Mouse Genome Monthly

M
G
S
C

Issue #4 April 2002

The Latest Progress From the
Mouse Genome Sequencing Consortium

NEW PUBLIC MOUSE ASSEMBLY RELEASED!!!

The sequencing of the genome of the strain C57BL/6J has reached another milestone. The whole genome shotgun phase has been completed and stands at approximately 7X sequence coverage. The whole genome shotgun data have been assembled and, as discussed below, the new assemblies have even better quality and contiguity than expected from the most optimistic prior predictions. The primary sequencing effort has now turned to the BAC tiling path. For the strategy being followed, please refer to *Genesis*. **31**:137-141 (2001).

Assembly of the mouse whole genome sequence data

The complete whole genome data set for the mouse genome has been assembled and looks outstanding. Actually, two assemblies were done, one by the Whitehead group led by David Jaffe using the ARACHNE program and the other by the Sanger Institute group led by Jim Mullikin using the PHUSION program. Both assemblies used the February 1, 2002 freeze of the data (ftp://ftp.ncbi.nlm.nih.gov/pub/TraceDB/mus_musculus/Feb_1_Freeze_Ti_List.gz).

The two assemblies were done using the same set of files; after removal of a relatively small number of reads (e.g. low quality or mislabeled), each assembly included about 33 million reads. As noted above, this corresponds to about 7X sequence coverage of the genome, based on trimmed reads and assuming a genome size of 2.6 Mb. Both assemblies were extremely good and were evenly matched by many criteria. In the end, the Mouse Genome Sequencing Consortium chose the ARACHNE assembly as the version with which to proceed for further analysis. The ARACHNE assembly was further refined to remove a few identified residual errors and to incorporate approximately 50 Mb of finished mouse sequence (NT contigs). This refined, composite assembly has been designated as the “MGSC Version 3” assembly

The assembly consists of sequence contigs (continuous stretches of assembled sequences), linked into “supercontigs” (also known as “scaffolds,” structures in which neighboring contigs are organized in their proper order and orientation using linking information provided by paired or mated sequences from the ends of genomic sub-clones), which in turn are linked into “ultracontigs” (structures in which neighboring supercontigs are organized into their proper order and orientation using linking information provided by the physical map of BAC clones independently assembled using restriction fragment patterns and the FPC program).

Figure 1 shows the distribution of ultracontigs across the chromosomes of the mouse genome. The MGSC version 3 assembly was placed on the genome using the MIT genetic map. This resulted in 123 mapped supercontigs that could be linked together into 89 ultracontigs using the fingerprint map. This core assembly constitutes 2.37 Gb of sequence and has only 69 unspanned gaps in the genome.

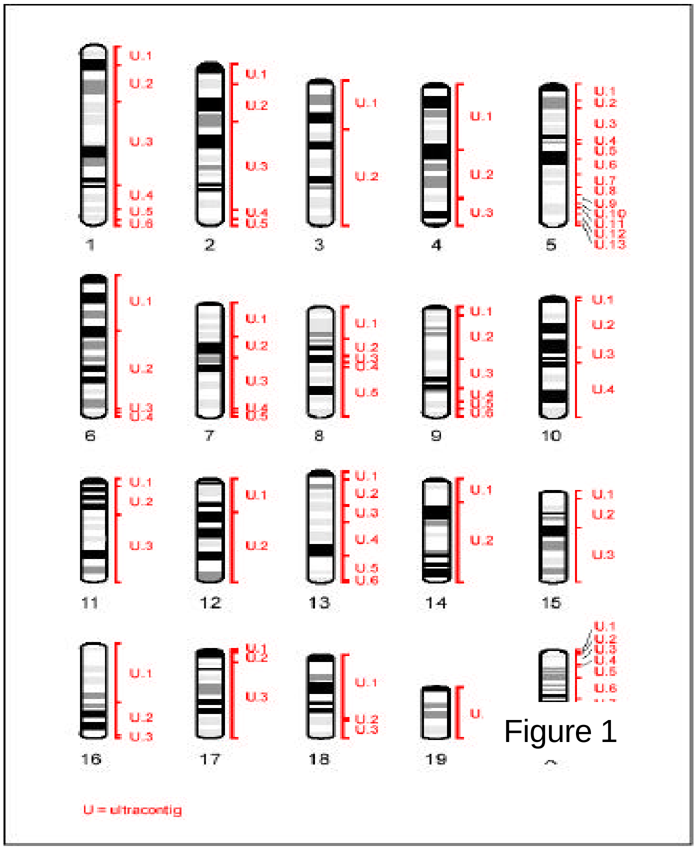


Figure 1

Figure 2 depicts a comparison of a representative 14.8 Mb supercontig from the MGSC version 3 assembly to 200 kb of finished sequence. Across the 200 kb of finished sequence shown, 10 contigs cover well over 95% of the sequence, with the average gap size being only 268 bp.

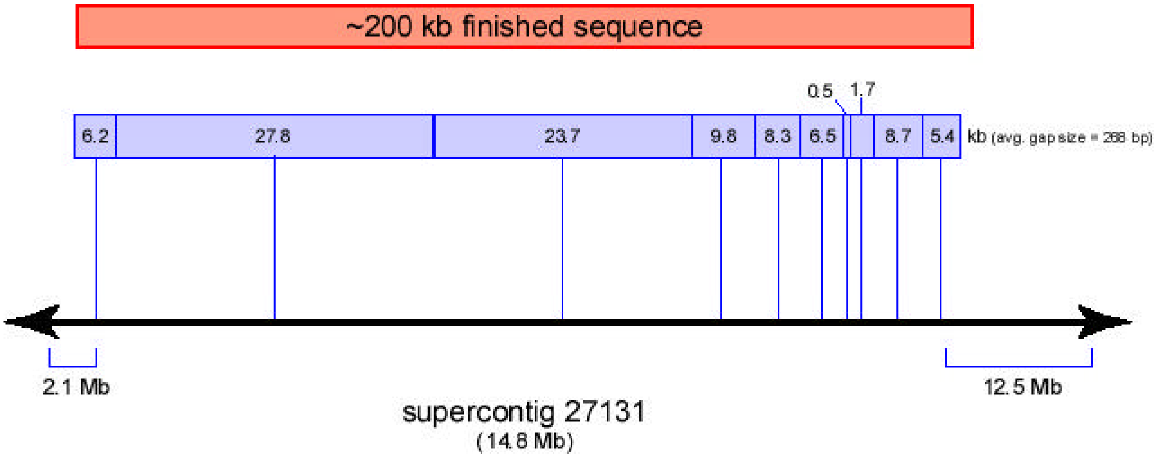


Figure 2

Some quantitative descriptors of the ARACHNE assembly include:

Contig Properties

Contig size (N50)	24.8 kb
Number of contigs	224,000
BAC ends in assembly	410,000
Total number of bases in the assembly	2.47 Gb

Supercontig Properties

Supercontig size (N50)	16.9 Mb
Number of supercontigs	44,500
Proportion of the assembly in	
100 largest supercontigs	79%
200 largest supercontigs	95%
300 largest supercontigs	96%
Summed length of supercontigs containing >= 1 BAC end	2.41 Gb

Coverage of the Genome

mRNA analysis coverage (by BLAT, from Jim Kent, UCSC)	
Coverage at 95% sequence identity	97%
Coverage at 99% sequence identity	90%
Gene-models built (by Deanna Church and colleagues, NCBI)	
mRNAs identified	96.8%
Refseqs identified	99.3%
Comparison to 10Mb of finished sequence (from David Jaffe, Whitehead)	96.1%

Global assembly accuracy. Genetic markers from the MIT genetic map (which has been extensively QC'ed, with known error rates) were examined for alignment with high stringency to supercontigs of length ≥ 1Mb. No marker hit more than one supercontig. A marker was declared "bad" if its chromosomal assignment disagreed with the majority assignment for the markers on its supercontig. 1.9% of the markers (from the pool) were declared bad by this criterion, suggesting very few global errors in the assembly.

Summary Table

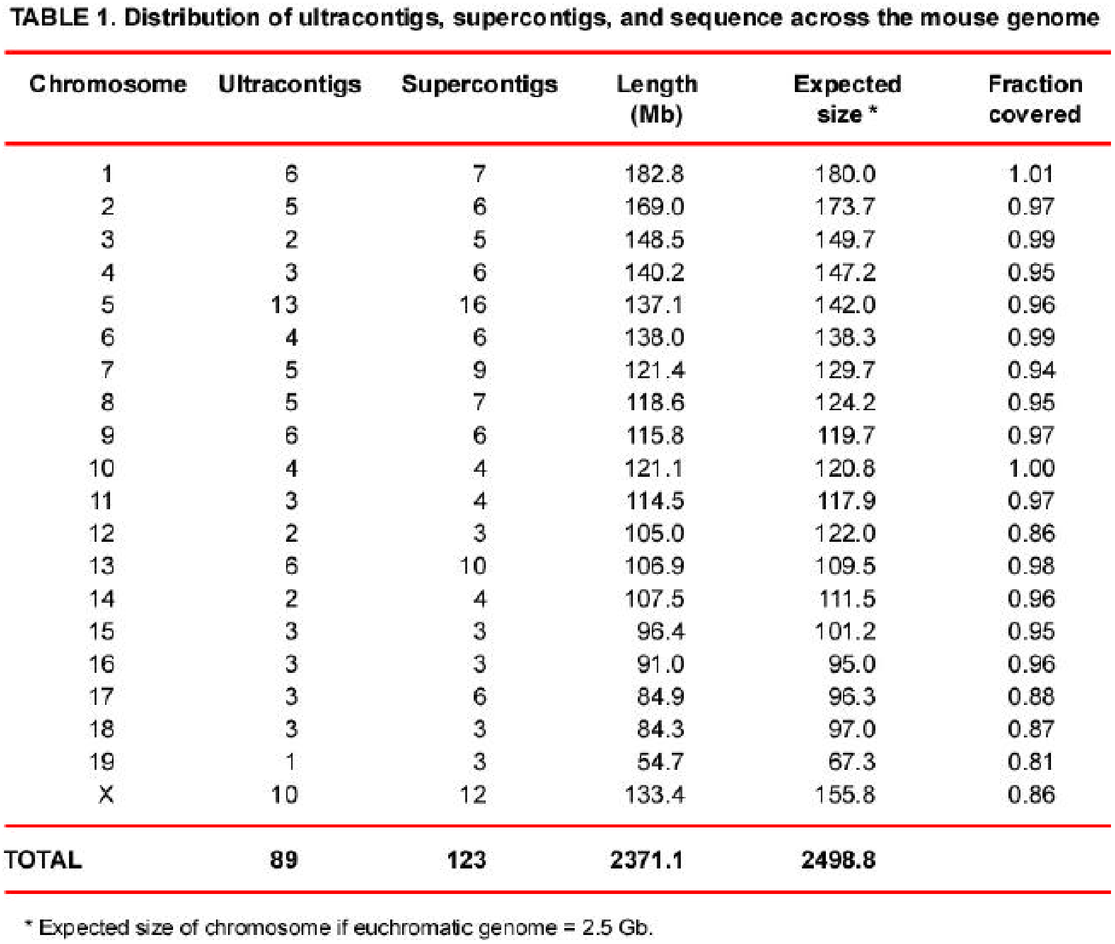


TABLE 1. Distribution of ultracontigs, supercontigs, and sequence across the mouse genome

Chromosome	Ultracontigs	Supercontigs	Length (Mb)	Expected size *	Fraction covered
1	6	7	182.8	180.0	1.01
2	5	6	169.0	173.7	0.97
3	2	5	148.5	149.7	0.99
4	3	6	140.2	147.2	0.95
5	13	16	137.1	142.0	0.96
6	4	6	138.0	138.3	0.99
7	5	9	121.4	129.7	0.94
8	5	7	118.6	124.2	0.95
9	6	6	115.8	119.7	0.97
10	4	4	121.1	120.8	1.00
11	3	4	114.5	117.9	0.97
12	2	3	105.0	122.0	0.86
13	6	10	106.9	109.5	0.98
14	2	4	107.5	111.5	0.96
15	3	3	96.4	101.2	0.95
16	3	3	91.0	95.0	0.96
17	3	6	84.9	96.3	0.88
18	3	3	84.3	97.0	0.87
19	1	3	54.7	67.3	0.81
X	10	12	133.4	155.8	0.86
TOTAL	**89**	**123**	**2371.1**	**2498.8**	

* Expected size of chromosome if euchromatic genome = 2.5 Gb.

Access to the MGSC version 3 assembly can be obtained in several ways:

The assembly can be downloaded from
ftp://ftp.ensembl.org/pub/assembly/mouse/mgsc_assembly_3.

The assembly contigs can be searched using SSAHA at
http://www.ensembl.org/Mus_musculus/ssahaview.

The assembly contigs can be searched using BLAST at
http://www.ncbi.nlm.nih.gov/genome/seq/MmBlast.html.

The chromosomally assigned supercontigs are being prepared for sequence searching. This capability should be available soon at the indicated sites.

The **PHUSION** assembly is also available for interested investigators at http://mouse.ensembl.org and can be searched at http://www.ncbi,nlm.nih.gov/genome/seq/MmBlast.html.

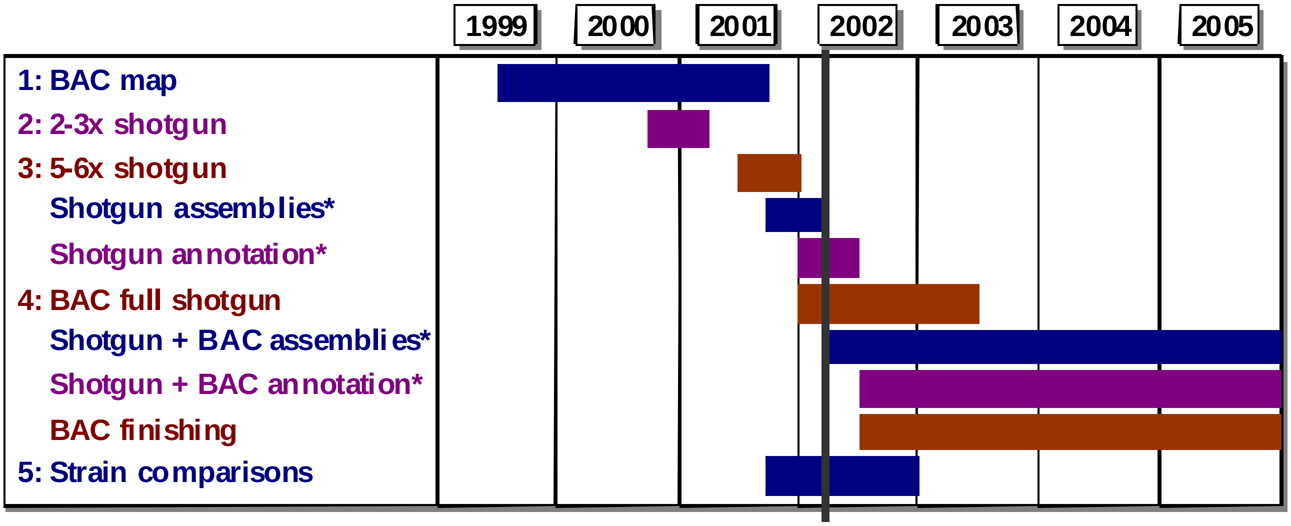
Summary of the current status of the mouse genome sequencing effort

- The whole genome shotgun phase has been completed. Sequence data corresponding to 6.4X (if the genome is 3 Gb) to 7.7X (if the genome is 2.5 Gb) sequence coverage have been collected (see January newsletter) and assembled (see above).

- A BAC map, based upon restriction fingerprinting and assembled by the FPC program is available (http://genome.wustl.edu/projects/mouse/index.php?fpc=1). As of March 14, 2002, the fingerprint map consisted of 296 contigs.

- Sequencing of a BAC-based tiling path has begun. As of March 29, 2002, 685.1 Mb of BAC-based "draft" quality had been deposited in the public databases (this is the total amount of mouse data in the HTGS division; it includes sequence from other strains and redundancy has not been accounted for). This is an increase of more than 90 Mb since January 4, 2002. In addition, 99.8 Mb of finished mouse DNA sequence are in the database, an increase of about 15 Mb since January.

- Annotation. Ensembl version 3.1.1 was released at the end of January 2002. The site (http://ensembl.mouse.org) presents version 1 of the mouse genome draft sequence, based on the pure whole genome shotgun data of ~4x coverage, frozen in October 2001. The whole genome shotgun assembly was aligned to the joint Sanger Institute/WUGSC BAC map (frozen in Sept 2001) to provide this assembly. Clone-based sequencing has not been incorporated. The assembly was run through the normal Ensembl pipeline to predict genes and other features of interest.



*Assemblies and annotation will be updated regularly.

Fosmid clones. An important component of the whole genome shotgun data set is a collection of 2.0 million reads from the ends of fosmid clones. The clones themselves may be a useful resource for finishing the sequence, as well as for other purposes. Arrangements are currently being made to deposit the fosmid clones in an archive from which they can be obtained.

New mouse BAC libraries. As noted in the January newsletter, the NIH has established a new program to increase BAC library-making capacity, and approved the construction of libraries from three additional mouse strains. Library making is currently under way. The strains from which new libraries are being constructed and the library producers are:

CAST/Ei	Pieter de Jong, Children's Hospital Research Institute (CHORI), Oakland
A/J	Pieter de Jong
DBA/2J	Rod Wing, Clemson University

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/mouse -- mouse genome resources at the NCBI (NOTE: This is a new URL).
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).