage 4. *BAC-based sequencing*
- necessary to produce a finished genome sequence with the smallest number of misassemblies and gaps. BAC finishing to 99.99% accuracy;
- further improvements in assembly and annotation.

Stage 5. *Sequence from additional mouse strains*
- to identify sequence variants (such as single nucleotide polymorphisms or SNPs);
- each 0.01-fold coverage of laboratory strains will provide roughly 10,000 SNPs;
- initially done for three strains (129S1/SvImJ, BALB/cByJ, C3H/HeJ). Additonal strains to be discussed with community.

Progress

The first two components are complete, with assembly and annotation efforts well underway. The current October 2001 assembly has contigs of ~6kb. A number of programs have been developed to align the mouse and human sequences, which have led to improved gene prediction/finding in the human. Analyses indicate that ~3% of the mouse genome is conserved at high specificity (50% of that gives exonic or 5'/3' UTR hits, 50% hit elsewhere in the human sequence). The Ensembl and UCSC browsers display all aligned sequences; users can find reads of interest and the assembled mini-contigs are also displayed.

Mouse Sequencing Liaison Group. Recently, NHGRI convened a group of leading mouse researchers to serve as a line of communication between the sequencing centers and the funding agencies, and the mouse research community.

The membership of the Mouse Sequencing Liaison Group is: **Wayne Frankel** (chair, The Jackson Laboratory), **Rudi Balling** (Institut für Biotechnologische Forschung - GBF), **Steve Brown** (MRC Mouse Genome Centre), **Maja Bucan** (University of Pennsylvania), **Sally Camper** (University of Michigan), **David Kingsley** (Stanford University), **Janet Rossant** (Mt. Sinai Hospital, Toronto), **Eddy Rubin** (Lawrence Berkeley Laboratory) and **Joseph Takahashi** (Northwestern University).

The goals of this group are to ensure that the mouse community is aware of the progress being made, knows how to access the data, knows the timetable for the project and can offer the sequencers and the funding agencies helpful advice on making the data as useful as possible to researchers. Maintaining good communication with the mouse community is a high priority for the mouse genome sequencing effort and the Liaison Group and the newsletter are intended to help to do so. At its first meeting, the Liaison Group recommended that this newsletter be produced and widely distributed. The Liaison Group welcomes your additional suggestions.

The Liaison Group meets every month or so by conference call, along with the Principal Investigators of the groups participating in the Mouse Genome Sequencing Consortium (MGSC: Eric Lander, Whitehead Institute Center for Genome Research; Bob Waterston and John McPherson, Washington University Genome Sequencing Center; Jane Rogers, Sanger Centre; and Ewan Birney, Ensembl; for more information about the consortium, see below), and staff from the funding agencies.

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://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

---

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).