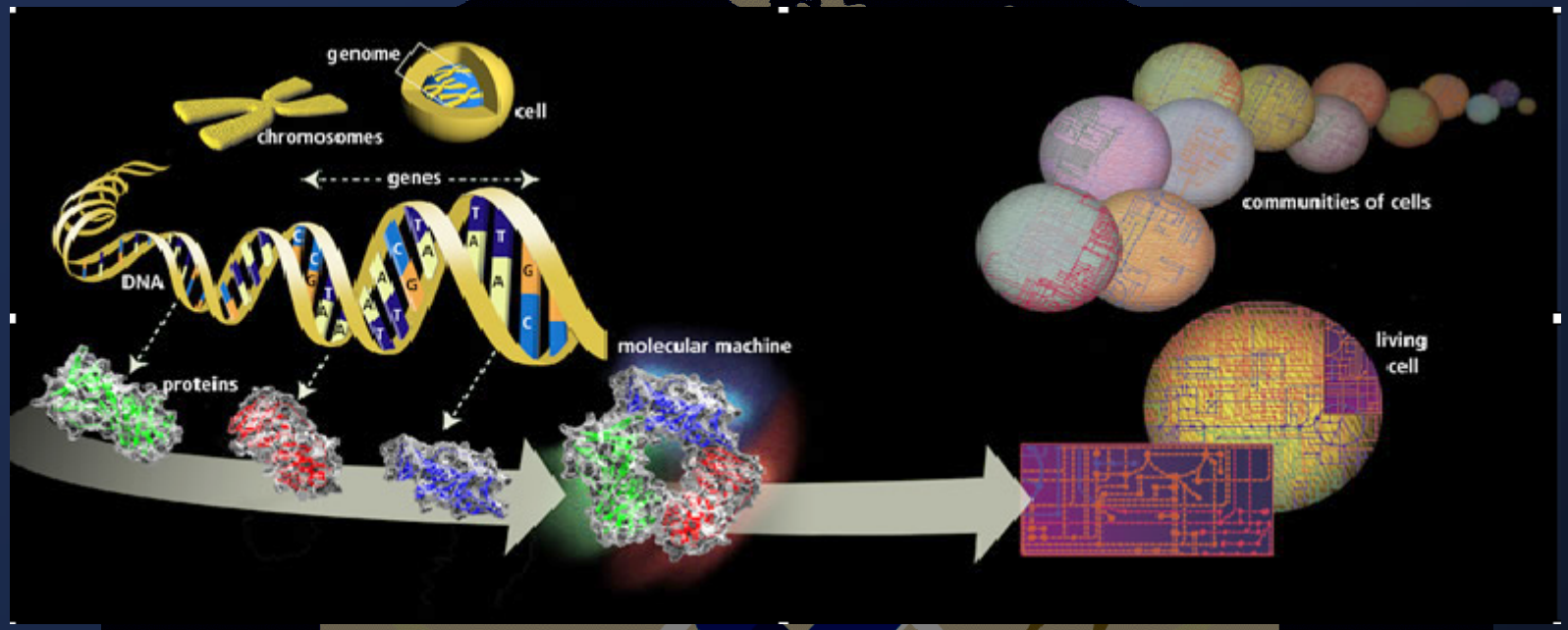


Cells contain **DNA**—the hereditary material of all living systems.

The **genome** is an organism's complete set of DNA and is organized into **chromosomes**.

DNA contains **genes** whose sequence specifies how and when to build proteins.

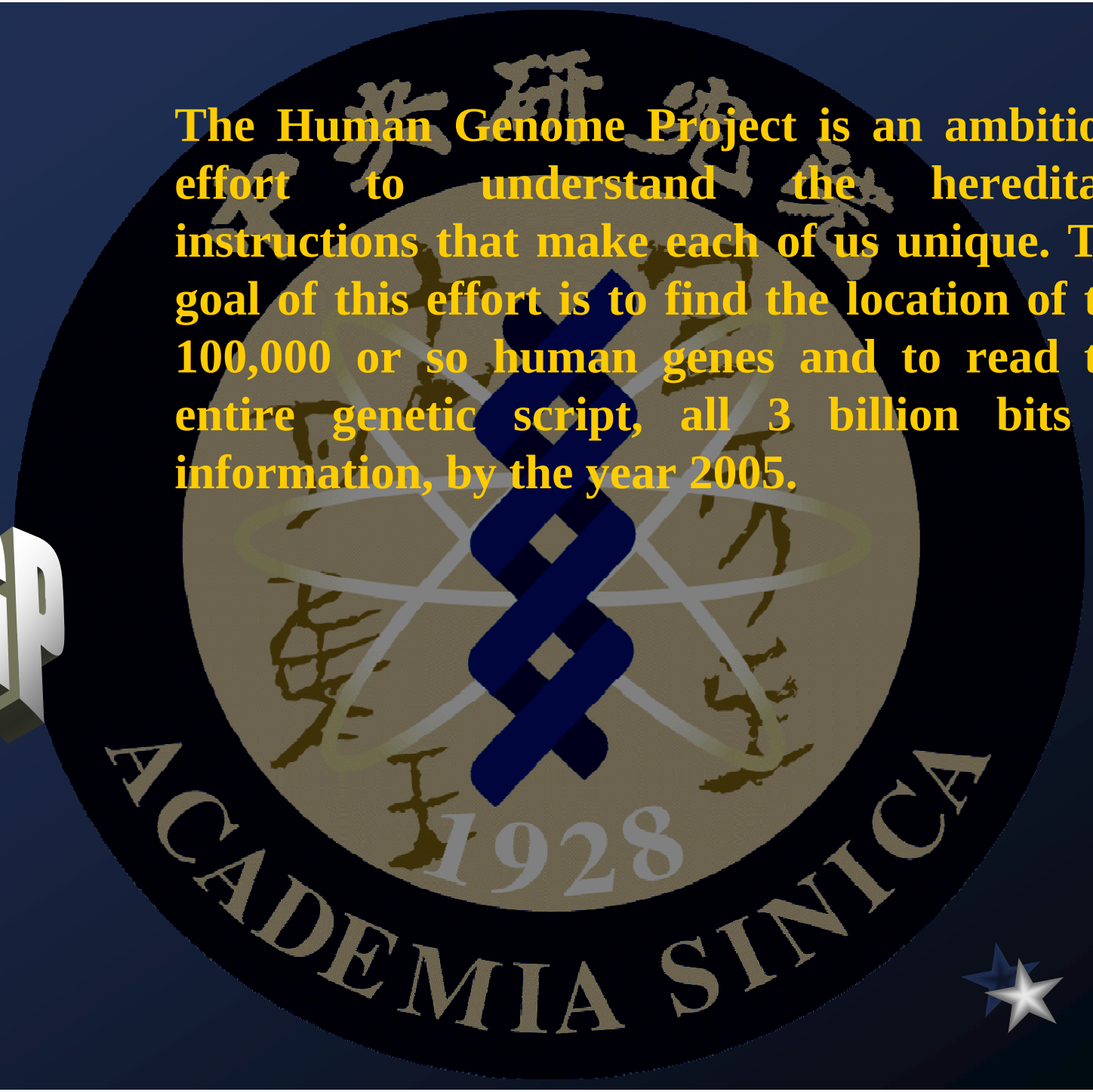
Proteins perform most essential life functions, often working together as molecular machines.

Molecular machines interact through complex, interconnected pathways and networks to make the cell come alive.

Communities of cells range from associations of microbes (each a single cell) to the hundred trillion cells in a human being.

- 1995 *Haemophilus influenzae (1.83 Mb; 1,742 genes)*
- 1996 *Saccharomyces cerevisae (12 Mb; 6,000 genes)*
- 1998 *Caenorhabditis elegans (97 Mb; 19,000 genes)*
- 2000 *Arabidopsis thaliana (115/125 Mb; 25,000 genes)*
- 2000 *Drosophila melanogaster (~120 Mb; 13,600 genes)*
- 2001 *Homo sapiens (90%; 2,900 Mb; ~30k genes)*
- 2002 *Mus musculus (96%; 2,500 Mb; ~30K genes)*
- 2002 *Oryza sativa L. ssp. indica (92%; 466 Mb; 46-56k genes)*
- 2002 *Oryza sativa L. ssp. Japonica (93%; 420 Mb; 32-50k genes)*
- 2002 *Fugu rubripes (95%; 365 Mb; 33,609 genes)*
- 2004 *H. sapiens (99% euchromatin; 2,850 Mb; 20,000-25,000 genes)*





The Human Genome Project is an ambitious effort to understand the hereditary instructions that make each of us unique. The goal of this effort is to find the location of the 100,000 or so human genes and to read the entire genetic script, all 3 billion bits of information, by the year 2005.





http://www.ornl.gov/TechResources/Human_Genome/

Human Genome Project

How does the human genome stack up?

Organism	Genome Size (Bases)	Estimated Genes
Human (*Homo sapiens*)	3 billion	30,000
Laboratory mouse (*M. musculus*)	2.6 billion	30,000
Mustard weed (*A. thaliana*)	100 million	25,000
Roundworm (*C. elegans*)	97 million	19,000
Fruit fly (*D. melanogaster*)	137 million	13,000
Yeast (*S. cerevisiae*)	12.1 million	6,000
Bacterium (*E. coli*)	4.6 million	3,200
Human immunodeficiency virus (HIV)	9700	9

International Human Genome Sequencing Consortium Announces "Working Draft" of Human Genome - - -

June 26, 2000.

The sixteen institutions that form the Human Genome Sequencing Consortium include:

1. Baylor College of Medicine, Houston, Texas, USA
2. Beijing Human Genome Center, Institute of Genetics, Chinese Academy of Sciences, Beijing, China
3. Gesellschaft fur Biotechnologische Forschung mbH, Braunschweig, Germany
4. Genoscope, Evry, France
5. Genome Therapeutics Corporation, Waltham, MA, USA
6. Institute for Molecular Biotechnology, Jena, Germany
7. Joint Genome Institute, U.S. Department of Energy, Walnut Creek, CA, USA
8. Keio University, Tokyo, Japan
9. Max Planck Institute for Molecular Genetics, Berlin, Germany
10. RIKEN Genomic Sciences Center, Saitama, Japan
11. The Sanger Centre, Hinxton, U.K.
12. Stanford DNA Sequencing and Technology Development Center, Palo Alto, CA, USA
13. University of Washington Genome Center, Seattle, WA, USA
14. University of Washington Multimegabase Sequencing Center, Seattle, WA, USA
15. Whitehead Institute for Biomedical Research, MIT, Cambridge, MA, USA
16. Washington University Genome Sequencing Center, St. Louis, MO, USA

In addition, two institutions played a key role in providing computational support and analysis for the Human Genome Project over the course of the past eighteen months. These include:

The National Center for Biotechnology Information at NIH
The European Bioinformatics Institute in Cambridge, UK
Scientists at the University of California, Santa Cruz, and Neomorphic, Inc. also assisted the assembly of the genome sequence across chromosomes,

CELERA
A PE Corporation Business

▸ **CeleraScience**

Accurate data. Powerful tools. Relevant news. CeleraScience delivers to your desktop a faster way to further scientific discovery.

▸ **CeleraCorporate**

A comprehensive source of information about Celera for those interested in learning more about our company.



The 14.8-billion bp DNA sequence was generated over 9 months from 27,271,853 high-quality sequence reads (5.11-fold coverage of the genome) from both ends of plasmid clones made from the DNA of five individuals.

26,588 protein-encoding genes plus 12,000 predicted.

THE GENOMIC SEQUENCE



Fifty years after the double helix, the reference DNA sequence of Homo sapiens is now available for downloading. Annotation of genes and other features is in progress.

Download Sequence Data

14 Apr 2003

The Human Genome Project announces completion of the DNA reference sequence of Homo sapiens.

Press releases:
NIH
Sanger Institute
Whitehead Institute

Reference Sequence Properties

Chrom. number	Reference accession	Sequence length	Determined bases*
1	NC_000001.4	245,203,898	218,712,898
2	NC_000002.5	243,315,028	237,043,673
3	NC_000003.5	199,411,731	193,607,218
4	NC_000004.5	191,610,523	186,580,523
5	NC_000005.4	180,967,295	177,524,972
6	NC_000006.5	170,740,541	166,880,540
7	NC_000007.7	158,431,299	154,546,299
8	NC_000008.5	145,908,738	141,694,337
9	NC_000009.5	134,505,819	115,187,714
10	NC_000010.4	135,480,874	130,710,865
11	NC_000011.4	134,978,784	130,709,420
12	NC_000012.5	133,464,434	129,328,332
13	NC_000013.5	114,151,656	95,511,656
14	NC_000014.4	105,311,216	87,191,216
15	NC_000015.4	100,114,055	81,117,055
16	NC_000016.4	89,995,999	79,890,791
17	NC_000017.5	81,691,216	77,480,855
18	NC_000018.4	77,753,510	74,534,531
19	NC_000019.5	63,790,860	55,780,860
20	NC_000020.5	63,644,868	59,424,990
21	NC_000021.3	46,976,537	33,924,742
22	NC_000022.4	49,476,972	34,352,051
X	NC_000023.4	152,634,166	147,686,664
Y	NC_000024.3	50,961,097	22,761,097
unplaced	various	25,263,157	25,062,835

Major Events in the U.S. Human Genome Project and Related Projects

The U.S. Human Genome Project, coordinated by DOE and NIH, is a multi-year effort to find all the genes on every chromosome in the human body and to determine their biochemical nature.





Acronyms listed below

2003

- Human Genome Project Completion, April 2003
- Human Chromosome 14 Finished - Chromosome 14 is the fourth chromosome to be completely sequenced.

2002

- Mouse Genome Sequencing Consortium publishes its draft sequence of mouse genome in the December 5, 2002, issue of Nature.
- International consortium led by the DOE Joint Genome Institute publishes draft sequence of *Fugu rubripes*.

2001

- Human Chromosome 20 Finished - Chromosome 20 is the third chromosome completely sequenced to the high quality specified by the Human Genome Project.

- **Publication of Initial Working Draft Sequence February 12, 2001**
  Special issues of *Science* (Feb. 16, 2001) and *Nature* (Feb. 15, 2001) contain the working draft of the human genome sequence. *Nature* papers include initial analysis of the descriptions of the sequence generated by the publicly sponsored Human Genome Project, while *Science* publications focus on the draft sequence reported by the private company, Celera Genomics. A press conference was held at 10 a.m., Monday, February 12, 2001, to discuss the landmark publications. Links for more information are:

2000



- HGP leaders and President Clinton announce the completion of a "working draft" DNA sequence of the human genome.
  - White House Press Conference: The Human Genome Project, June 26, 2000 (Total Running Time: 00:41:23; Bandwidth: 33 Kbps)
  - Press briefing and remarks
  - An Interview with Ari Patrinos, Director U.S. DOE Human Genome Program
    - Part One: Reaction to President Clinton's Announcement of the Completion of a Draft Sequence of the Human Genome
    - Part Two: Origins of the Human Genome Project, NIH Collaboration, and the Private Sector Role
    - Part Three: Application of Genome Discoveries, Next Steps in the Human Genome Project, and Ethical Considerations
  - Craig Venter (head of Celera Genomics), Ari Patrinos (director of DOE Human Genome Program and Biological and Environmental Research Program), and Francis Collins (director, NIH National Human Genome Research Institute).



- International research consortium publishes chromosome 21 genome, the smallest human chromosome and the second to be completely sequenced.
- DOE researchers announce completion of chromosomes 5, 16, and 19 draft sequence.
  - Elbert Branscomb, Director, Joint Genome Institute
  - Trevor Hawkins, Director of Sequencing, Joint Genome Institute
- International collaborators publish genome of fruit fly *Drosophila melanogaster*.

1999

- First Human Chromosome Completely Sequenced! On December 1, researchers in the Human Genome Project announced the complete sequencing of the DNA making up human chromosome 22.
- Joint Genome Institute sequencing facility opens in Walnut Creek, CA.
- Major Drug Firms Create Public SNP Consortium
- The Billion Base Pair Celebration November 23, 1999. Bruce Alberts, President, National Academy of Sciences and early planner of the Genome Project; Francis Collins, Director, NHGRI; Secretary of HHS, Donna Shalala; Secretary of DOE, Bill Richardson. (Total Running Time: 01:09:45; Bandwidth: 146 Kbps)
- HGP advances goal for obtaining a draft sequence of the entire human genome from 2001 to 2000.

1998

- Hospital for Sick Children, Toronto, Ontario, to continue GDB data collection, curation.
- *Caenorhabditis elegans* genome sequence completed.
- DOE and NIH reveal new five-year plan for HGP, predict project completion by 2003.
- JGI exceeds sequencing goal, achieves 20 Mb for FY 1998.
- GeneMap'98 containing 30,000 markers released.
- Incyte Pharmaceuticals announces plans to sequence human genome in 2 years.
- *Mycobacterium tuberculosis* bacterium sequenced.
- Celera Genomics formed to sequence much of human genome in 3 years using HGP-generated resources.
- DOE funds production BAC end sequencing projects
- Largest-ever ELSI meeting attended by over 800 from diverse disciplines and sponsored by DOE; Whitehead Institute; and the American Society of Law, Medicine, and Ethics.

1997

- NIH NCHGR becomes National Human Genome Research Institute (NHGRI).
- *Escherichia coli* genome sequence completed.
- Second large-scale sequencing strategy meeting held in Bermuda. (see also summary)
- High-resolution physical maps of chromosomes X and 7 completed.
- DOE-NIH Task Force on Genetic Testing releases final report and recommendations.
- DOE forms Joint Genome Institute for implementing high-throughput activities at DOE human genome centers, initially in sequencing and functional genomics.
- UNESCO adopts *Universal Declaration on the Human Genome and Human Rights*

Acronyms

1996

- *Methanococcus jannaschii* genome sequenced; confirms existence of third major branch of life on earth.
- DOE initiates 6 pilot projects on BAC end sequencing.
- Health Care Portability and Accountability Act prohibits use of genetic information in certain health-insurance eligibility decisions, requires DHHS to enforce health-information privacy provisions.
- HGP Participants Agree on Sequencing Data Release Policies Bermuda Conference
- DOE and NCHGR issue guidelines on use of human subjects for large-scale sequencing projects.
- *Saccharomyces cerevisiae* (yeast) genome sequence completed by international consortium.
- Sequence of the human T-cell receptor region completed.
- Wellcome Trust sponsors large-scale sequencing strategy meeting for international coordination of human genome sequencing.

1995

- LANL and LLNL announce high-resolution physical maps of chromosome 16 and chromosome 19, respectively.
- Moderate-resolution maps of chromosomes 3, 11, 12, and 22 maps published.
- Physical map with over 15,000 STS markers published.
- First (nonviral) whole genome sequenced (for the bacterium *Haemophilus influenzae*).
- Sequence of smallest bacterium, *Mycoplasma genitalium*, completed; provides a model of the minimum number of genes needed for independent existence.
- EEOC guidelines extend ADA employment protection to cover discrimination based on genetic information related to illness, disease, or other conditions.

Acronyms

1994

- Genetic-mapping 5-year goal achieved 1 year ahead of schedule.
- Completion of second-generation DNA clone libraries representing each human chromosome by LLNL and LBNL.
- Genetic Privacy Act, first U.S. HGP legislative product, proposed to regulate collection, analysis, storage, and use of DNA samples and genetic information obtained from them; endorsed by ELSI Working Group.
- DOE MGP launched; spin-off of HGP.
- LLNL chromosome paints commercialized.
- SBH technologies from ANL commercialized.
- DOE HGP Information Web site activated for public and researchers.

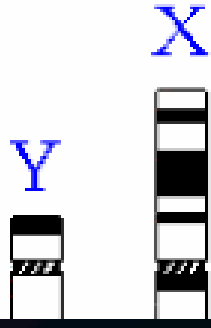
1993

- International IMAGE Consortium established to coordinate efficient mapping and sequencing of gene-representing cDNAs.
    - *The Scientist* 13[4]:17, Feb. 15, 1999 Hot Papers In Genomics:
      G. Lennon, C. Auffray, M. Polymeropoulos, M.B. Soares, "The I.M.A.G.E. Consortium: An Integrated Molecular Analysis of Genomes and Their Expression," Genomics, 33:151–2, 1996. (Cited in more than 290 papers since publication)
- DOE-NIH ELSI Working Group's Task Force on Genetic and Insurance Information releases recommendations.
- DOE and NIH revise 5-year goals [*Science* **262**, 43-46 (Oct. 1, 1993)].
- French Généthon provides mega-YACs to the genome community.
- IOM releases U.S. HGP-funded report, "Assessing Genetic Risks."
- LBNL implements novel transposon-mediated chromosome-sequencing system.
- GRAIL sequence-interpretation service provides Internet access at ORNL.

Acronyms

1992

- Low-resolution genetic linkage map of entire human genome published.
- Guidelines for data release and resource sharing announced by DOE and NIH.



1991

- Human chromosome mapping data repository, GDB, established.

Acronyms

1990

- DOE and NIH present joint 5-year U.S. HGP plan to Congress. The 15-year project formally begins.
- Projects begun to mark gene sites on chromosome maps as sites of mRNA expression.
- Research and development begun for efficient production of more stable, large-insert BACs.

Acronyms

1989

- DNA STSs recommended to correlate diverse types of DNA clones.
- DOE and NIH establish Joint ELSI Working Group.

1988

- Reports by congressional OTA and NAS NRC committees recommend concerted genome research program.
- HUGO founded by scientists to coordinate efforts internationally.
- First annual Cold Spring Harbor Laboratory meeting on human genome mapping and sequencing.
- DOE and NIH sign MOU outlining plans for cooperation on genome research.
- Telomere (chromosome end) sequence having implications for aging and cancer research is identified at LANL.

Acronyms

1987

- Congressionally chartered DOE advisory committee, HERAC, recommends a 15-year, multidisciplinary, scientific, and technological undertaking to map and sequence the human genome. DOE designates multidisciplinary human genome centers.
- NIH NIGMS begins funding of genome projects.

Acronyms

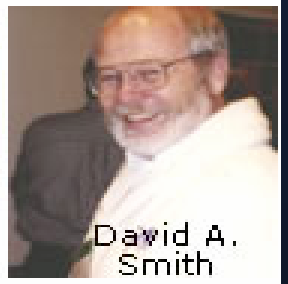
1986

- Following the Santa Fe conference, DOE OHER announces Human Genome Initiative. With $5.3 million, pilot projects begin at DOE national laboratories to develop critical resources and technologies.

1985

- Robert Sinsheimer holds meeting on human genome sequencing at University of California, Santa Cruz.
- At OHER, Charles DeLisi and David A. Smith commission the first Santa Fe conference to assess the feasibility of a Human Genome Initiative.


Charles DeLisi


David A. Smith

Acronyms

1984

- DOE OHER and ICPEMC cosponsor Alta, Utah, conference highlighting the growing role of recombinant DNA technologies. OTA incorporates Alta proceedings into report acknowledging value of human genome reference sequence.

Acronyms

1983

- LANL and LLNL begin production of DNA clone (cosmid) libraries representing single chromosomes.

Chromosome 22



heterochromatin
draft sequence

click on finished sequence to view the contig components

> 1000 kb
250 – 1000 kb
< 250 kb

Download the entire sequence!

STS Maps

Map	STSs	Type	Ref
GeneMap'99	565	RH	2
Whitehead	274	YAC	3
Stanford	182	RH	4
Genethon	67	Link	5

Sequencing is Complete!

Chromosome 22 is the smallest human chromosome and the first to be declared finished, according to a recent report by Dunham et al.(1).

For more in-depth information, see the 2 Dec 1999 issue of *Nature* or Nature On-Line:
Article
Features



The following genome sequencing laboratories have made major contributions to the sequencing of chromosome 22:
The Sanger Center
The University of Oklahoma
Keio University
Washington University

Euchromatic size:	34,491 kb
Finished sequence:	33,464 kb
Percent finished:	97 %
Number of contigs:	12

Disease Genes

Some diseases associated with chromosome 22 mutations

DiGeorge syndrome
Cat eye syndrome
Ewing sarcoma
Neurofibromatosis, type2
Velocardiofacial syndrome

Chromosome 22 Disease List

This table lists diseases and syndromes which are known to involve human chromosome 22. Links are provided to Online Mendelian Inheritance in Man (OMIM) for detailed descriptions of the conditions. Where a specific gene is known to have a primary involvement in the aetiology of the disease, a link is provided to the sequence of that gene. Where no gene has been identified as being causally involved this is indicated by "unkown".

Syndrome	Gene	OMIM	Position	Bases from centromere
Pheochromocytoma (deletions)	*unknown*	171300	22q	*unknown*
Meningioma	*unknown*	156100	22q	*unknown*
Cat eye syndrome	*unknown*	115470	22q11	*unknown*
Hyperprolinemia type 1	PRODH	239500	22q11.2	2798458
DiGeorge syndrome	*unknown*	188400	22q11	3346428
Bernard-Soulier syndrome, type B	GP1BB	138720	22q11.2	3590839
Thrombofilia due to heparin cofactor 2 def.	HCF2	142360	22q11	4723507
Breakpoint cluster region (CML)	BCR	151410	22q11.21	7113166
Malignant rhabdoid tumor	SMARCB1	601607	22q11	7717354
Glutathionuria	GGT1	231950	22q11-12	8614453
Breast cancer, t(11:22) associated	*unknown*	600048	22q11	*unknown*
Cataract, cerulean, type 2	CRYBB2	123620	22q11.2-12.2	9205376
Opitz G/BBB syndrome, autosomal dominant	*unknown*	145410	22q11-12	*unknown*
Schizophrenia 4	*unknown*	600850	22q11-q13	*unknown*
Ewing sarcoma breakpoint region 1	EWS	133450	22q12	13209321
Amyotrophic lateral sclerosis, susceptibility to	NEFH	162230	22q12.2	13421288

Genomics and Its Impact on Science and Society:

The Human Genome Project and Beyond

U.S. Department of Energy Genome Programs
http://doegenomes.org

Human Genome Project

Goals:

- identify all the approximate 30,000 genes in human DNA,
- determine the sequences of the 3 billion chemical base pairs that make up human DNA,
- store this information in databases,
- improve tools for data analysis,
- transfer related technologies to the private sector, and
- address the ethical, legal, and social issues (ELSI) that may arise from the project.

Milestones:

- 1990: Project initiated as joint effort of U.S. Department of Energy and the National Institutes of Health
- June 2000: Completion of a working draft of the entire human genome
- February 2001: Analyses of the working draft are published
- April 2003: HGP sequencing is completed and Project is declared finished two years ahead of schedule

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

Specific Goals (Phase II 1998-2003)

Area	Goals 1993-98	Status as of Oct. 1998	Goals 1998-2003
Genetic map	Average 2- to 5-cM resolution	1 cM map published Sept. 1994	Completed
Physical map	Map 30,000 STSs	52,000 STSs mapped	Completed
DNA sequence	Complete 80 Mb for all organisms by 1998	180 Mb human plus 111 Mb nonhuman	Finish 1/3 of human sequence by end of 2001 Working draft of remainder by end of 2001 Complete human sequence by end of 2003
Sequencing technology	Evolutionary improvements and innovative technologies	90 Mb/year capacity at ~$0.50 per base Capillary array electrophoresis validated Microfabrication feasible	Integrate and automate to achieve 500 Mb/year at $\leq$$0.25 per base Support innovation
Human sequence variation	Not a goal	-	100,000 mapped SNPs
Gene identification	Develop technology	30,000 ESTs mapped	Develop technology Full-length cDNAs
Functional analysis	Not a goal	-	Develop genomic-scale technologies
Model organisms	*E. coli*: complete sequence	Published Sept. 1997	-
	Yeast: complete sequence	Released Apr. 1996	-
	C. elegans: most of sequence	80% complete	Complete Dec. 1998
	Drosophila: begin sequencing	9% done	Sequence by 2002
	Mouse: map 10,000 STSs	12,000 STSs mapped	Develop extensive genomic resources Lay basis for finishing sequence by 2005 Produce working draft before 2005

Bioinformatics

Specific Goals (Phase I 1993-1998)

Genetic Map

Complete the 2-5 cM map by 1995
Develop technology for rapid genotyping
Develop markers that are easier to use
Develop new mapping technologies

Physical Map

Complete an STS map of the human genome at a resolution of 100 kb

DNA Sequencing

Develop efficient approaches to sequencing one- to several- megabase regions of DNA of high biological interest.
Develop technology for high throughput sequencing, focusing on systems integration of all steps from template preparation to data analysis.
Build up sequencing capacity to a collective rate of 50 Mb per year by the end of the period. This rate should result in an aggregate of 80 Mb of DNA sequence completed by the end of FY 1998.

Bioinformatics

Specific Goals (Phase I 1993-1998)

Gene Identification

Develop efficient methods of identifying genes and for placement of known genes on physical maps or sequenced DNA.

Technology Development

Substantially expand support of innovative technological developments as well as improvements in current technology for DNA sequencing and to meet the needs of the Human Genome Project as a whole.

Model Organisms

Finish an STS map of the mouse at 300 Kb resolution
Finish the sequence of the E. coli and S. cerevisiae genomes by 1998 or earlier
Continue sequencing C. elegans and Drosophila genomes, with the aim of bringing C. elegans to near completion by 1998
Sequence selected segments of mouse DNA side by side with corresponding human DNA in areas of high biological interest

Bioinformatics

Specific Goals (Phase I 1993-1998)

Informatics

Continue to create, develop and operate databases and database tools for easy access to data, including effective tools and standards for data exchange and links among databases

Consolidate, distribute and continue to develop effective software for large-scale genome projects

Continue to develop tools for comparing and interpreting genome information

Ethical, Legal and Social Implications (ELSI)

Continue to identify and define issues and develop policy options to address them

Develop and disseminate policy options regarding genetic testing services with widespread potential use

Foster greater acceptance of human genetic variation

Enhance and expand public and professional education that is sensitive to sociocultural and psychological issues

Training

Continue to encourage training of scientists in interdisciplinary sciences related to genome research

Bioinformatics

Specific Goals (Phase I 1993-1998)

Technology Transfer

Encourage and enhance technology transfer both into and out of centers of genome research

Outreach

Cooperate with those who would set up distribution centers for genome materials. Share all information and materials within 6 months of their development. This should be accomplished by submission to public databases or repositories, or both, where appropriate.

Specific Goals (Phase II 1998-2003)

Area	Goals 1993-98	Status as of Oct. 1998	Goals 1998-2003
Genetic map	Average 2- to 5-cM resolution	1 cM map published Sept. 1994	Completed
Physical map	Map 30,000 STSs	52,000 STSs mapped	Completed
DNA sequence	Complete 80 Mb for all organisms by 1998	180 Mb human plus 111 Mb nonhuman	Finish 1/3 of human sequence by end of 2001 Working draft of remainder by end of 2001 Complete human sequence by end of 2003
Sequencing technology	Evolutionary improvements and innovative technologies	90 Mb/year capacity at ~$0.50 per base Capillary array electrophoresis validated Microfabrication feasible	Integrate and automate to achieve 500 Mb/year at $\leq$$0.25 per base Support innovation
Human sequence variation	Not a goal	-	100,000 mapped SNPs
Gene identification	Develop technology	30,000 ESTs mapped	Develop technology Full-length cDNAs
Functional analysis	Not a goal	-	Develop genomic-scale technologies
Model organisms	*E. coli*: complete sequence	Published Sept. 1997	-
	Yeast: complete sequence	Released Apr. 1996	-
	C. elegans: most of sequence	80% complete	Complete Dec. 1998
	Drosophila: begin sequencing	9% done	Sequence by 2002
	Mouse: map 10,000 STSs	12,000 STSs mapped	Develop extensive genomic resources Lay basis for finishing sequence by 2005 Produce working draft before 2005

Bioinformatics

Specific Goals (Phase II 1998-2003)

Goal 1--The Human DNA Sequence

a) Finish the complete human genome sequence by the end of 2003.

b) Finish one-third of the human DNA sequence by the end of 2001.

c) Achieve coverage of at least 90% of the genome in a working draft based on mapped clones by the end of 2001.

d) Make the sequence totally and freely accessible.

Bioinformatics

Specific Goals (Phase II 1998-2003)

Goal 2--Sequencing Technology

a) Continue to increase the throughput and reduce the cost of current sequencing technology.

b) Support research on novel technologies that can lead to significant improvements in sequencing technology.

c) Develop effective methods for the advanced development and introduction of new sequencing technologies into the sequencing process.

Bioinformatics

Specific Goals (Phase II 1998-2003)

Goal 3--Human Genome Sequence Variation

a) Develop technologies for rapid, large-scale identification or scoring, or both, of SNPs and other DNA sequence variants.

b) Identify common variants in the coding regions of the majority of identified genes during this 5-year period.

c) Create an SNP map of at least 100,000 markers.

d) Develop the intellectual foundations for studies of sequence variation.

e) Create public resources of DNA samples and cell lines.

Bioinformatics

Specific Goals (Phase II 1998-2003)

Goal 4--Technology for Functional Genomics

a) Develop cDNA resources.

b) Support research on methods for studying functions of non-protein-coding sequences.

c) Develop technology for comprehensive analysis of gene expression.

d) Improve methods for genome-wide mutagenesis.

e) Develop technology for global protein analysis.

Bioinformatics

Specific Goals (Phase II 1998-2003)

Goal 5--Comparative Genomics

a) Complete the sequence of the C. elegans genome in 1998.

b) Complete the sequence of the Drosophila genome by 2002.

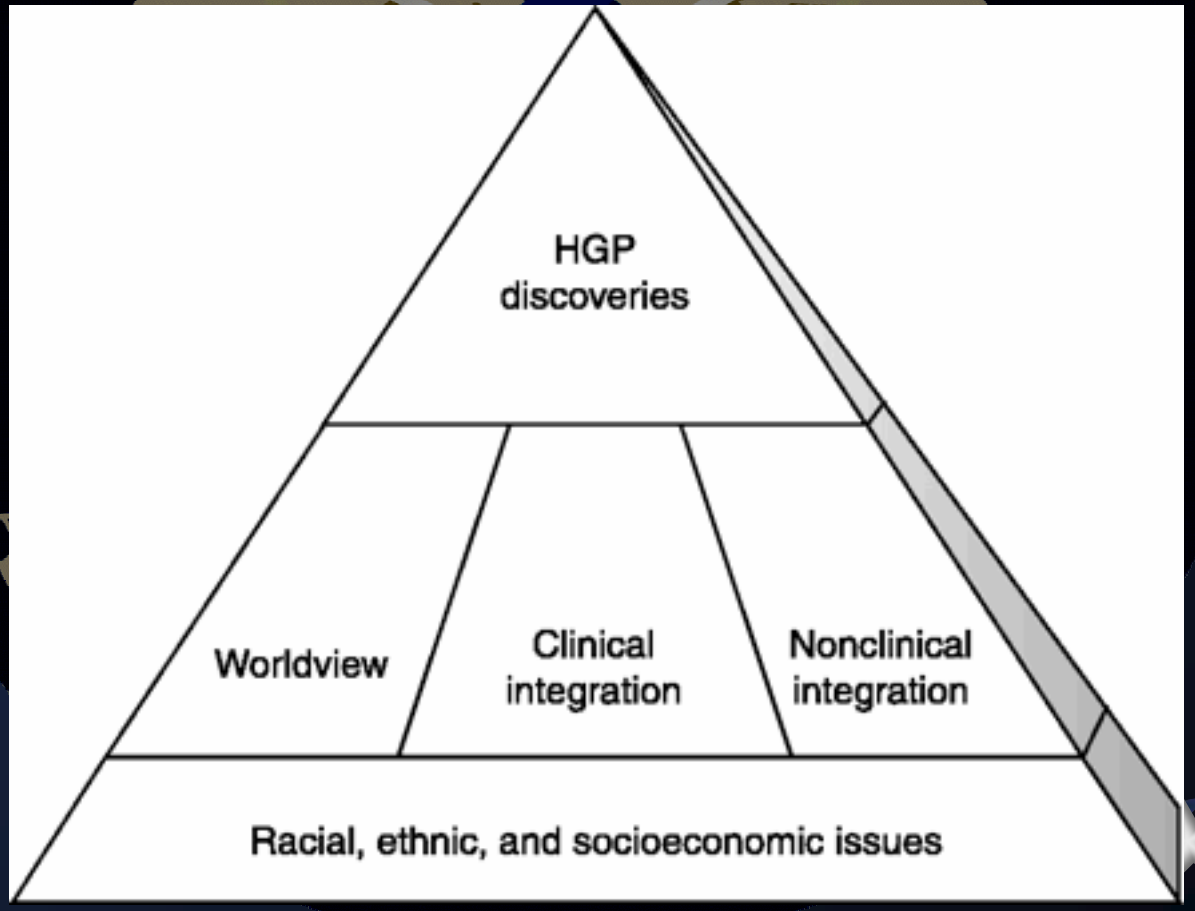
c) The mouse genome.

1) Develop physical and genetic mapping resources.
2) Develop additional cDNA resources.
3) Complete the sequence of the mouse genome by 2005.

d) Identify other model organisms that can make major contributions to the understanding of the human genome and support appropriate genomic studies.

Bioinformatics

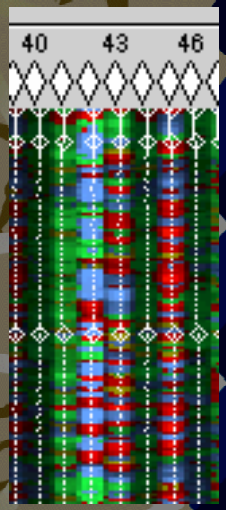
Goal 6--Ethical, Legal, and Social Implications (ELSI)

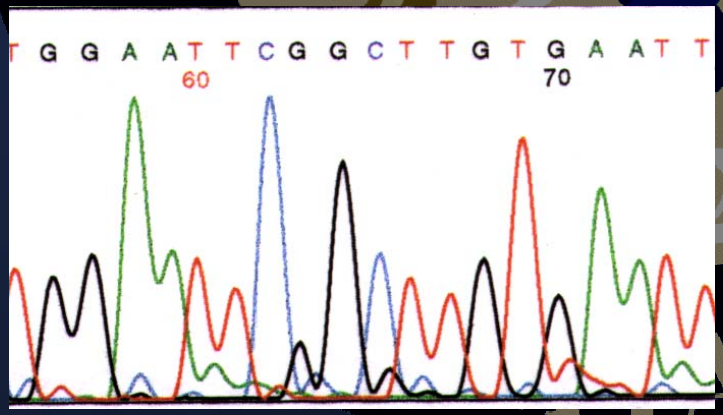


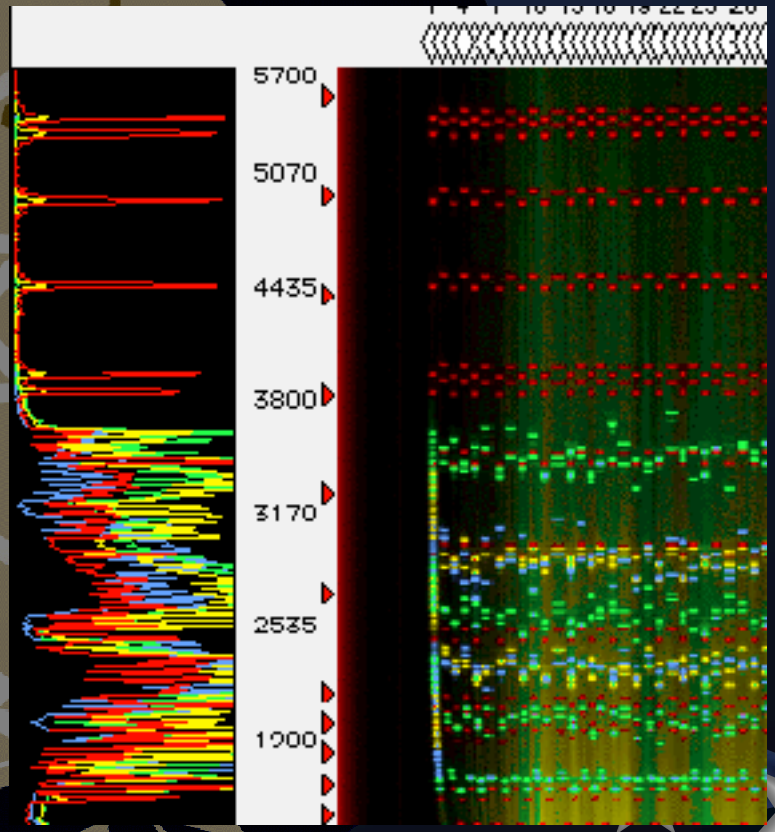
Bioinformatics



Automatic sequencer







Bioinformatics

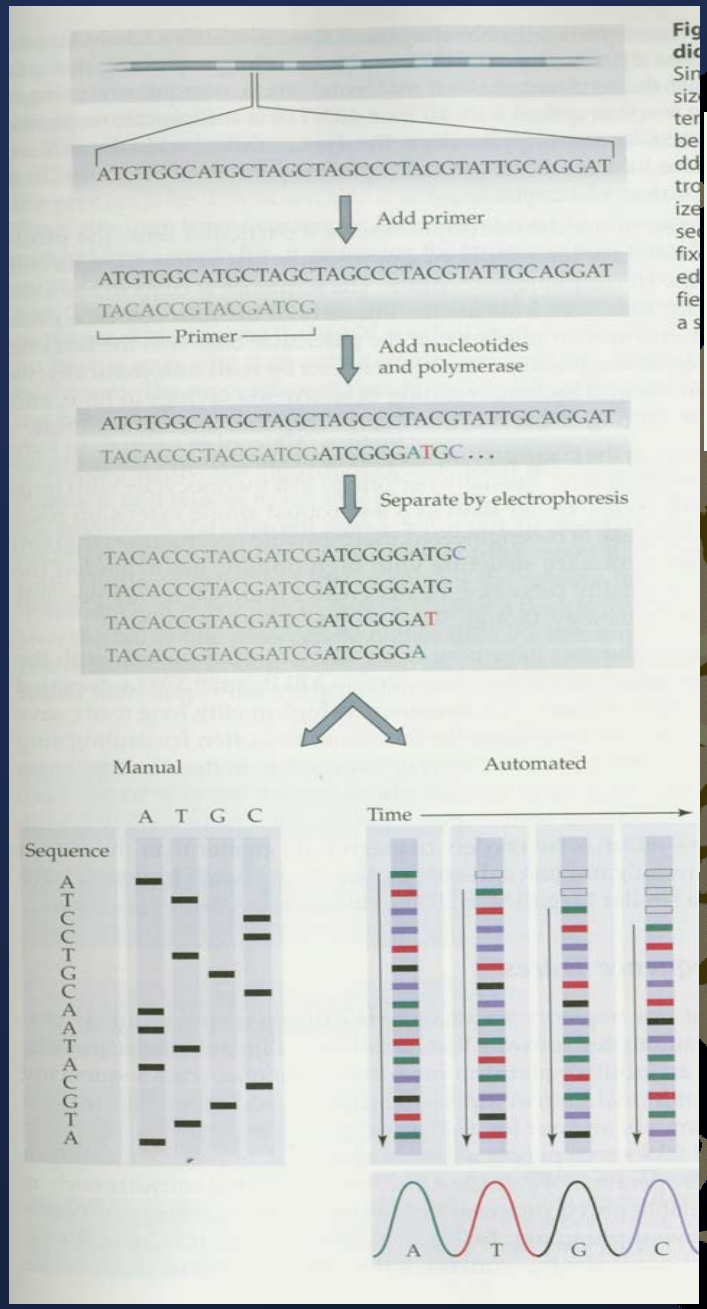
The dawn of sequencing:

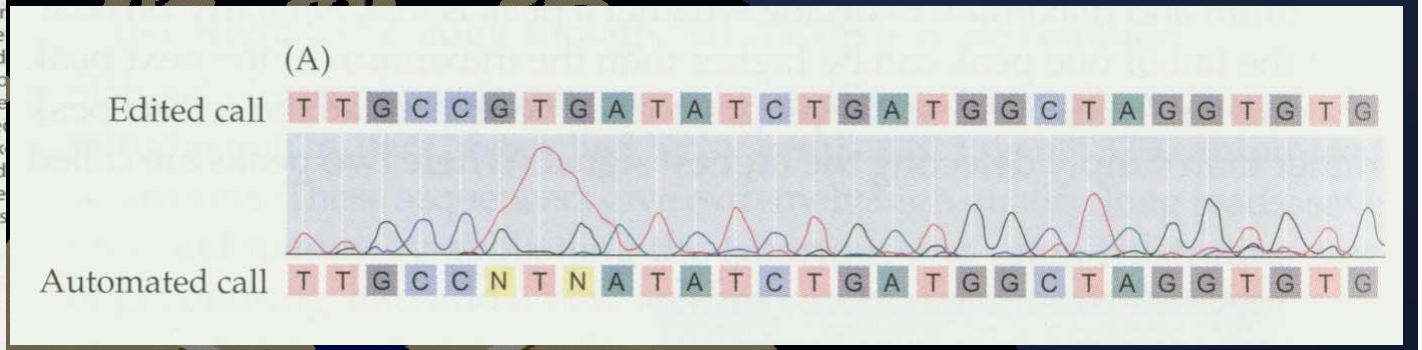
Protein sequencing-

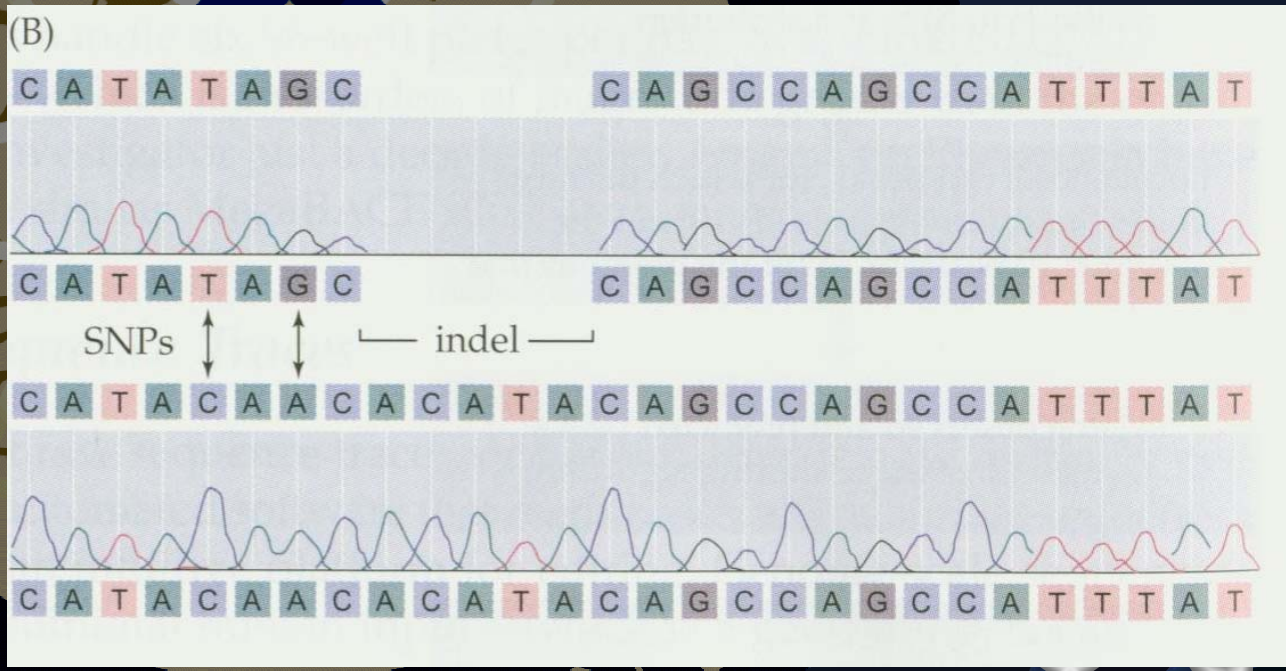
The earliest techniques were based on methods for separation of proteins and peptides, coupled with methods for identification and quantification of amino acids. Initially, the majority of protein sequences were obtained by the manual process of sequential Edman degradation - dansylation. Advanced in mass spectrometry allowed significant progress.

Nucleic acid sequencing-

The first technique were only applicable to RNA. Fast, and efficient DNA sequencing had emerged using chain termination (Sanger) and chemical degradation (Maxam and Gilbert) approaches.







A Primer of Genome Science by Gibson and Muse, 2002. Sinauer Associates Inc.

Bioinformatics

The Growth of GenBank sequence database in the past 10 years.

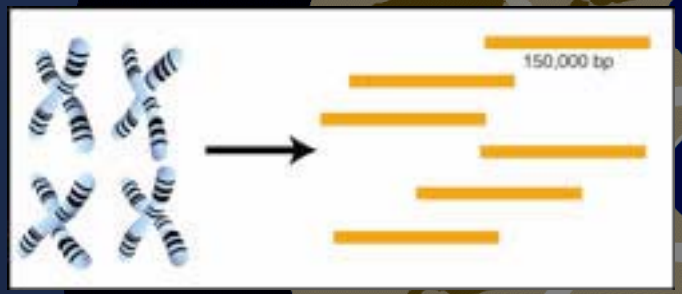
Release	Year	Base pairs	Entries
58	88	24,690,876	21,248
62	89	37,183,950	31,229
66	90	51,306,092	41,057
70	91	77,337,678	58,952
74	92	120,242,234	97,084
80	93	163,802,597	150,744
86	94	230,485,928	237,775
92	95	425,860,958	620,765
98	96	730,552,938	1,114,581
104	97	1,258,290,513	1,891,953
110	98	2,162,067,871	3,043,729
115	99	4,653,932,745	5,354,511

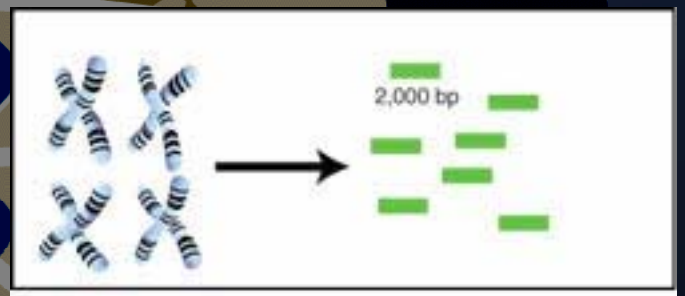
BAC to BAC Sequencing

The BAC to BAC approach first creates a crude physical map of the whole genome before sequencing the DNA. Constructing a map requires cutting the chromosomes into large pieces and figuring out the order of these big chunks of DNA before taking a closer look and sequencing all the fragments.

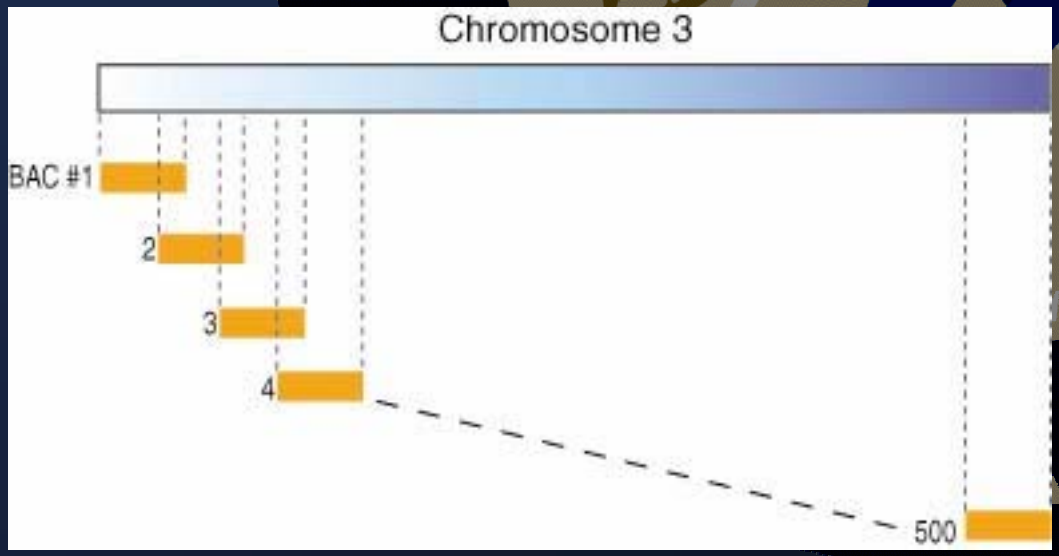
Whole Genome Shotgun Sequencing

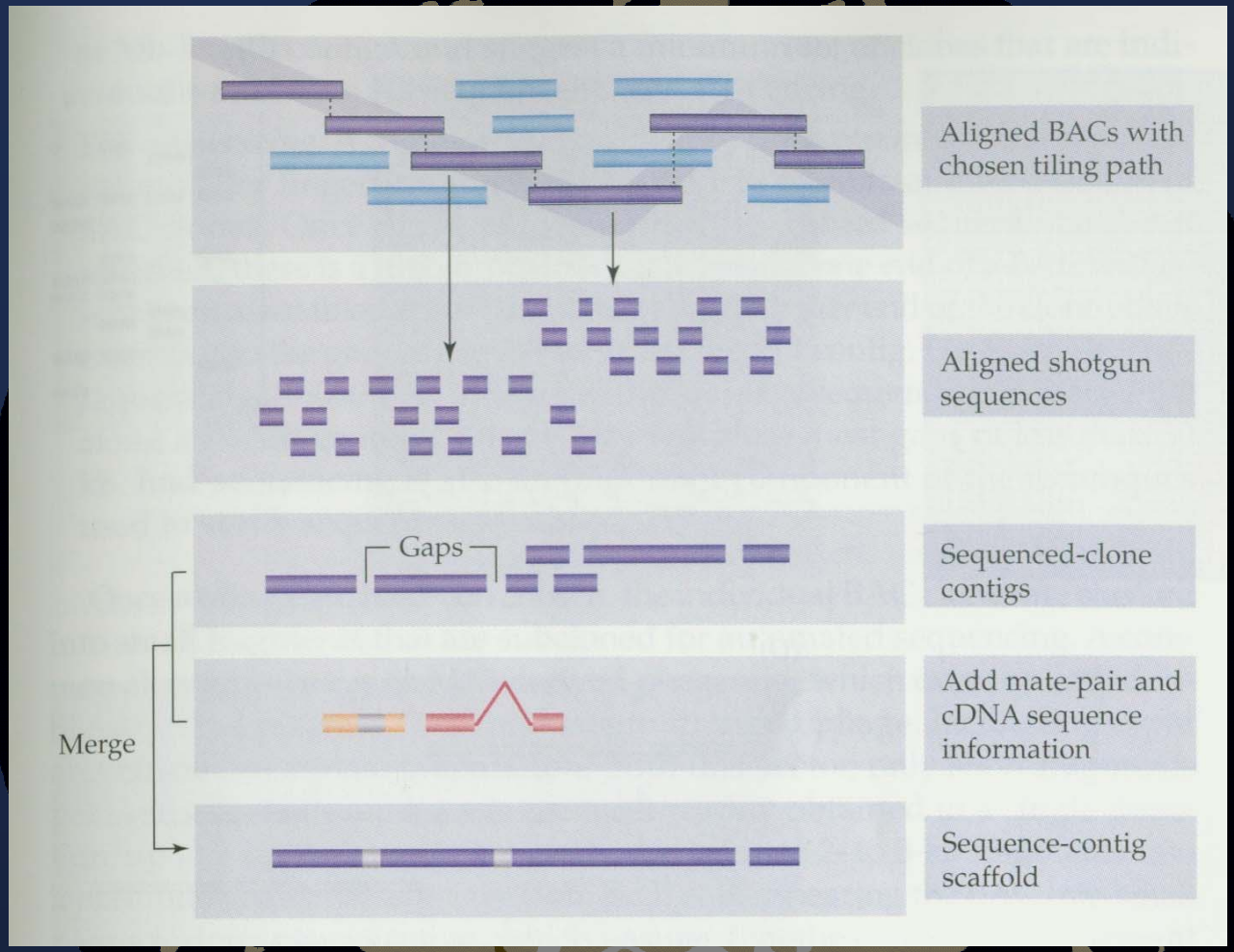
The shotgun sequencing method goes straight to the job of decoding, bypassing the need for a physical map. Therefore, it is much faster.



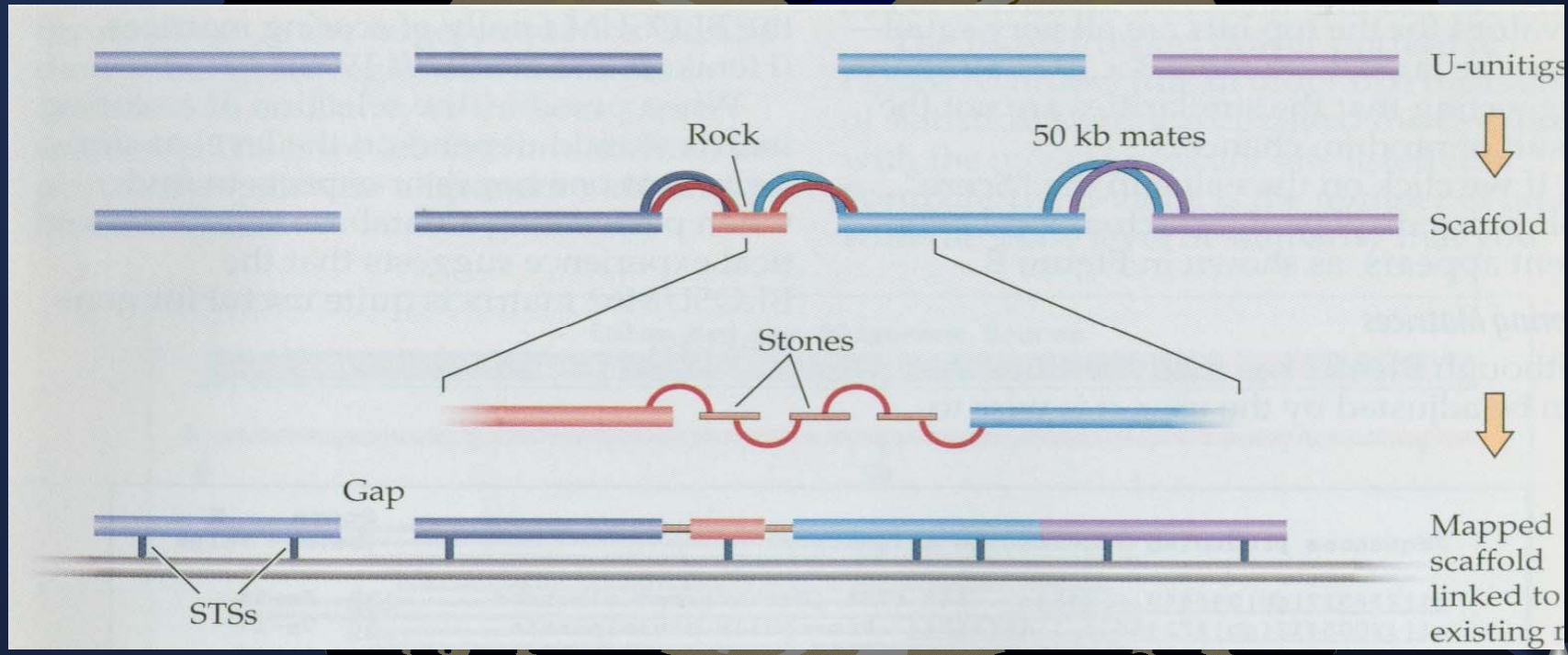






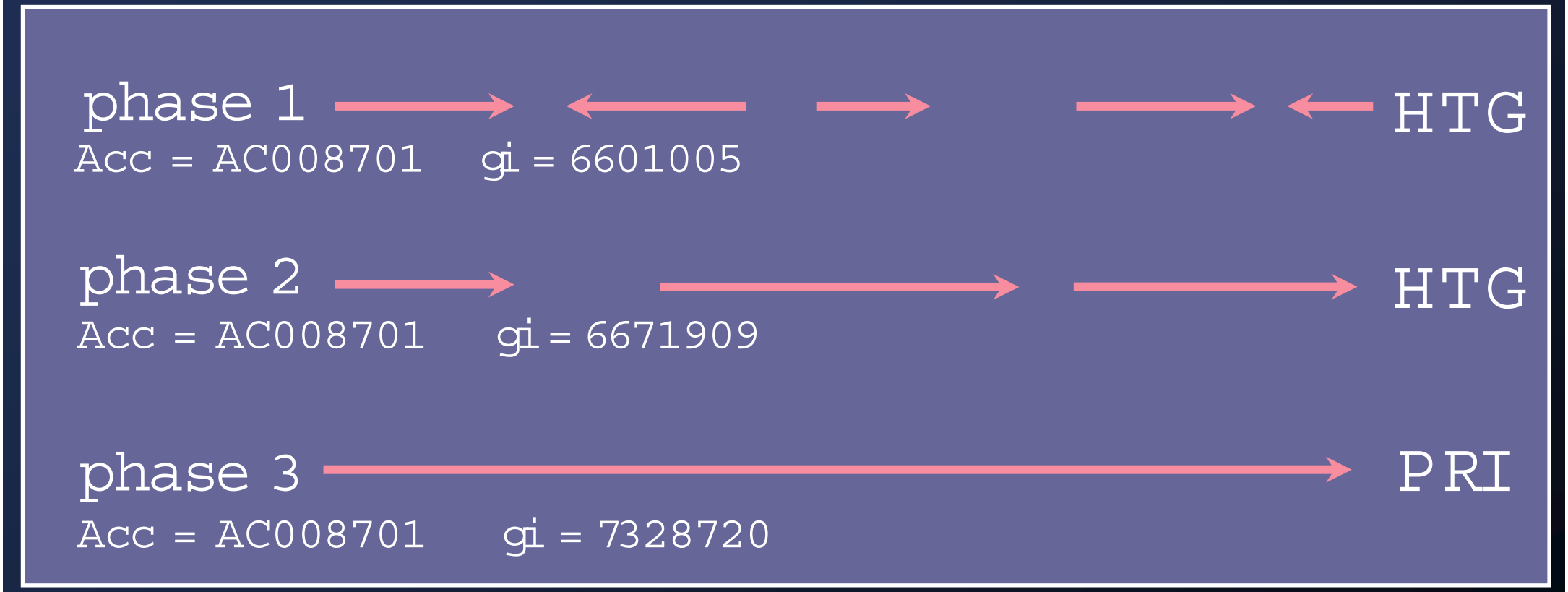


A Primer of Genome Science by Gibson and Muse, 2002. Sinauer Associates Inc.



A Primer of Genome Science by Gibson and Muse, 2002. Sinauer Associates Inc.

HTG Division:
High Throughput Genome Records



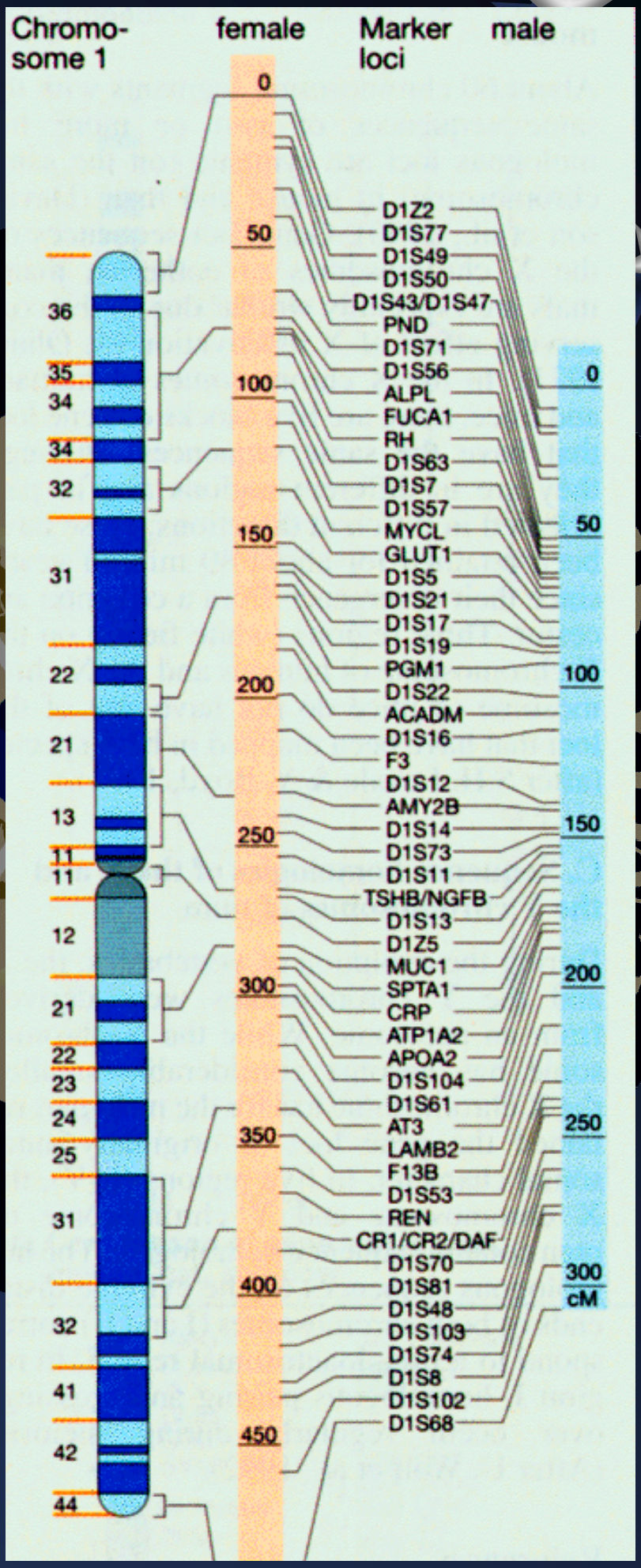
40,000 to > 350,000 bp

Physical Mapping of the Genome

This page briefly presents some techniques used for the physical mapping of the genome. This is a very active domain, and we do not pretend to be exhaustive. For more details, see the References, or have a look at the related sites.

http://www.pasteur.fr/recherche/unites/biophyadn/e-mapping.html

Bioinformatics



Techniques not using FISH

None of today's techniques allows for a precise positioning of the probes down to one nucleotidic base. It is thus necessary to use overlaping clones, that is, clones with a common part. Covering of a region of the genome can then be done by a set of partially overlaping clones, also called a contig (for contiguous clones).

Building a contig of clones for a given region is thus the first step of physical mapping. Basically, one picks up clones out of a clones librairy or bank obtained by systematic cloning of pieces resulting from the enzymatic digestion of the whole genome. These clones are chosen when they are positive for markers specific of the studied region, and have to be organised by physical mapping: one thus obtains a minimal continuous string of overlaoing clones which can eventually be sequenced.

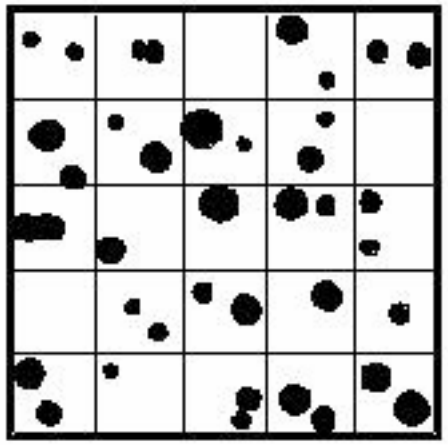
One method used for organising these clones uses a matrix organisation of these clones on a membrane, which are then hybridized to one anothers, and detected by autoradiography. A combinatorial analysis then unambiguously gives the order of the overlaping clones.

Overlap Detection by Hybridization

1. Select a set of clones as probes.



2. Hybridize the probes against the clones.



	probe1	probe2	probe3	probe4	probe5	probe6
clone1	0	1	1	C	1	0
clone2	1	0	1	C	1	0
clone3	1	0	C	C	1	1
clone4	0	1	C	C	0	0
clone5	1	0	C	C	0	1
clone6	1	0	1	C	0	0
clone7	0	0	C	1	0	1

Techniques using FISH

The above contig organisation does not give a precise information on the size of the overlap between two clones, and even less on their distance (or gap between clones).

Different techniques have been developed in the last few years, in order to measure more precisely the respective position of clones onto a partially linearized DNA fiber. All these techniques use FISH (or fluorescente *in situ* hybridization), which was initially developped for detecting probes on chromosomes in metaphase.

They all consist in first preparing cells on a slide, and in more or less completely decondensing and/or releasing the DNA out of the cell nucleus. /p>

The hybridization of this DNA with labelled probes followed by detection with fluorecent systems (see FISH) allows the direct observation by epifluorescence microscopy of the relative position of the probes. The precision of these observations depend on the technique (see the articles cited in reference for more details), but is impaired by a rather low reproducibility:

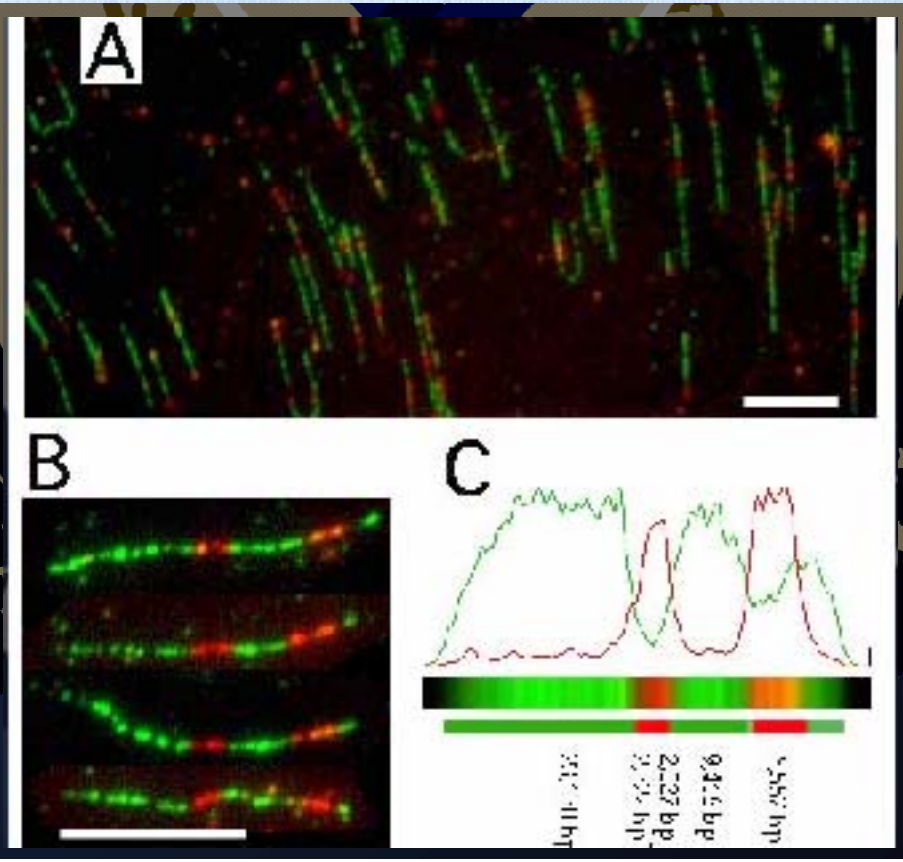
- Halo method(cf [1]).
- extended chromatin (cf [2]).
- extended fiber (cf [3] et [4]).

Molecular combing and FISH

The method developed in collaboration with Joe W. Gray's group (University of California) improves these measurements both qualitatively and quantitatively.

It is based on molecular combing, which allows a regular stretching of the target DNA (cf [5]). Protocols used in standard fluorescent hybridization assays work perfectly in these new conditions. This technique yields physical mapping results which precision is in principle only limited by the optical resolution, i.e. roughly 0.5 microns, corresponding to approximately 1000 bases (1 kb).

The following figure (cf [6]) gives an example of the first results obtained with this technique:

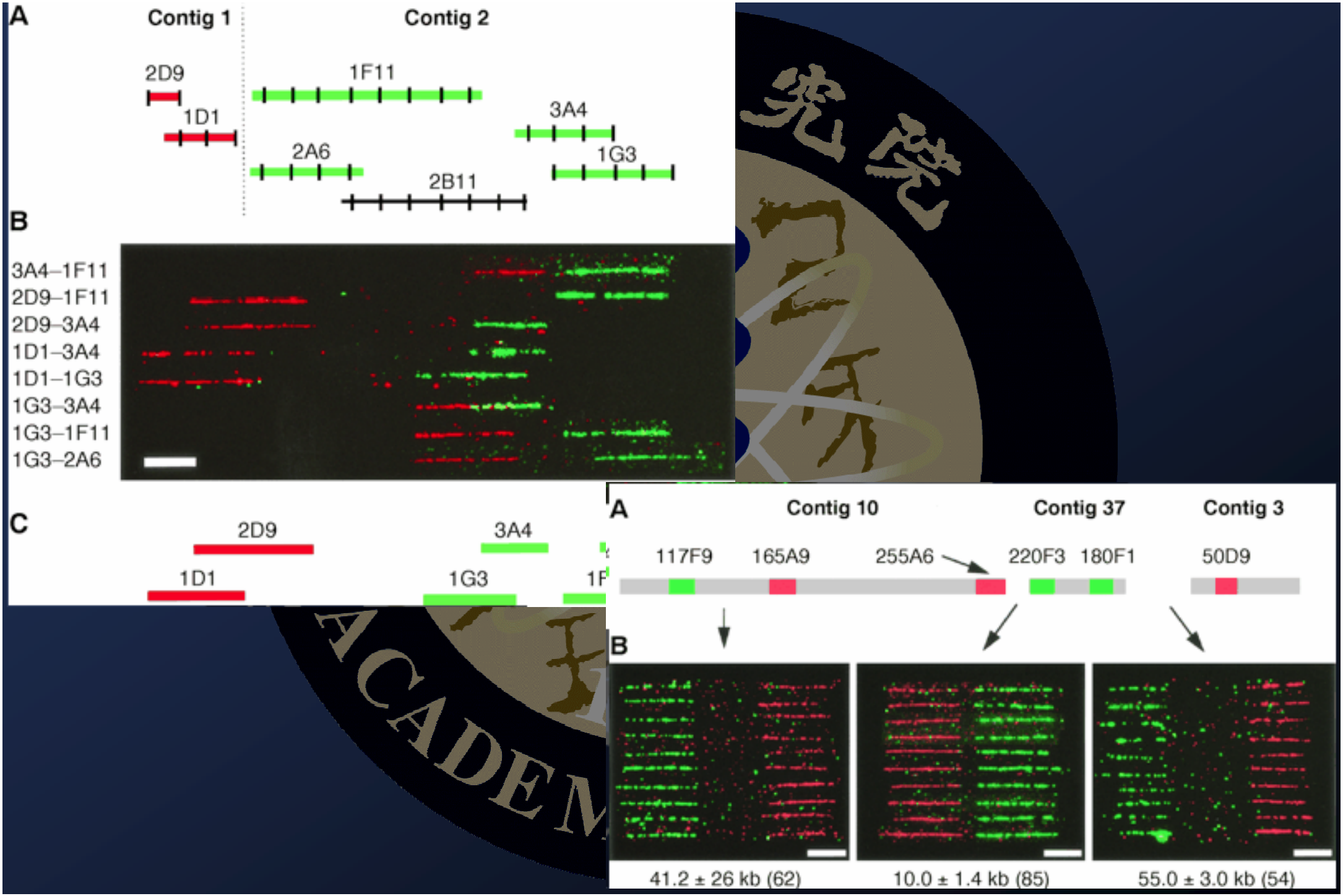


High-Resolution Physical Mapping

With Françoise Fougerousse, from Jacques Beckmann's group at the Généthon, we looked for the transcriptional orientation of the calpain 3 gene, located on chromosome 15q15.2. Mutations of this gene are responsible for a limb girdle muscular distrophy (LGMD 2A) [7].

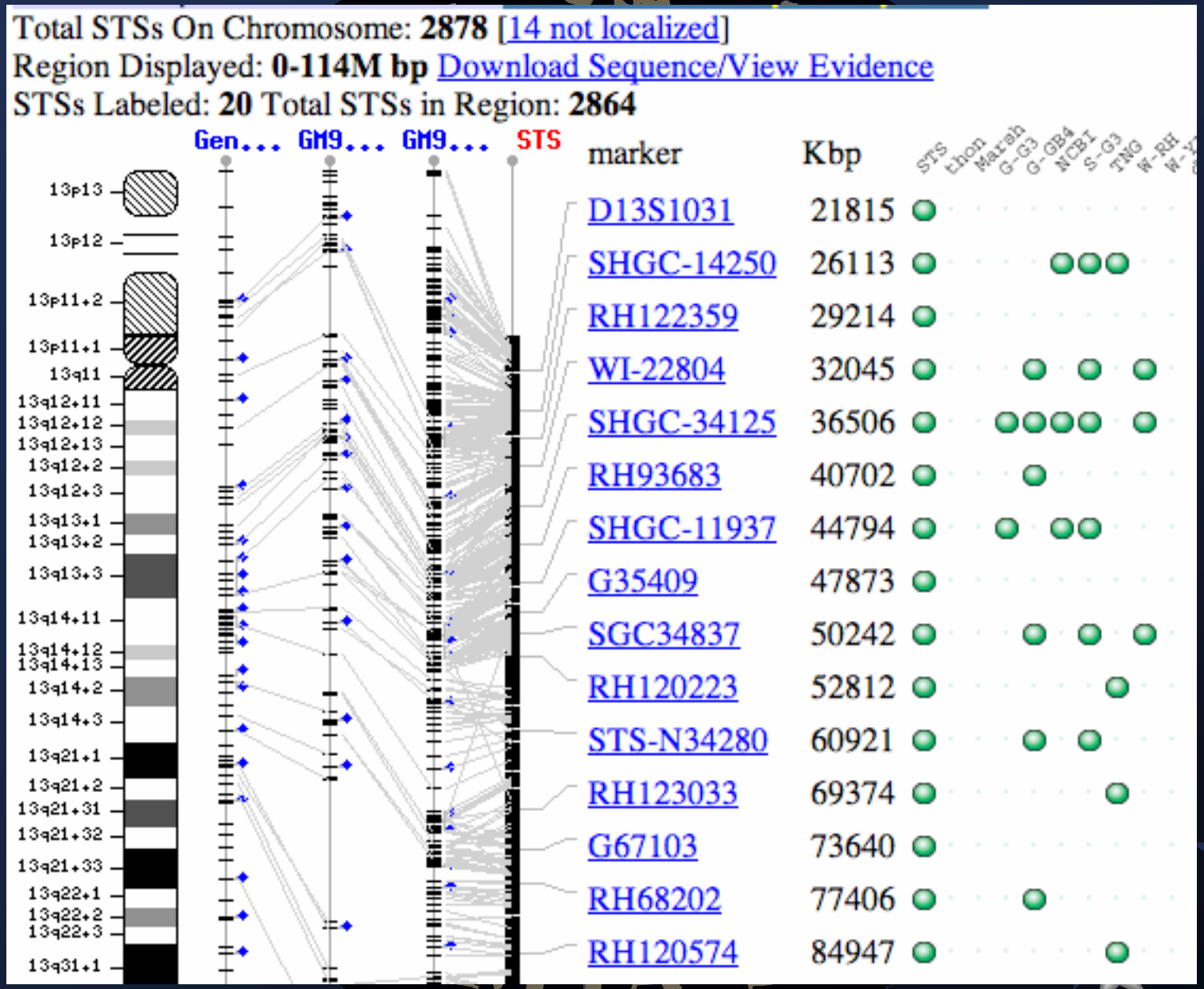
We used cosmids from two contigs which had already ben oriented using STS (Sequenced Tag Site) markers. One of the contig contained all the exons of the gene (contig 2 on next figure), but was of unknown orientation with respect to the centromere of the chromosome. Cosmids were hybridized by pairs on combed DNA from a yeast strain containing a 1.6 Mb human genomic clone of the region (YAC 774G4). The high density of DNA fibers balances the useless DNA of the yeast (12.6 Mb) which plays no role in the mapping study, but which was combed with the YAC, in order to avoid the tedious separation step of the YAC.

In each experiment, cosmids were labelled differently, in order to detect them with different fluorescent antibody systems (red, in case antibodies coupled to Texas Red were used, green when antibodies coupled to FITC were used). The direct measurement of sizes and distances on a few tens pairs of signals allowed the straightforward construction of a precise map of the region. This map is depicted on the following figure, and extend over 200 kb with a precision of a few kb. (From reference [7]):


A
Contig 1
Contig 2
2D9
1D1
1F11
2A6
2B11
3A4
1G3
B
3A4–1F11
2D9–1F11
2D9–3A4
1D1–3A4
1D1–1G3
1G3–3A4
1G3–1F11
1G3–2A6
C
2D9
1D1
3A4
1G3
A
Contig 10
Contig 37
Contig 3
117F9
165A9
255A6
220F3
180F1
50D9
B
41.2 ± 26 kb (62)
10.0 ± 1.4 kb (85)
55.0 ± 3.0 kb (54)

Sequence tagged site (STS): A short (200 to 500 base pair) DNA sequence that occurs but once in the genome and whose location and base sequence are known. STSs are detectable by polymerase chain reaction (PCR), are useful for localizing and orienting the mapping and sequence data, and serve as landmarks on the physical map of a genome.

Random amplified polymorphic DNAs (RAPDs) is a method of PCR where arbitrarily chosen 10 base primers are used to search for variation in DNA (Williams et al., 1990). RAPDs data can contain artifacts (Ellsworth et al. 1993), and are not fully reproduceable (Paterson 1996). However, RAPDs have been be used to generate large numbers of genetic markers useful for linakge mapping quickly and cheaply (Antolin et al., 1996).


Total STSs On Chromosome: 2878 [14 not localized]
Region Displayed: 0-114M bp Download Sequence/View Evidence
STSs Labeled: 20 Total STSs in Region: 2864
Gen... GM9... GM9... STS
marker
Kbp
STS
Lhon
Marsh
G-G3
G-GB4
NCBI
S-G3
TNG
W-RH
13p13
13p12
13p11.2
13p11.1
13q11
13q12.11
13q12.12
13q12.13
13q12.2
13q12.3
13q13.1
13q13.2
13q13.3
13q14.11
13q14.12
13q14.13
13q14.2
13q14.3
13q21.1
13q21.2
13q21.31
13q21.32
13q21.33
13q22.1
13q22.2
13q22.3
13q31.1
D13S1031 21815
SHGC-14250 26113
RH122359 29214
WI-22804 32045
SHGC-34125 36506
RH93683 40702
SHGC-11937 44794
G35409 47873
SGC34837 50242
RH120223 52812
STS-N34280 60921
RH123033 69374
G67103 73640
RH68202 77406
RH120574 84947

What does the draft human genome sequence tell us?

By the Numbers

- The human genome contains 3 billion chemical nucleotide bases (A, C, T, and G).
- The average gene consists of 3000 bases, but sizes vary greatly, with the largest known human gene being dystrophin at 2.4 million bases.
- The total number of genes is estimated at around 30,000--much lower than previous estimates of 80,000 to 140,000.
- Almost all (99.9%) nucleotide bases are exactly the same in all people.
- The functions are unknown for over 50% of discovered genes.

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

What does the draft human genome sequence tell us?

How It's Arranged

• The human genome's gene-dense "urban centers" are predominantly composed of the DNA building blocks G and C.

• In contrast, the gene-poor "deserts" are rich in the DNA building blocks A and T. GC- and AT-rich regions usually can be seen through a microscope as light and dark bands on chromosomes.

• Genes appear to be concentrated in random areas along the genome, with vast expanses of noncoding DNA between.

• Stretches of up to 30,000 C and G bases repeating over and over often occur adjacent to gene-rich areas, forming a barrier between the genes and the "junk DNA." These CpG islands are believed to help regulate gene activity.

• Chromosome 1 has the most genes (2968), and the Y chromosome has the fewest (231).

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

What does the draft human genome sequence tell us?

The Wheat from the Chaff

- Less than 2% of the genome codes for proteins.

- Repeated sequences that do not code for proteins ("junk DNA") make up at least 50% of the human genome.

- Repetitive sequences are thought to have no direct functions, but they shed light on chromosome structure and dynamics. Over time, these repeats reshape the genome by rearranging it, creating entirely new genes, and modifying and reshuffling existing genes.

- The human genome has a much greater portion (50%) of repeat sequences than the mustard weed (11%), the worm (7%), and the fly (3%).

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

What does the draft human genome sequence tell us?

How the Human Compares with Other Organisms

- Unlike the human's seemingly random distribution of gene-rich areas, many other organisms' genomes are more uniform, with genes evenly spaced throughout.

- Humans have on average three times as many kinds of proteins as the fly or worm because of mRNA transcript "alternative splicing" and chemical modifications to the proteins. This process can yield different protein products from the same gene.

- Humans share most of the same protein families with worms, flies, and plants; but the number of gene family members has expanded in humans, especially in proteins involved in development and immunity.

- Although humans appear to have stopped accumulating repeated DNA over 50 million years ago, there seems to be no such decline in rodents. This may account for some of the fundamental differences between hominids and rodents, although gene estimates are similar in these species. Scientists have proposed many theories to explain evolutionary contrasts between humans and other organisms, including those of life span, litter sizes, inbreeding, and genetic drift.

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

What does the draft human genome sequence tell us?

Variations and Mutations

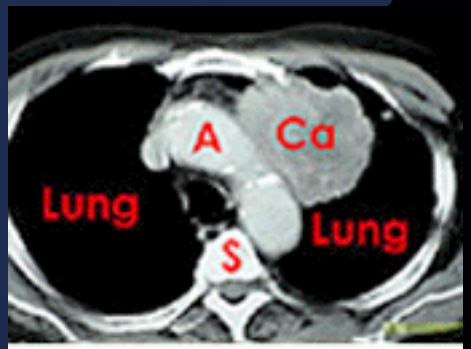
- Scientists have identified about 3 million locations where single-base DNA differences (SNPs) occur in humans. This information promises to revolutionize the processes of finding chromosomal locations for disease-associated sequences and tracing human history.

- The ratio of germline (sperm or egg cell) mutations is 2:1 in males vs females. Researchers point to several reasons for the higher mutation rate in the male germline, including the greater number of cell divisions required for sperm formation than for eggs.

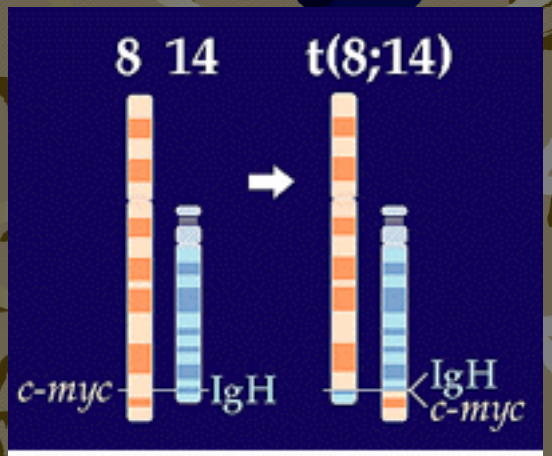
U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

NCBI Genes and disease Map

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y



CT scan showing lung cancer. [Image credit: Pat Connelly, Miami Valley Hospital, Dayton, OH, USA.]



In Burkitt lymphoma, Myc, which is normally found on chromosome 8, is transferred to chromosome 14. This is known as chromosome translocation and can be characteristic of a cancer type. [image credit: Gregory Schuler, NCBI, NLM, NIH.]

Cancer
uncontrolled cell division

Immune system
fighting invaders

NEW Metabolism
food and energy

NEW Muscle and bone
movement and growth

Nervous system
mind and body

Signals
cellular messages

NEW **Genome View**

OB
on chromosome 7

Databases

PubMed
the literature

LocusLink
collection of gene-related information

OMIM
catalog of human genes and disorders

Information

Weight-control information network
from the National Institute of Diabetes and Digestive and Kidney Diseases, NIH

OBESITY is an excess of body fat that frequently results in a significant impairment of health. Doctors generally agree that men with more than 25% body fat and women with more than 30% are obese. Obesity is a known risk factor for chronic diseases including heart disease, diabetes, high blood pressure, stroke and some forms of cancer. Evidence suggests that obesity has more than one cause: genetic, environmental, psychological and other factors may all play a part.

The hormone leptin, produced by adipocytes (fat cells), was discovered about three years ago in mice. Subsequently the human Ob gene was mapped to chromosome 7. Leptin is thought to act as a lipostat: as the amount of fat stored in adipocytes rises, leptin is released into the blood and signals to the brain that the body has enough to eat. However, most overweight people have high levels of leptin in their bloodstream, indicating that other molecules also effect feelings of satiety and contribute to the regulation of body weight.

The discovery of leptin has initiated a flurry of research into the molecular basis of weight control. A whole network of signals contributes to weight homeostasis, and other key players are being discovered on an ongoing basis. Mice have proved to be an extremely useful model for human obesity, and have helped to begin to unravel the components that contribute to maintaining body weight. Since the market for effective weight-reducing therapies is enormous, drug companies are working alongside basic scientists to find possible drug targets among the tangle of molecules that control body weight.



A mouse with the Obese (Ob) mutation and a normal mouse. [Image credit: Jeff Friedman, Rockfeller University, New York, NY, USA. Reprinted from Science, with permission.]

Future Challenges: What We Still Don't Know

- Gene number, exact locations, and functions
- Gene regulation
- DNA sequence organization
- Chromosomal structure and organization
- Noncoding DNA types, amount, distribution, information content, and functions
- Coordination of gene expression, protein synthesis, and post-translational events
- Interaction of proteins in complex molecular machines
- Predicted vs experimentally determined gene function
- Evolutionary conservation among organisms
- Protein conservation (structure and function)
- Proteomes (total protein content and function) in organisms
- Correlation of SNPs (single-base DNA variations among individuals) with health and disease
- Disease-susceptibility prediction based on gene sequence variation
- Genes involved in complex traits and multigene diseases
- Complex systems biology including microbial consortia useful for environmental restoration
- Developmental genetics, genomics

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

Anticipated Benefits of Genome Research

Molecular Medicine

- improve diagnosis of disease
- detect genetic predispositions to disease
- create drugs based on molecular information
- use gene therapy and control systems as drugs
- design “custom drugs” (pharmacogenomics) based on individual genetic profiles

Microbial Genomics

- rapidly detect and treat pathogens (disease-causing microbes) in clinical practice
- develop new energy sources (biofuels)
- monitor environments to detect pollutants
- protect citizenry from biological and chemical warfare
- clean up toxic waste safely and efficiently

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

Anticipated Benefits of Genome Research-cont.

Risk Assessment

• evaluate the health risks faced by individuals who may be exposed to radiation (including low levels in industrial areas) and to cancer-causing chemicals and toxins

Bioarchaeology, Anthropology, Evolution, and Human Migration

• study evolution through germline mutations in lineages
• study migration of different population groups based on maternal inheritance
• study mutations on the Y chromosome to trace lineage and migration of males
• compare breakpoints in the evolution of mutations with ages of populations and historical events

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

Anticipated Benefits of Genome Research-cont.

DNA Identification (Forensics)

- identify potential suspects whose DNA may match evidence left at crime scenes
- exonerate persons wrongly accused of crimes
- identify crime and catastrophe victims
- establish paternity and other family relationships
- identify endangered and protected species as an aid to wildlife officials (could be used for prosecuting poachers)
- detect bacteria and other organisms that may pollute air, water, soil, and food
- match organ donors with recipients in transplant programs
- determine pedigree for seed or livestock breeds
- authenticate consumables such as caviar and wine

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

Anticipated Benefits of Genome Research-cont.

Agriculture, Livestock Breeding, and Bioprocessing

- grow disease-, insect-, and drought-resistant crops
- breed healthier, more productive, disease-resistant farm animals
- grow more nutritious produce
- develop biopesticides
- incorporate edible vaccines incorporated into food products
- develop new environmental cleanup uses for plants like tobacco

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

Medicine and the New Genetics

Gene Testing ***Pharmacogenomics*** ***Gene Therapy***

Anticipated Benefits:

- improved diagnosis of disease
- earlier detection of genetic predispositions to disease
- rational drug design
- gene therapy and control systems for drugs
- personalized, custom drugs



U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003



ELSI: Ethical, Legal, and Social Issues

- **Privacy and confidentiality of genetic information**.
- **Fairness in the use of genetic information** by insurers, employers, courts, schools, adoption agencies, and the military, among others.
- **Psychological impact, stigmatization, and discrimination** due to an individual's genetic differences.
- **Reproductive issues** including adequate and informed consent and use of genetic information in reproductive decision making.
- **Clinical issues** including the education of doctors and other health-service providers, people identified with genetic conditions, and the general public about capabilities, limitations, and social risks; and implementation of standards and quality−control measures.

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

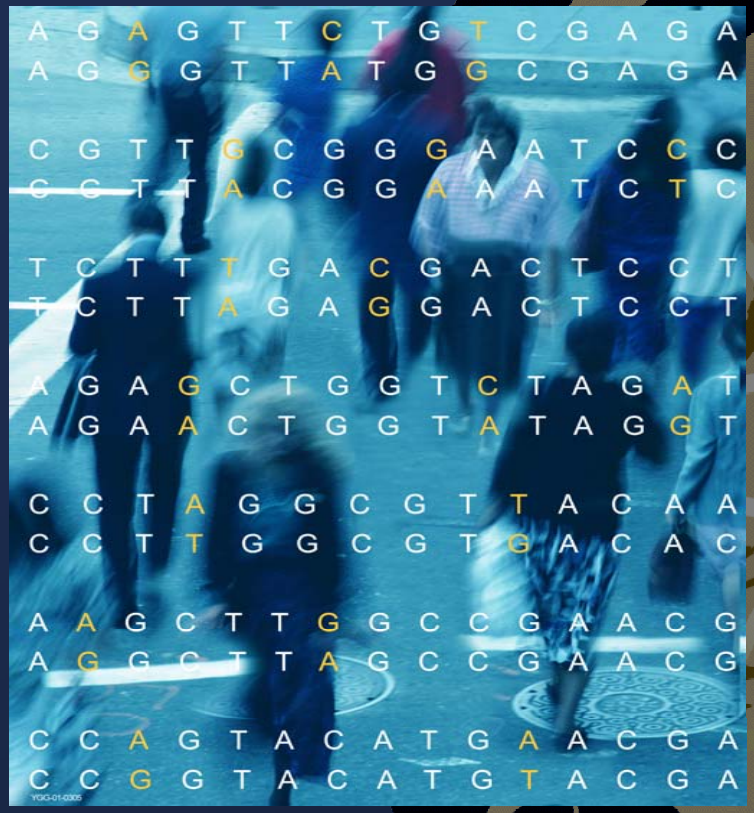


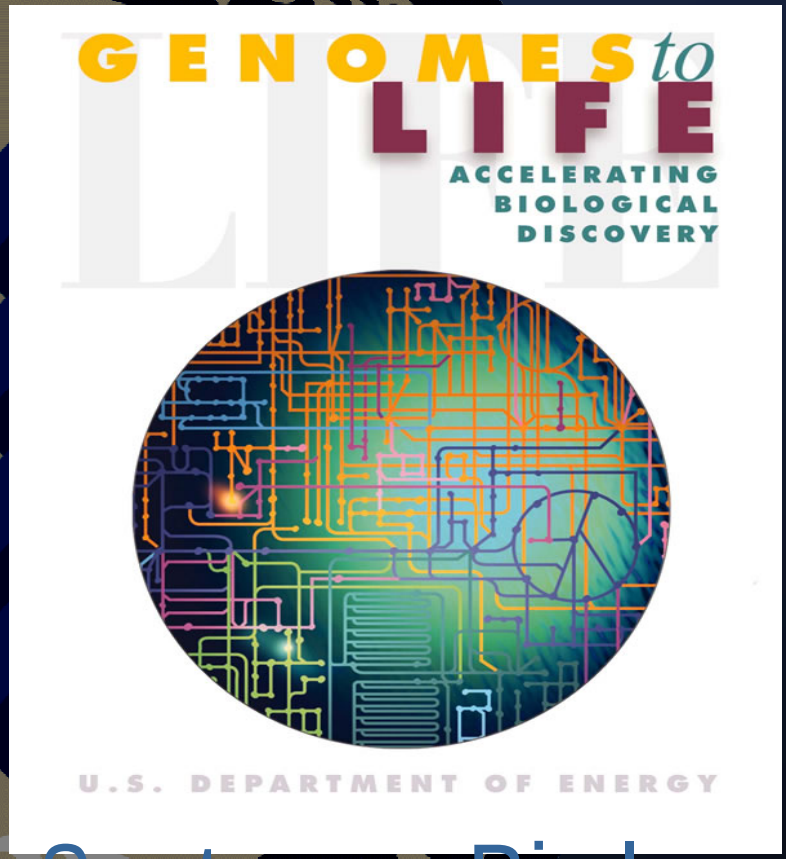
Issues (cont.)

- **Uncertainties associated with gene tests for susceptibilities and complex conditions** (e.g., heart disease, diabetes, and Alzheimer's disease).
- **Fairness in access to advanced genomic technologies**.
- **Conceptual and philosophical implications** regarding human responsibility, free will vs genetic determinism, and concepts of health and disease.
- **Health and environmental issues** concerning genetically modified (GM) foods and microbes.
- **Commercialization of products** including property rights (patents, copyrights, and trade secrets) and accessibility of data and materials.

U.S. Department of Energy Genome Programs, Genomics and Its Impact on Science and Society, 2003

Beyond the HGP: What's Next?





HapMap

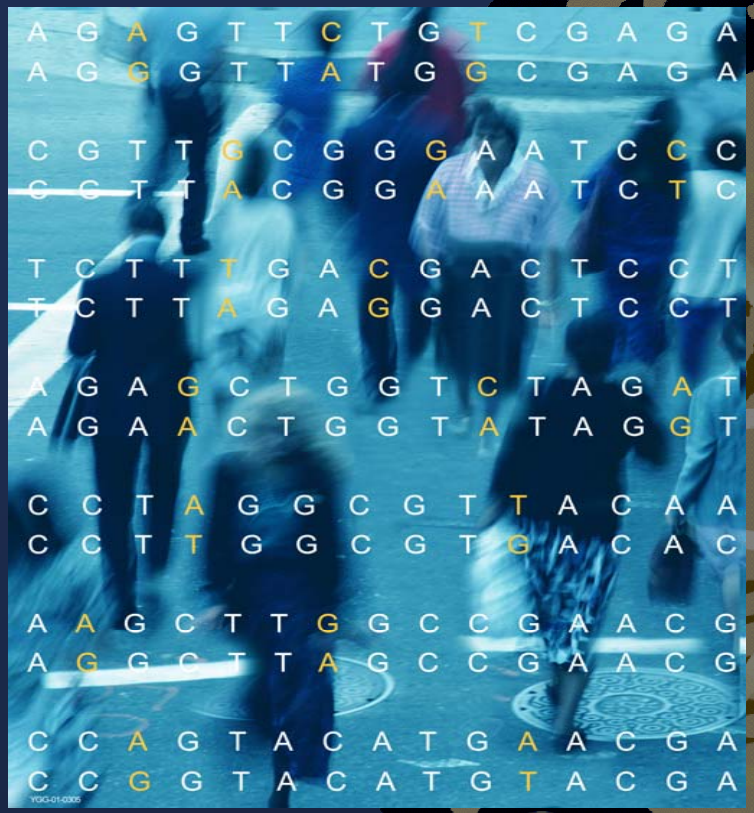
Chart genetic variation within the human genome

Systems Biology

Exploring Microbial Genomes for Energy and the Environment

HapMap

An NIH program to chart genetic variation within the human genome

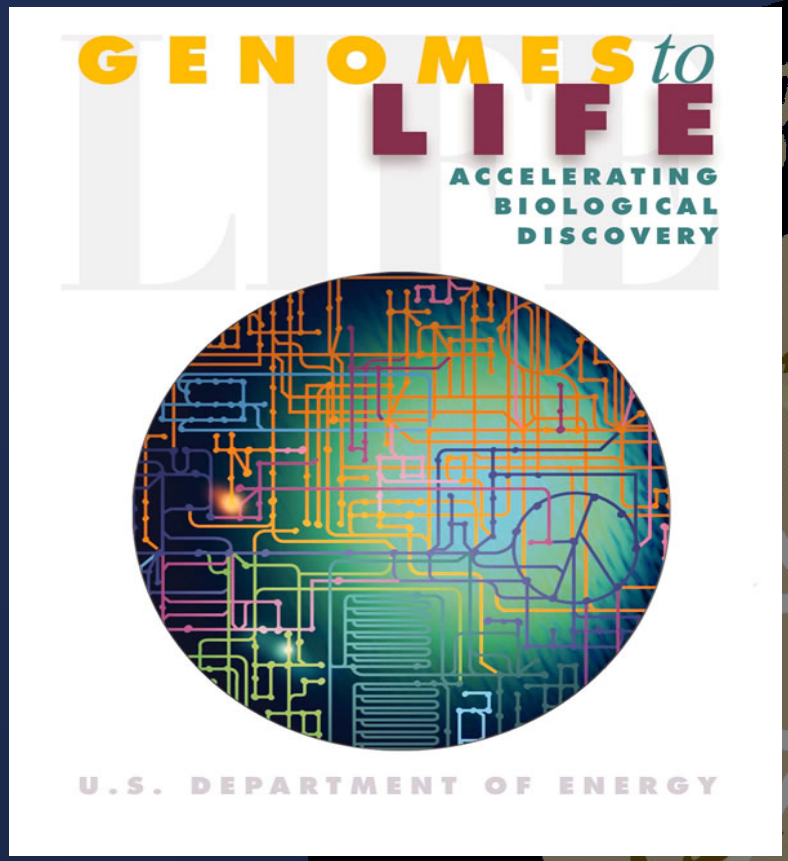


- Begun in 2002, the project is a 3-year effort to construct a map of the patterns of SNPs (single nucleotide polymorphisms) that occur across populations in Africa, Asia, and the United States.
- Consortium of researchers from six countries
- Researchers hope that dramatically decreasing the number of individual SNPs to be scanned will provide a shortcut for identifying the DNA regions associated with common complex diseases
- Map may also be useful in understanding how genetic variation contributes to responses in environmental factors

Haplotype list for HapSet DBMHC|ABDR:

Prototype display

Hap id	Hap name	Hap Alleles of Marker Set
1	ABDR-1	C-C-A-G-G-T-A-G-C-G-C-G-A-A-G-C-A-T-T-C-T-G-T-A-G-T-A-C-G-A
2	ABDR-2	C-C-A-G-G-T-A-G-C-G-C-T-C-T-A-A-G-C-A-A-C-T-G-G-C-G-A-C-G-A-G
3	ABDR-3	C-C-C-G-G-T-G-G-T-A-C-G-A-A-G-C-A-C-A-A-T-C-G-G-C-G-A-A-C-A-C
4	ABDR-4	C-C-C-G-G-T-G-G-T-A-C-G-A-A-G-C-A-C-A-A-T-C-G-T-A-G-T-A-C-G-A
5	ABDR-5	C-C-C-G-G-T-G-G-T-A-C-G-A-G-G-T-A-T-C-T-C-T-A-G-C-A-G-T-C-G-T
6	ABDR-6	C-C-C-G-G-T-G-G-T-A-C-G-A-G-G-T-A-T-C-T-C-T-A-G-C-G-A-C-G-A-G
7	ABDR-7	C-T-C-G-G-T-G-G-C-G-C-G-A-G-G-T-G-T-A-T-C-T-A-T-A-A-A-C-G-G-C
8	ABDR-8	C-T-C-G-G-T-G-G-C-G-C-T-C-T-A-A-G-C-A-G-C-T-G-T-A-G-A-T-C-G-C
9	ABDR-9	C-T-C-G-G-T-G-G-C-G-C-T-C-T-A-T-G-C-A-A-C-T-G-G-C-G-A-C-G-A-G
10	ABDR-10	G-A-C-G-G-T-G-A-T-A-T-G-A-G-G-T-A-T-C-T-C-T-A-G-C-G-T-A-C-G-C
11	ABDR-11	G-A-C-G-G-T-G-A-T-A-T-T-C-T-A-G-G-C-T-T-T-C-G-G-A-A-A-C-G-A-C
12	ABDR-12	G-C-C-G-G-T-C-G-C-G-C-G-A-G-G-C-A-T-A-T-C-T-A-T-A-G-A-T-C-G-C
13	ABDR-13	G-C-C-G-G-T-C-G-C-G-C-T-C-T-A-A-G-C-A-G-C-T-G-T-A-A-A-C-G-G-C
14	ABDR-14	G-T-A-T-C-G-C-A-T-A-T-G-A-A-G-C-A-C-A-A-T-C-G-G-A-A-A-C-G-A-C
15	ABDR-15	G-T-A-T-C-G-C-A-T-A-T-G-A-G-G-T-G-T-A-T-C-T-A-G-A-A-A-C-G-A-C
16	ABDR-16	G-T-A-T-C-G-C-A-T-A-T-G-A-G-G-T-G-T-A-T-C-T-A-G-C-A-G-T-C-G-T
17	ABDR-17	G-T-A-T-C-G-C-A-T-A-T-G-C-G-G-A-G-C-C-G-T-C-A-A-A-A-T-T-G-G-A
18	ABDR-18	G-T-A-T-C-G-C-A-T-A-T-G-C-G-G-C-G-C-C-G-T-C-A-G-A-A-A-C-G-A-C
19	ABDR-19	G-T-A-T-C-G-C-A-T-A-T-T-C-T-A-A-G-C-A-G-C-T-G-T-A-G-A-T-C-G-C
20	ABDR-20	G-T-A-T-C-G-C-A-T-A-T-T-C-T-A-G-G-C-T-T-T-C-G-G-A-A-A-C-G-A-C



Genomes to Life:
A DOE Systems Biology Program

Exploring Microbial Genomes for Energy and the Environment

Goals

- identify the protein machines that carry out critical life functions
- characterize the gene regulatory networks that control these machines
- characterize the functional repertoire of complex microbial communities in their natural environments
- develop the computational capabilities to integrate and understand these data and begin to model complex biological systems

GTL Applications in Energy Security and Global Climate Change

Within a Decade

Long Term

2020

Develop knowledge base for cost-effective cleanup strategies

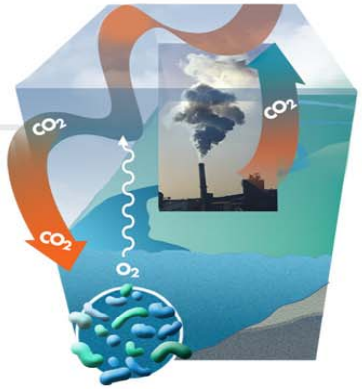
Save billions of dollars in toxic waste cleanup and disposal



2040

Understand earth's natural carbon cycle and design strategies for enhanced carbon capture

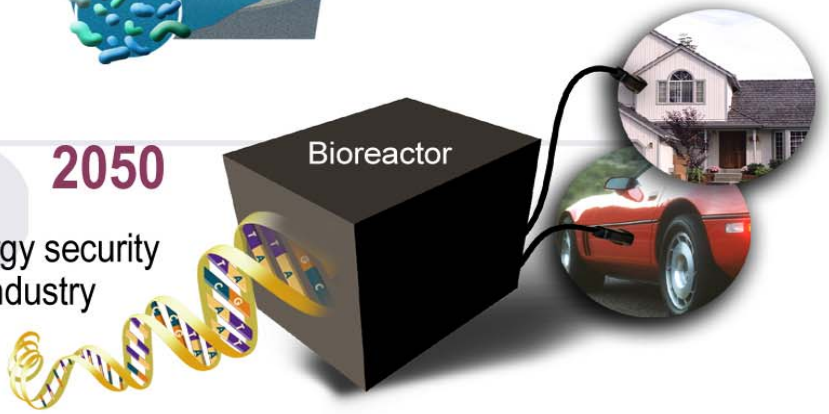
Help stabilize atmospheric carbon dioxide to counter global warming



2050

Increase biological sources of fuels and electricity

Contribute to U.S. energy security
- Biohydrogen-based industry in place



YGG-01-0521R7

THE GENOMIC SEQUENCE



Fifty years after the double helix, the reference DNA sequence of Homo sapiens is now available for downloading. Annotation of genes and other features is in progress.

Download Sequence Data

14 Apr 2003

The Human Genome Project announces completion of the DNA reference sequence of Homo sapiens.

Press releases:
NIH
Sanger Institute
Whitehead Institute

Reference Sequence Properties

Chrom. number	Reference accession	Sequence length	Determined bases*
1	NC_000001.4	245,203,898	218,712,898
2	NC_000002.5	243,315,028	237,043,673
3	NC_000003.5	199,411,731	193,607,218
4	NC_000004.5	191,610,523	186,580,523
5	NC_000005.4	180,967,295	177,524,972
6	NC_000006.5	170,740,541	166,880,540
7	NC_000007.7	158,431,299	154,546,299
8	NC_000008.5	145,908,738	141,694,337
9	NC_000009.5	134,505,819	115,187,714
10	NC_000010.4	135,480,874	130,710,865
11	NC_000011.4	134,978,784	130,709,420
12	NC_000012.5	133,464,434	129,328,332
13	NC_000013.5	114,151,656	95,511,656
14	NC_000014.4	105,311,216	87,191,216
15	NC_000015.4	100,114,055	81,117,055
16	NC_000016.4	89,995,999	79,890,791
17	NC_000017.5	81,691,216	77,480,855
18	NC_000018.4	77,753,510	74,534,531
19	NC_000019.5	63,790,860	55,780,860
20	NC_000020.5	63,644,868	59,424,990
21	NC_000021.3	46,976,537	33,924,742
22	NC_000022.4	49,476,972	34,352,051
X	NC_000023.4	152,634,166	147,686,664
Y	NC_000024.3	50,961,097	22,761,097
unplaced	various	25,263,157	25,062,835

UCSC Genome Bioinformatics

Genome Browser - Blat Search - Table Browser - FAQ - User Guide

Genome
Human

Browser
Blat
Tables
Downloads
Mirrors
Archives
Credits
Cite Us
Jobs
Links

About the UCSC Genome Bioinformatics Site

This site contains working drafts for the human, mouse, and rat genomes.

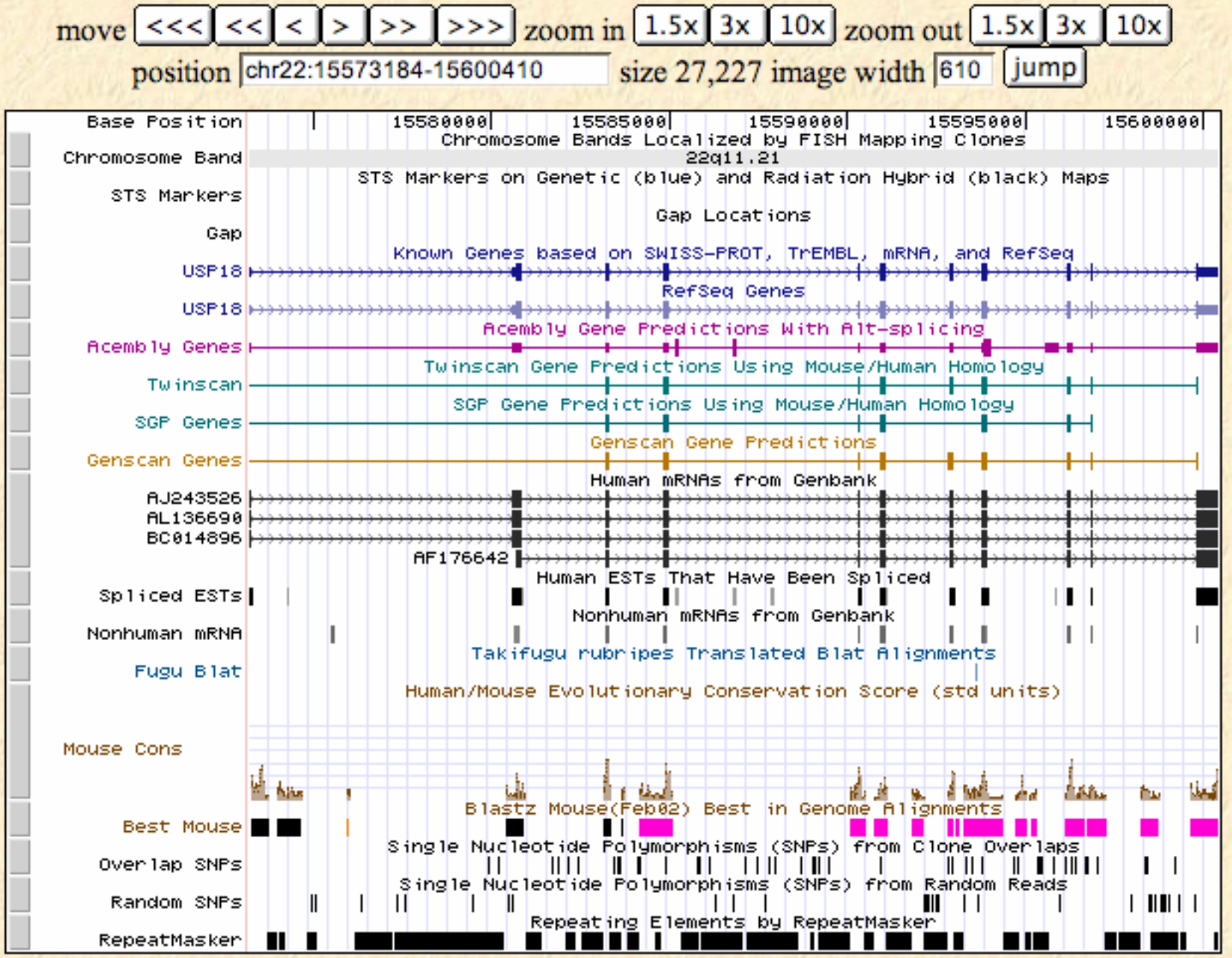
The human genome is over 95% complete. Approximately 90% of the sequence is in a highly accurate 'finished' state. The other 10% is merely 'draft' quality. Some care must be taken when interpreting draft regions, but these are still very useful to the working scientist.

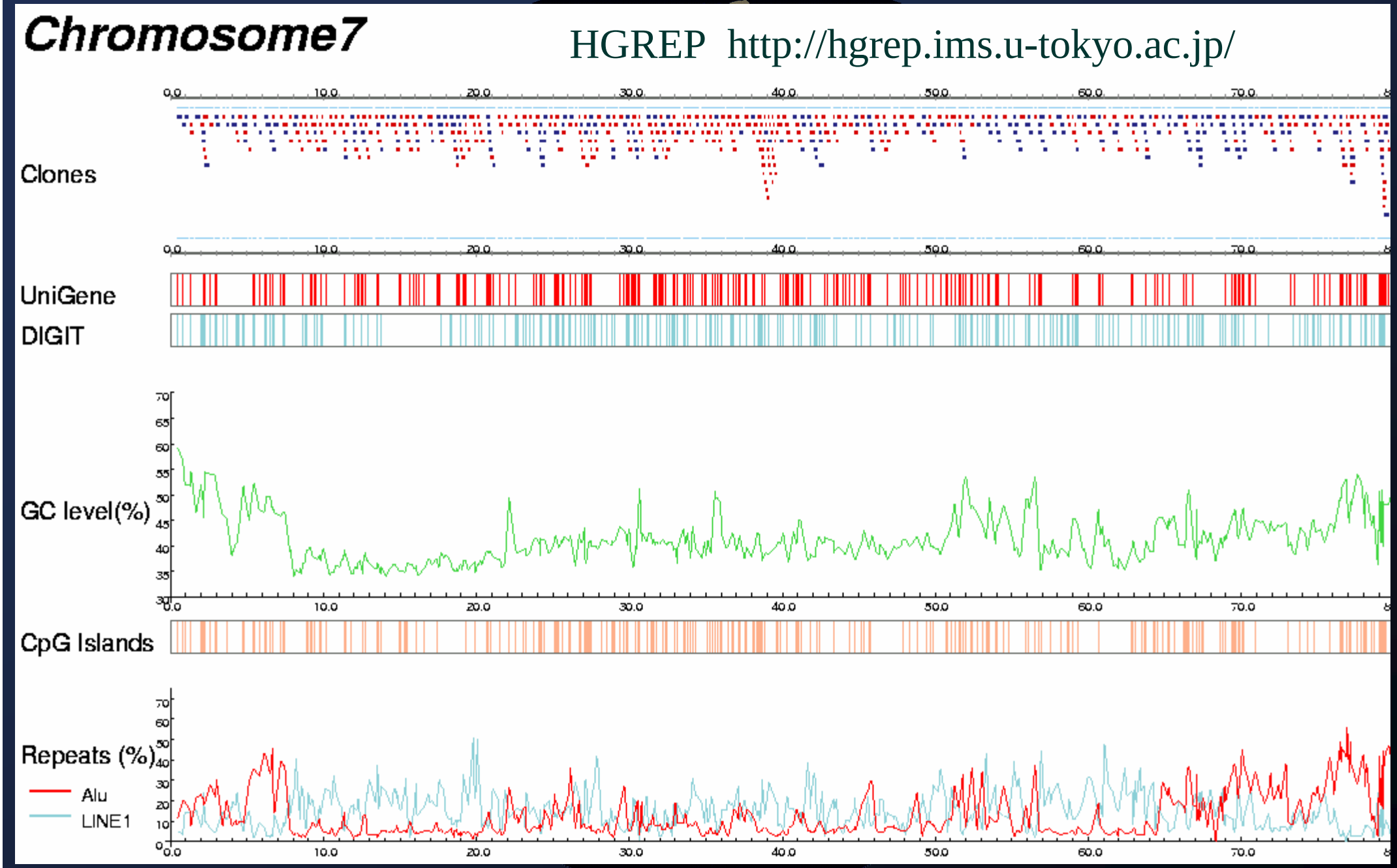
The mouse genome is largely the result of a full genome shotgun assembly. While there are many gaps in the sequence, the order and orientation of the contigs on either side of a gap are known in almost all cases. We estimate that 90-96% of the mouse genome is present in the assembly. We encourage you to explore the working drafts with the genome browser, which displays the work of many annotators worldwide.

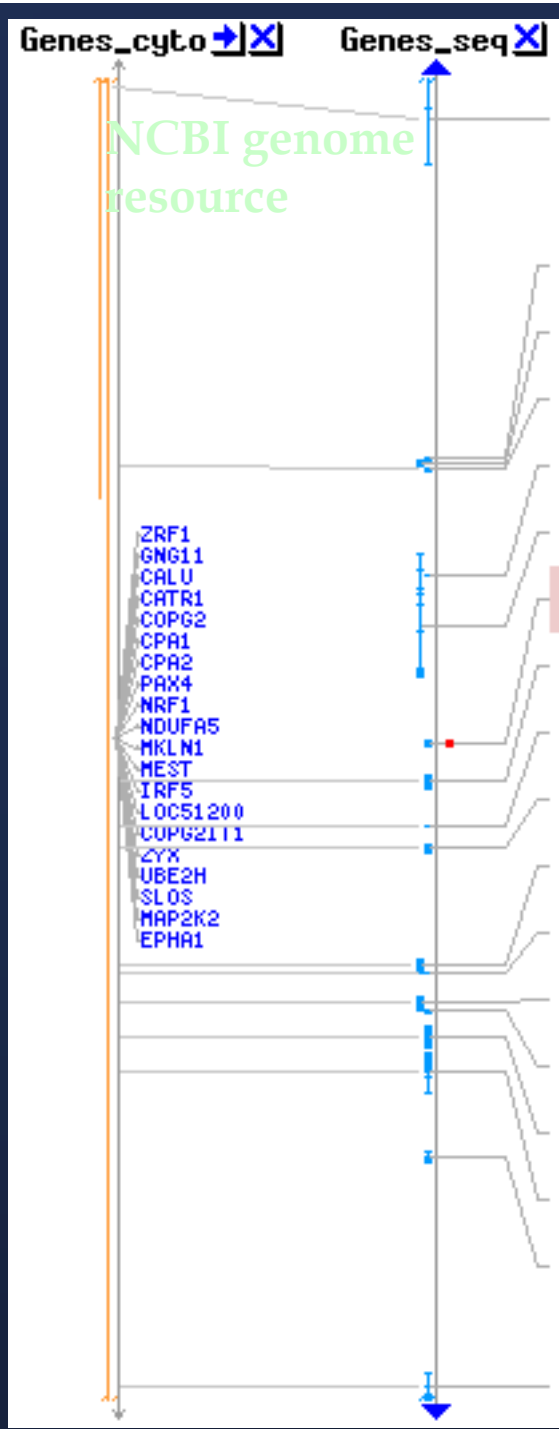
The Genome Browser and the data it displays are freely available for academic, nonprofit, and personal use with a few exceptions. Check the README.txt file in the download directory of a particular assembly to view the use restrictions specific to that release. Interactive use is limited to a maximum of 10,000 hits/day or 10 hits/minute.

http://genome.ucsc.edu/

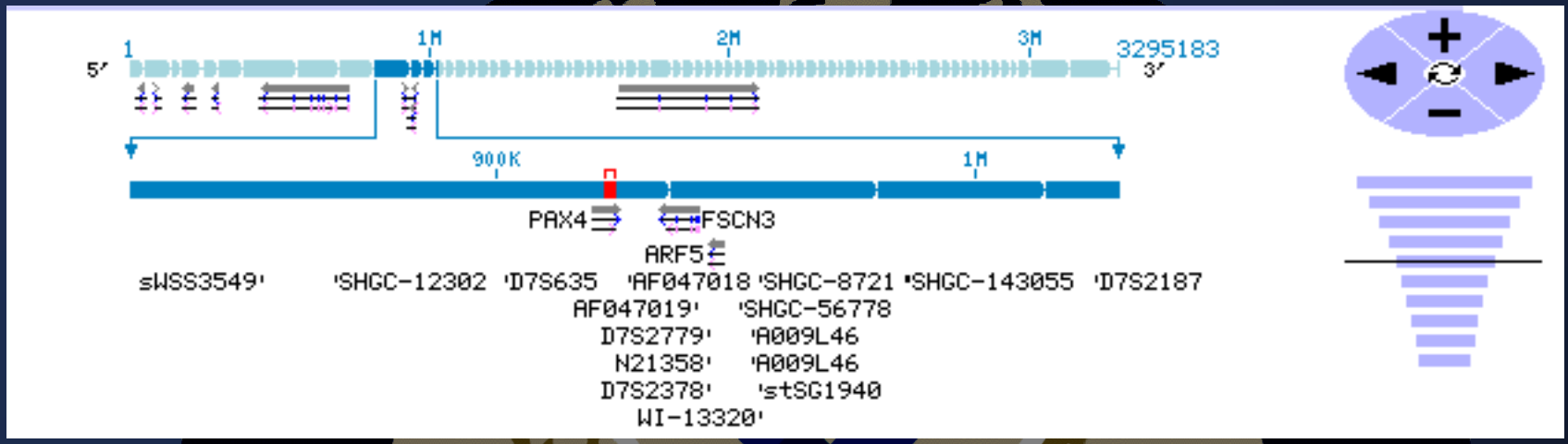
UCSC Genome Browser on Human Nov. 2002 Freeze

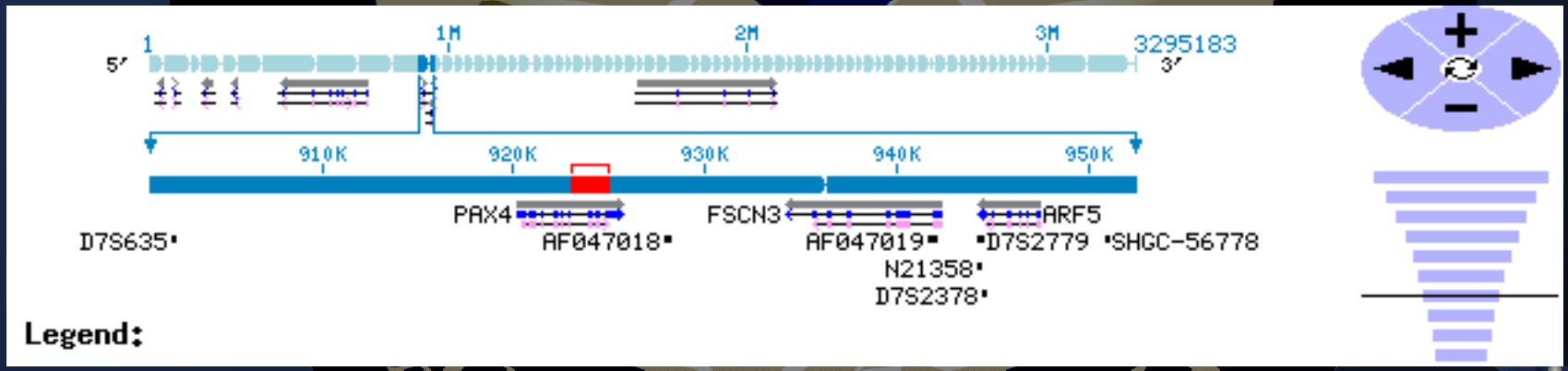



Chromosome7
HGREP http://hgrep.ims.u-tokyo.ac.jp/
Clones
UniGene
DIGIT
GC level(%)
CpG Islands
Repeats (%)
Alu
LINE1
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0

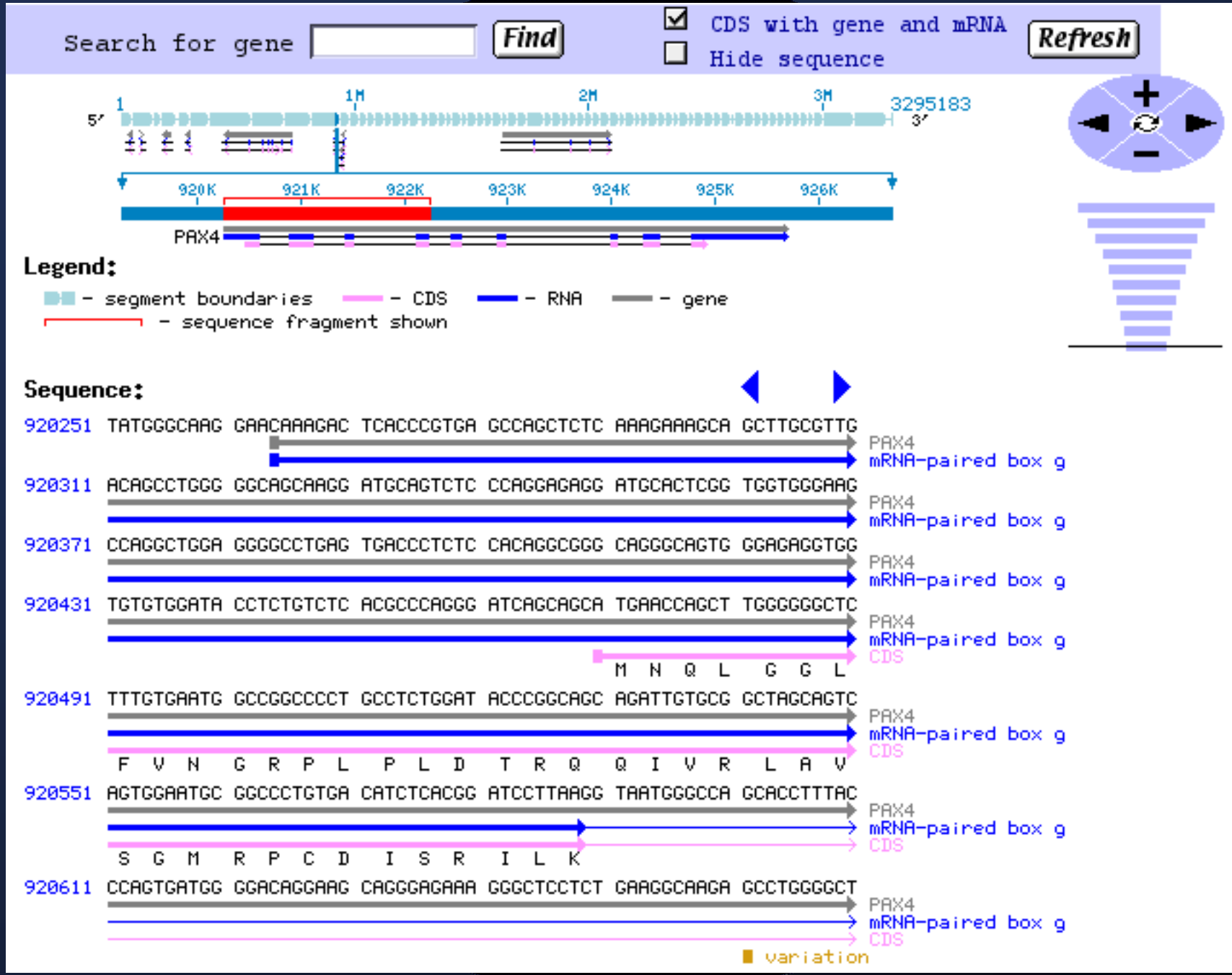


symbol	orient.	links		cyto.	full name
GRM8	-	av	sv	7q31.3-q32.1	glutamate receptor, metabotropic
ARF5	+	av	sv	7q31.3	ADP-ribosylation factor 5
FSCN3	+	av	sv	7q31.3	fascin (Strongylocentrotus purpu
PAX4	-	av	sv	7q32	paired box gene 4
LOC65544	-	av	sv	7	similar to CD2 antigen (H. sapie
p100	+	av	sv	7q31.3	EBNA-2 co-activator (100kD)
LEP	+	av	sv	7q31.3	leptin (murine obesity homolog)
FLJ10377	+	av	sv	7q22.1-7q32.3	hypothetical protein FLJ10377
HIG2	-	av	sv	7q22.1-7q32.3	hypoxia-inducible protein 2
IMPDH1	+	av	sv	7q31.3-q32	IMP (inosine monophosphate) dehy
CALU	+	av	sv	7q32	calumenin
OPN1SW	-	av	sv	7q31.3-q32	opsin 1 (cone pigments), short-w
FLNC	+	av	sv	7q32-q35	filamin C, gamma (actin-binding
LOC63220	+	av	sv	7	similar to ATPase, vacuolar, 14
IRF5	+	av	sv	7q32	interferon regulatory factor 5
TRN-SR	-	av	sv	7	transportin-SR
SMOH	+	av	sv	7q32.3	smoothened (Drosophila) homolog
NRF1	+	av	sv	7q32	nuclear respiratory factor 1


1
1M
2M
3M
3295183
5′
3′
900K
1M
PAX4
FSCN3
ARF5
sWSS3549'
'SHGC-12302
'D7S635
'AF047018
'SHGC-8721
'SHGC-143055
'D7S2187
AF047019'
'SHGC-56778
D7S2779'
'A009L46
N21358'
'A009L46
D7S2378'
'stSG1940
WI-13320'

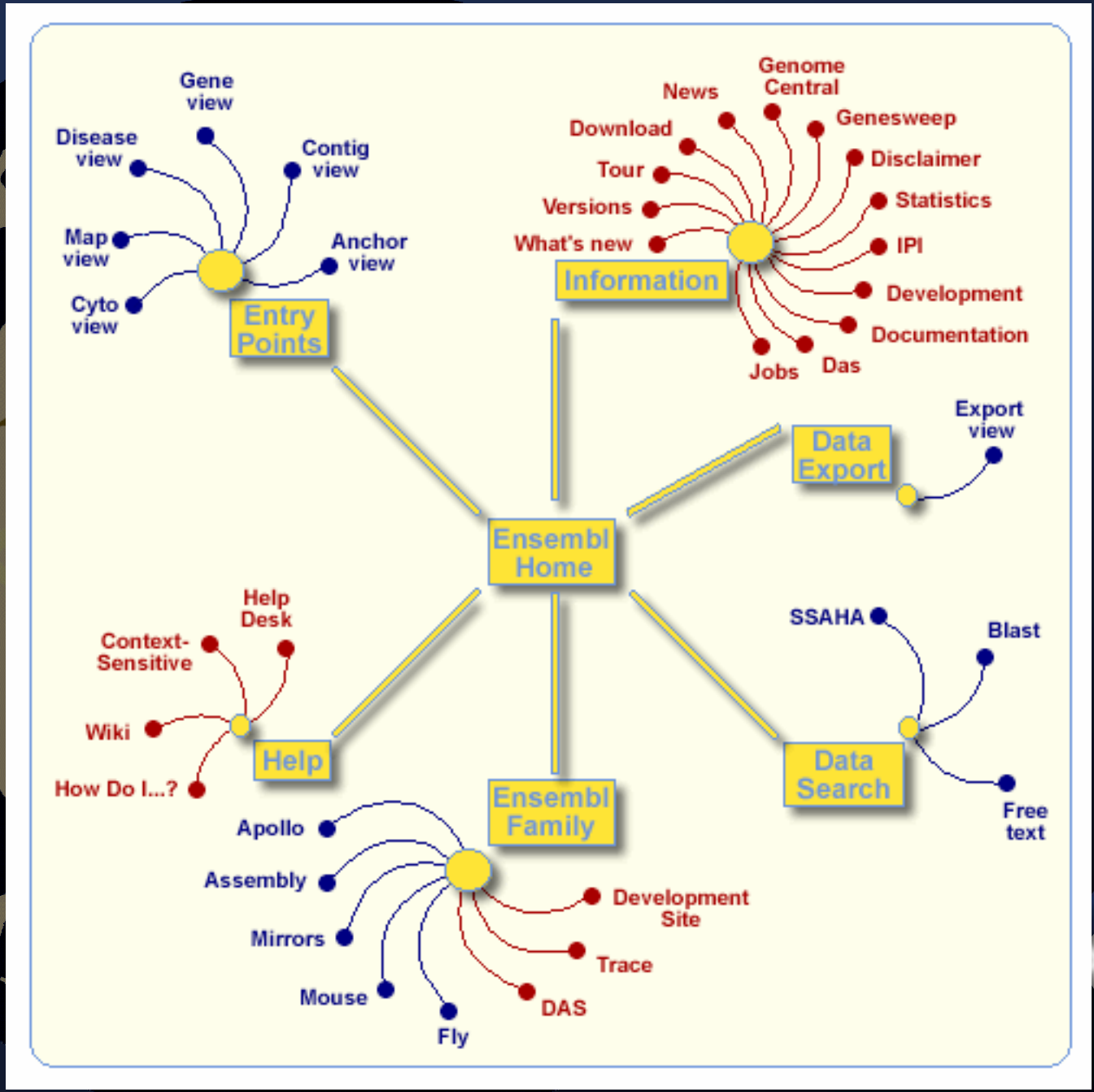

1
1M
2M
3M
3295183
5′
3′
910K
920K
930K
940K
950K
PAX4
FSCN3
ARF5
D7S635'
AF047018'
AF047019'
'D7S2779
'SHGC-56778
N21358'
D7S2378'

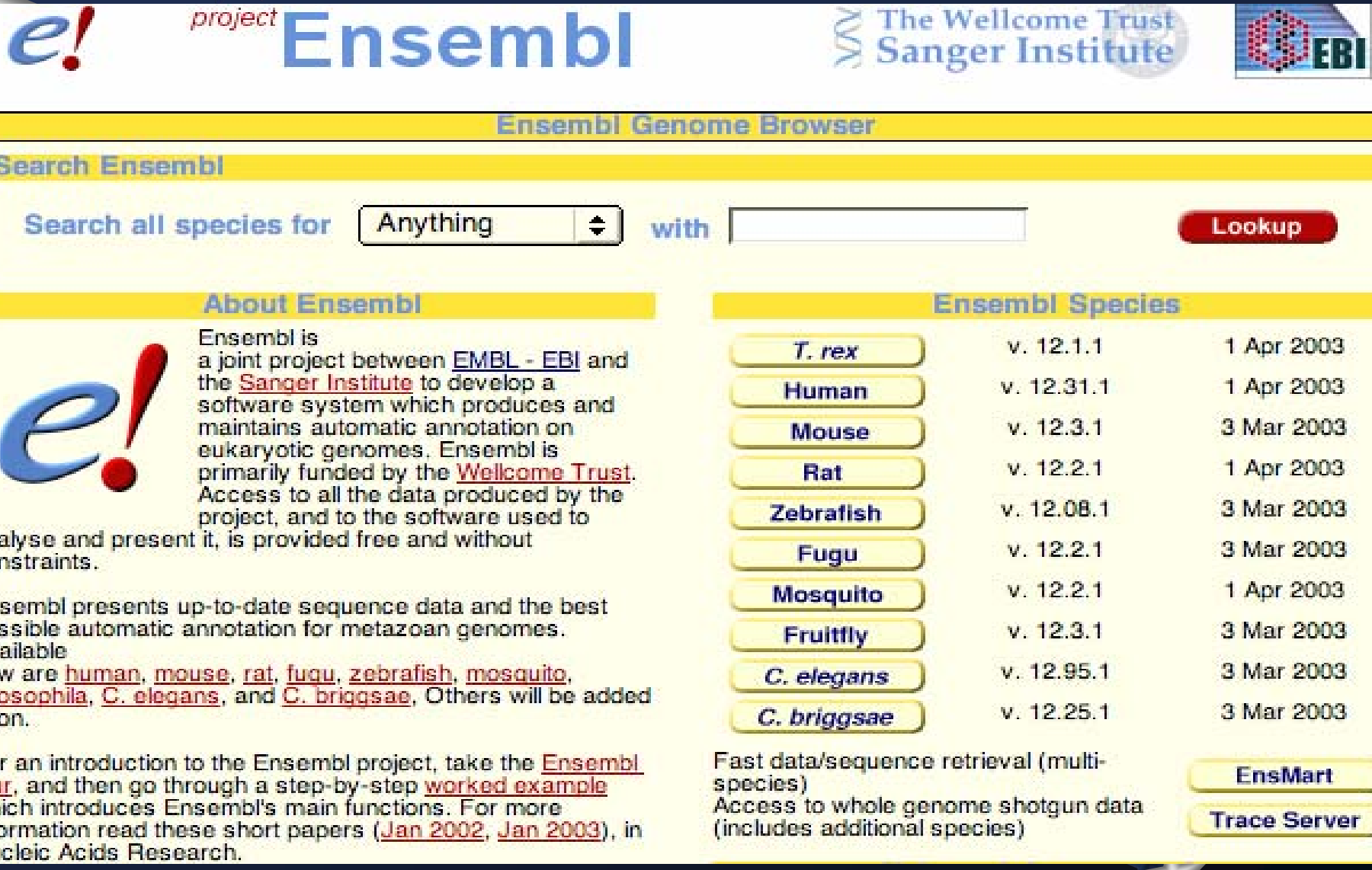
Legend:


Search for gene
Find
CDS with gene and mRNA
Hide sequence
Refresh
1
1M
2M
3M
3295183
5′
3′
920K
921K
922K
923K
924K
925K
926K
PAX4
Legend:
- segment boundaries
- CDS
- RNA
- gene
- sequence fragment shown
Sequence:
920251 TATGGGCAAG GAACAAAGAC TCACCCGTGA GCCAGCTCTC AAAGAAAGCA GCTTGCGTTG
920311 ACAGCCTGGG GGCAGCAAGG ATGCAGTCTC CCAGGAGAGG ATGCACTCGG TGGTGGGAAG
920371 CCAGGCTGGA GGGGCCTGAG TGACCCTCTC CACAGGCGGG CAGGGCAGTG GGAGAGGTGG
920431 TGTGTGGATA CCTCTGTCTC ACGCCCAGGG ATCAGCAGCA TGAACCAGCT TGGGGGGCTC
M N Q L G G L
920491 TTTGTGAATG GCCGGCCCCT GCCTCTGGAT ACCCGGCAGC AGATTGTGCG GCTAGCAGTC
F V N G R P L P L D T R Q Q I V R L A V
920551 AGTGGAATGC GGCCCTGTGA CATCTCACGG ATCCTTAAGG TAATGGGCCA GCACCTTTAC
S G M R P C D I S R I L K
920611 CCAGTGATGG GGACAGGAAG CAGGGAGAAA GGGCTCCTCT GAAGGCAAGA GCCTGGGGCT
PAX4
mRNA-paired box g
CDS
variation

Ensembl Site Map

www.ensembl.org




project Ensembl
The Wellcome Trust Sanger Institute
EBI
Ensembl Genome Browser
Search Ensembl
Search all species for
Anything
with
Lookup
About Ensembl
Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust. Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints.
Ensembl presents up-to-date sequence data and the best possible automatic annotation for metazoan genomes. Available now are human, mouse, rat, fugu, zebrafish, mosquito, Drosophila, C. elegans, and C. briggsae, Others will be added soon.
For an introduction to the Ensembl project, take the Ensembl tour, and then go through a step-by-step worked example which introduces Ensembl's main functions. For more information read these short papers (Jan 2002, Jan 2003), in Nucleic Acids Research.
Ensembl Species
T. rex v. 12.1.1 1 Apr 2003
Human v. 12.31.1 1 Apr 2003
Mouse v. 12.3.1 3 Mar 2003
Rat v. 12.2.1 1 Apr 2003
Zebrafish v. 12.08.1 3 Mar 2003
Fugu v. 12.2.1 3 Mar 2003
Mosquito v. 12.2.1 1 Apr 2003
Fruitfly v. 12.3.1 3 Mar 2003
C. elegans v. 12.95.1 3 Mar 2003
C. briggsae v. 12.25.1 3 Mar 2003
Fast data/sequence retrieval (multi-species)
EnsMart
Access to whole genome shotgun data (includes additional species)
Trace Server

中央研究院

ACADEMIA SINICA

1928



T. rex assembly

The world's first annotated T.rex assembly is now available at Ensembl!

Click for more information

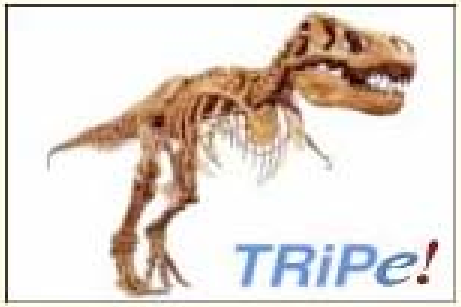
Ensembl **T. rex**

The Wellcome Trust Sanger Institute

EBI

Tyrannosaurus rex Genome Browser

About TRiPE



The T.rex International

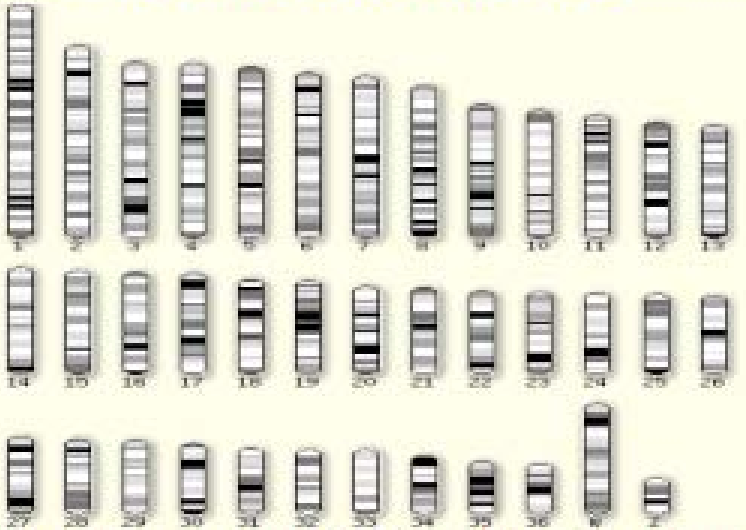
Palaeontonomics Experiment is a consortium set up to investigate the remarkable finds made at the Wellcome Trust Genome Campus during building excavations in 2002.

Ensembl works closely with both TRiPE and other T.rex sequencing groups to provide an integrated resource.

T.rex Genome Assembly (1 Apr 2003)

This site proudly presents the first T.rex draft assembly. This is based principally on whole genome shotgun of around 7x coverage from the sample GRR/ARG, recovered from the Hinxton Campus. The assembly consists of large supercontigs aligned to chromosomes

Browse a Chromosome



Note that the apparent blurring of the karyotype is caused by DNA degradation due to the 65 million years the sample spent underground.

The Discovery



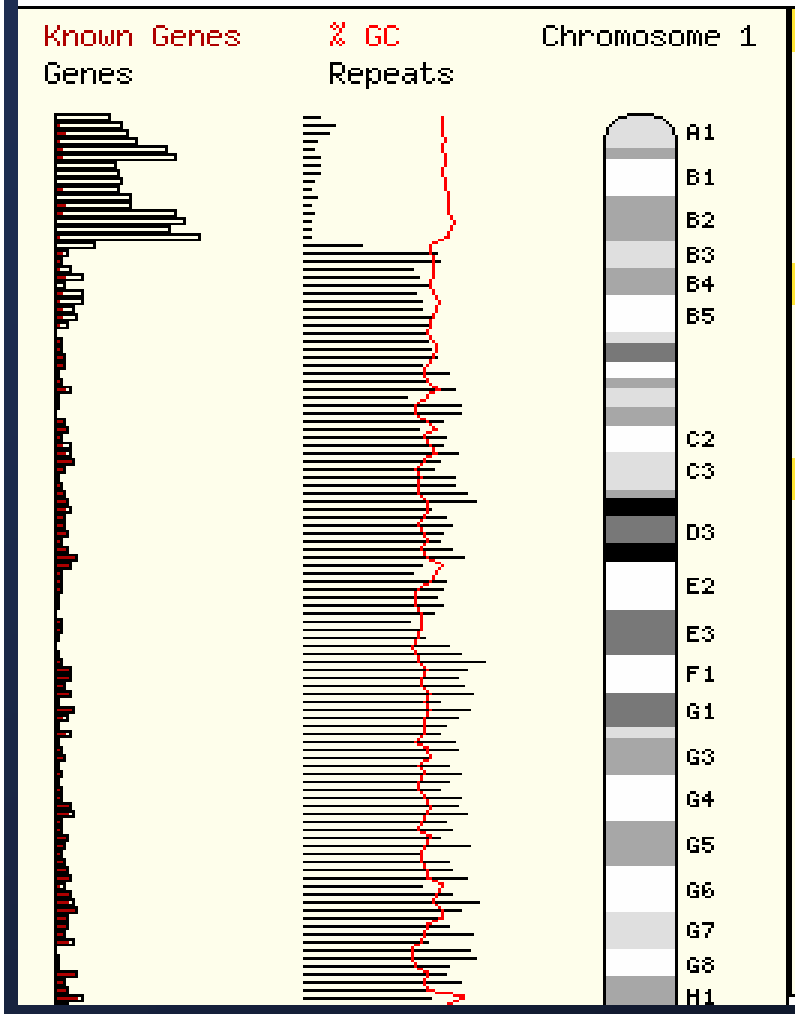
"Last year's discovery of long-lost dinosaur remains in Cambridgeshire astonished the scientific community. The remains, most notably those of a Tyrannosaurus rex, were found on the Wellcome Trust Genome Campus whilst preparations

Ensembl **T. rex** *MapView*

The Wellc Sanger

Home ▸ T.rex ▲ What's New ▲ Export Data ▲ Docs ▲

Find All 1 Lookup [e.g. 12, W]



Chromosome 1

Length: 130,638,654 bps
Known Ensembl genes: 888
Novel Ensembl genes: 1,165

Change Chromosome

Chromosome: 1 Lookup

Jump to Contigview

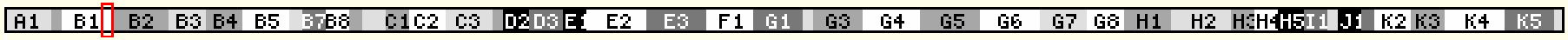
Click anywhere on the chromosome ideogram or one of the f
plots to jump to a contig-level view of features at that point. Alt
jump to contigview between any two markers on this ch

Between:
and:
Lookup

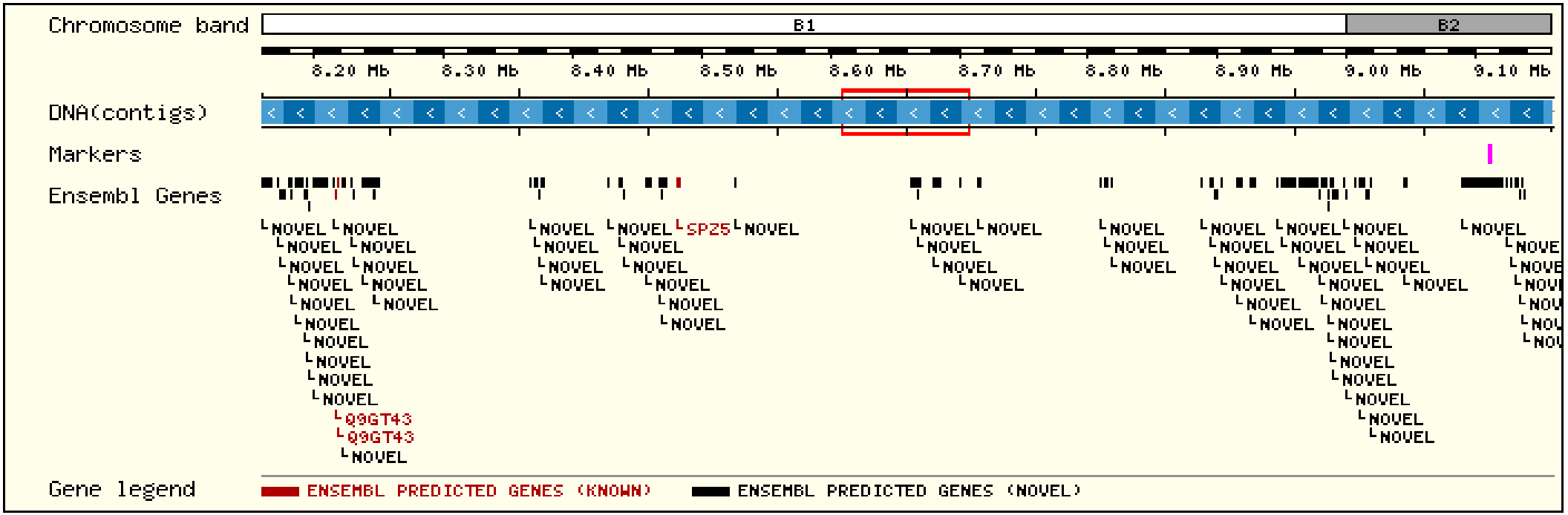
Display contig-level view between any two featu

Chromosome 1

Chr 1

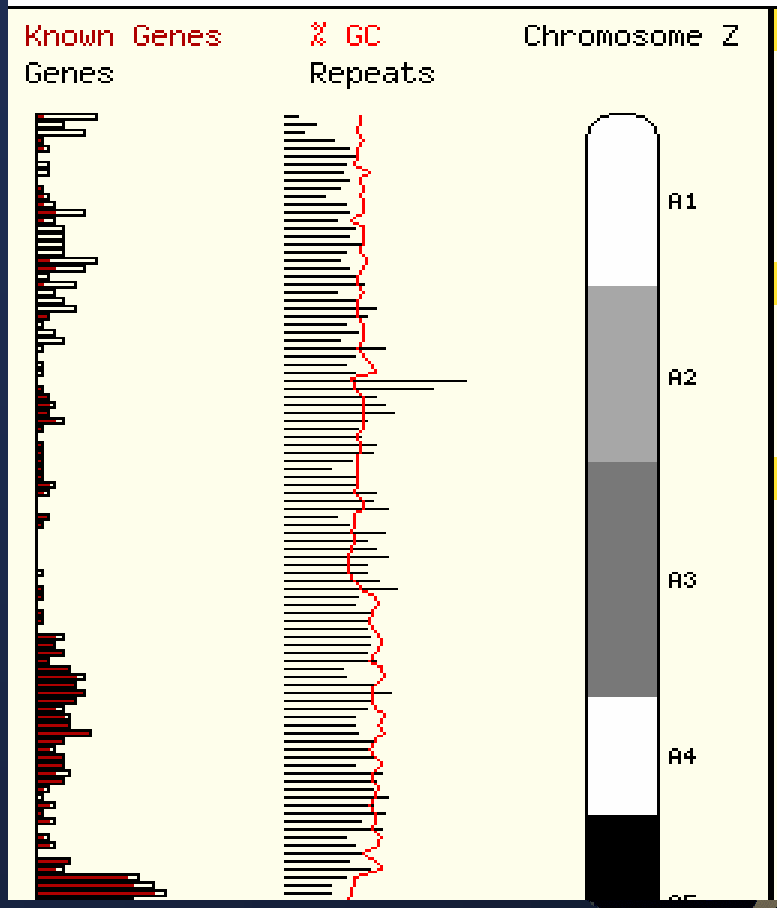


Overview



Home ▸ T.rex ▲ What's New ▲ Export Data ▲ Docs ▲

Find All Z Lookup [e.g. 12, W]



Chromosome Z

Length: 20,503,547
Known Ensembl genes: 743
Novel Ensembl genes: 122

Change Chromosome

Chromosome: Z Lookup

Jump to Contigview

Click anywhere on the chromosome ideogram or one of th
plots to jump to a contig-level view of features at that point.
jump to contigview between any two markers on thi

Between:
and:

Loo



T. rex assembly withdrawn

It has come
to our attention that the *Tyrannosaurus rex* spp *hinxtonii* genome released yesterday, 1 April, was badly contaminated. Further examination of the sequence indicates the presence of large amounts of human, anopheles, c.elegans, c.briggsae, and drosophila DNA. In fact, pretty much the only organism whose DNA appears to be absent from the sample is T.rex itself.

It is unclear how this contamination occurred - one theory is that fragments of all these organisms were present under the toenail of the toe the tissue sample was taken from. We can only apologise for the error, and assure you that all other Ensembl data are not affected by the T.rex errors, and are of the highest quality.

In any case, we are sad to announce that, under the circumstances, we feel we have no option but to withdraw the T.rex assembly.

The team here are very appreciative of the many comments we have received on the T.rex release, and hope you will not be too disappointed that it turned out to be of little scientific value.

On behalf of the Ensembl team, and the TRiPE consortium

April 2 2003

MGSC Mouse

The Wellcome Trust Sanger Institute

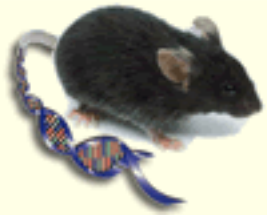
EBI

Mouse Genome Server

Search Select Lookup

Browse Chromosome 1 Start 1 End 100000 Lookup

About MGSC (3.1.1 details)



MGSC

The Mouse Genome Sequencing Consortium is a joint project between The Whitehead Institute/MIT Center for Genome Research, The Washington University Genome Sequencing Center, The Wellcome Trust Sanger Institute and EMBL - EBI to provide the Mouse genome sequence to the world. We work closely with other Mouse groups to provide an integrated resource [details].

Mouse Genome Assembly v 1

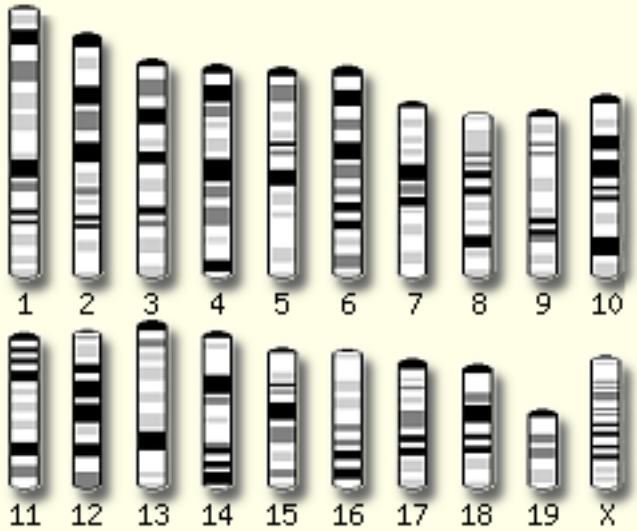
This site presents v1 of the mouse genome draft sequence, based on pure whole genome shotgun data of ~4x coverage, frozen in October 2001. The WGS assembly was aligned to the joint Sanger Institute/WUGSC BAC map (frozen in Sept 2001 - details) to provide this assembly. Clone-based sequencing has not been incorporated. The assembly was run through the normal Ensembl pipeline to predict genes and other features of interest. (statistics)

Deeper (5.5x) shotgun coverage, improved assembly software,

Other Species

Fly Zebrafish Mosquito Human

Browse a Chromosome



Ensembl Links and Site Map

News

2nd December 2002

SNP data update

The SNP database has been updated to dbSNP 108 data.

The affected databases are snp, lite, and mart databases.
New versions are:

- mus_musculus_snp_9_3a
- mus_musculus_lite_9_3a
- ensembl_mart_9_2

For further details please see What's New.

Ensembl Mouse Statistics

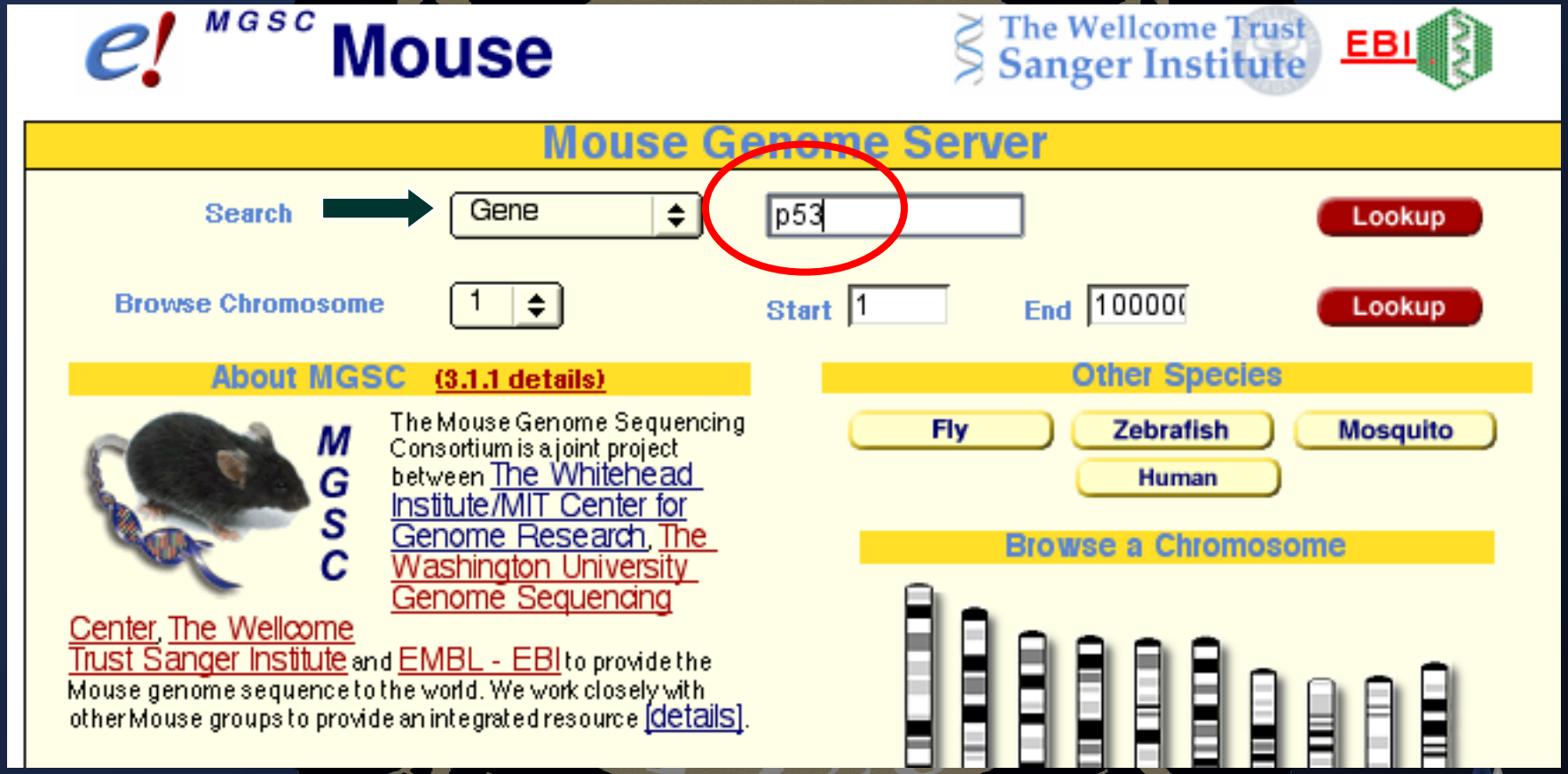
Last Update: 2-12-2002

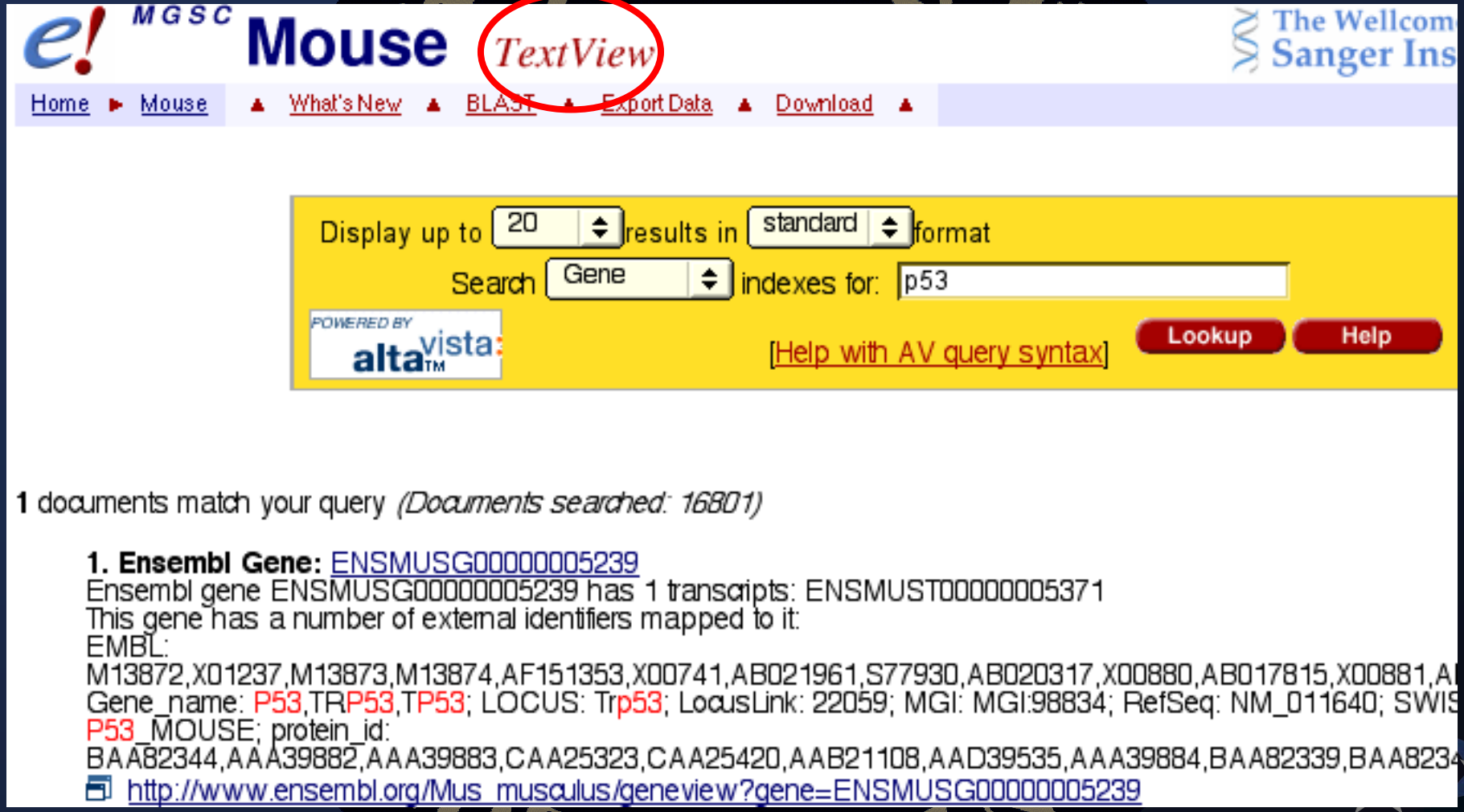
Ensembl gene predictions: 22444
Genscan gene predictions: 110379
Ensembl gene exons: 191275
Ensembl gene transcripts: 28097
Contigs: 2740
Clones: 2740
Base Pairs: 2726795854
Golden Path Length: 2726795854

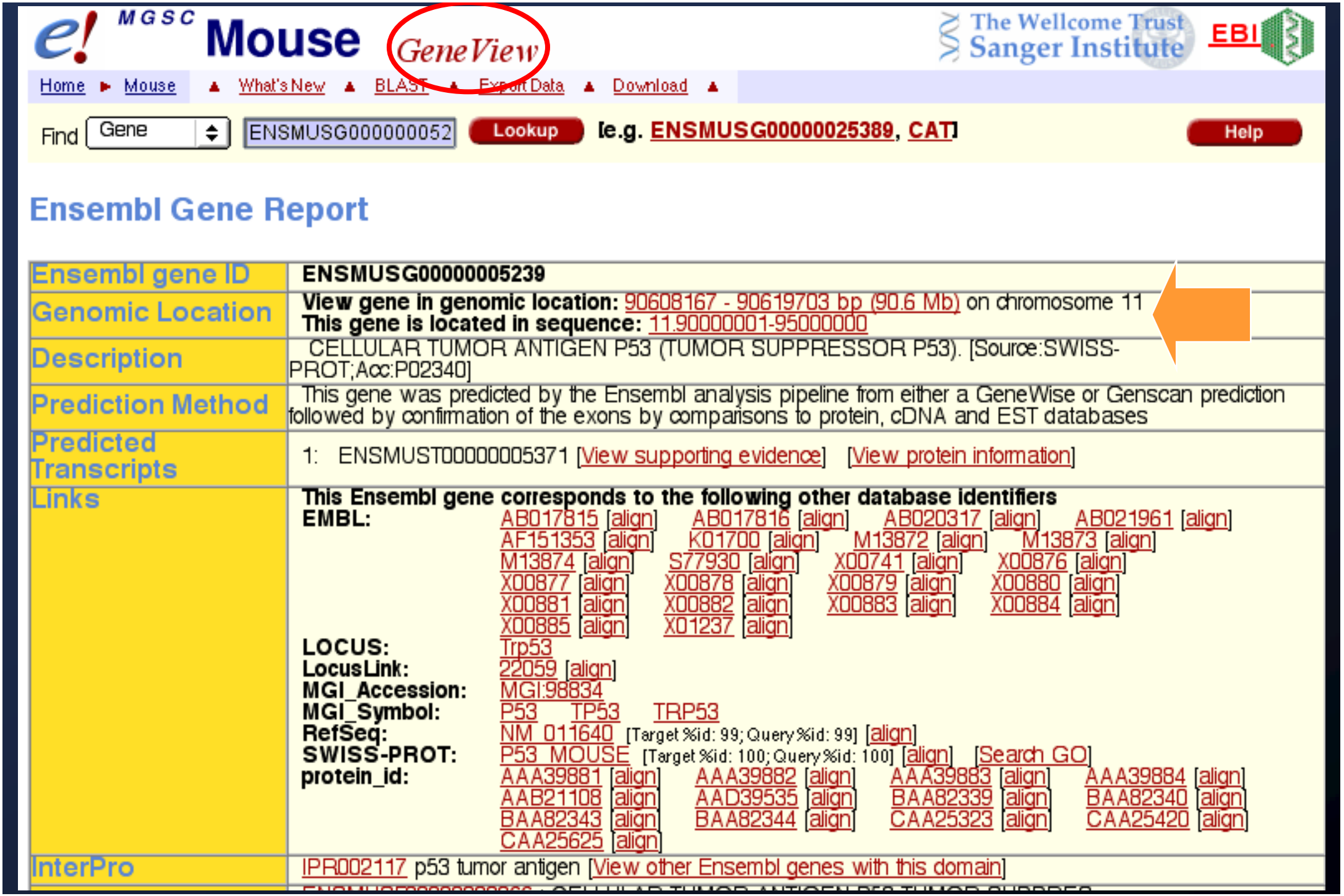
InterPro Top 40 Domains

The 40 most common InterPro domains found in the Ensembl Mouse gene set can be viewed here.




MGSC Mouse
The Wellcome Trust Sanger Institute
EBI
Mouse Genome Server
Search
Gene
p53
Lookup
Browse Chromosome
1
Start
1
End
100000
Lookup
About MGSC (3.1.1 details)
MGSC
The Mouse Genome Sequencing Consortium is a joint project between The Whitehead Institute/MIT Center for Genome Research, The Washington University Genome Sequencing Center, The Wellcome Trust Sanger Institute and EMBL - EBI to provide the Mouse genome sequence to the world. We work closely with other Mouse groups to provide an integrated resource [details].
Other Species
Fly
Zebrafish
Mosquito
Human
Browse a Chromosome


MGSC
Mouse
TextView
The Wellcom
Sanger Ins
Home ▶ Mouse ▲ What's New ▲ BLAST ▲ Export Data ▲ Download ▲
Display up to 20 results in standard format
Search Gene indexes for: p53
POWERED BY altavista:
[Help with AV query syntax]
Lookup
Help
1 documents match your query (Documents searched: 16801)
1. Ensembl Gene: ENSMUSG00000005239
Ensembl gene ENSMUSG00000005239 has 1 transcripts: ENSMUST00000005371
This gene has a number of external identifiers mapped to it:
EMBL:
M13872,X01237,M13873,M13874,AF151353,X00741,AB021961,S77930,AB020317,X00880,AB017815,X00881,A
Gene_name: P53,TRP53,TP53; LOCUS: Trp53; LocusLink: 22059; MGI: MGI:98834; RefSeq: NM_011640; SWIS
P53_MOUSE; protein_id:
BAA82344,AAA39882,AAA39883,CAA25323,CAA25420,AAB21108,AAD39535,AAA39884,BAA82339,BAA823
http://www.ensembl.org/Mus_musculus/geneview?gene=ENSMUSG00000005239



MGSC Mouse GeneView

The Wellcome Trust Sanger Institute EBI

Home ▸ Mouse ▲ What's New ▲ BLAST ▲ Export Data ▲ Download ▲

Find Gene ENSMUSG000000052 Lookup [e.g. ENSMUSG00000025389, CAT] Help

Ensembl Gene Report

Ensembl gene ID	**ENSMUSG00000005239**
Genomic Location	**View gene in genomic location:** 90608167 - 90619703 bp (90.6 Mb) on chromosome 11 **This gene is located in sequence:** 11.90000001-95000000
Description	CELLULAR TUMOR ANTIGEN P53 (TUMOR SUPPRESSOR P53). [Source:SWISS-PROT;Acc:P02340]
Prediction Method	This gene was predicted by the Ensembl analysis pipeline from either a GeneWise or Genscan prediction followed by confirmation of the exons by comparisons to protein, cDNA and EST databases
Predicted Transcripts	1: ENSMUST00000005371 [View supporting evidence] [View protein information]
Links	**This Ensembl gene corresponds to the following other database identifiers** **EMBL:** AB017815 [align] AB017816 [align] AB020317 [align] AB021961 [align] AF151353 [align] K01700 [align] M13872 [align] M13873 [align] M13874 [align] S77930 [align] X00741 [align] X00876 [align] X00877 [align] X00878 [align] X00879 [align] X00880 [align] X00881 [align] X00882 [align] X00883 [align] X00884 [align] X00885 [align] X01237 [align] **LOCUS:** Trp53 **LocusLink:** 22059 [align] **MGI_Accession:** MGI:98834 **MGI_Symbol:** P53 TP53 TRP53 **RefSeq:** NM_011640 [Target %id: 99; Query %id: 99] [align] **SWISS-PROT:** P53_MOUSE [Target %id: 100; Query %id: 100] [align] [Search GO] **protein_id:** AAA39881 [align] AAA39882 [align] AAA39883 [align] AAA39884 [align] AAB21108 [align] AAD39535 [align] BAA82339 [align] BAA82340 [align] BAA82343 [align] BAA82344 [align] CAA25323 [align] CAA25420 [align] CAA25625 [align]
InterPro	IPR002117 p53 tumor antigen [View other Ensembl genes with this domain]

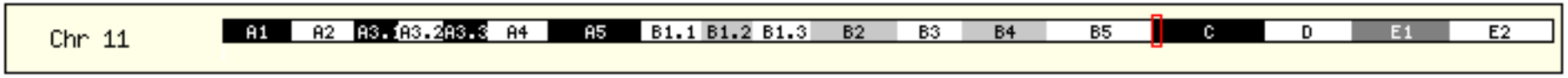
MGSC **Mouse** *ContigView*

The Wellcome Trust Sanger Institute EBI

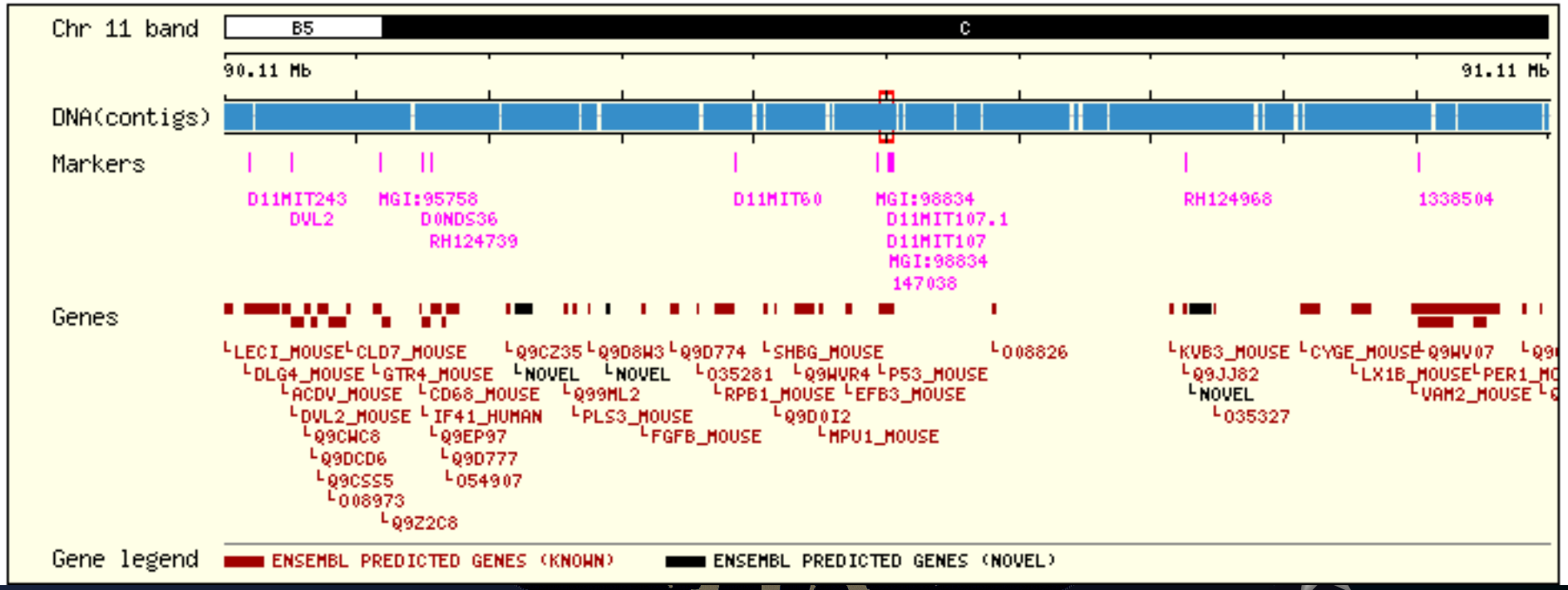
Home ▸ Mouse ▲ What's New ▲ BLAST ▲ Export Data ▲ Download ▲

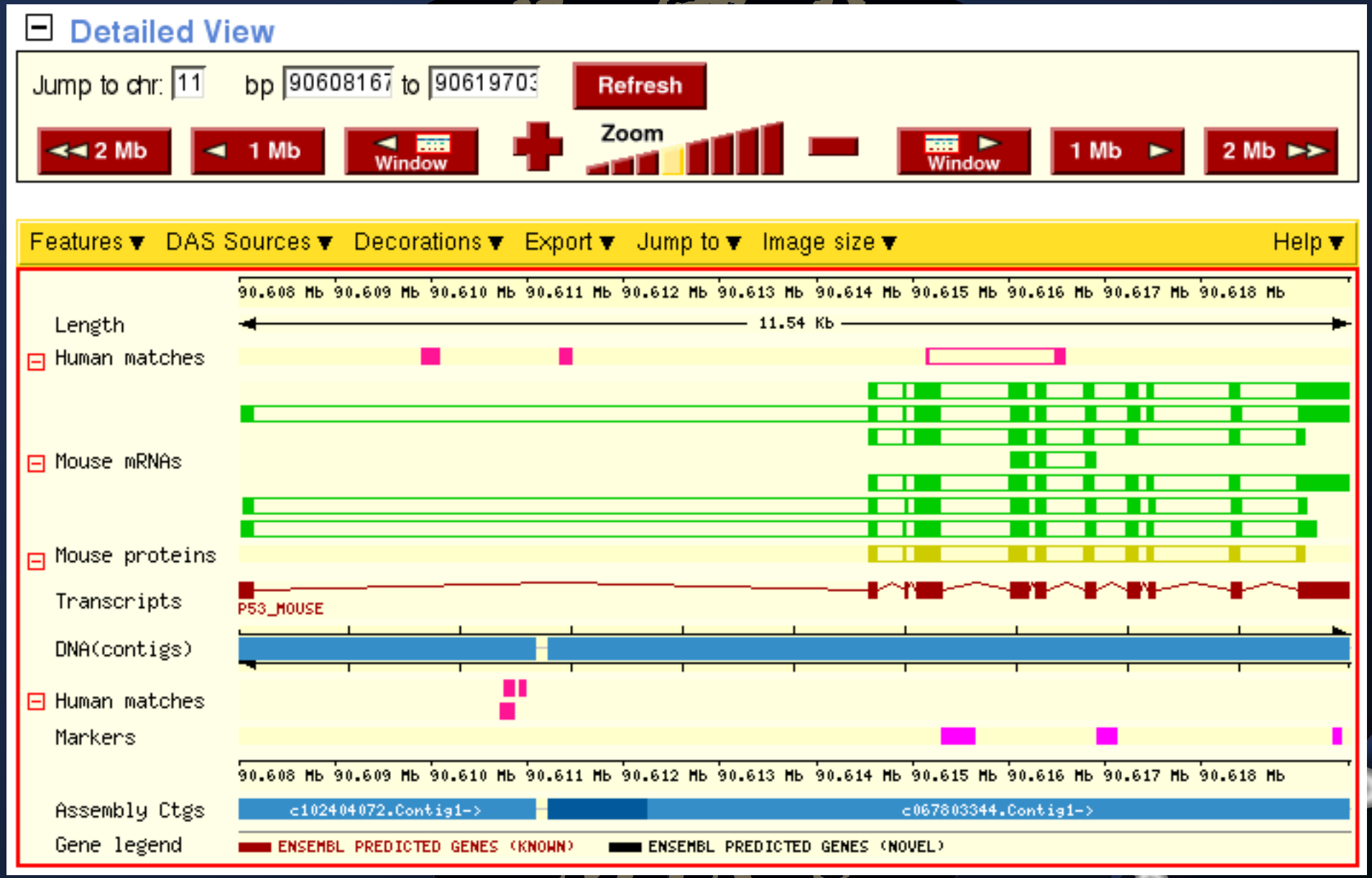
Find Sequence 11 Lookup [e.g. 1, CAT] Help

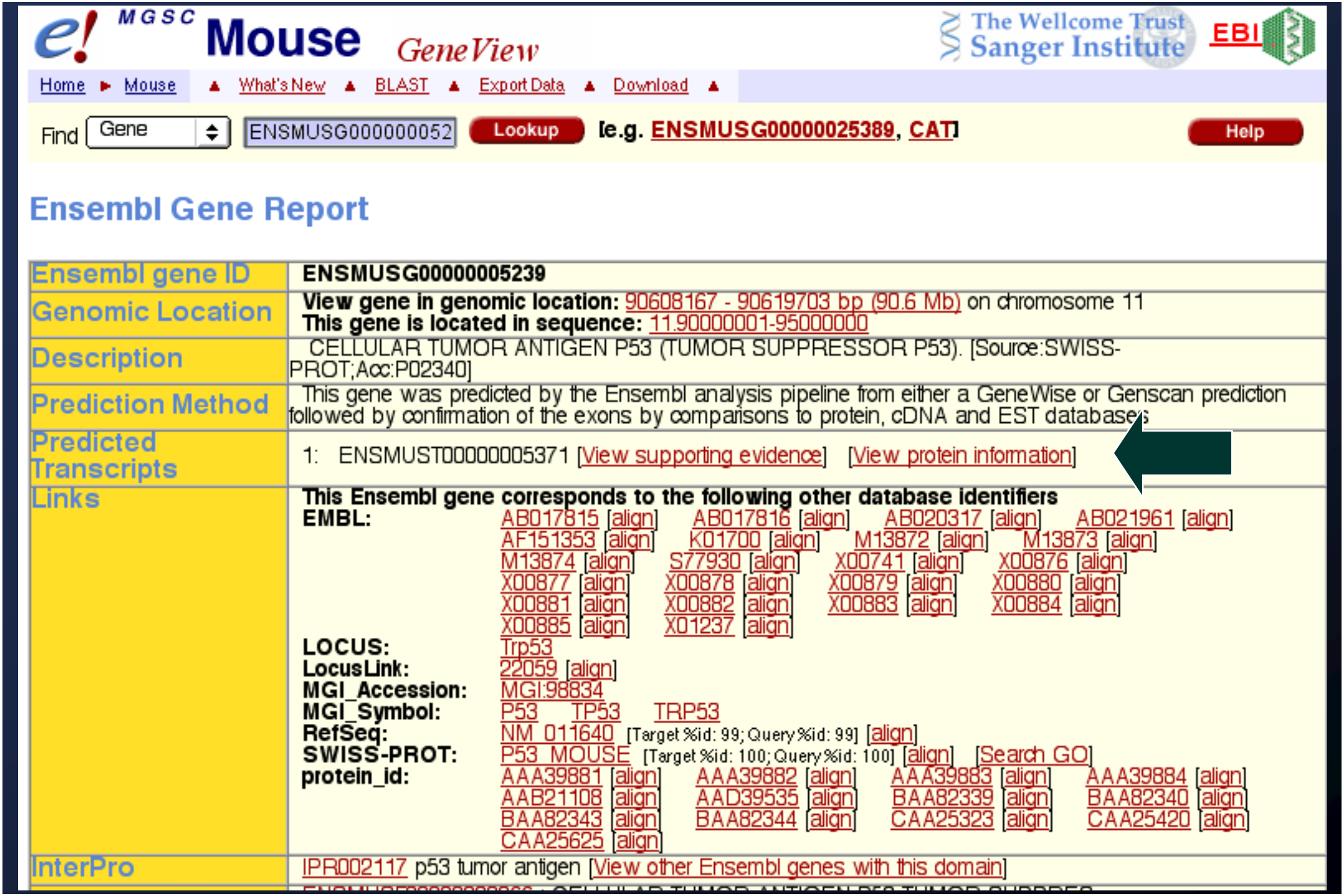
Chromosome 11



Overview




Detailed View
Jump to chr: 11 bp 90608167 to 90619703
Refresh
2 Mb
1 Mb
Window
Zoom
Window
1 Mb
2 Mb
Features
DAS Sources
Decorations
Export
Jump to
Image size
Help
90.608 Mb 90.609 Mb 90.610 Mb 90.611 Mb 90.612 Mb 90.613 Mb 90.614 Mb 90.615 Mb 90.616 Mb 90.617 Mb 90.618 Mb
Length
11.54 Kb
Human matches
Mouse mRNAs
Mouse proteins
Transcripts
P53_MOUSE
DNA(contigs)
Human matches
Markers
90.608 Mb 90.609 Mb 90.610 Mb 90.611 Mb 90.612 Mb 90.613 Mb 90.614 Mb 90.615 Mb 90.616 Mb 90.617 Mb 90.618 Mb
Assembly Ctgs
c102404072.Contig1->
c067803344.Contig1->
Gene legend
ENSEMBL PREDICTED GENES (KNOWN)
ENSEMBL PREDICTED GENES (NOVEL)

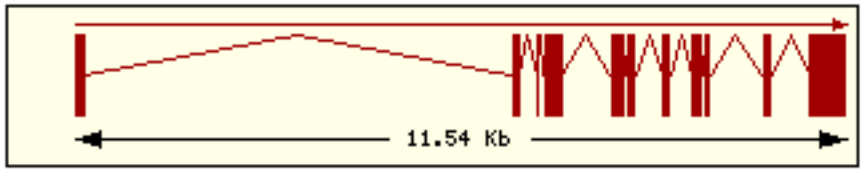

MGSC Mouse GeneView
The Wellcome Trust Sanger Institute
EBI
Home ▶ Mouse ▲ What's New ▲ BLAST ▲ Export Data ▲ Download ▲
Find Gene ENSMUSG00000052 Lookup [e.g. ENSMUSG00000025389, CAT] Help
Ensembl Gene Report
Ensembl gene ID
ENSMUSG00000005239
Genomic Location
View gene in genomic location: 90608167 - 90619703 bp (90.6 Mb) on chromosome 11
This gene is located in sequence: 11.90000001-95000000
Description
CELLULAR TUMOR ANTIGEN P53 (TUMOR SUPPRESSOR P53). [Source:SWISS-PROT;Acc:P02340]
Prediction Method
This gene was predicted by the Ensembl analysis pipeline from either a GeneWise or Genscan prediction followed by confirmation of the exons by comparisons to protein, cDNA and EST databases
Predicted Transcripts
1: ENSMUST00000005371 [View supporting evidence] [View protein information]
Links
This Ensembl gene corresponds to the following other database identifiers
EMBL: AB017815 [align] AB017816 [align] AB020317 [align] AB021961 [align] AF151353 [align] K01700 [align] M13872 [align] M13873 [align] M13874 [align] S77930 [align] X00741 [align] X00876 [align] X00877 [align] X00878 [align] X00879 [align] X00880 [align] X00881 [align] X00882 [align] X00883 [align] X00884 [align] X00885 [align] X01237 [align]
LOCUS: Trp53
LocusLink: 22059 [align]
MGI_Accession: MGI:98834
MGI_Symbol: P53 TP53 TRP53
RefSeq: NM_011640 [Target %id: 99; Query %id: 99] [align]
SWISS-PROT: P53_MOUSE [Target %id: 100; Query %id: 100] [align] [Search GO]
protein_id: AAA39881 [align] AAA39882 [align] AAA39883 [align] AAA39884 [align] AAB21108 [align] AAD39535 [align] BAA82339 [align] BAA82340 [align] BAA82343 [align] BAA82344 [align] CAA25323 [align] CAA25420 [align] CAA25625 [align]
InterPro
IPR002117 p53 tumor antigen [View other Ensembl genes with this domain]

Transcript 1: ENSMUST00000005371

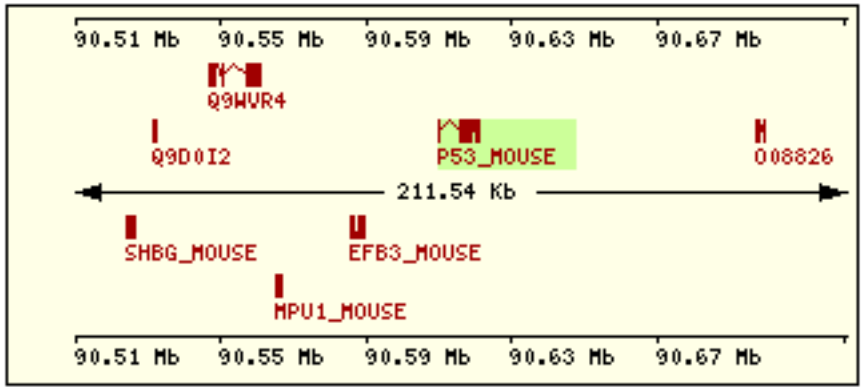
Transcript cDNA Sequence **Total length:** 1772 bp **No. Exons:** 11

>ENSMUST00000005371
TTTCCCCTCCCACGTGCTCACCCTGGCTAAAGTTCTGTAGCTTCAGTTCA
TTGGGACCATCCTGGCTGTAGGTAGCGACTACAGTTAGGGGGCACCTAGC
ATTCAGGCCCTCATCCTCCTCCTTCCCAGCAGGGTGTCACGCTTCTCCGA
AGACTGGNNATGACTGCCATGGAGGAGTCACAGTCGGATATCAGCCTCGA
GCTCCCTCTGAGCCAGGAGACATTTTCAGGCTTATGGAAACTACTTCCTC
CAGAAGATATCCTGCCATCACCTCACTGCATGGACGATCTGTTGCTGCCC
CAGGATGTTGAGGAGTTTTTTGAAGGCCCAAGTGAAGCCCTCCGAGTGTC
AGGAGCTCCTGCAGCACAGGACCCTGTCACCGAGACCCCTGGGCCAGTGG
CCCCTGCCCCAGCCACTCCATGGCCCCTGTCATCTTTTGTCCCTTCTCAA
AAAACTTACCAGGGCAACTATGGCTTCCACCTGGGCTTCCTGCAGTCTGG
GACAGCCAAGTCTGTTATGTGCACGTACTCTCCTCCCCTCAATAAGCTAT
TCTGCCAGCTGGCGAAGACGTGCCCTGTGCAGTTGTGGGTCAGCGCCACA
CCTCCAGCTGGGAGCCGTGTCCGCGCCATGGCCATCTACAAGAAGTCACA
GCACATGACGGAGGTCGTGAGACGCTGCCCCCACCATGAGCGCTGCTCCG
ATGGTGATGGCCTGGCTCCTCCCCAGCATCTTATCCGGGTGGAAGGAAAT
TTGTATCCCGAGTATCTGGAAGACAGGCAGACTTTTCGCCACAGCGTGGT
GGTACCTTATGAGCCACCCGAGGCCGGCTCTGAGTATACCACCATCCACT
ACAAGTACATGTGTAATAGCTCCTGCATGGGGGGCATGAACCGCCGACCT
ATCCTTACCATCATCACACTGGAAGACTCCAGTGGGAACCTTCTGGGACG
GGACAGCTTTGAGGTTCGTGTTTGTGCCTGCCCTGGGAGAGACCGCCGTA
CAGAAGAAGAAAATTTCCGCAAAAAGGAAGTCCTTTGCCCTGAACTGCCC
CCAGGGAGCGCAAAGAGAGCGCTGCCCACCTGCACAAGCGCCTCTCCCCC
GCAAAAGAAAAAACCACTTGATGGAGAGTATTTCACCCTCAAGATCCGCG
GGCGTAAACGCTTCGAGATGTTCCGGGAGCTGAATGAGGCCTTAGAGTTA
AAGGATGCCCATGCTACAGAGGAGTCTGGAGACAGCAGGGCTCACTCCAG
CTACCTGAAGACCAAGAAGGGCCAGTCTACTTCCCGCCATAAAAAAACAA
TGGTCAAGAAAGTGGGGCCTGACTCAGACTGACTGCCTNTGCATCCCGTC
CCCATCACCAGCCTCCCCCTCTCCTTGCTGTTTTATGACTTCAGGGCTGA
GACACAATCCTCCCGGTCCCTTCTGCTGCCTTTTTTACCTTGTAGCTAGG
GCTCAGCCCCCTCTCTGAGTAGTGGTTCCTGGCCCAAGTTGGGGAATAGG
TTGATAGTTGTCAGGTCTCTGCTGGCCCAGCGAAATTCTATCCAGCCAGT
TGTTGGACCCTGGCACCTACAATGAAATCTCACCCTACCCCACACCCTGT
AAGATTCTATCTTGGGCCCTCATAGGGTCCATATCCTCCAGGGCCTACTT
TCCTTCCATTCTGCAAAGCCTGTCTGCATTTATCCACCCCCCACCCTGTC
TCCCTCTTTTTTTTTTTTTTACCCCTTTTTATATATCAATTTCCTATTTT
ACAATAAAATTTTGTTATCACTTA

Exon structure



Transcript neighbourhood



View Evidence View Protein

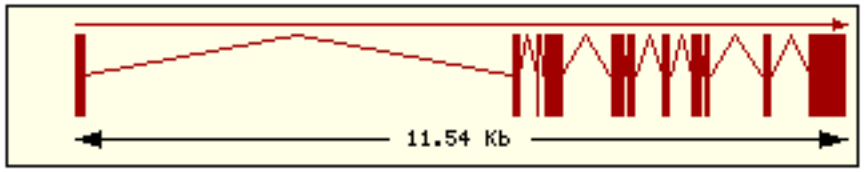
Data Library	Sequence	Definition	Exons 1 2 3 4 5 6 7 8 9 10 11
UniSTS	160672		
mouse_mrna	MMP53R	MMP53R X01237.1 Mouse mRNA for cellular tumour antigen p53	
mouse_mrna	MMP53A	MMP53A M13872.1 Mouse p53 mRNA, complete cds, clone pcD53.	
mouse_mrna	MMKP53M	MMKP53M K01700.1 Mouse p53 cellular tumor antigen, mRNA.	
mouse_mrna	MMP53C	MMP53C M13874.1 Mouse p53 mRNA, complete cds, clone p53-m8.	
mouse_mrna	BC005448	BC005448 BC005448.1 Mus musculus, transformation related protein 53, clone MGC:6174 IMAGE:3599812, mRNA, complete cds.	
mouse_mrna	AB017815	AB017815 AB017815.1 Mus musculus mRNA for p53, complete cds.	
mouse_mrna	AF151353	AF151353 AF151353.1 Mus musculus tumor suppressor p53 (p53) mRNA, complete cds.	
mouse_mrna	AB020317	AB020317 AB020317.1 Mus musculus mRNA for p53, complete cds.	
mouse_mrna	AF161020	AF161020 AF161020.1 Mus musculus tumor supressor p53 (p53) mRNA, complete cds.	
mouse_mrna	MMU297973	MMU297973 AJ297973.1 Mus musculus mRNA for transformation related protein 53 (Trp53 gene)	
mouse_mrna	MSP53A	MSP53A D63399.1 Mus sp. p53 mRNA.	
mouse_mrna	MM59757	MM59757 U59757.1 Mus musculus p53 cellular tumor antigen (p53) mRNA, partial cds.	

Transcript 1: ENSMUST00000005371

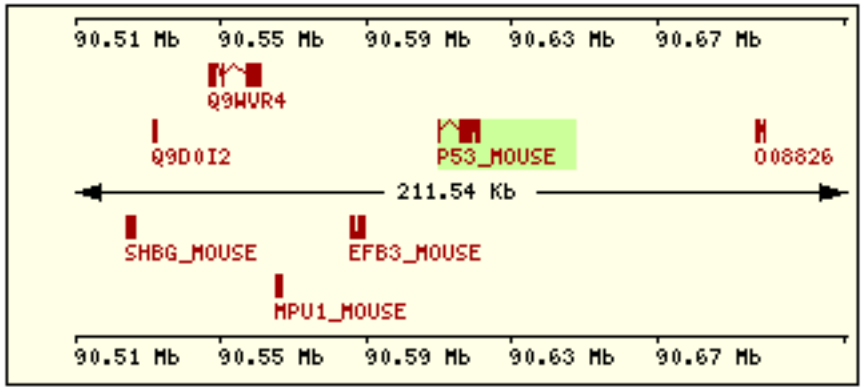
Transcript cDNA Sequence **Total length:** 1772 bp **No. Exons:** 11

>ENSMUST00000005371
TTTCCCCTCCCACGTGCTCACCCTGGCTAAAGTTCTGTAGCTTCAGTTCA
TTGGGACCATCCTGGCTGTAGGTAGCGACTACAGTTAGGGGGCACCTAGC
ATTCAGGCCCTCATCCTCCTCCTTCCCAGCAGGGTGTCACGCTTCTCCGA
AGACTGGNNATGACTGCCATGGAGGAGTCACAGTCGGATATCAGCCTCGA
GCTCCCTCTGAGCCAGGAGACATTTTCAGGCTTATGGAAACTACTTCCTC
CAGAAGATATCCTGCCATCACCTCACTGCATGGACGATCTGTTGCTGCCC
CAGGATGTTGAGGAGTTTTTTGAAGGCCCAAGTGAAGCCCTCCGAGTGTC
AGGAGCTCCTGCAGCACAGGACCCTGTCACCGAGACCCCTGGGCCAGTGG
CCCCTGCCCCAGCCACTCCATGGCCCCTGTCATCTTTTGTCCCTTCTCAA
AAAACTTACCAGGGCAACTATGGCTTCCACCTGGGCTTCCTGCAGTCTGG
GACAGCCAAGTCTGTTATGTGCACGTACTCTCCTCCCCTCAATAAGCTAT
TCTGCCAGCTGGCGAAGACGTGCCCTGTGCAGTTGTGGGTCAGCGCCACA
CCTCCAGCTGGGAGCCGTGTCCGCGCCATGGCCATCTACAAGAAGTCACA
GCACATGACGGAGGTCGTGAGACGCTGCCCCCACCATGAGCGCTGCTCCG
ATGGTGATGGCCTGGCTCCTCCCCAGCATCTTATCCGGGTGGAAGGAAAT
TTGTATCCCGAGTATCTGGAAGACAGGCAGACTTTTCGCCACAGCGTGGT
GGTACCTTATGAGCCACCCGAGGCCGGCTCTGAGTATACCACCATCCACT
ACAAGTACATGTGTAATAGCTCCTGCATGGGGGGCATGAACCGCCGACCT
ATCCTTACCATCATCACACTGGAAGACTCCAGTGGGAACCTTCTGGGACG
GGACAGCTTTGAGGTTCGTGTTTGTGCCTGCCCTGGGAGAGACCGCCGTA
CAGAAGAAGAAAATTTCCGCAAAAAGGAAGTCCTTTGCCCTGAACTGCCC
CCAGGGAGCGCAAAGAGAGCGCTGCCCACCTGCACAAGCGCCTCTCCCCC
GCAAAAGAAAAAACCACTTGATGGAGAGTATTTCACCCTCAAGATCCGCG
GGCGTAAACGCTTCGAGATGTTCCGGGAGCTGAATGAGGCCTTAGAGTTA
AAGGATGCCCATGCTACAGAGGAGTCTGGAGACAGCAGGGCTCACTCCAG
CTACCTGAAGACCAAGAAGGGCCAGTCTACTTCCCGCCATAAAAAAACAA
TGGTCAAGAAAGTGGGGCCTGACTCAGACTGACTGCCTNTGCATCCCGTC
CCCATCACCAGCCTCCCCCTCTCCTTGCTGTTTTATGACTTCAGGGCTGA
GACACAATCCTCCCGGTCCCTTCTGCTGCCTTTTTTACCTTGTAGCTAGG
GCTCAGCCCCCTCTCTGAGTAGTGGTTCCTGGCCCAAGTTGGGGAATAGG
TTGATAGTTGTCAGGTCTCTGCTGGCCCAGCGAAATTCTATCCAGCCAGT
TGTTGGACCCTGGCACCTACAATGAAATCTCACCCTACCCCACACCCTGT
AAGATTCTATCTTGGGCCCTCATAGGGTCCATATCCTCCAGGGCCTACTT
TCCTTCCATTCTGCAAAGCCTGTCTGCATTTATCCACCCCCCACCCTGTC
TCCCTCTTTTTTTTTTTTTTACCCCTTTTTATATATCAATTTCCTATTTT
ACAATAAAATTTTGTTATCACTTA

Exon structure

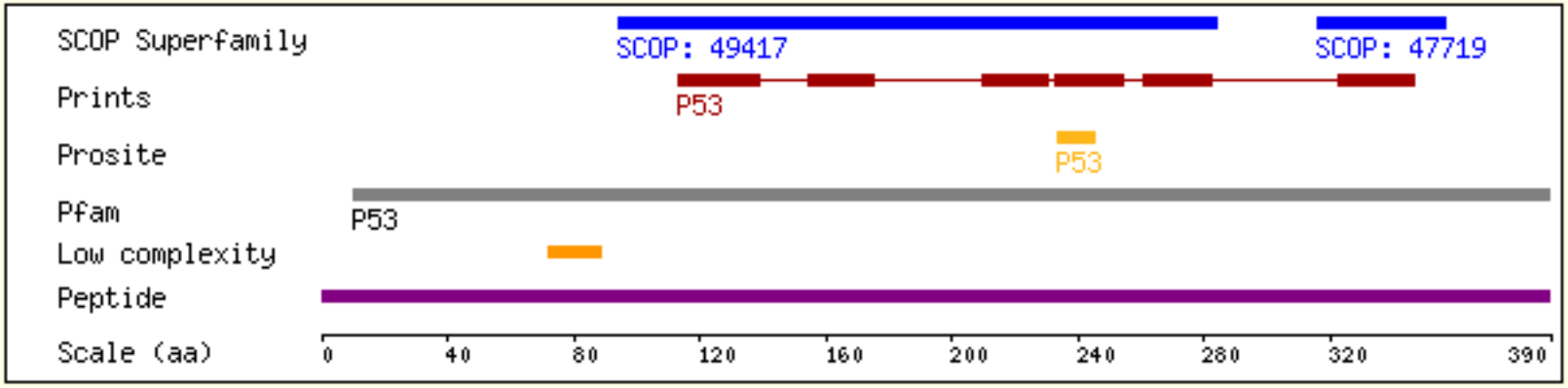


Transcript neighbourhood



View Evidence | View Protein

Ensembl Protein Report

Ensembl Protein	**ENSMUSP00000005371**
Ensembl Gene	This protein is a product of Ensembl gene ENSMUSG00000005239 [Supporting evidence]
Description	CELLULAR TUMOR ANTIGEN P53 (TUMOR SUPPRESSOR P53). [Source:SWISS-PROT;Acc:P02340]
Prediction Method	This protein was predicted by the Ensembl analysis pipeline from either a GeneWise or Genscan prediction followed by confirmation of the exons by comparisons to protein, cDNA and EST databases
InterPro	IPR002117 p53 tumor antigen [View other Ensembl genes with this domain]
Protein Family	ENSMUSF00000000366 : CELLULAR TUMOR ANTIGEN P53 TUMOR SUPPRES This cluster contains 1 Ensembl gene member(s)
Protein structure	

Peptide : ENSMUSP00000005371

Peptide sequence

>ENSMUSP00000005371
MTAMEESQSDISLELPLSQETFSGLWKLLPPEDILPSPHCMDDLLLPQDVEEFFEGPSEALRVSGAPAAQ
DPVTETPGPVAPAPATPWPLSSFVPSQKTYQGNYGFHLGFLQSGTAKSVMCTYSPPLNKLFCQLAKTCPV
QLWVSATPPAGSRVRAMAIYKKSQHMTEVVRRCPHHERCSDGDGLAPPQHLIRVEGNLYPEYLEDRQTFR
HSVVVPYEPPEAGSEYTTIHYKYMCNSSCMGGMNRRPILTIITLEDSSGNLLGRDSFEVRVCACPGRDRR
TEEENFRKKEVLCPELPPGSAKRALPTCTSASPPQKKKPLDGEYFTLKIRGRKRFEMFRELNEALELKDA
HATEESGDSRAHSSYLKTKKGQSTSRHKKTMVKKVGPDSD

Peptide properties

Residues: 390
MW: 43458.31
Avg. Res. Wt.: 111.432
Charge: 4.5
pI: 7.2470

View Transcript Info

This peptide corresponds to the following identifiers with the percent identity specified:

SWISS-PROT: P53_MOUSE [Target % identity : 100; Query % identity : 100]
RefSeq: NM_011640 [Target % identity : 99; Query % identity : 99]

Domains				
Prediction method	**Accession number**	**Start**	**End**	**Description**
prints	PR00386	113	139	P53
prints	PR00386	155	176	P53
prints	PR00386	210	231	P53
prints	PR00386	233	255	P53
prints	PR00386	261	283	P53
prints	PR00386	323	347	P53
Pfam	PF00870	10	390	P53
prosite	PS00348	234	246	P53
superfamily	47719	316	357	p53
superfamily	49417	94	284	p53-like

MGSC Mouse GeneView

The Wellcome Trust Sanger Institute EBI

Home ▸ Mouse ▲ What's New ▲ BLAST ▲ Export Data ▲ Download ▲

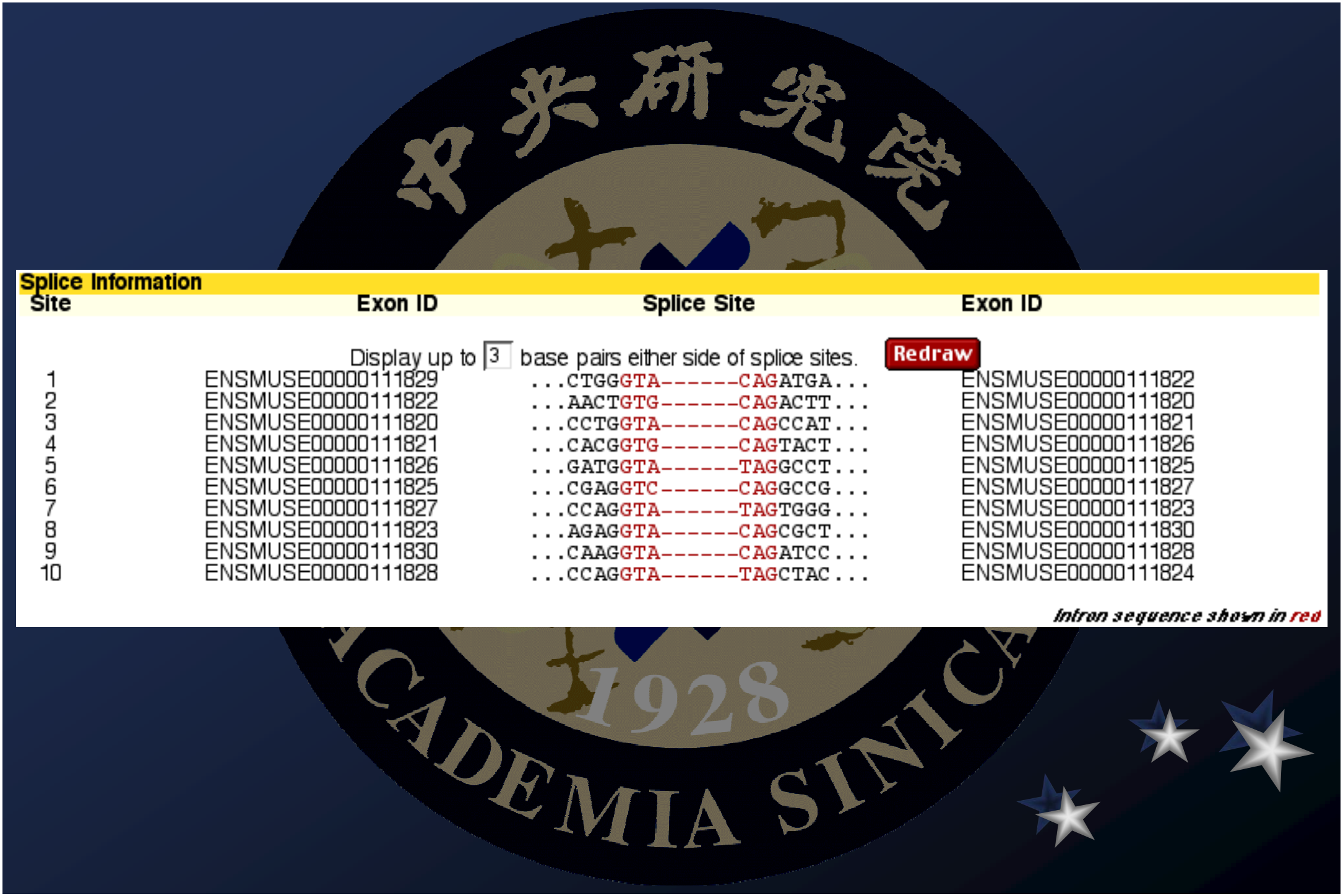
Find Gene ENSMUSG000000052 Lookup [e.g. ENSMUSG00000025389, CAT] Help

Ensembl Gene Report

Ensembl gene ID	**ENSMUSG00000005239**
Genomic Location	**View gene in genomic location:** 90608167 - 90619703 bp (90.6 Mb) on chromosome 11 **This gene is located in sequence:** 11.90000001-95000000
Description	CELLULAR TUMOR ANTIGEN P53 (TUMOR SUPPRESSOR P53). [Source:SWISS-PROT;Acc:P02340]
Prediction Method	This gene was predicted by the Ensembl analysis pipeline from either a GeneWise or Genscan prediction followed by confirmation of the exons by comparisons to protein, cDNA and EST databases
Predicted Transcripts	1: ENSMUST00000005371 [View supporting evidence] [View protein information]
Links	**This Ensembl gene corresponds to the following other database identifiers** **EMBL:** AB017815 [align] AB017816 [align] AB020317 [align] AB021961 [align] AF151353 [align] K01700 [align] M13872 [align] M13873 [align] M13874 [align] S77930 [align] X00741 [align] X00876 [align] X00877 [align] X00878 [align] X00879 [align] X00880 [align] X00881 [align] X00882 [align] X00883 [align] X00884 [align] X00885 [align] X01237 [align] **LOCUS:** Trp53 **LocusLink:** 22059 [align] **MGI_Accession:** MGI:98834 **MGI_Symbol:** P53 TP53 TRP53 **RefSeq:** NM_011640 [Target %id: 99; Query %id: 99] [align] **SWISS-PROT:** P53_MOUSE [Target %id: 100; Query %id: 100] [align] [Search GO] **protein_id:** AAA39881 [align] AAA39882 [align] AAA39883 [align] AAA39884 [align] AAB21108 [align] AAD39535 [align] BAA82339 [align] BAA82340 [align] BAA82343 [align] BAA82344 [align] CAA25323 [align] CAA25420 [align] CAA25625 [align]
InterPro	IPR002117 p53 tumor antigen [View other Ensembl genes with this domain]

Exon Information

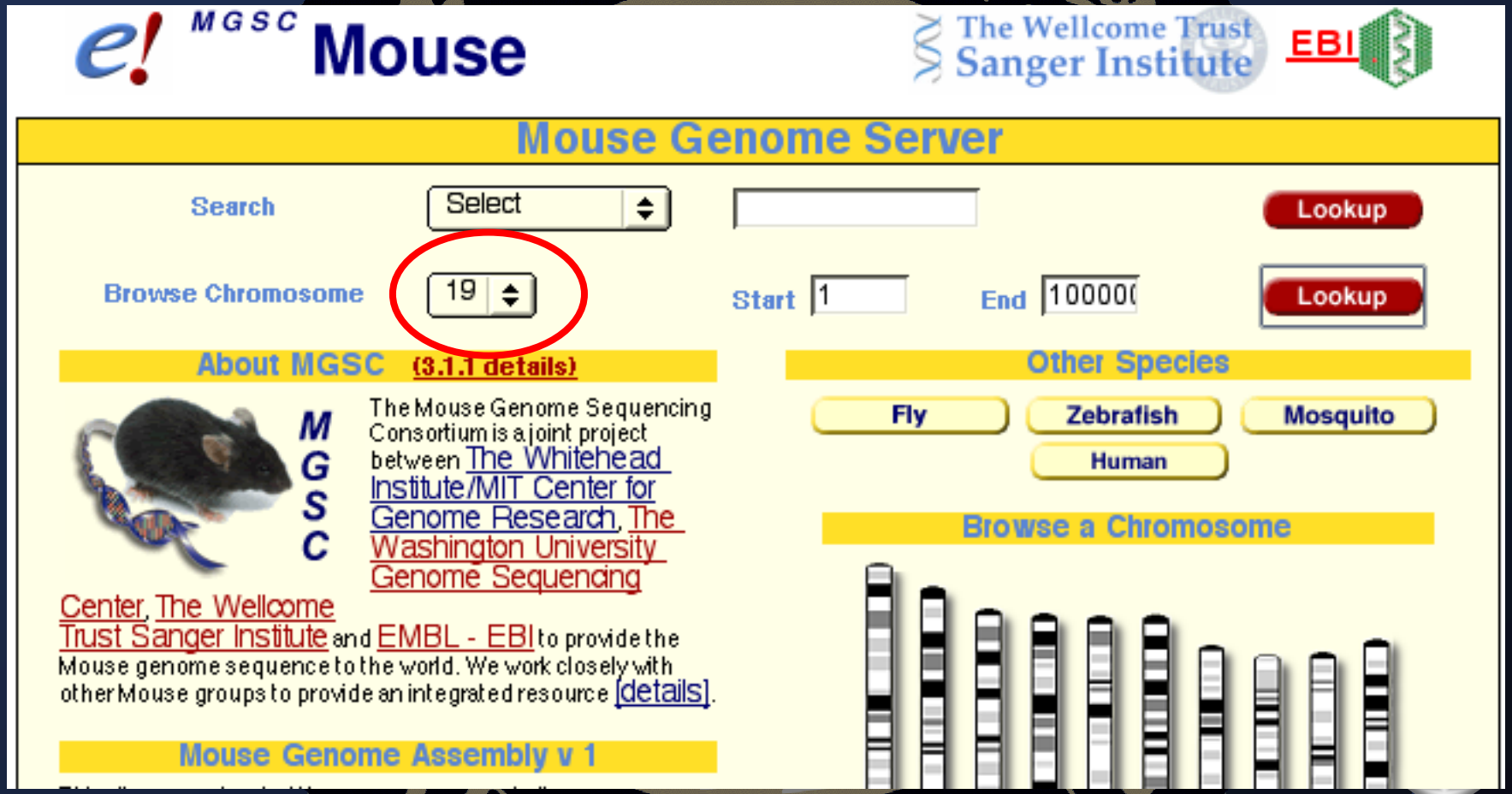
No.	Exon ID	Contig	Strand	Start	End	Length	Exon Sequence
1	ENSMUSE00000111829	11.90000001-95000000	1	608167	608323	157 bp	TTTCCCCTCCCACGTGCTCACCCTGGCTAAAGTTCTGTAGCTTCAGTTCA TTGGGACCATCCTGGCTGTAGGTAGCGACTACAGTTAGGGGGCACCTAGC ATTCAGGCCCTCATCCTCCTCCTTCCCAGCAGGGTGTCACGCTTCTCCGA AGACTGG
2	ENSMUSE00000111822	11.90000001-95000000	1	614721	614803	83 bp	ATGACTGCCATGGAGGAGTCACAGTCGGATATCAGCCTCGAGCTCCCTCT GAGCCAGGAGACATTTTCAGGCTTATGGAAACT
3	ENSMUSE00000111820	11.90000001-95000000	1	615087	615108	22 bp	ACTTCCTCCAGAAGATATCCTG
4	ENSMUSE00000111821	11.90000001-95000000	1	615202	615462	261 bp	CCATCACCTCACTGCATGGACGATCTGTTGCTGCCCCAGGATGTTGAGGA GTTTTTTGAAGGCCCAAGTGAAGCCCTCCGAGTGTCAGGAGCTCCTGCAG CACAGGACCCTGTCACCGAGACCCCTGGGCCAGTGGCCCCTGCCCCAGCC ACTCCATGGCCCCTGTCATCTTTTGTCCCTTCTCAAAAAACTTACCAGGG CAACTATGGCTTCCACCTGGGCTTCCTGCAGTCTGGGACAGCCAAGTCTG TTATGTGCACG
5	ENSMUSE00000111826	11.90000001-95000000	1	616194	616377	184 bp	TACTCTCCTCCCCTCAATAAGCTATTCTGCCAGCTGGCGAAGACGTGCCC TGTGCAGTTGTGGGTCAGCGCCACACCTCCAGCTGGGAGCCGTGTCCGCG CCATGGCCATCTACAAGAAGTCACAGCACATGACGGAGGTCGTGAGACGC TGCCCCCACCATGAGCGCTGCTCCGATGGTGATG
6	ENSMUSE00000111825	11.90000001-95000000	1	616456	616568	113 bp	GCCTGGCTCCTCCCCAGCATCTTATCCGGGTGGAAGGAAATTTGTATCCC GAGTATCTGGAAGACAGGCAGACTTTTCGCCACAGCGTGGTGGTACCTTA TGAGCCACCCGAG
7	ENSMUSE00000111827	11.90000001-95000000	1	616971	617080	110 bp	GCCGGCTCTGAGTATACCACCATCCACTACAAGTACATGTGTAATAGCTC CTGCATGGGGGGCATGAACCGCCGACCTATCCTTACCATCATCACACTGG AAGACTCCAG
8	ENSMUSE00000111823	11.90000001-95000000	1	617403	617539	137 bp	TGGGAACCTTCTGGGACGGGACAGCTTTGAGGTTCGTGTTTGTGCCTGCC CTGGGAGAGACCGCCGTACAGAAGAAGAAAATTTCCGCAAAAAGGAAGTC CTTTGCCCTGAACTGCCCCCAGGGAGCGCAAAGAGAG
9	ENSMUSE00000111830	11.90000001-95000000	1	617619	617692	74 bp	CGCTGCCCACCTGCACAAGCGCCTCTCCCCCGCAAAAGAAAAAACCACTT GATGGAGAGTATTTCACCCTCAAG
10	ENSMUSE00000111828	11.90000001-95000000	1	618486	618592	107 bp	ATCCGCGGGCGTAAACGCTTCGAGATGTTCCGGGAGCTGAATGAGGCCTT AGAGTTAAAGGATGCCCATGCTACAGAGGAGTCTGGAGACAGCAGGGCTC ACTCCAG
11	ENSMUSE00000111824	11.90000001-95000000	1	619180	619703	524 bp	CTACCTGAAGACCAAGAAGGGCCAGTCTACTTCCCGCCATAAAAAAACAA TGGTCAAGAAAGTGGGGCCTGACTCAGACTGACTGCCTNTGCATCCCGTC CCCATCACCAGCCTCCCCCTCTCCTTGCTGTTTTATGACTTCAGGGCTGA GACACAATCCTCCCGGTCCCTTCTGCTGCCTTTTTTACCTTGTAGCTAGG GCTCAGCCCCCTCTCTGAGTAGTGGTTCCTGGCCCAAGTTGGGGAATAGG TTGATAGTTGTCAGGTCTCTGCTGGCCCAGCGAAATTCTATCCAGCCAGT TGTTGGACCCTGGCACCTACAATGAAATCTCACCCTACCCCACACCCTGT AAGATTCTATCTTGGGCCCTCATAGGGTCCATATCCTCCAGGGCCTACTT TCCTTCCATTCTGCAAAGCCTGTCTGCATTTATCCACCCCCCACCCTGTC TCCCTCTTTTTTTTTTTTTTTACCCCTTTTTATATATCAATTTCCTATTTT ACAATAAAATTTTGTTATCACTTA

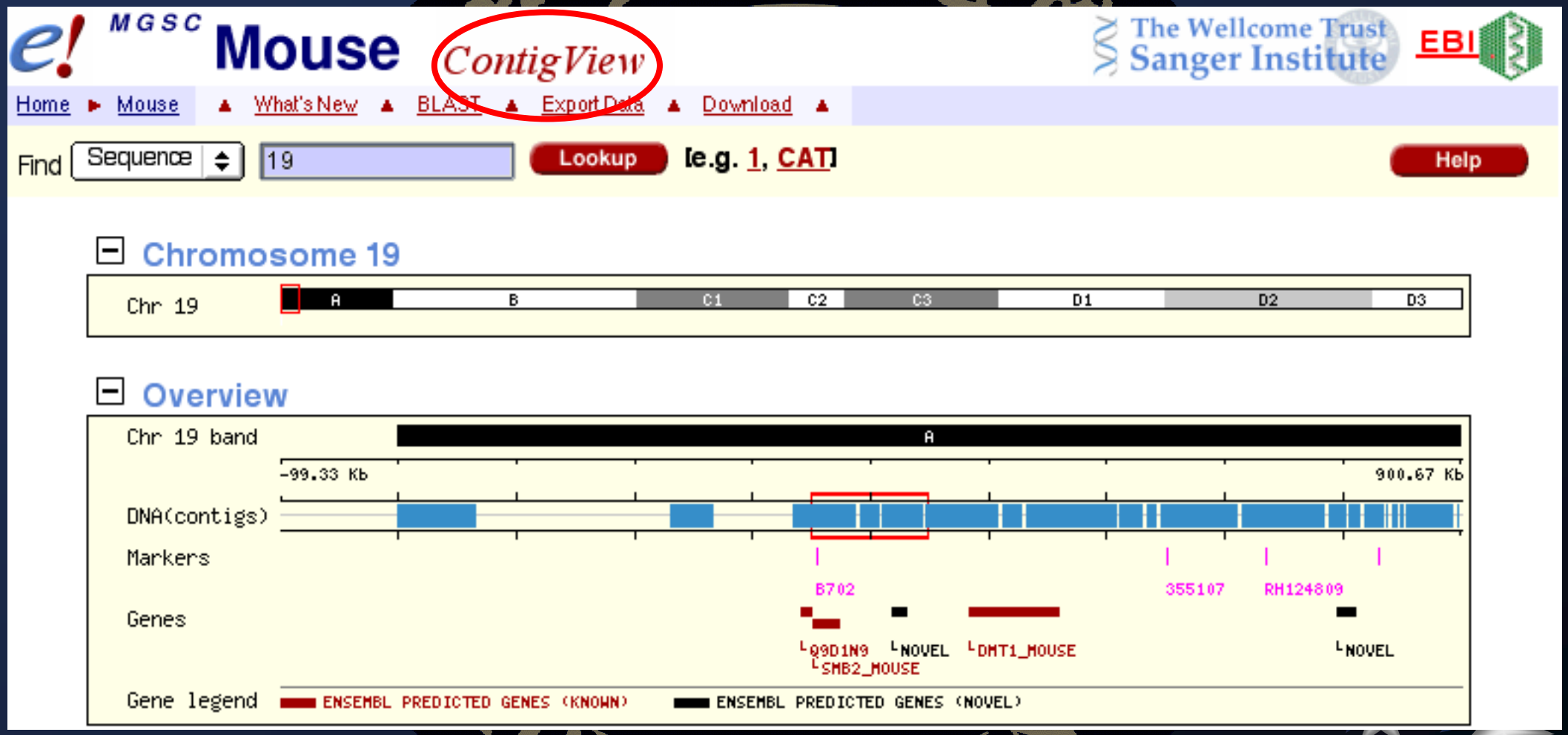


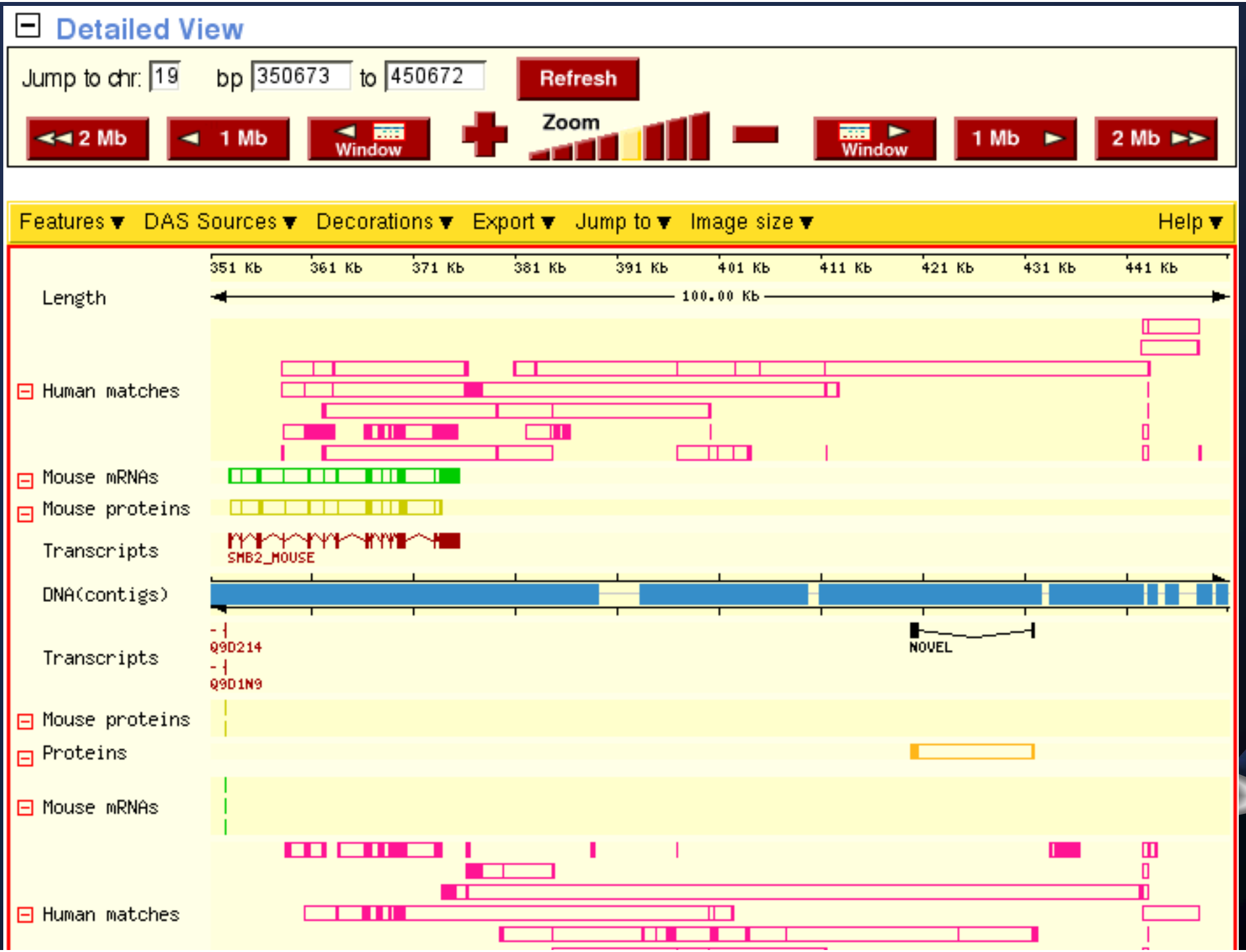
Splice Information

Site	Exon ID	Splice Site	Exon ID
		Display up to 3 base pairs either side of splice sites. Redraw	
1	ENSMUSE00000111829	...CTGGGTA------CAGATGA...	ENSMUSE00000111822
2	ENSMUSE00000111822	...AACTGTG------CAGACTT...	ENSMUSE00000111820
3	ENSMUSE00000111820	...CCTGGTA------CAGCCAT...	ENSMUSE00000111821
4	ENSMUSE00000111821	...CACGGTG------CAGTACT...	ENSMUSE00000111826
5	ENSMUSE00000111826	...GATGGTA------TAGGCCT...	ENSMUSE00000111825
6	ENSMUSE00000111825	...CGAGGTC------CAGGCCG...	ENSMUSE00000111827
7	ENSMUSE00000111827	...CCAGGTA------TAGTGGG...	ENSMUSE00000111823
8	ENSMUSE00000111823	...AGAGGTA------CAGCGCT...	ENSMUSE00000111830
9	ENSMUSE00000111830	...CAAGGTA------CAGATCC...	ENSMUSE00000111828
10	ENSMUSE00000111828	...CCAGGTA------TAGCTAC...	ENSMUSE00000111824

Intron sequence shown in red


MGSC Mouse
The Wellcome Trust Sanger Institute
EBI
Mouse Genome Server
Search
Select
Lookup
Browse Chromosome
19
Start
1
End
100000
Lookup
About MGSC
(3.1.1 details)
MGSC
The Mouse Genome Sequencing Consortium is a joint project between The Whitehead Institute/MIT Center for Genome Research, The Washington University Genome Sequencing Center, The Wellcome Trust Sanger Institute and EMBL - EBI to provide the Mouse genome sequence to the world. We work closely with other Mouse groups to provide an integrated resource [details].
Mouse Genome Assembly v 1
Other Species
Fly
Zebrafish
Mosquito
Human
Browse a Chromosome


MGSC
Mouse
ContigView
The Wellcome Trust Sanger Institute
EBI
Home ▶ Mouse ▲ What's New ▲ BLAST ▲ Export Data ▲ Download ▲
Find Sequence 19 Lookup [e.g. 1, CAT] Help
Chromosome 19
Chr 19
A B C1 C2 C3 D1 D2 D3
Overview
Chr 19 band
A
-99.33 Kb
900.67 Kb
DNA(contigs)
Markers
B702
355107
RH124809
Genes
Q9D1N9
SMB2_MOUSE
NOVEL
DMT1_MOUSE
NOVEL
Gene legend
ENSEMBL PREDICTED GENES (KNOWN)
ENSEMBL PREDICTED GENES (NOVEL)

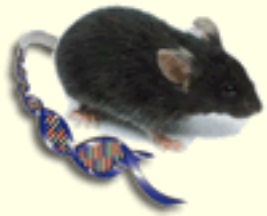

Detailed View
Jump to chr: 19 bp 350673 to 450672
Refresh
2 Mb
1 Mb
Window
Zoom
Window
1 Mb
2 Mb
Features
DAS Sources
Decorations
Export
Jump to
Image size
Help
351 Kb
361 Kb
371 Kb
381 Kb
391 Kb
401 Kb
411 Kb
421 Kb
431 Kb
441 Kb
Length
100.00 Kb
Human matches
Mouse mRNAs
Mouse proteins
Transcripts
SMB2_MOUSE
DNA(contigs)
Transcripts
Q9D214
Q9D1N9
NOVEL
Mouse proteins
Proteins
Mouse mRNAs
Human matches

Mouse Genome Server

Search Select Lookup

Browse Chromosome 19 Start 1 End 100000 Lookup

About MGSC (3.1.1 details)



MGSC

The Mouse Genome Sequencing Consortium is a joint project between The Whitehead Institute/MIT Center for Genome Research, The Washington University Genome Sequencing Center, The Wellcome Trust Sanger Institute and EMBL - EBI to provide the Mouse genome sequence to the world. We work closely with other Mouse groups to provide an integrated resource [details].

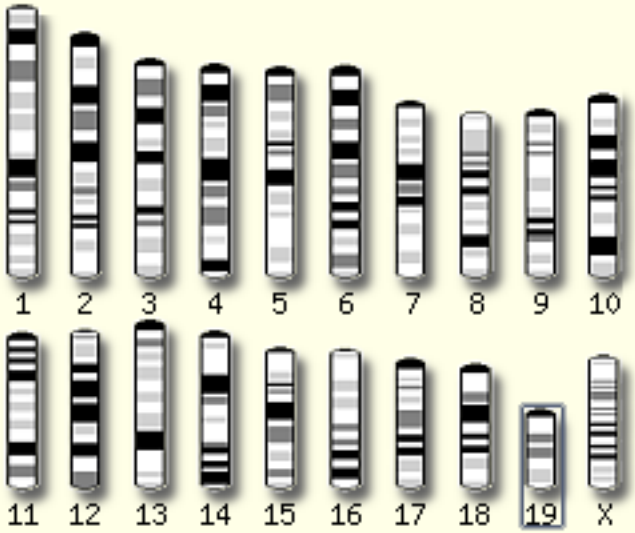
Mouse Genome Assembly v 1

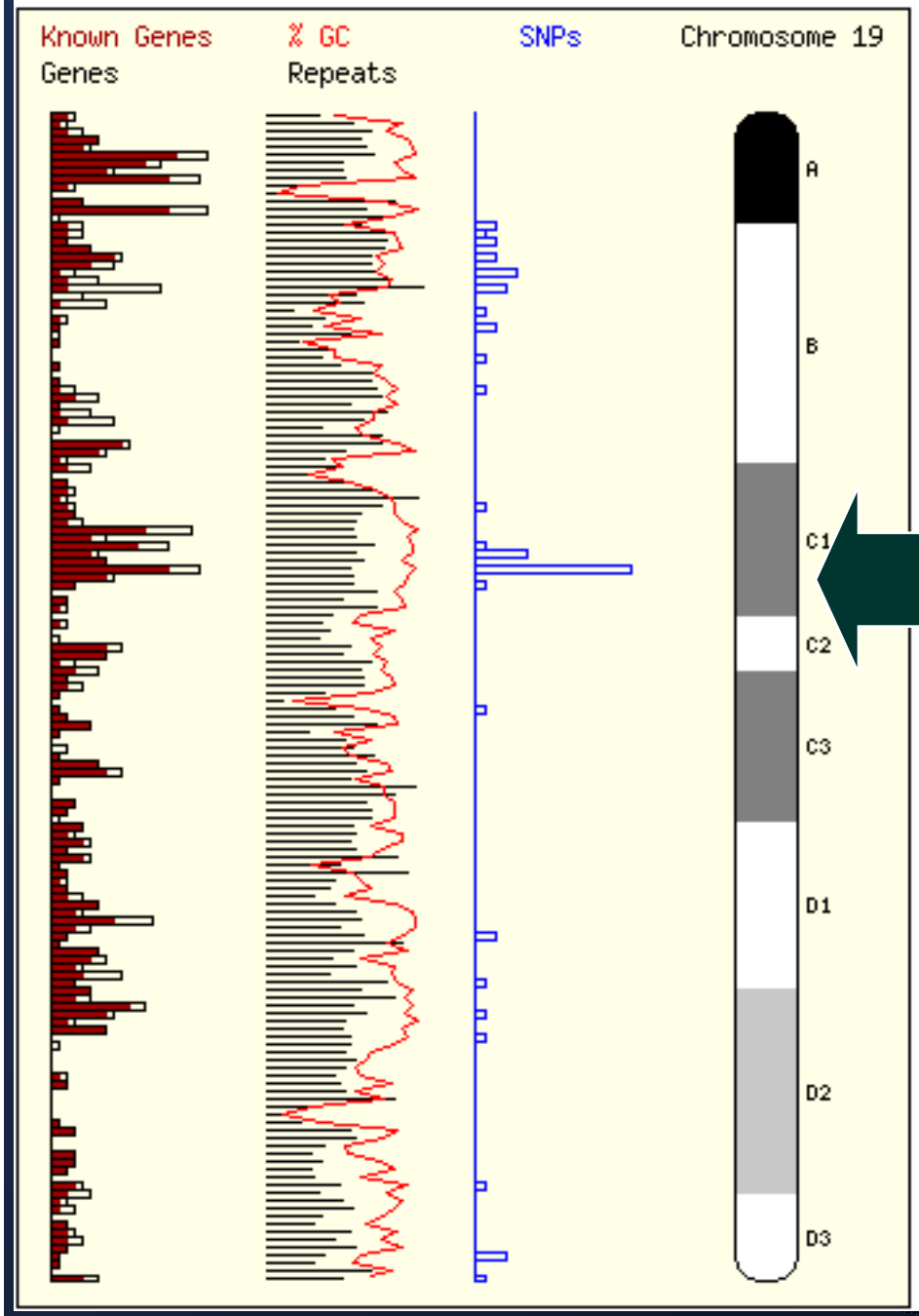
This site presents v1 of the mouse genome draft sequence, based on pure whole genome shotgun data of ~4x coverage, frozen in October 2001. The WGS assembly was aligned to the joint Sanger Institute/WUGSC BAC map (frozen in Sept 2001 - details) to provide this assembly. Clone-based sequencing has not been incorporated. The assembly was run through the normal Ensembl pipeline to predict genes and other features of interest. (statistics)

Other Species

Fly Zebrafish Mosquito Human

Browse a Chromosome





Chromosome 19

Known Ensembl Genes: 400 **SNPs:** 53
Novel Ensembl Genes: 165 **Length:** 65625882 bp

Change Chromosome

Chromosome: 19 Lookup

Jump to Contigview

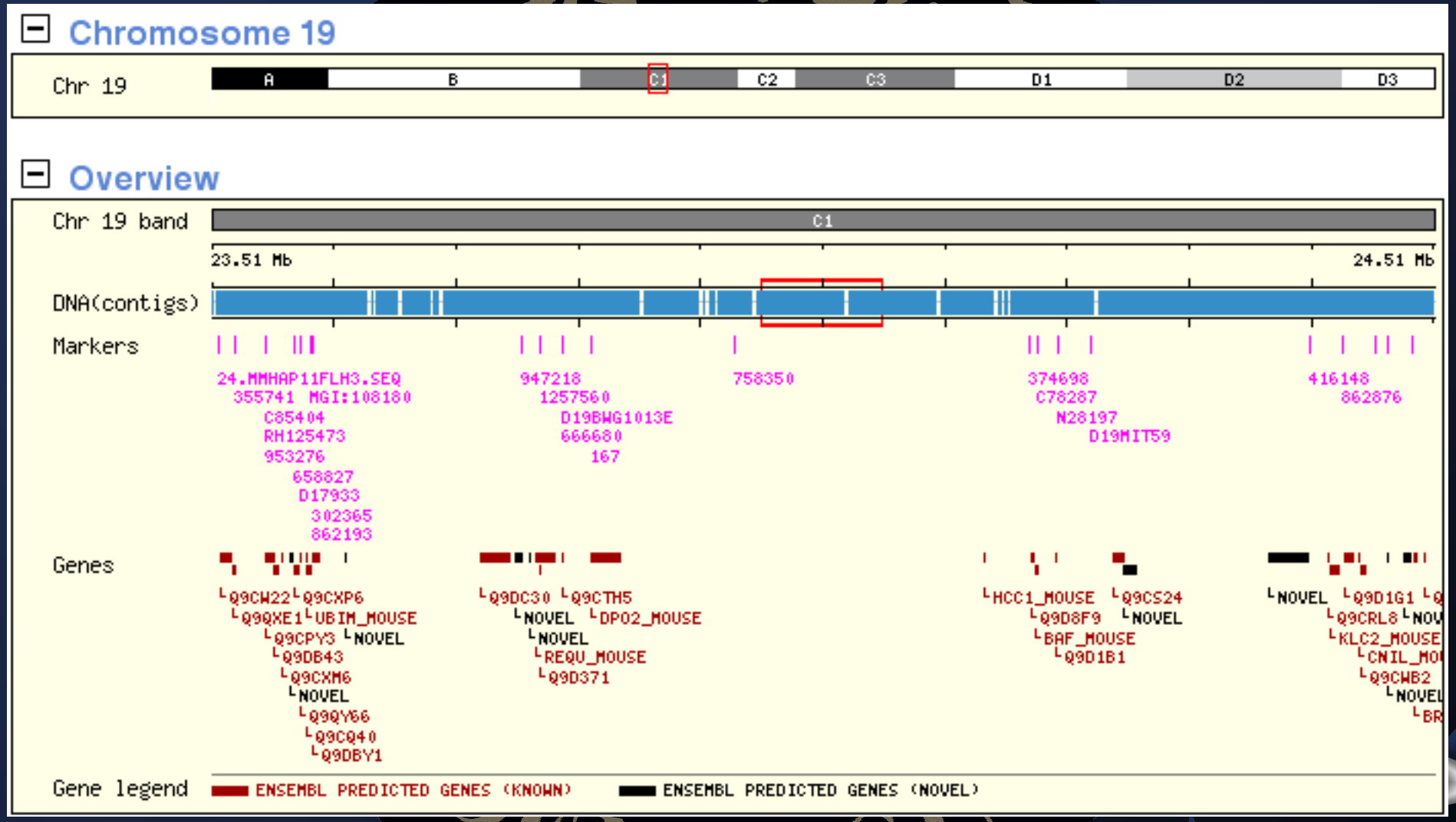
Click anywhere on the chromosome ideogram or one of the feature distribution plots to jump to a contig-level view of features at that point. Alternatively, you can jump to contigview between any two markers on this chromosome:

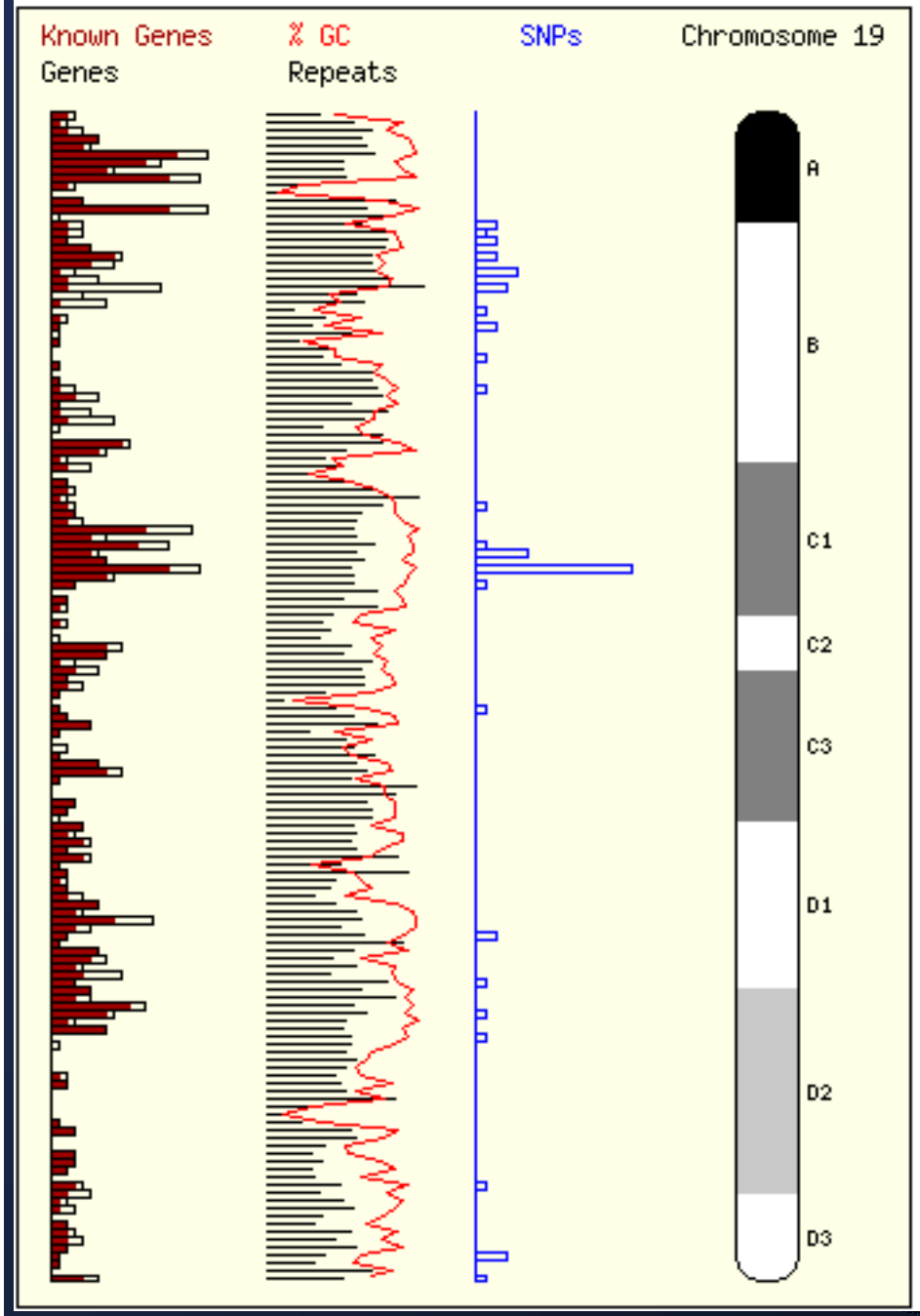
Between:
and:
Lookup

Display contig-level view between any two features.

Synteny

View Mouse Chr 19 vs Human Lookup


Chromosome 19
Chr 19
A
B
C1
C2
C3
D1
D2
D3
Overview
Chr 19 band
C1
23.51 Mb
24.51 Mb
DNA(contigs)
Markers
24.MMHAP11FLH3.SEQ
355741 MGI:108180
C85404
RH125473
953276
658827
D17933
302365
862193
947218
1257560
D19BWG1013E
666680
167
758350
374698
C78287
N28197
D19MIT59
416148
862876
Genes
Q9CW22 Q9CXP6
Q9QXE1 UBIM_MOUSE
Q9CPY3 NOVEL
Q9DB43
Q9CXM6
NOVEL
Q9QY66
Q9CQ40
Q9DBY1
Q9DC30 Q9CTH5
NOVEL DPO2_MOUSE
NOVEL
REQU_MOUSE
Q9D371
HCC1_MOUSE Q9CS24
Q9D8F9 NOVEL
BAF_MOUSE
Q9D1B1
NOVEL Q9D1G1
Q9CRL8
KLC2_MOUSE
Q9CWB2
NOVEL
Gene legend
ENSEMBL PREDICTED GENES (KNOWN)
ENSEMBL PREDICTED GENES (NOVEL)



Chromosome 19

Known Ensembl Genes: 400 **SNPs:** 53
Novel Ensembl Genes: 165 **Length:** 65625882 bp

Change Chromosome

Chromosome: 19 Lookup

Jump to Contigview

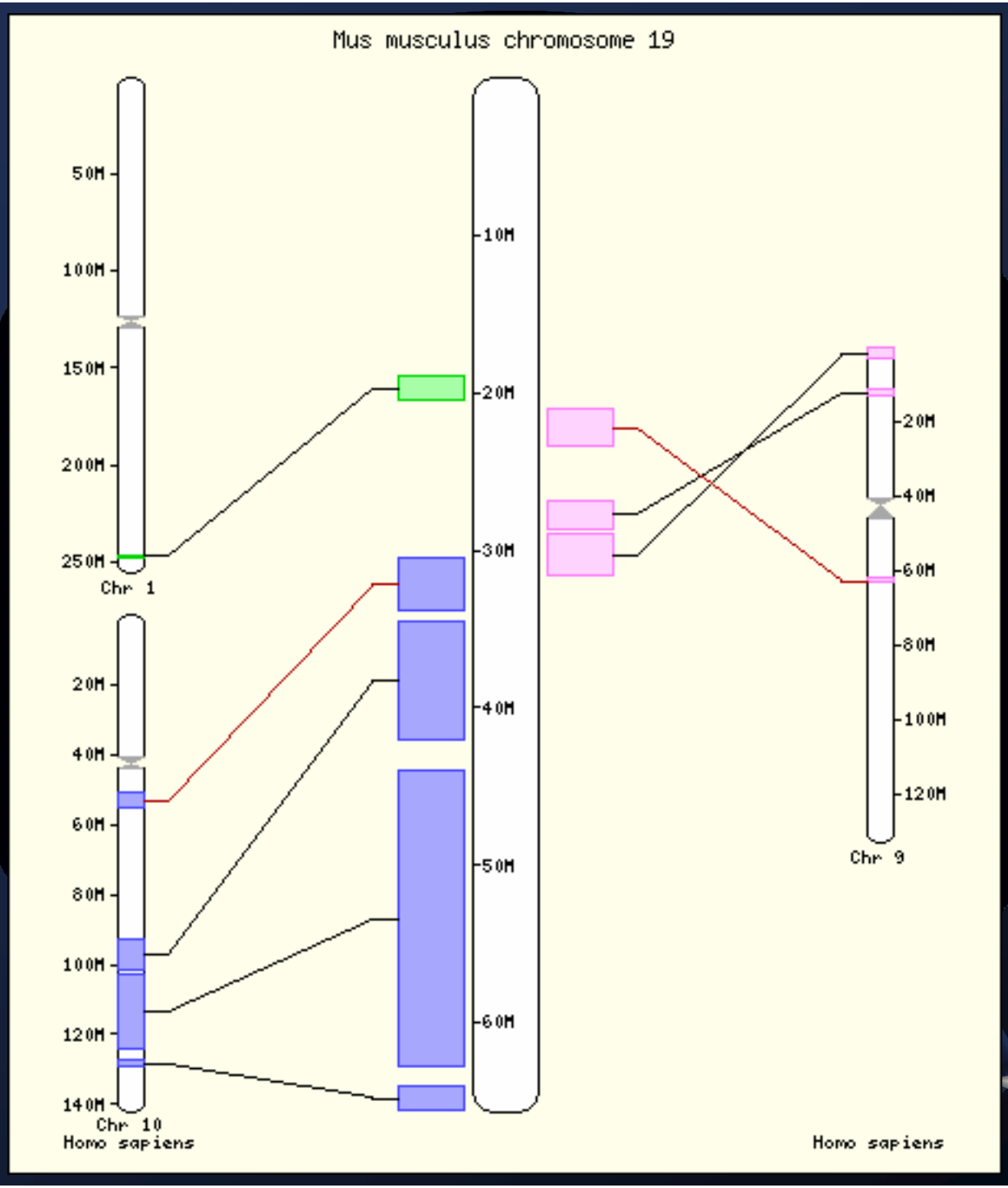
Click anywhere on the chromosome ideogram or one of the feature distribution plots to jump to a contig-level view of features at that point. Alternatively, you can jump to contigview between any two markers on this chromosome:

Between:
and: Lookup

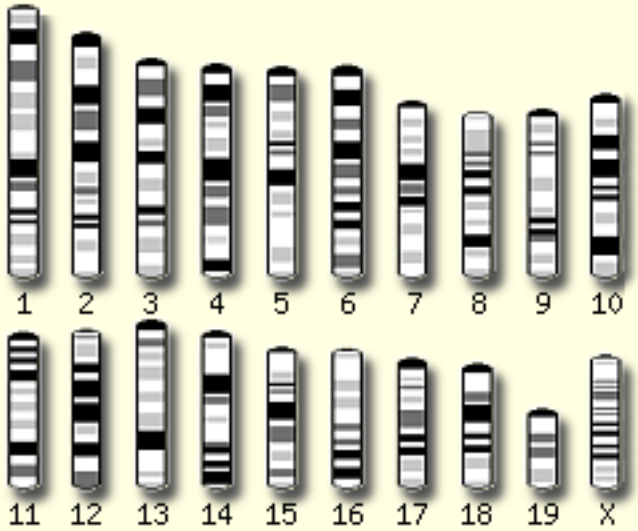
Display contig-level view between any two features.

Synteny

View Mouse Chr 19 vs Human Lookup


Mus musculus chromosome 19
50M
100M
150M
200M
250M
Chr 1
20M
40M
60M
80M
100M
120M
140M
Chr 10
Homo sapiens
10M
20M
30M
40M
50M
60M
20M
40M
60M
80M
100M
120M
Chr 9
Homo sapiens

Browse a Chromosome

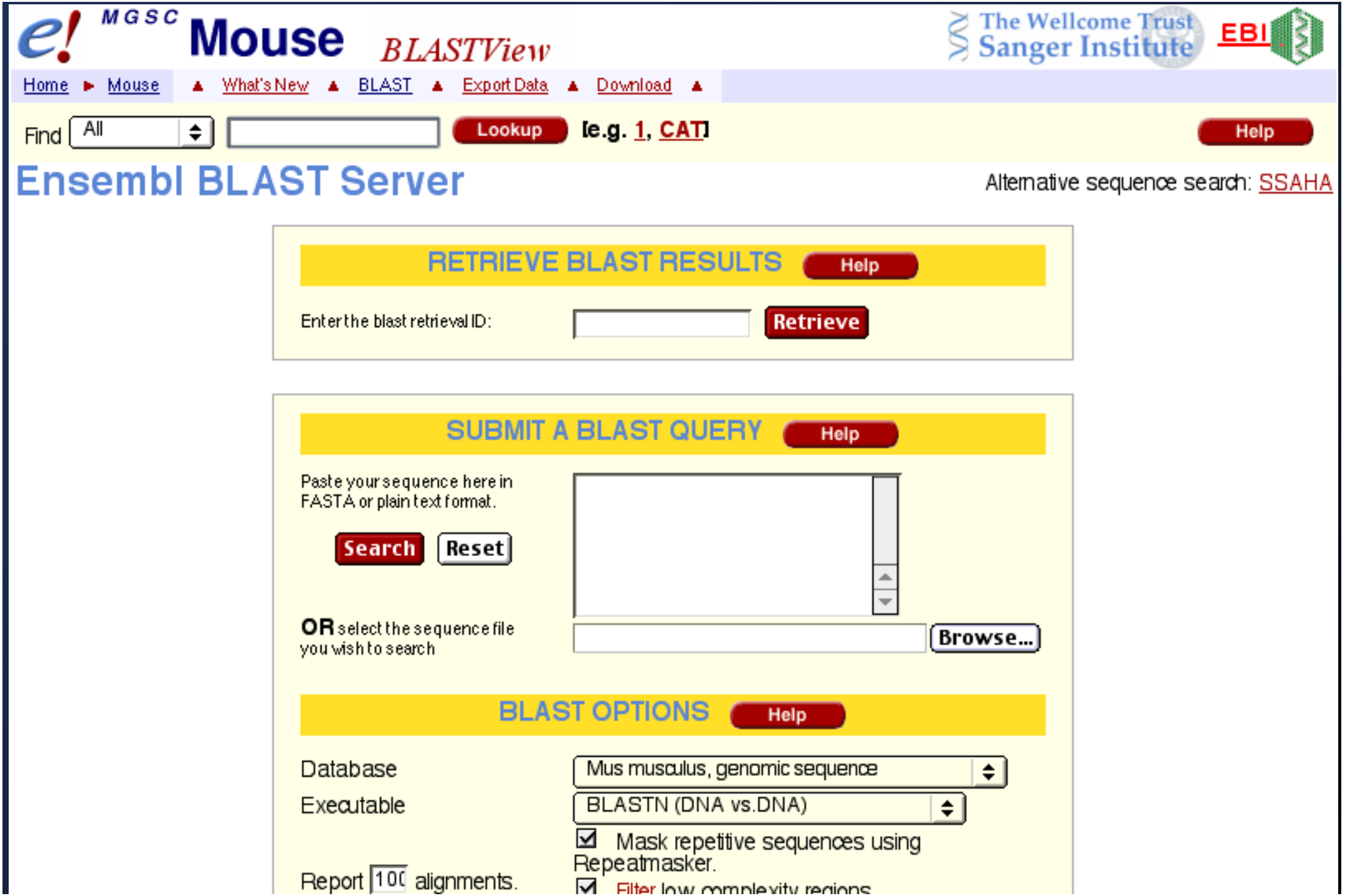


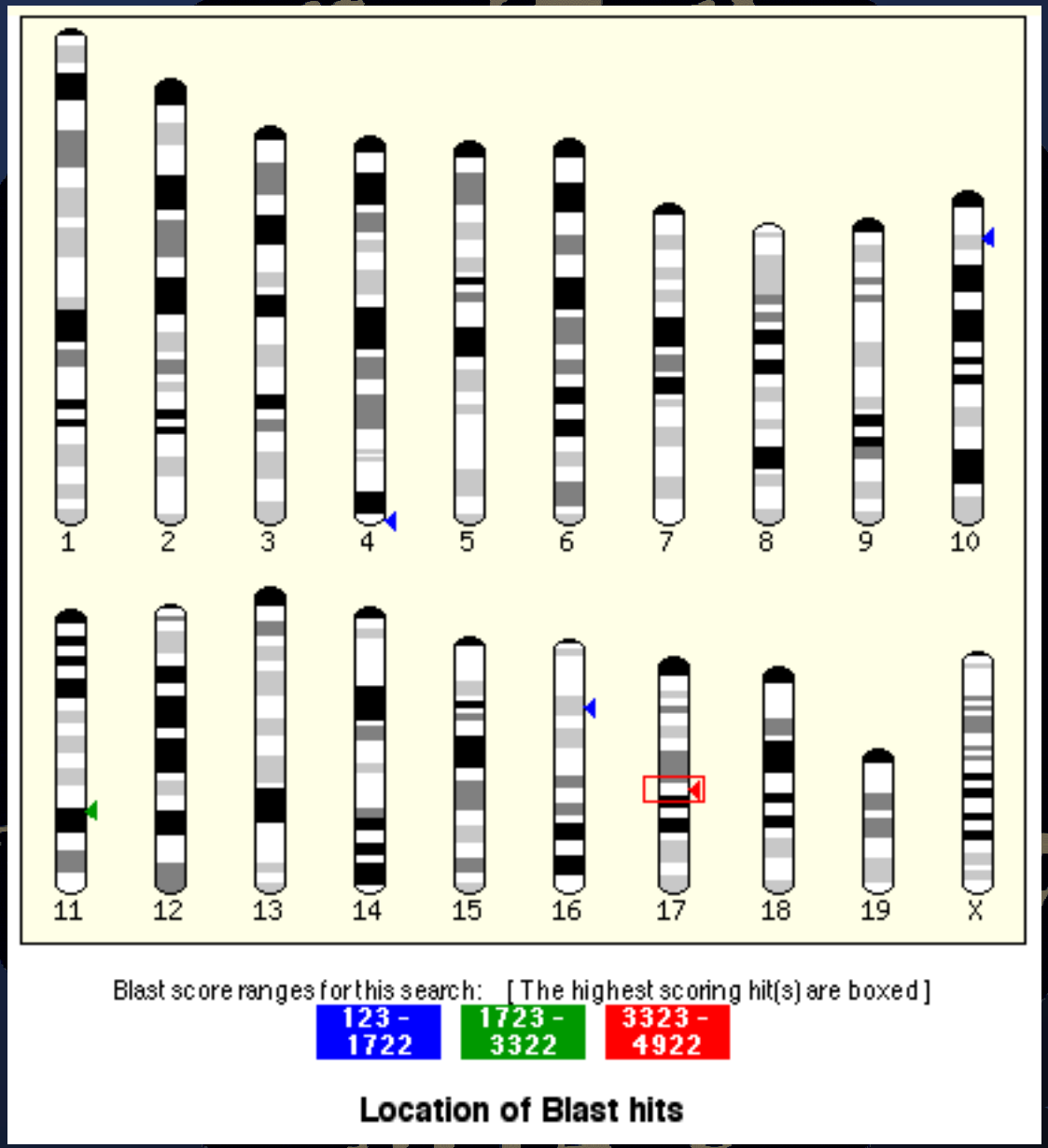
Ensembl Links and Site Map

- What's New
- Blast
- SSAHA
- Download
- Export Data



Site Map


MGSC Mouse BLASTView
The Wellcome Trust Sanger Institute
EBI
Home ▶ Mouse ▲ What's New ▲ BLAST ▲ Export Data ▲ Download ▲
Find All Lookup [e.g. 1, CAT] Help
Ensembl BLAST Server
Alternative sequence search: SSAHA
RETRIEVE BLAST RESULTS Help
Enter the blast retrieval ID: Retrieve
SUBMIT A BLAST QUERY Help
Paste your sequence here in FASTA or plain text format.
Search Reset
OR select the sequence file you wish to search
Browse...
BLAST OPTIONS Help
Database Mus musculus, genomic sequence
Executable BLASTN (DNA vs.DNA)
Mask repetitive sequences using Repeatmasker.
Report 100 alignments.
Filter low complexity regions


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
X
Blast score ranges for this search: [The highest scoring hit(s) are boxed]
123 - 1722
1723 - 3322
3323 - 4922
Location of Blast hits

View standard format BLAST results

```
  BLASTN 2.0a13MP-WashU [10-Jun-1997] [Build 23:08:16 Jun 10 1997]
Query= /
                                        (1774 letters)
Database: ensembl/Mus_musculus.latestgp.fa
                          677 sequences; 3320754623 total letters
```

Sequences producing significant alignments:		High Score	Smallest Sum Probability P(N)	N
11	11.90000001-95000000	2628	0.0	8
17	17.55000001-60000000	4922	3.4e-274	2
16	16.30000001-35000000	644	6.0e-38	3
10	10.20000001-25000000	594	4.4e-14	2
4	4.170000001-175000000	338	1.6e-10	2

```
BLASTN 2.0a13MP-WashU [10-Jun-1997] [Build 23:08:16 Jun 10 1997]
Query= /
          (1774 letters)
Database: ensembl/Mus_musculus.latestgp.fa
          677 sequences; 3320754623 total letters

>17.55000001-60000000
                      Length: 5,000,000

  Plus Strand HSPs:

 Score = 4922 (738.5 bits), Expect = 3.4e-274, Sum P(2) = 3.4e-274
 Identities = 1054/1114 (95%), Positives = 1054/1114 (95%), Strand = Plus / Plus

Query: 80      TACAGTTAGGGGGCACCTAGCATTCAGGCCCTCATCCTCCTCCTTCCCAGCAGGGTGTCA 139
                TA AGTTAGGGGGCACCTAGCATTCAGGCCCTCATCCTCCTCCTTCCCAGCAGGGTGTCA
Sbjct: 4807867 TAGAGTTAGGGGGCACCTAGCATTCAGGCCCTCATCCTCCTCCTTCCCAGCAGGGTGTCA 4807926

Query: 140     CGCTTCTCCGAAGACTGGNNATGACTGCCATGGAGGAGTCACAGTCGGATATCAGCCTCG 199
                CGCTTC CCGAAGACTGG  ATGACTGCCATGGAGGAGTCACAGTCGGATATCAGCCTCG
Sbjct: 4807927 CGCTTC-CCGAAGACTGG--ATGACTGCCATGGAGGAGTCACAGTCGGATATCAGCCTCG 4807983

Query: 200     AGCTCCCTCTGAGCCAGGAGACATTTTCAGGCTTATGGAAACTACTTCCTCCAGAAGATA 259
                AGCTCCCTCTGAG CAGGAGACATTTTCAGGCTTATGGAAACTACTTCCTCCAGAAGATA
Sbjct: 4807984 AGCTCCCTCTGAGTCAGGAGACATTTTCAGGCTTATGGAAACTACTTCCTCCAGAAGATA 4808046

Query: 260     TCCTGCCATCACCTCACTGCATGGACGATCTGTTGCTGCCCCAGGATGTTGAGGAGTTTT 319
                TCCTGCCATCACCTCACT CATGGACGATCTGTTGCTGCCCCAGGATGTTGAGGAGTTTT
Sbjct: 4808047 TCCTGCCATCACCTCACTCCATGGACGATCTGTTGCTGCCCCAGGATGTTGAGGAGTTTT 4808106

Query: 320     TTGAAGGCCCAAGTGAAGCCCTCCGAGTGTCAGGAGCTCCTGCAGCACAGGACCCTGTCA 379
                TTGAAGGCCCAAGTGAAGCCCTCC AGTGT AGGAGCTCCTGCAGCACAGGACCCTGTCA
Sbjct: 4808107 TTGAAGGCCCAAGTGAAGCCCTCCAAGTGTTAGGAGCTCCTGCAGCACAGGACCCTGTCA 4808166

Query: 380     CCGAGACCCCTGGGCCAGTGGCCCCTGCCCCAGCCACTCCATGGCCCCTGTCATCTTTTG 439
                C GAGACCCCTGGGCCAGTGGCCCCTGCCCCAGC ACTCCATGGCCCC GTCATCTTTTG
Sbjct: 4808167 CTGAGACCCCTGGGCCAGTGGCCCCTGCCCCAGCTACTCCATGGCCCCCGTCATCTTTTG 4808226

Query: 440     TCCCTTCTCAAAAAACTTACCAGGGCAACTATGGCTTCCACCTGGGCTTCCTGCAGTCTG 499
                TCCCTTCTCAAAAAACTTACCAGGGCAACTATGGCTTCCACCTGGGCTTCCTGCAGTCTG
Sbjct: 4808227 TCCCTTCTCAAAAAACTTACCAGGGCAACTATGGCTTCCACCTGGGCTTCCTGCAGTCTG 4808286
```

```
BLASTN 2.0a13MP-WashU [10-Jun-1997] [Build 23:08:16 Jun 10 1997]

Reference:  Gish, Warren (1994-1997).  unpublished.
Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J.
Lipman (1990).  Basic local alignment search tool.  J. Mol. Biol. 215:403-10.

Notice:  this program and its default parameter settings are optimized to find
nearly identical sequences rapidly.  To identify weak similarities encoded in
nucleic acid, use BLASTX, TBLASTN or TBLASTX.

Query=  /
        (1774 letters)

Database: Mus_musculus.latestgp.fa
           677 sequences; 3,320,754,623 total letters.
Searching....10....20....30....40....50....60....70....80....90....100% done
```

Sequences producing High-scoring Segment Pairs:	High Score	Smallest Sum Probability P(N)	N
11.90000001-95000000	2628	0.0	8
17.55000001-60000000	4922	3.4e-274	2
16.30000001-35000000	644	6.0e-38	3
10.20000001-25000000	594	4.4e-14	2
4.170000001-175000000	338	1.6e-10	2

```
>11.90000001-95000000
            Length = 5,000,000

  Plus Strand HSPs:

 Score = 2628 (394.3 bits), Expect = 0.0, Sum P(8) = 0.0
 Identities = 530/536 (98%), Positives = 530/536 (98%), Strand = Plus / Plus

Query:   1239 GGCTCACTCCAGCTACCTGAAGACCAAGAAGGGCCAGTCTACTTCCCGCCATAAAAAAAC 1298
              | ||  ||  ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 619168 GTCTTCCTATAGCTACCTGAAGACCAAGAAGGGCCAGTCTACTTCCCGCCATAAAAAAAC 619227
```

Ensembl SSAHA Search Server

Paste your **DNA** sequence(s) here in FASTA or plain text format.

Search Reset

OR select the sequence file you wish to search

Browse...

SSAHA OPTIONS Help

Choose no. bp to display either side of SSAHA matches in contig displays (**human golden path data searches only**).

50Kbp (SSAHA context)

Select the data source for the SSAHA search.

Ensembl Current (based on MGSC v1)

Only display matches containing at least

28bp matching bases.

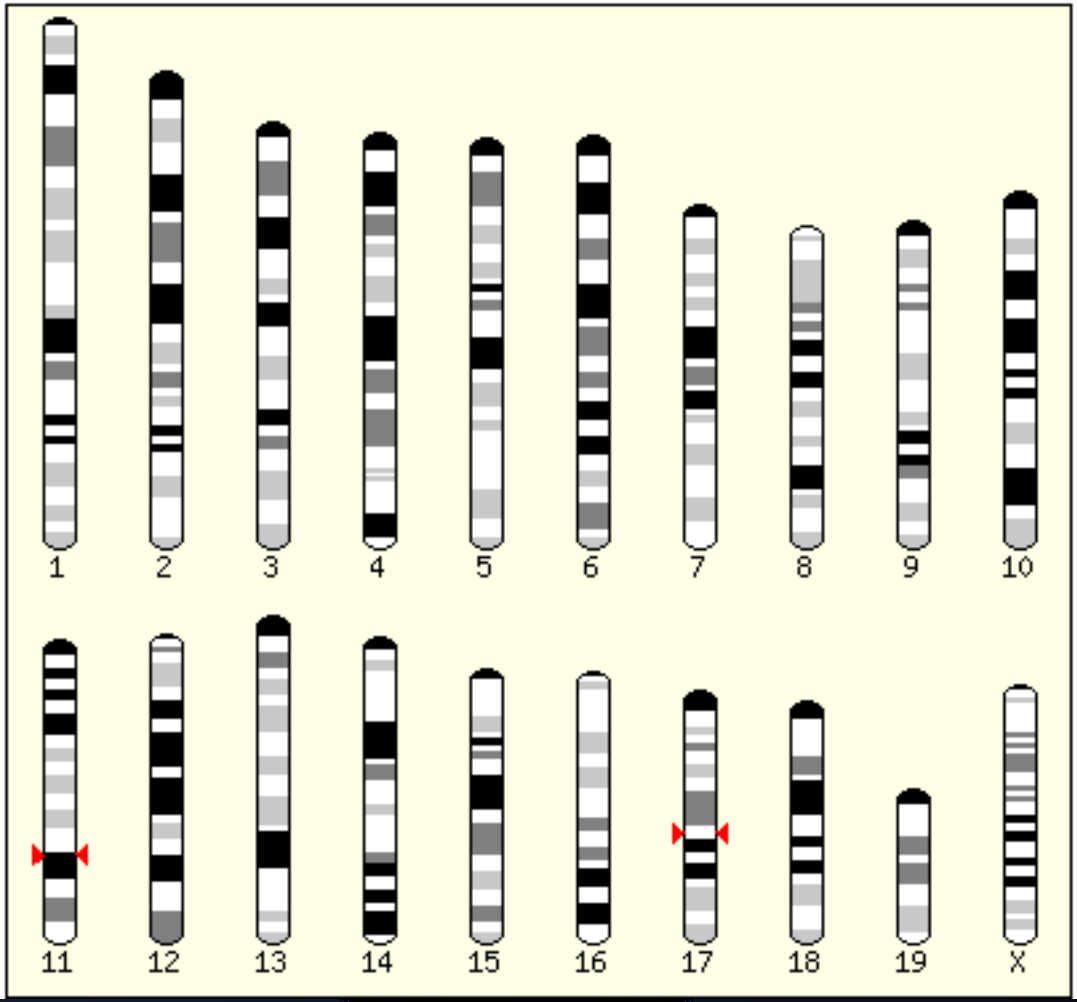
For each query sequence, display

top 50 matches.

ADVANCED SSAHA OPTIONS Help

Ensembl SSAHA Server: Ensembl Current (based on MGSC v1)

- Found a total of 24 matches among 2 sequences in the database.
- For each query sequence, will show the top 50 matches containing 28 or more matching bases.
- Matches may contain gaps of up to 14 bases in length and up to 1 insertion/deletion between successive hits.
- Query data was scanned for tandem repeating motifs of length 1 or less.
- Algorithm has ignored any words in the query that occur in the database with a frequency of 10000 times more than the expected value for a random database of the same size.



Location of SSAHA hits

Result #	Chromosome	Query Details Name	Start	End	Subject Details Name	Start	End	Direction	Match Details # bases	% identity	
1	11	ENSMUST00000005371	**1250**	**1767**	11.90000001-95000000	**619179**	**619696**	FF	518	100.00	[align]
2	11	ENSMUST00000005371	**266**	**517**	11.90000001-95000000	**615203**	**615454**	FF	252	100.00	[align]
3	11	ENSMUST00000005371	**529**	**710**	11.90000001-95000000	**616197**	**616378**	FF	182	100.00	[align]
4	11	ENSMUST00000005371	**9**	**148**	11.90000001-95000000	**608175**	**608314**	FF	140	100.00	[align]
5	17	ENSMUST00000005371	**167**	**334**	17.55000001-60000000	**4807951**	**4808118**	FF	140	83.33	[align]
6	11	ENSMUST00000005371	**931**	**1056**	11.90000001-95000000	**617401**	**617526**	FF	126	100.00	[align]
7	11	ENSMUST00000005371	**717**	**814**	11.90000001-95000000	**616463**	**616560**	FF	98	100.00	[align]
8	11	ENSMUST00000005371	**833**	**930**	11.90000001-95000000	**616981**	**617078**	FF	98	100.00	[align]
9	11	ENSMUST00000005371	**1151**	**1248**	11.90000001-95000000	**618493**	**618590**	FF	98	100.00	[align]
10	17	ENSMUST00000005371	**536**	**633**	17.55000001-60000000	**4808315**	**4808412**	FF	84	85.71	[align]

<Back to top>

```
Result #              : 1
Query name            : ENSMUST00000005371
Start in query        : 1250
End in query          : 1767
Subject Name          : 11.90000001-95000000
Start in subject      : 619179
End in subject        : 619696
Direction             : FF
Number of bases       : 518
Percentage identity: 100.00

alignment score: 517
Q:000001250 GCTACCTGAAGACCAAGAAGGGCCAGTCTACTTCCCGCCATAAAAAAACAATGGTCAAGAAAGTGGGG
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
S:000619179 GCTACCTGAAGACCAAGAAGGGCCAGTCTACTTCCCGCCATAAAAAAACAATGGTCAAGAAAGTGGGG

Q:000001318 CCTGACTCAGACTGACTGCCTNTGCATCCCGTCCCCATCACCAGCCTCCCCCTCTCCTTGCTGTTTTA
            ||||||||||||||||||||x||||||||||||||||||||||||||||||||||||||||||||||
S:000619247 CCTGACTCAGACTGACTGCCTNTGCATCCCGTCCCCATCACCAGCCTCCCCCTCTCCTTGCTGTTTTA

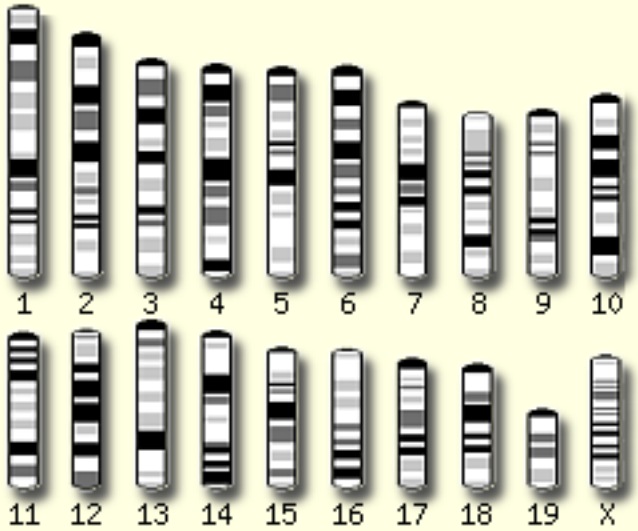
Q:000001386 TGACTTCAGGGCTGAGACACAATCCTCCCGGTCCCTTCTGCTGCCTTTTTTACCTTGTAGCTAGGGCT
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
S:000619315 TGACTTCAGGGCTGAGACACAATCCTCCCGGTCCCTTCTGCTGCCTTTTTTACCTTGTAGCTAGGGCT

Q:000001454 CAGCCCCCTCTCTGAGTAGTGGTTCCTGGCCCAAGTTGGGGAATAGGTTGATAGTTGTCAGGTCTCTG
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
S:000619383 CAGCCCCCTCTCTGAGTAGTGGTTCCTGGCCCAAGTTGGGGAATAGGTTGATAGTTGTCAGGTCTCTG

Q:000001522 CTGGCCCAGCGAAATTCTATCCAGCCAGTTGTTGGACCCTGGCACCTACAATGAAATCTCACCCTACC
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
S:000619451 CTGGCCCAGCGAAATTCTATCCAGCCAGTTGTTGGACCCTGGCACCTACAATGAAATCTCACCCTACC

Q:000001590 CCACACCCTGTAAGATTCTATCTTGGGCCCTCATAGGGTCCATATCCTCCAGGGCCTACTTTCCTTCC
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
S:000619519 CCACACCCTGTAAGATTCTATCTTGGGCCCTCATAGGGTCCATATCCTCCAGGGCCTACTTTCCTTCC
```

Browse a Chromosome



Ensembl Links and Site Map

- What's New
- Blast
- SSAHA
- Download
- Export Data



Site Map

Ensembl Mouse Download and Documentation

Ensembl Mouse Downloads

Overview

Before downloading and installing the Ensembl data and software we strongly encourage you to first read the documentation.

- ensembl_tutorial.pdf
- Ensembl Website Install instructions (pdf)

The software which provides Perl-based Object access to the database is available via anonymous CVS (see website install docs for details)

- WebCVS

Ensembl Mouse Data Information

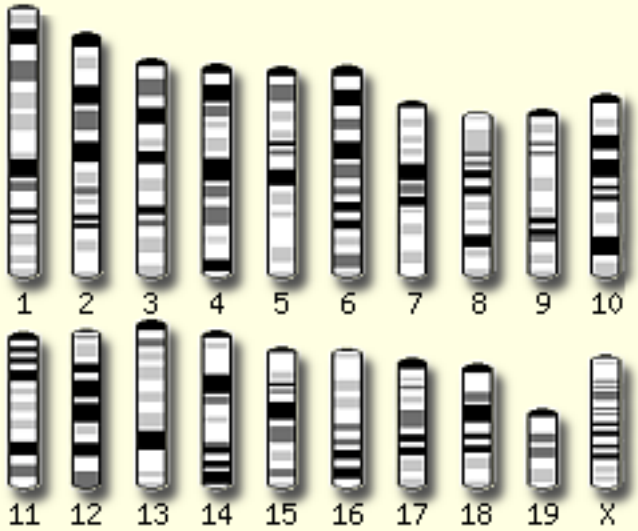
The first MGSC assembly release of the Ensembl data for the mouse genome is ready. This release is based on pure whole genome shotgun data of ~4x coverage, frozen in October 2001. The WGS assembly was aligned to the joint Sanger/St Louis BAC map (frozen in Sept 2001 - details) to provide this assembly. BAC-based sequencing has not been incorporated. The assembly was run through the normal Ensembl pipeline to predict genes and other features of interest. (statistics)
All the distributed gene predictions are confirmed by similarity to either proteins or cDNA sequences.

Mouse Data files are available from:

ftp://ftp.ensembl.org/pub/current_mouse/data/

- fasta/ - gzipped fasta format cDNA, DNA, and peptide dumps
- flatfiles/ - gzipped EMBL and Genbank dumps of Ensembl data
- mysql/ - SQL dumps of Ensembl databases, from mysql, (should be portable to any SQL database)

Browse a Chromosome



Ensembl Links and Site Map

- What's New
- Blast
- SSAHA
- Download
- Export Data



Site Map

Flat File | FASTA | **Gene List** | Feature List | SNP List | Image

Select genes to export

Region

Chromosome: 1

- ○ Bands from: --- to ---
- ○ Contigs from: ____ to ____
- ○ Markers from: ____ to ____
- ◉ Bases from: ____ to ____
- ○ Entire Chromosome
- ○ Entire Genome

Restrict selection

Include:	○ Known genes only	○ Exclude Known genes	◉ Both
Include:	○ Transmembrane domains only	○ Exclude Transmembrane domains	◉ Both
Include:	○ Low-complexity domains only	○ Exclude Low-complexity domains	◉ Both
Include:	○ Signal domains only	○ Exclude Signal domains	◉ Both

Include only those genes with Family ID: ____

Include Gene data

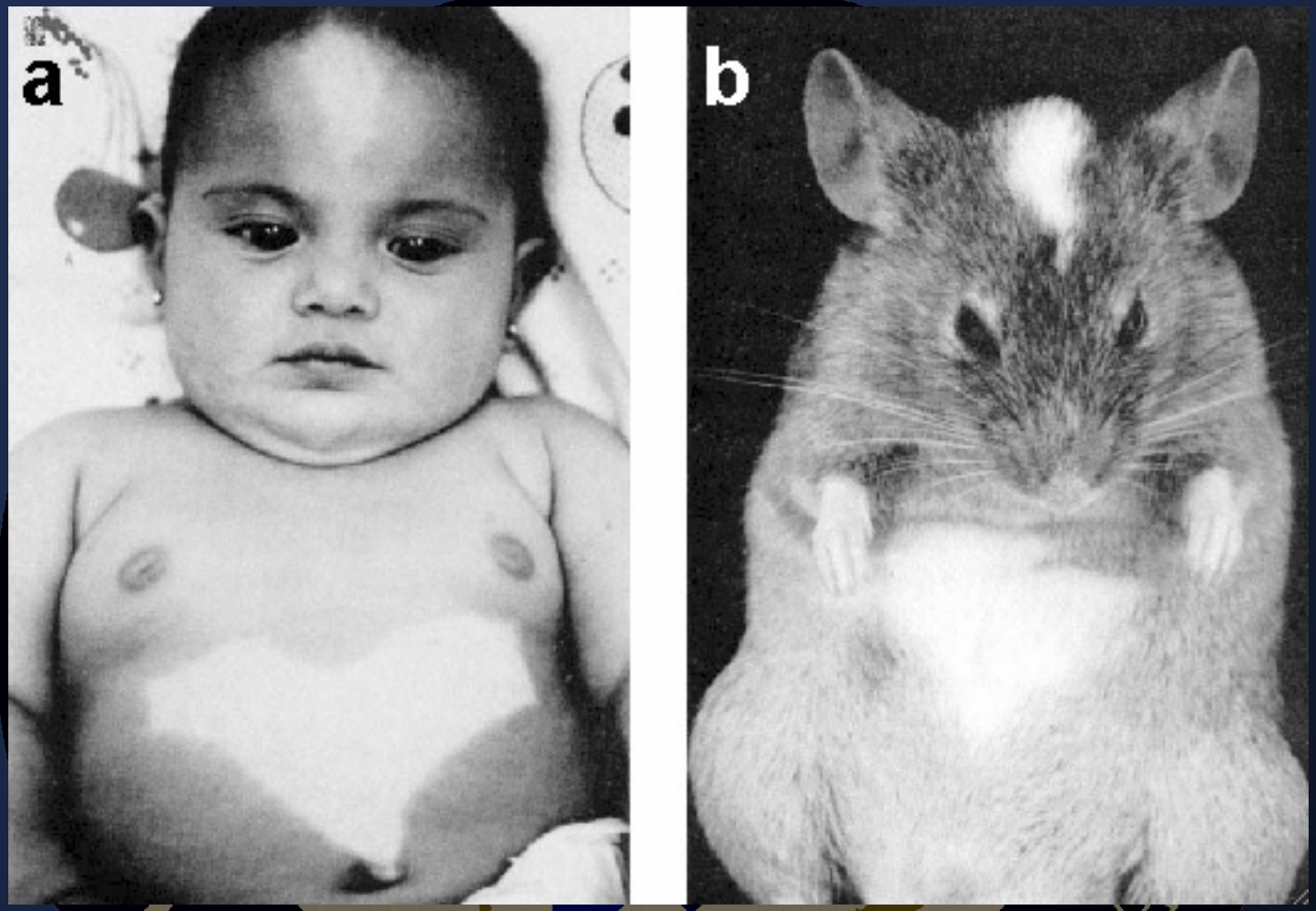
Include the following data

- ☐ Chromosome
- ☐ Chromosome start
- ☐ Chromosome end
- ☐ Chromosome band
- ☐ Chromosome_strand
- ☑ Ensembl gene ID
- ☐ Known gene external id

Nature **420**, 520 - 562 (2002) December 5 2002

Initial sequencing and comparative analysis of the mouse genome

Mouse Genome Sequencing Consortium



In the roughly 75 million years since the divergence of the human and mouse lineages, the process of evolution has altered their genome sequences and caused them to diverge by nearly one substitution for every two nucleotides (see below) as well as by deletion and insertion.

Distribution of quality scores in the draft mouse genome sequence

Quality score	Bases in draft genome sequence with this score (%)	Disagreement with finished sequence (%) *
0–9	0.3	9.800
10–19	0.6	2.500
20–29	1.0	0.280
30–39	2.3	0.045
>40	95.8	0.005

The draft sequence was generated by assembling about sevenfold sequence coverage from female mice of the C57BL/6J strain (referred to below as B6). The assembly contains about 96% of the sequence of the euchromatic genome (excluding chromosome Y) in sequence contigs linked together into large units, usually larger than 50 megabases (Mb).

Table 1 **Distribution of sequence reads**

Insert size (kb)*	Vector	Reads (millions)				Bases† (billions)		Sequence coverage‡		Physical coverage§
		All	Used	Paired	Assembled	Total	>Phred20	Total	>Phred20	
2	Plasmid	3.8	3.7	3.1	2.9	1.8	1.5	0.71	0.61	1.2
4	Plasmid	31.3	24.7	22.1	21.5	14.7	12.6	5.89	5.03	17.7
6	Plasmid	1.2	1.0	0.8	0.8	0.5	0.5	0.22	0.19	1.0
10	Plasmid	2.5	2.4	2.1	1.7	1.3	1.0	0.52	0.42	4.3
40	Fosmid	2.1	1.3	1.2	1.1	0.6	0.5	0.26	0.21	9.3
150–200	BAC	0.4	0.4	0.4	0.4	0.2	0.2	0.09	0.07	13.7
Other‖	Plasmid	0.07	0.05	0.03	0.04	0.03	0.03	0.01	0.01	0.02
Total		41.4	33.6	29.7	28.4	19.2	16.3	7.68	6.53	47.2

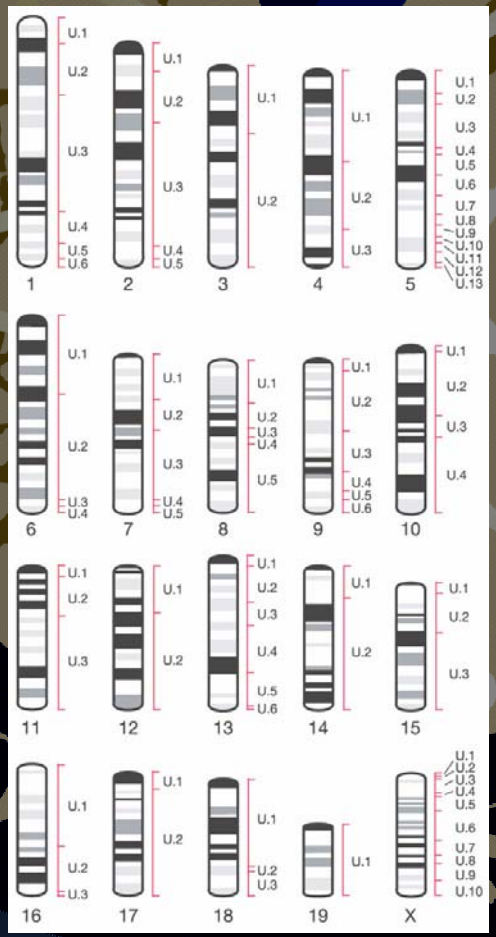
Centre	Reads (millions)				Bases† (billions)		Sequence coverage‡		Physical coverage§
	All	Used	Paired	Assembled	Total	>Phred20	Total	>Phred20	
Whitehead Institute	22.2	18.0	15.9	15.7	10.7	9.2	4.28	3.68	21.3
Washington University	11.5	8.3	7.5	7.1	4.7	3.9	1.87	1.57	5.9
Sanger Institute	6.7	6.3	5.4	4.7	3.3	2.7	1.31	1.09	5.3
University of Utah	0.6	0.6	0.5	0.5	0.3	0.3	0.13	0.11	1.0
The Institute for Genomic Research	0.5	0.4	0.4	0.4	0.2	0.2	0.09	0.08	13.7
Total	41.4	33.6	29.7	28.4	19.2	16.3	7.68	6.53	47.2

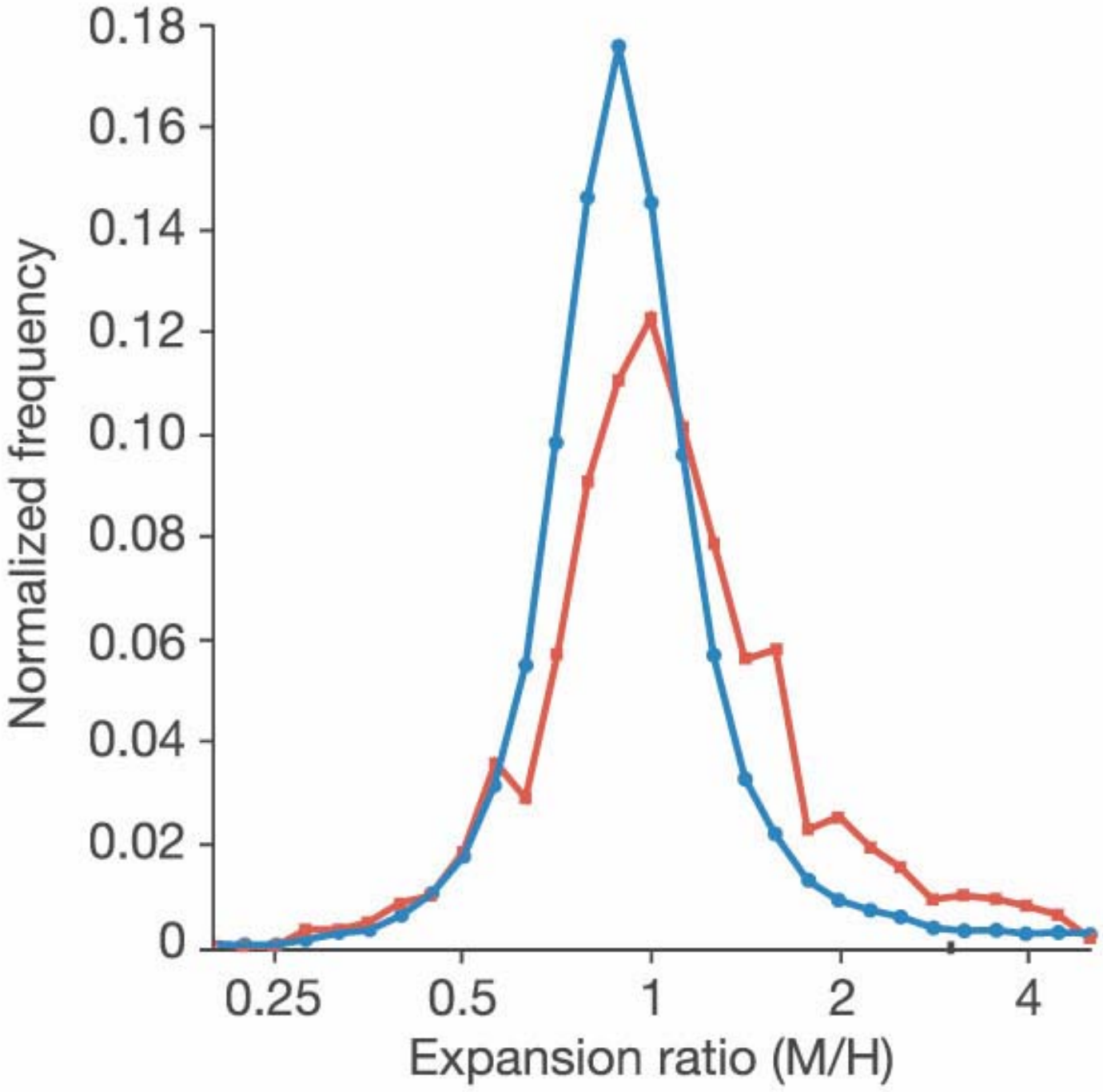
Table 2 **Basic statistics of the MGSCv3 assembly**

Features	Number	N50 length (kb)*	Bases (Gb)	Bases plus gaps (Gb)	Percentage of gen
All anchored contigs†	176,471	25.9	2.372	2.372	94.9
All anchored supercontigs	377	18,600	2.372	2.477	99.1
All ultracontigs	88	50,600	2.372	2.493	99.7
Unanchored contigs‡	48,242	2.3	0.106	0.106	–
Largest 200 supercontigs	200	18,700	2.352	2.455	98.2
Largest 100 supercontigs	100	22,900	1.955	2.039	81.6

*Not including gaps.

a total of 88 ultracontigs with an N50 length of 50.6 Mb (exclusive of gaps) contained 95.7% of the assembled sequence





Size ratio of mouse to human for orthologous 100-kb windows

The mouse genome is about 14% smaller than the human genome (2.5 Gb compared with 2.9 Gb). The difference probably reflects a higher rate of deletion in the mouse lineage.

Table 3 **Mouse chromosome size estimates**

Chromosome	Actual bases in sequence (Mb)	Ultracontigs (Mb)		Gaps within supercontigs		Gaps between supercontigs					Total estimated size (Mb)‡
		Number	N50 size	Number	Mb	Captured by additional read pairs		Captured by fingerprint contigs*		Uncaptured†	
						Number	Mb	Number	Mb	Number	
All	2,372	88	52.7	176,094	104.5	252	14.0	37	2.30	68	2,493
1	183	6	52.7	13,178	7.8	16	1.1	1	0.32	5	192
2	169	5	111.1	12,141	6.5	4	0.1	1	0.20	4	176
3	149	2	108.9	10,630	6.8	17	0.7	3	0.16	1	157
4	140	3	83.1	10,745	6.3	14	0.4	3	0.26	2	147
5	137	13	17.8	11,288	6.7	11	0.5	3	0.11	12	144
6	138	4	91.4	10,021	6.6	19	1.1	2	0.26	3	146
7	122	5	45.1	9,484	5.7	55	3.4	4	0.12	4	131
8	119	5	35.0	9,186	6.1	7	0.2	2	0.12	4	125
9	116	6	26.8	8,479	4.5	6	0.6	1	0.06	5	121
10	121	4	50.4	9,490	5.4	9	0.6	0	0	3	127
11	115	3	80.4	8,681	4.3	2	0.0	1	0.05	2	119
12	105	2	77.4	7,577	4.0	27	1.2	2	0.00	1	110
13	107	6	28.0	7,910	4.7	13	0.8	4	0.19	5	113
14	107	2	93.6	7,605	4.0	10	0.5	2	0.12	1	112
15	96	3	65.3	7,025	4.3	2	0.1	0	0	2	100
16	91	3	62.3	6,695	4.4	1	0.0	0	0	2	95
17	85	2	80.8	6,584	3.7	17	1.2	4	0.19	1	90
18	84	3	73.5	6,192	3.2	2	0.0	0	0	2	87
19	55	1	57.7	3,934	2.4	7	0.6	2	0.12	0	58
X	134	10	19.9	9,249	7.0	13	0.8	2	0.00	9	142

*These gaps had fingerprint contigs spanning them. The size for 18 out of 37 were estimated using conserved synteny to determine the size of the region in the human genome. The remaining gaps were arbitrarily given the average size of the assessed gaps (59 kb), adjusted to reflect the 16% difference in genome size.

†Uncaptured gaps were estimated by mouse–human synteny to have a total size of 5 Mb. However, because some of these gaps are due to repetitive expansions in mouse (absent in human), the actual total for the uncaptured gaps is probably substantially higher. For example, one large uncaptured gap on chromosome 1 (the Sp-100rs region) is roughly 6 Mb (see text).

‡Omitting centromeres and telomeres. These would add, on average, approximately 8 Mb per chromosome, or about 160 Mb to the genome. Also omitting uncaptured gaps between supercontigs.

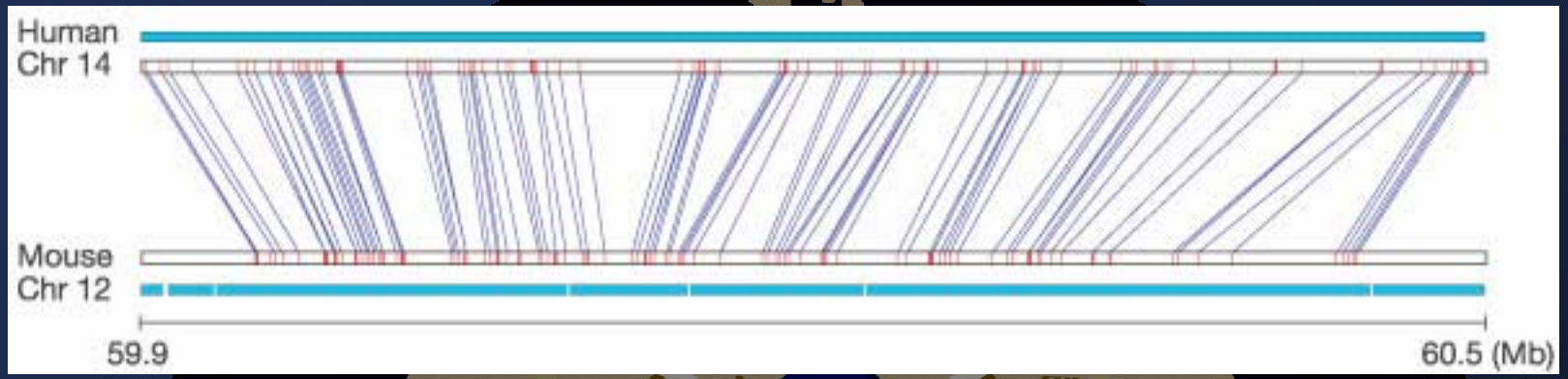
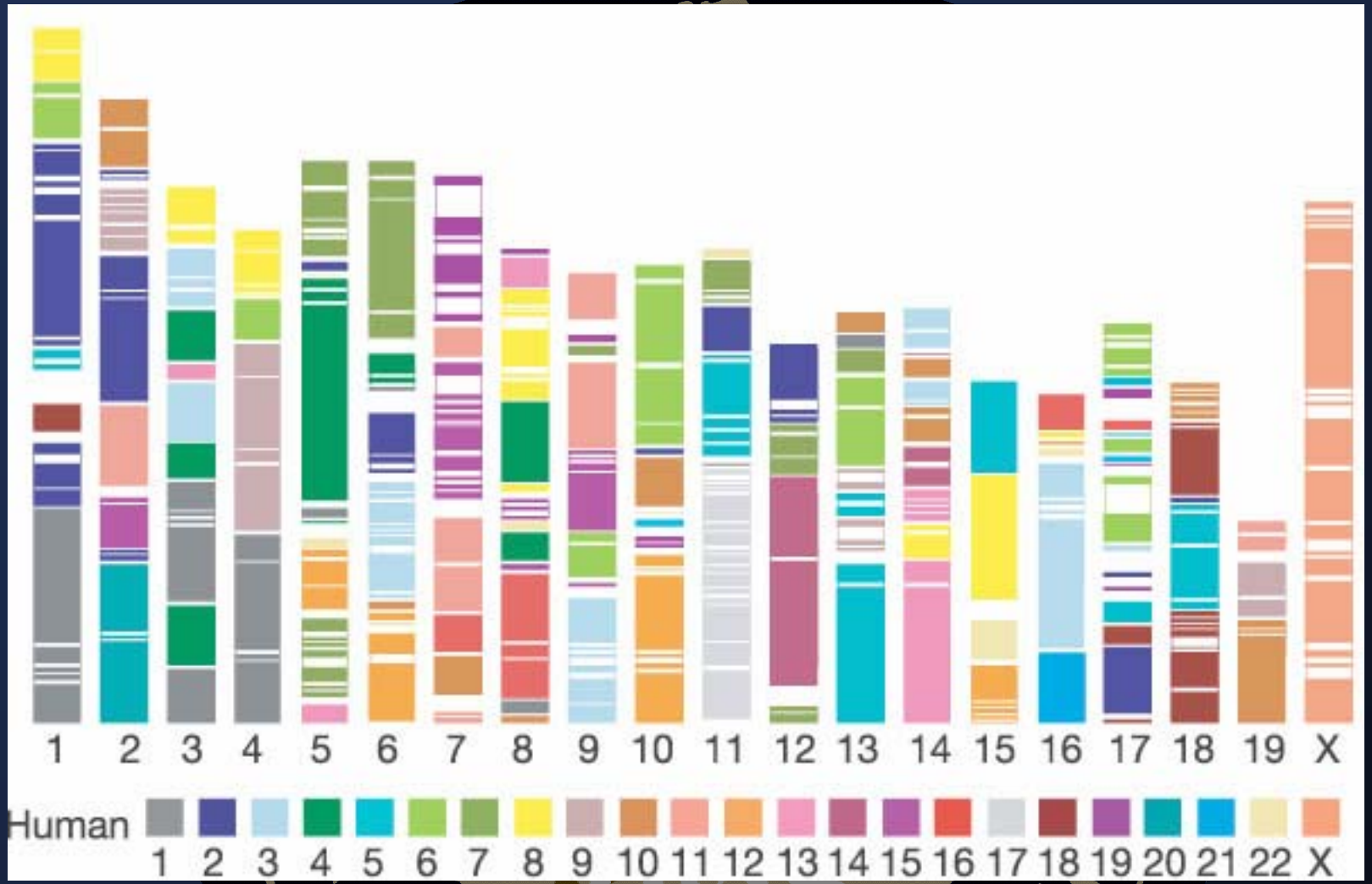


Figure 2 Conservation of synteny between human and mouse. We detected 558,000 highly conserved, reciprocally unique landmarks within the mouse and human genomes, which can be joined into conserved syntenic segments and blocks (defined in text). A typical 510-kb segment of mouse chromosome 12 that shares common ancestry with a 600-kb section of human chromosome 14 is shown. Blue lines connect the reciprocal unique matches in the two genomes. The cyan bars represent sequence coverage in each of the two genomes for the regions. In general, the landmarks in the mouse genome are more closely spaced, reflecting the 14% smaller overall genome size.


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X
Human
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X

Over 90% of the mouse and human genomes can be partitioned into corresponding regions of conserved synteny, reflecting segments in which the gene order in the most recent common ancestor has been conserved in both species.

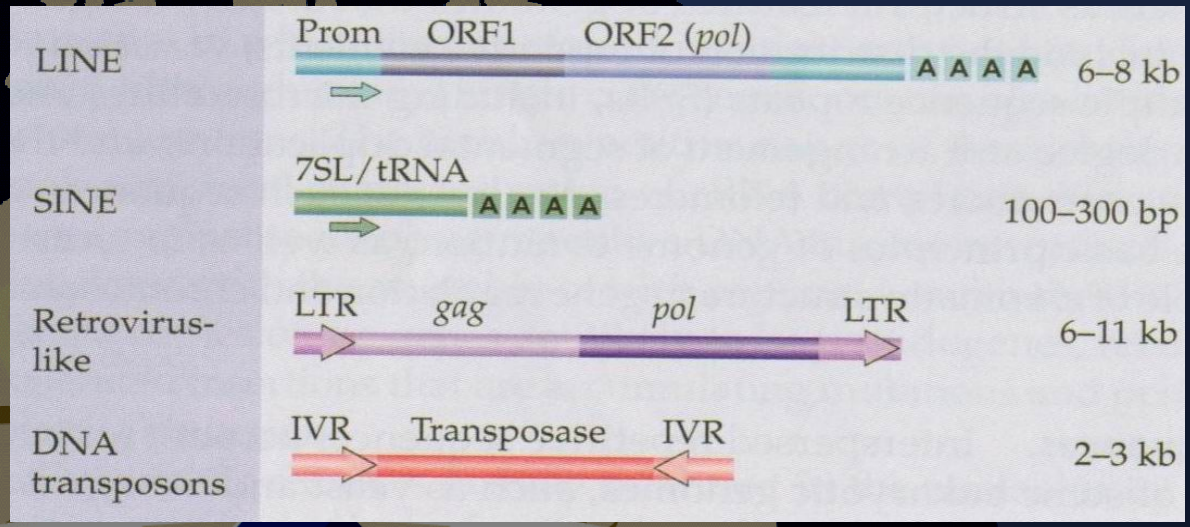
At the nucleotide level, approximately 40% of the human genome can be aligned to the mouse genome. These sequences seem to represent most of the orthologous sequences that remain in both lineages from the common ancestor, with the rest likely to have been deleted in one or both genomes.

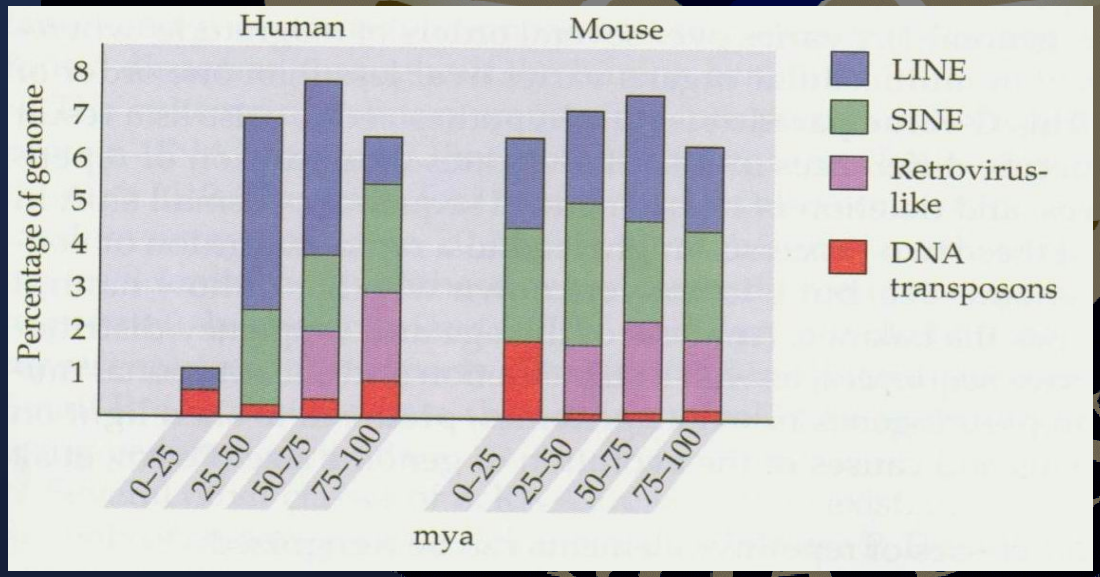
The single most prevalent feature of mammalian genomes is their repetitive sequences, most of which are interspersed repeats representing 'fossils' of transposable elements. Transposable elements are a principal force in reshaping the genome, and their fossils thus provide powerful reporters for measuring evolutionary forces acting on the genome.

All mammals have essentially the same four classes of transposable elements: (1) the autonomous long interspersed nucleotide element (LINE)-like elements; (2) the LINE-dependent, short RNA-derived short interspersed nucleotide elements (SINEs); (3) retrovirus-like elements with long terminal repeats (LTRs); and (4) DNA transposons. The first three classes procreate by reverse transcription of an RNA intermediate (retroposition), whereas DNA transposons move by a cut-and-paste mechanism of DNA sequence

Long interspersed nuclear elements

Short interspersed nuclear elements





A Primer of Genome Science by Gibson and Muse, 2002. Sinauer Associates Inc.

Table SI 6. Definition of Ancestral Repeats for the Genome Evolution Section.

Name	Family	Class
5S	rRNA	rRNA
7SK	RNA	RNA
7SLRNA	srpRNA	srpRNA
BLACKJACK	DNA	AcHobo
Charlie1	DNA	MER1_type
Charlie1a	DNA	MER1_type
Charlie1b	DNA	MER1_type
Charlie2	DNA	MER1_type
Charlie2a	DNA	MER1_type
Charlie2b	DNA	MER1_type
Charlie4	DNA	MER1_type
Charlie4a	DNA	MER1_type
Charlie7	DNA	MER1_type
Charlie8	DNA	MER1_type
Charlie9	DNA	MER1_type
ERVL	LTR	ERVL
ERVL-E	LTR	ERVL
HAL1	LINE	L1

Table SI 4. Germline mouse mutations caused by viral retrotransposon insertion.

Locus	Mutant mouse name	Strain of origin	Reference
	Early transposon (ETn)		
Adcy^{brl}	barreless	ICR	1
Cacng2^{stg}	stargazer	A/J	2
Cacng2^{stg3J}	stargazer-3J	BALB/cJ	3
Clc1^{adr}	myotonia	A2G	4
Tnfrfs6^{lpr}	lymphoproliferative response	MRL/MpJ	5
Edar$^{Dl\text{-}slk}$	sleek	STOCK Rb(8.12)5Bnr *Tyrp1b m*	6
Fbxw4^{Dac}	dactylaplasia	SM/Ckc	7
Fbxw4$^{Dac\text{-}2J}$	dactylaplasia-2J	MRL/MpJ	7
Fign^{fi}	fidget	unknown/mixed	8
Foxn1$^{nu\text{-}BC}$	nude-British Columbia	STOCK *Spna1*$^{sph\text{-}2Bc}$	9
Gli3^{pdn}	extra-toes	Jcl:ICR	10
Hk1^{dea}	downeast anemia	A/J	11
H2^{Q8}	MHC expression variant	strain variant	12
Mip	major intrinsic protein of eye lens fiber	strain variant	13
Mip$^{Cat\text{-}Fr}$	Dominant cataract (Fraser)	A/J	
Muted^{mu}	muted	STOCK t	14
Lep^{ob2J}	obese	SM/Ckc-*Fbxw4*^{Dac}	15
Tyr$^{c\text{-}BC3}$	albino-British Columbia 3	STOCK *Spna1*$^{sph\text{-}2Bc}$	9

Intracisternal A-type Particle (IAP)

A^{vy}	viable yellow	C3H/HeJ	16
A^{hvy}	hypervariable yellow	C3H/HeSnJ	17
A^{iapy}	agouti IAP yellow	B6C3F1/J	18
Atrn^{mg}	mahogany	C3H x Swiss stock	19
Atrn^{mg-L}	mahogany-Leicester	C3H/HeJ	19
Axin^{Fu}	fused	unspecified	20
Axin^{Kb}	knobbly	unspecified	20
Dab1^{scm}	scrambler	Dc/Le	21
Eya1^{bor}	branchio otorenal syndrome	C3H/HeJ	22
Gus^{Mps2J}	mucopolysaccharidosis-3J	C3H/HeOuJ	23
Hps1^{ep}	pale ear	C3HeB/FeJ	24
Hps3^{coa6J}	cocoa-6J	C3H/HeJ	25
Lamb3^{Iap}	Herlitz junctional epidermolysis bullosa	C3H	26
Pitp1b^{vb}	vibrator	DBA/2J	27
Reln^{Alb2J}	reelin-Albany-2	C3H/HeJ	28
Spna1$^{sph-Dem}$	spherocytosis-Demant	CcS3/Dem	29
Tyr^{c-1R}	albino-1R	(101/Rl x C3H/Rl)F1	30
Usp14^{ax}	ataxia	CBA/J	31

Murine leukemia virus (MLV)

Abcb1a	multiple drug resistance-3	strain variant	32
Aif^{Hq}	harlequin	CF1	33
Myo5a^{d}	dilute	fancy mice	34
hr	hairless	unknown	35
Pde6b^{rd1}	retinal-degeneration 1	fancy mice	36

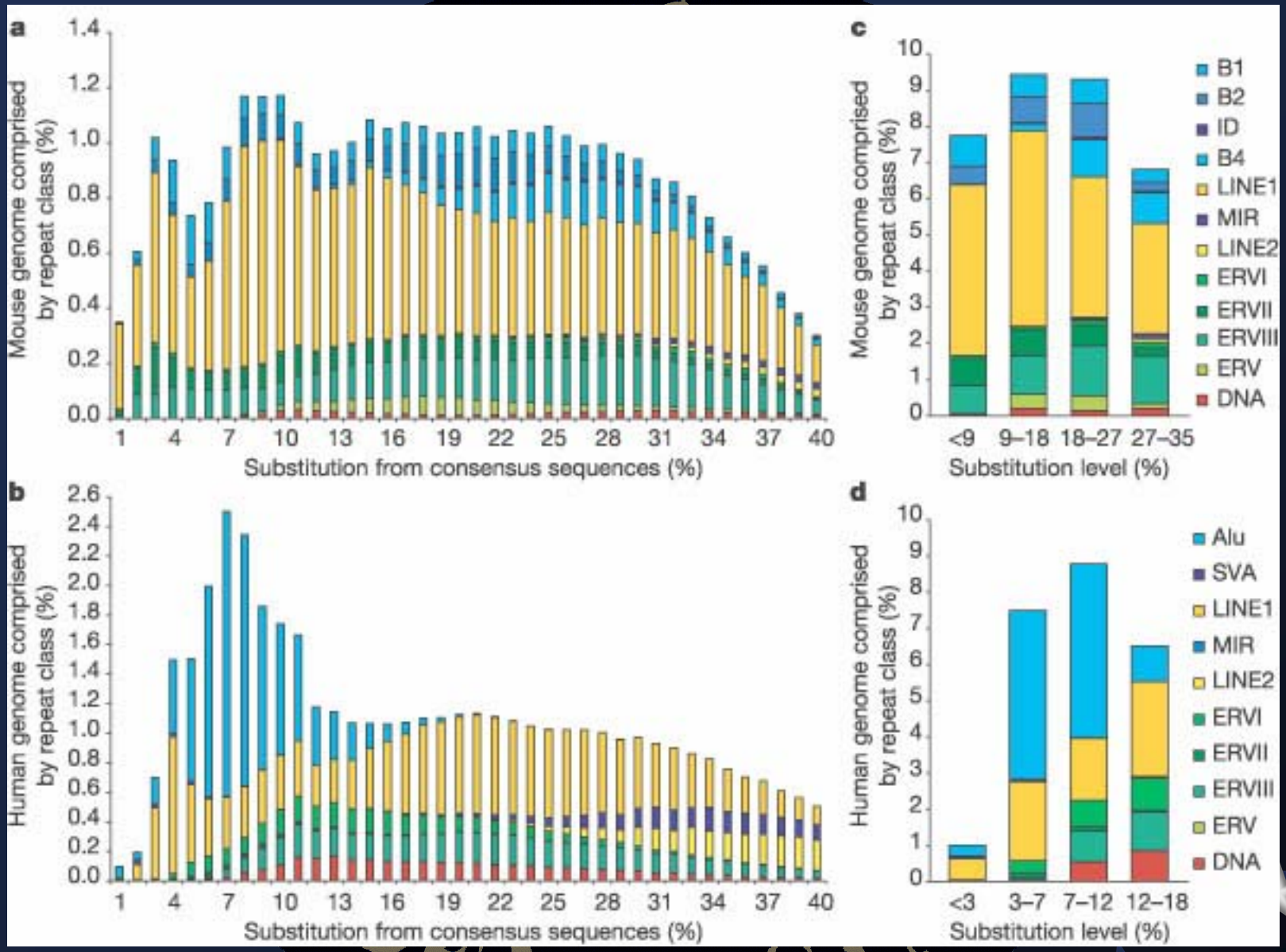
Virus-like 30 (VL30)

a	non-agouti	fancy mice	37

Table 5 **Composition of interspersed repeats in the mouse genome**

	Mouse				Human	
	Thousands of copies	Length occupied (Mb)	Fraction of genome (%)	Lineage specific (%)	Fraction of genome (%)	Lineage specific (%)
LINEs	660	475.3	19.20	16.46	20.99	7.94
LINE1	599	464.8	18.78	16.46	17.37	7.94
LINE2	53	9.4	0.38	–	3.30	–
L3/CR1	8	1.2	0.05	–	0.32	–
SINEs	1,498	202.9	8.22	7.63	13.64	10.74
B1 (Alu)	564	67.3	2.66	2.66	10.74	10.74
B2	348	59.6	2.39	2.39	–	–
B4/RSINE	391	57.1	2.36	2.36	–	–
ID	79	5.3	0.25	0.25	–	–
MIR/MIR3	115	14.1	0.57	–	2.90	–
LTR elements	631	244.3	9.87	8.72	8.55	4.09
ERV_classI	34	16.8	0.68	0.58	2.92	2.02
ERV_classII	127	79.1	3.14	3.14	0.30	0.30
ERV_classIII	37	14.0	0.58	0.32	1.55	0.19
MaLRs (III)	388	112.2	4.82	4.02	3.78	1.58
DNA elements	112	21.8	0.88	0.36	3.03	1.00
Charlie	82	15.2	0.62	0.35	1.41	0.14
Other hATs	8	1.6	0.06	–	0.31	–
Tigger	24	4.4	0.17	–	1.06	0.76
Mariner	1	0.2	0.01	0.01	0.10	0.07
Unclassified	26	9.2	0.38	0.37	0.15	0.14
Total	2,926	953.6	38.55	33.53	46.36	24.05
Small RNAs	19	1.5	0.06	0.04	0.04	0.02
Satellites	7	0.7	0.30	NA	0.34	NA
Simple repeats	960	56.1	2.27	NA	0.87	NA

The two right columns show the fractions in the human genome (excluding chromosome Y) for comparison. These and all other interspersed repeat-related data are based on RepeatMasker analysis (version July 2002, sensitive settings, RepBase release 5.3) of the February 2002 mouse and June 2002 human draft assemblies. Each repeat family in the RepeatMasker library was determined to be either order-specific or 'ancestral repeats' present at orthologous sites, usually on the basis of the average divergence level of the interspersed repeat family copies. For elements with an average divergence of 15–19% in human, copies were checked to be present or absent at mouse orthologous sites, to have inserted in known primate-specific repeats, or to have inserts of known mammalian-wide elements. No mammalian-wide repeats were found in the mouse genome that were not already known from the human genome. NA, not applicable.



Age distribution of interspersed repeats in the mouse and human genomes.

Mouse has accumulated more new repetitive sequence than human

Lineage-specific versus ancestral repeats

Despite marked differences in the activity of transposable elements between mouse and human, similar types of repeat sequences have accumulated in the corresponding genomic regions in both species. The correlation is stronger than can be explained simply by local (G+C) content and points to additional factors influencing how the genome is moulded by transposons.

Table 8 **Density of short SSRs in mouse compared with other mammals**

Cutoff (%)	All 1–5-bp unit SSRs including those in IRs					Excluding SSRs within IRs and IR tails		
	Mouse (%)	Human (%)	Ratio*	Dog (%)	Pig (%)	Mouse (%)	Human (%)	Ratio*
5	0.82	0.25	3.24	0.38	0.25	0.64	0.15	4.41
10	1.35	0.46	2.91	0.72	0.50	1.03	0.25	4.19
15	1.86	0.68	2.73	1.14	0.72	1.38	0.37	3.77
None	2.67	1.19	2.24	1.90	1.27	1.99	0.64	3.10

The cutoff indicates the maximum level of imperfect copies allowed, with 'none' indicating that all SSRs are recognized by RepeatMasker. To determine which SSRs were spawned from within IRs or IR tails, the locations of SSRs were overlapped with the RepeatMasker output. We excluded those SSRs overlapped by IRs and those resembling unmasked tails; that is, occurring more than one-third of the time adjacent to the same type of IR. IR, interspersed repeat; SSR, simple sequence repeat.

* Ratio was determined by dividing mouse percentage by human percentage.

Table 9 **Frequency of different SSRs in mouse and human**

Simple sequence repeat (SSR)	Including SSRs in IRs					Excluding SSRs from SINE and LINE tails and within IRs	
	Fraction of mouse genome (bp Mb^{-1})	Average no. units per SSR (mouse)	Average no. units per SSR (human)	Frequency (number SSRs per Mb)	Frequency ratio (mouse/human)	Frequency (number per Mb)	Frequency ratio (mouse/human)
A	555	26.7	25.0	20.8	0.4	10.4	1.6
C	11	22.7	25.2	0.5	6.2	0.2	4.6
AC	3070	20.8	18.1	73.8	3.5	62.1	3.2
AG	1365	24.3	15.7	28.1	6.9	26.7	7.5
AT	370	21.2	17.8	8.7	1.9	8.6	2.2
CG	4	11.6	11.2	0.2	10.5	0.1	11.1
AAC	148	9.5	8.6	5.2	1.7	3.9	2.8
AAG	152	29.1	15.9	1.7	11.7	1.6	14.5
AAT	147	12.2	9.7	4.0	0.9	2.9	1.9
ACC	53	11.2	9.6	1.6	6.2	1.3	6.4
AGC	30	11.3	8.8	0.9	3.0	0.8	3.3
AGG	46	12.2	8.9	1.3	5.1	1.1	6.4
AAAC	465	6.7	6.2	17.5	2.2	10.3	4.5
AAAG	203	16.2	10.2	3.1	2.8	2.3	5.6
AAAT	346	8.2	7.1	10.5	0.8	5.8	2.0
AACC	36	8.8	7.2	1.0	8.3	0.6	6.1
AAGG	122	13.5	12.3	2.3	4.9	2.0	6.7
AATG	41	7.8	6.7	1.3	1.1	1.0	1.6
ACAG	70	7.4	6.5	2.4	21.1	1.8	21.2
ACAT	85	10.3	8.4	2.1	6.0	1.7	8.6
AGAT	282	16.2	12.8	4.4	4.0	3.7	4.0
AGGG	39	8.2	6.6	1.2	6.2	1.1	8.8
AAAAC	369	5.9	5.0	12.4	1.5	7.2	2.9

Frequency and density of monomeric and dimeric SSRs and the most common trimeric and tetrameric SSRs. The numbers are for >20-bp-long SSRs containing <10% substitutions and indels. The two right columns show the density of each repeat, excluding those SSRs spawned from the tails of SINEs and LINEs or inherently part of other IRs. For this, SSRs were counted in a default (-m −s) RepeatMasker run.

Overall, mouse has 2.25–3.25-fold more short SSRs (1–5 bp unit) than human.

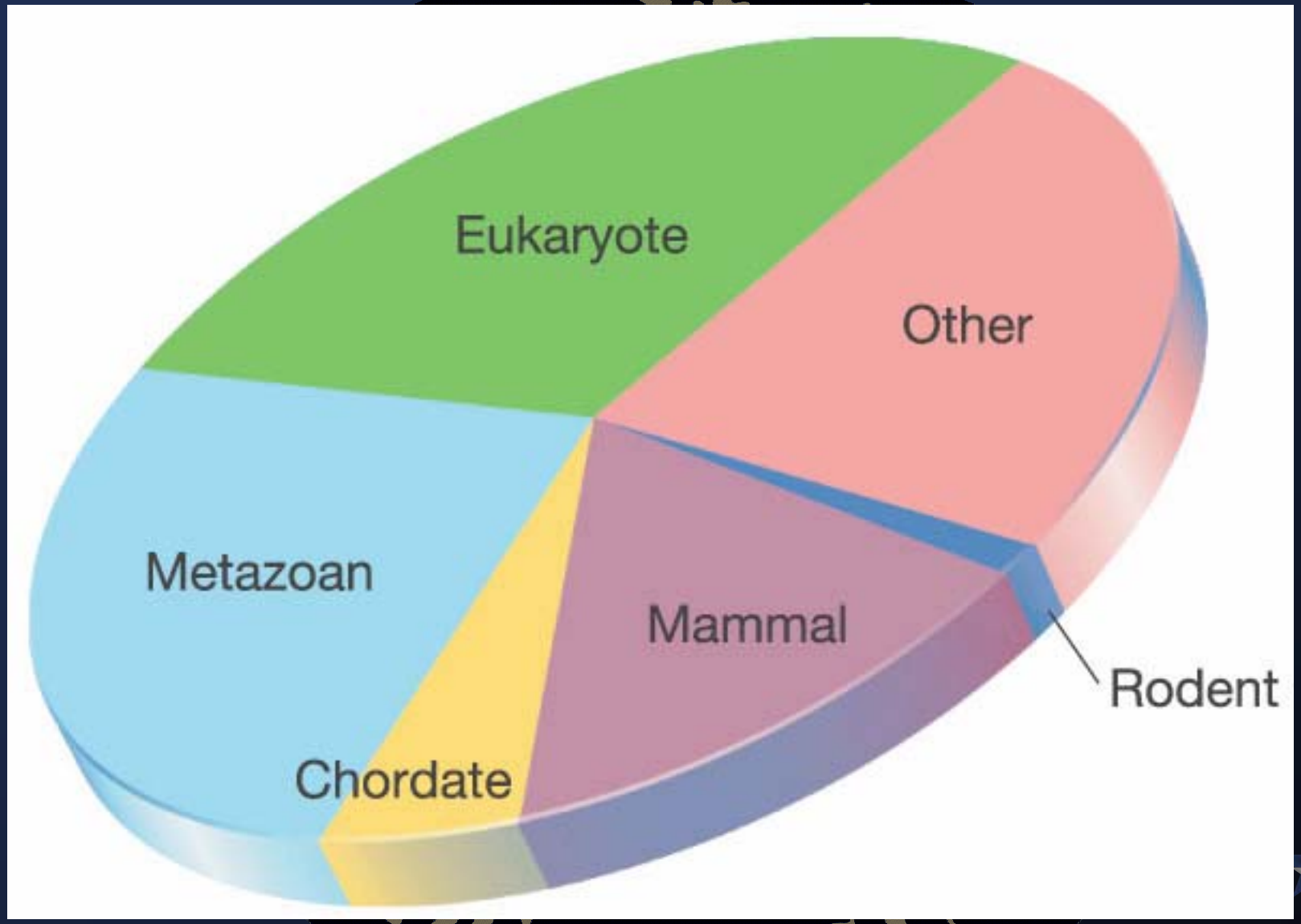
Apart from the absolute number of SSRs, there are also some marked differences in the frequency of certain SSR classes (Table 9). The most extreme is the tetramer (ACAG)*n*, which is 20-fold more common in mouse than human

We constructed catalogues of human and mouse gene predictions on the basis of available experimental evidence. The main computational tool was the Ensembl gene prediction pipeline augmented with the Genie gene prediction pipeline. Briefly, the Ensembl system uses three tiers of input. First, known protein-coding cDNAs are mapped onto the genome. Second, additional protein-coding genes are predicted on the basis of similarity to proteins in any organism using the GeneWise program. Third, *de novo* gene predictions from the GENSCAN program that are supported by experimental evidence (such as ESTs) are considered. These three strands of evidence are reconciled into a single gene catalogue by using heuristics to merge overlapping predictions, detect pseudogenes and discard misassemblies. These results are then augmented by using conservative predictions from the Genie system, which predicts gene structures in the genomic regions delimited by paired 5' and 3' ESTs on the basis of cDNA and EST information from the region.

Table 10 **Gene count in human and mouse genomes**

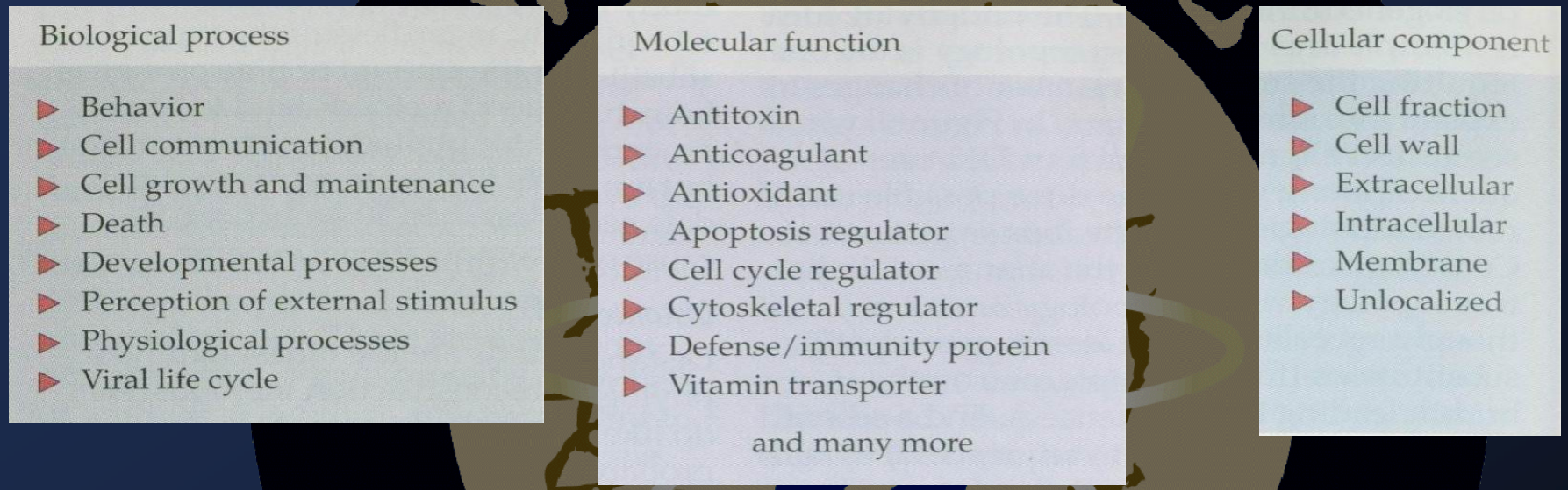
Genome feature	Human		Mouse	
	Initial (Feb. 2001)	Current (Sept. 2002)	Initial* (this paper)	Extended† (this paper)
Predicted transcripts	44,860	27,048	28,097	29,201
Predicted genes	31,778	22,808	22,444	22,011
Known cDNAs	14,882	17,152	13,591	12,226
New predictions	16,896	5,656	8,853	9,785
Mean exons/transcript‡	4.2 (3)	8.7 (6)	8.2 (6)	8.4 (6)
Total predicted exons	170,211	198,889	191,290	213,562

*Without RIKEN cDNA set.
†With RIKEN cDNA set.
‡Median values are in parentheses.

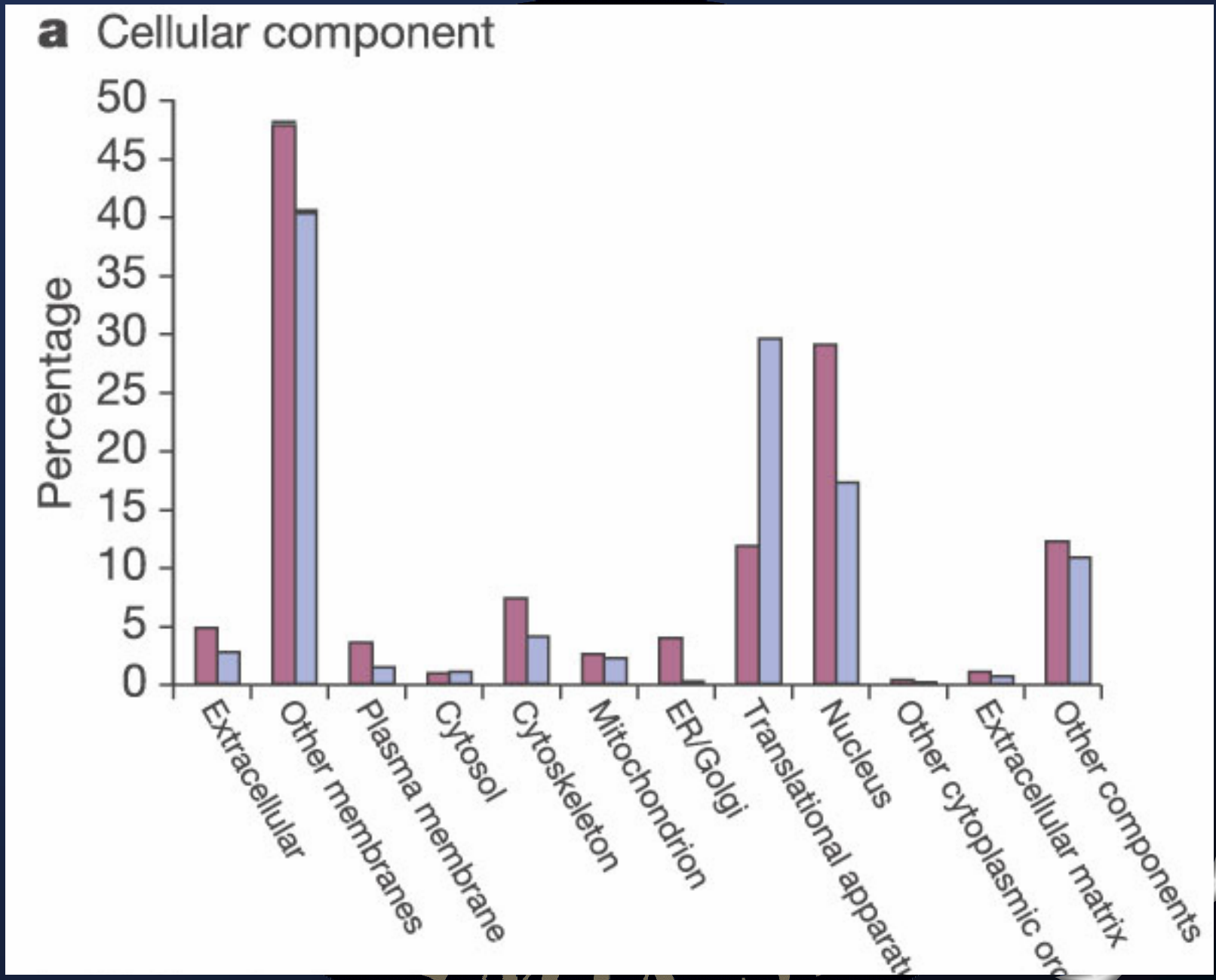


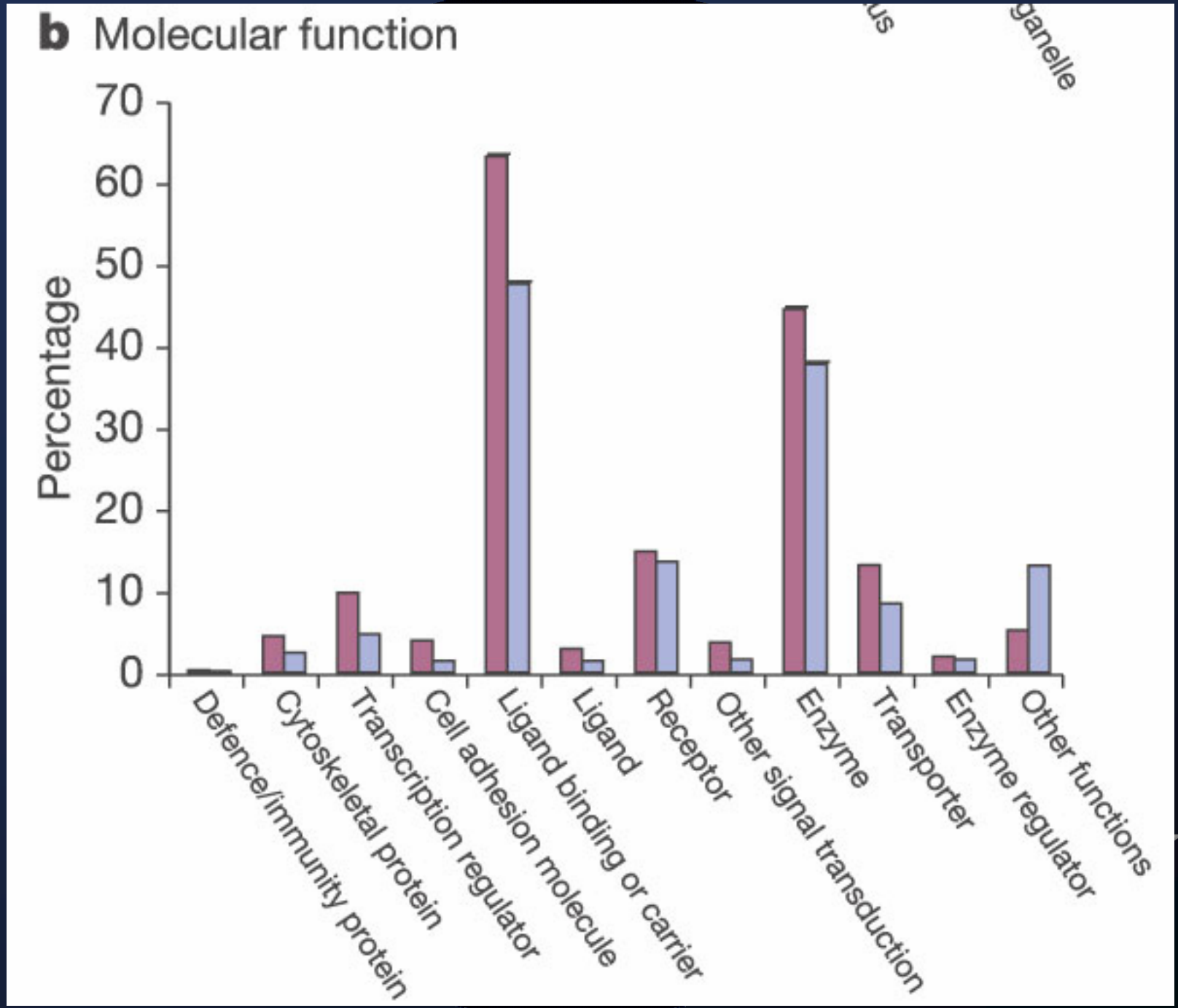
Taxonomic breakdown of homologues of mouse proteins according to taxonomic range

Gene Ontology: GO



A Primer of Genome Science by Gibson and Muse, 2002. Sinauer Associates Inc.


a Cellular component
Percentage
50
45
40
35
30
25
20
15
10
5
0
Extracellular
Other membranes
Plasma membrane
Cytosol
Cytoskeleton
Mitochondrion
ER/Golgi
Translational apparat
Nucleus
Other cytoplasmic or
Extracellular matrix
Other components


b Molecular function
Percentage
70
60
50
40
30
20
10
0
Defence/immunity protein
Cytoskeletal protein
Transcription regulator
Cell adhesion molecule
Ligand binding or carrier
Ligand
Receptor
Other signal transduction
Enzyme
Transporter
Enzyme regulator
Other functions

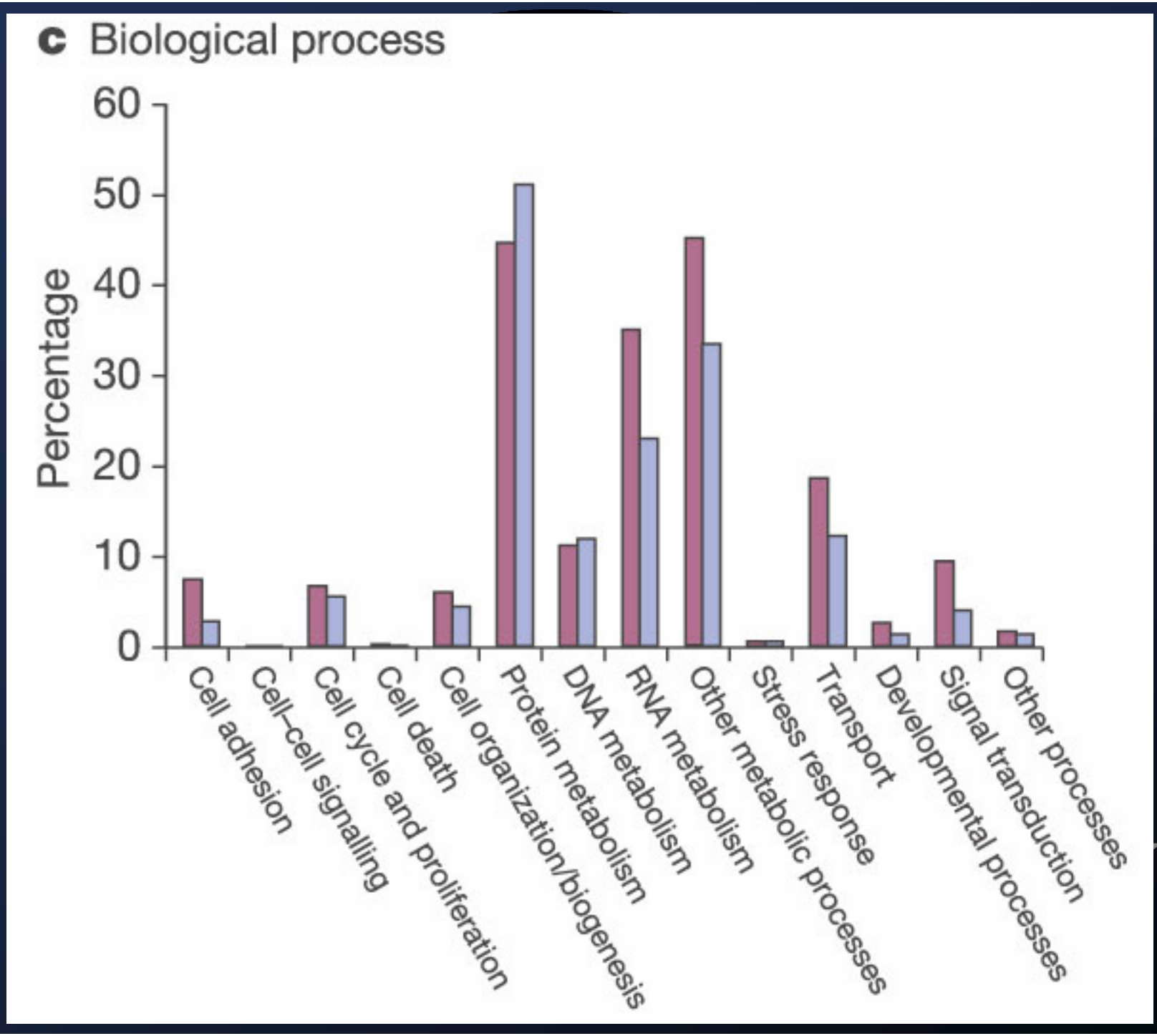

c Biological process
Percentage
60
50
40
30
20
10
0
Cell adhesion
Cell–cell signalling
Cell cycle and proliferation
Cell death
Cell organization/biogenesis
Protein metabolism
DNA metabolism
RNA metabolism
Other metabolism
Stress response
Transport
Developmental processes
Signal transduction
Other processes

Table 11 **Domain-based and family-based InterPro analysis**

InterPro	Name	*M. musculus* (%)	*H. sapiens* (%)	*T. rubripes* (fish) (%)	*C. elegans* (nematode) (%)	*D. melanogaster* (insect) (%)
000276	**Rhodopsin-like GPCR superfamily**	**3.4 (1)**	**2.8 (2)**	1.6 (4)	2.0 (4)	0.6 (16)
000822	Zn-finger, C2H2 type	3.1 (2)	3.0 (1)	1.7 (3)	1.0 (10)	2.4 (1)
003006	Immunoglobulin/major histocompatibility complex	2.7 (3)	2.8 (3)	1.7 (2)	0.4 (32)	1.0 (6)
000719	Eukaryotic protein kinase	2.1 (4)	2.0 (4)	2.1 (1)	2.2 (2)	1.6 (3)
003593	ATPase	1.7 (5)	1.4 (5)	1.1 (5)	1.3 (8)	1.8 (2)
000504	**RNA-binding region RNP-1 (RNA recognition motif)**	**1.4 (6)**	**1.0 (10)**	0.7 (18)	0.6 (19)	0.9 (8)
004244	***L1 transposable element***	**1.4 (7)**	**0.7 (20)**	0.0 (772)	0.0 (–)	0.0 (–)
001680	G-protein beta WD-40 repeat	1.3 (8)	1.1 (7)	1.0 (7)	0.7 (16)	1.3 (5)
001841	Zn-finger, RING	1.3 (9)	1.1 (8)	0.9 (11)	0.8 (12)	0.8 (10)
000477	**RNA-directed DNA polymerase (reverse transcriptase)**	**1.3 (10)**	**0.7 (19)**	0.8 (14)	0.3 (42)	0.1 (163)
001849	Pleckstrin-like domain	1.2 (11)	1.0 (9)	1.0 (6)	0.4 (35)	0.5 (24)
001611	Leucine-rich repeat	1.2 (12)	0.9 (13)	0.9 (10)	0.3 (47)	0.9 (9)
001356	Homeobox	1.2 (13)	0.9 (12)	1.0 (8)	0.5 (21)	0.8 (11)
001909	KRAB box	1.1 (14)	1.1 (6)	0.0 (–)	0.0 (–)	0.0 (–)
002048	Calcium-binding EF-hand	1.1 (15)	0.9 (15)	0.8 (16)	0.4 (30)	0.7 (13)
002110	Ankyrin	1.0 (16)	1.0 (11)	0.8 (15)	0.5 (24)	0.6 (19)
001452	SH3 domain	1.0 (17)	0.8 (16)	0.8 (12)	0.3 (48)	0.5 (23)
000561	EGF-like domain	1.0 (18)	0.9 (14)	0.9 (9)	0.5 (22)	0.5 (27)
001584	***Integrase, catalytic domain***	**0.9 (19)**	**0.0 (412)**	0.2 (61)	0.1 (134)	0.0 (490)
003961	Fibronectin, type III	0.9 (20)	0.8 (17)	0.8 (13)	0.2 (58)	0.5 (26)
005225	Small GTP-binding protein domain	0.8 (21)	0.7 (18)	0.7 (17)	0.4 (29)	0.6 (18)
000210	BTB/POZ domain	0.8 (22)	0.6 (21)	0.6 (20)	0.8 (13)	0.5 (22)
001440	TPR repeat	0.7 (23)	0.6 (23)	0.5 (24)	0.3 (45)	0.6 (17)
001478	PDZ/DHR/GLGF domain	0.7 (24)	0.6 (22)	0.7 (19)	0.3 (43)	0.5 (25)
000008	C2 domain	0.6 (25)	0.6 (25)	0.6 (22)	0.2 (59)	0.3 (37)
000636	Cation channel, non-ligand gated	0.6 (26)	0.5 (26)	0.6 (21)	0.4 (31)	0.4 (32)
001650	Helicase, C-terminal	0.6 (27)	0.4 (31)	0.4 (32)	0.4 (27)	0.5 (21)
000980	SH2 domain	0.5 (28)	0.5 (28)	0.4 (26)	0.4 (37)	0.2 (51)
001092	Basic helix-loop-helix dimerization domain bHLH	0.5 (29)	0.4 (34)	0.4 (25)	0.2 (73)	0.4 (28)
001254	Serine protease, trypsin family	0.5 (30)	0.5 (29)	0.4 (27)	0.1 (202)	1.5 (4)
003308	***Integrase, N-terminal zinc binding***	**0.5 (31)**	**0.0 (1,297)**	0.0 (–)	0.0 (–)	0.0 (–)
000379	Esterase/lipase/thioesterase, active site	0.5 (32)	0.4 (33)	0.4 (30)	0.7 (15)	1.0 (7)
000626	**Ubiquitin domain**	**0.5 (33)**	**0.2 (61)**	0.1 (104)	0.1 (98)	0.2 (58)
004822	Histone-fold/TFIID-TAF/NF-Y domain	0.5 (34)	0.4 (30)	0.2 (82)	0.5 (26)	0.1 (155)
000387	Tyrosine-specific protein phosphatase and dual-specificity protein phosphatase	0.5 (35)	0.4 (32)	0.4 (28)	0.7 (17)	0.2 (47)
002156	***RNase H***	**0.5 (36)**	**0.0 (599)**	0.0 (297)	0.0 (538)	0.0 (957)

Table 12 **K_A, K_S, K_A/K_S and pairwise percentage amino acid identities for 1:1 mouse-human orthologues**

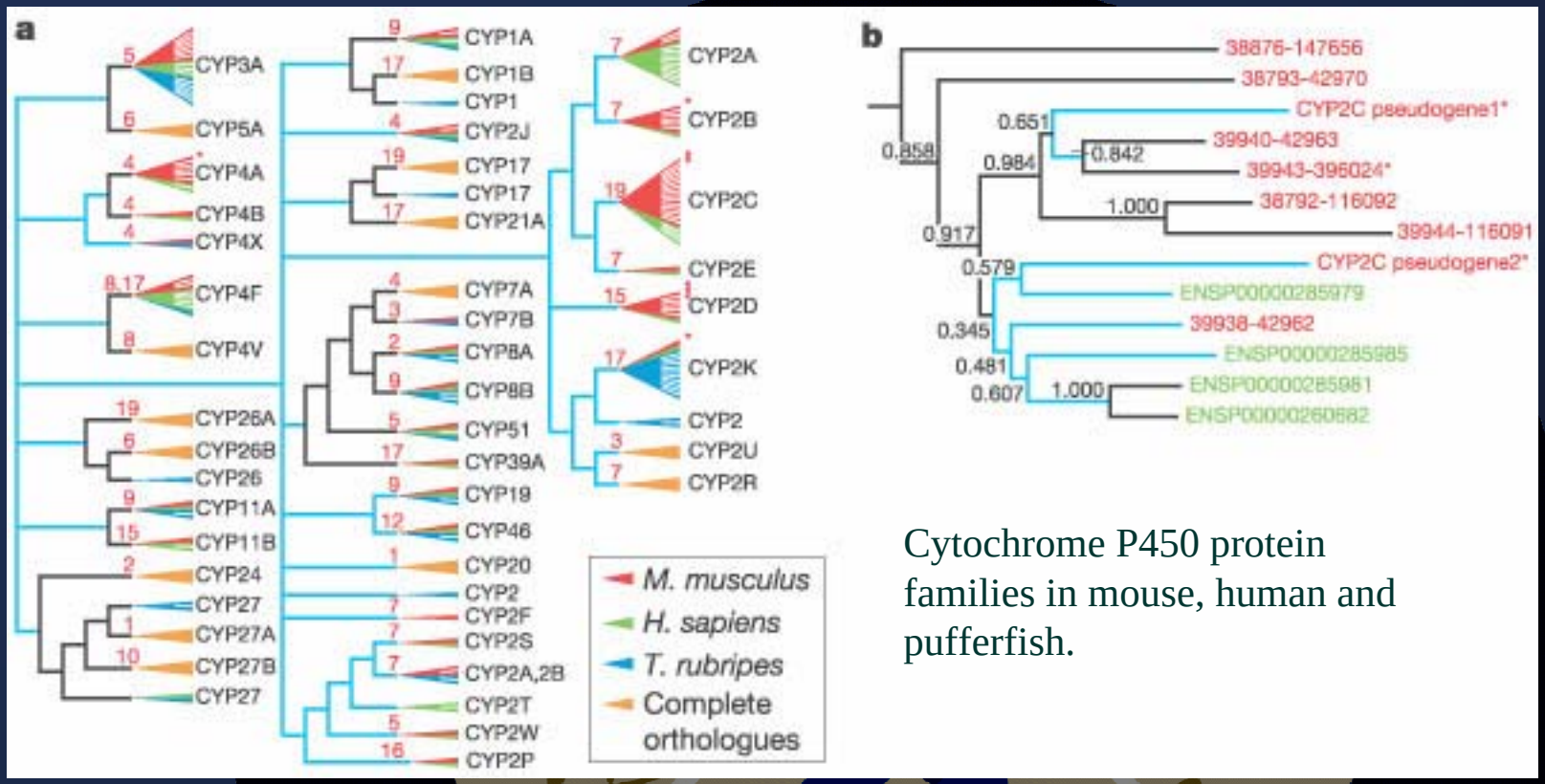
Orthologue regions	Amino acid identity (%); median (*n*); 16–83%	K_A; median (*n*); 16–83%	K_S; median (*n*); 16–83%	K_A/K_S; median (*n*); 16–83%
Aligned positions in the full-length protein	78.5; (12,845); 51.0–93.1	0.071; (12,845); 0.019–0.181	0.602; (12,845); 0.414–0.981	0.115; (12,615); 0.036–0.275
Domain-containing protein regions*	93.5; (8,280); 80.6–99.1	0.032; (8,280); 0.004–0.111	0.601; (8,280); 0.390–1.072	0.061; (7,399); 0.015–0.178
Domain-free protein regions*	71.1; (12,782); 37.8–90.9	0.090; (12,615); 0.022–0.237	0.586; (12,614); 0.383–0.986	0.155; (12,035); 0.048–0.370
All predicted domains*	95.1; (17,735); 80.6–100.0	0.024; (17,735); 0–0.108	0.627; (17,735); 0.345–1.309	0.062; (13,391); 0.016–0.201
Catalytic domains†	96.6; (1,982); 86.1–100.0	0.015; (1,982); 0–0.065	0.578; (1,982); 0.346–0.979	0.033; (1,646); 0.009–0.115
Non-catalytic domains†	94.9; (15,753); 80.0–100.0	0.026; (15,753); 0–0.114	0.635; (15,753); 0.345–1.352	0.068; (11,745); 0.018–0.213
Secreted domains†	88.9; (3,901); 75.4–97.5	0.058; (3,901); 0.012–0.147	0.694; (3,901); 0.414–1.357	0.091; (3,537); 0.023–0.241
Cytoplasmic domains†	96.7; (5,795); 87.1–100	0.015; (5,795); 0–0.064	0.587; (5,795); 0.331–1.152	0.041; (4,300); 0.012–0.133
Nuclear domains†	98.6; (3,757); 85.7–100.0	0.008; (3,757); 0–0.077	0.655; (3,757); 0.302–1.696	0.050; (2,103); 0.011–0.185

*Domains predicted by Pfam and SMART.
†Domains predicted by SMART.

synonymous [Ks]

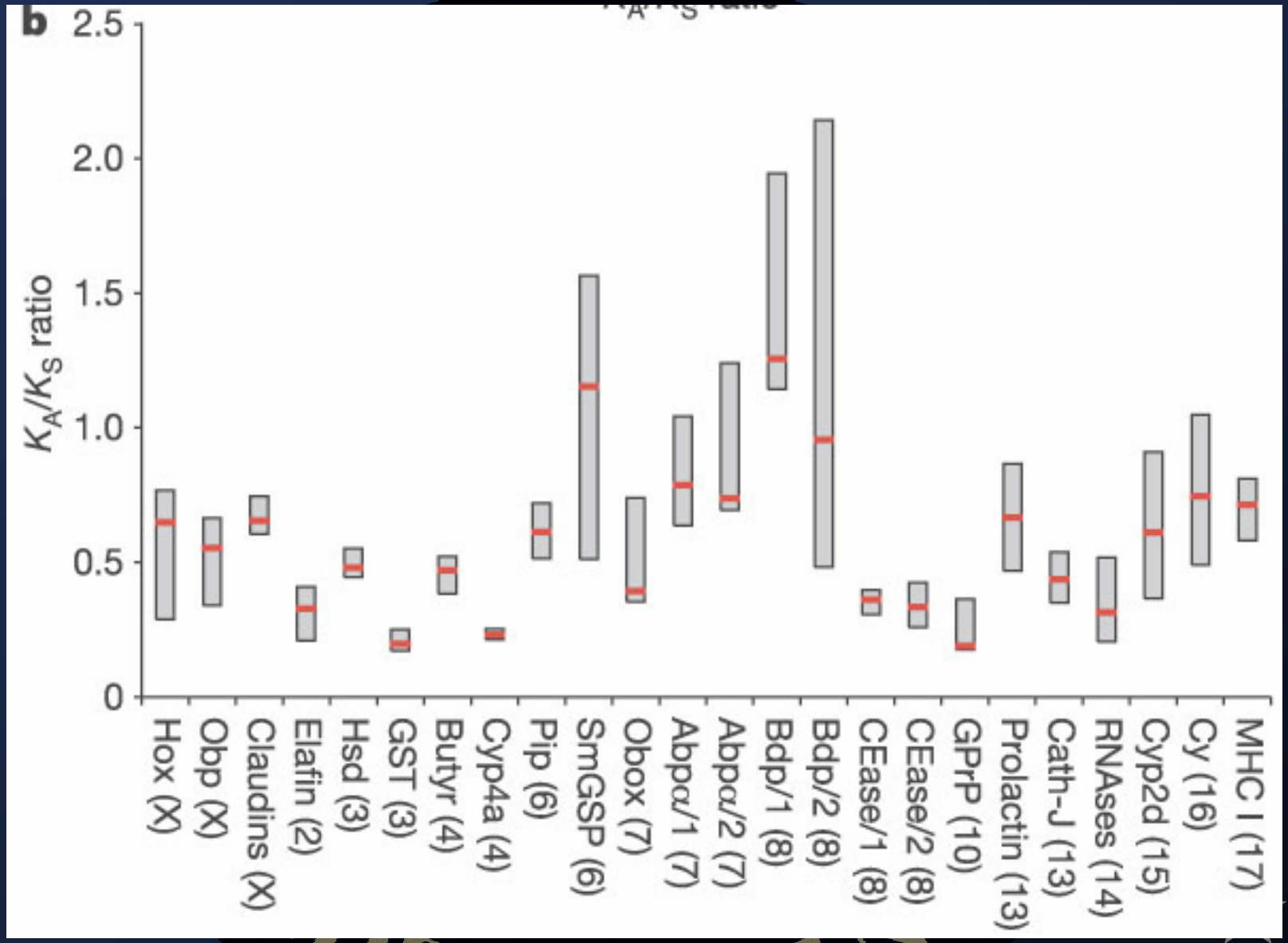
nonsynonymous [Ka]

For the 12,845 pairs of mouse–human 1:1 orthologues, 70.1% of the residues were identical. The median amino acid identity was 78.5% and the median *KA/KS* ratio was 0.115



Cytochrome P450 protein families in mouse, human and pufferfish.

This gene family is moderately but significantly expanded in mouse (84 genes) relative to human (63 genes). By comparing the cytochrome P450 gene families from mouse, human and pufferfish (*Takifugu rubripes*), we found clear expansions in four subfamilies (Cyp2b, Cyp2c, Cyp2d and Cyp4a) in mouse relative to human. These occur in local gene clusters that also contain unprocessed pseudogenes. The expansions appear to be associated, in part, with gender differences in the metabolism of androgens and xenobiotics.



Protein evolution

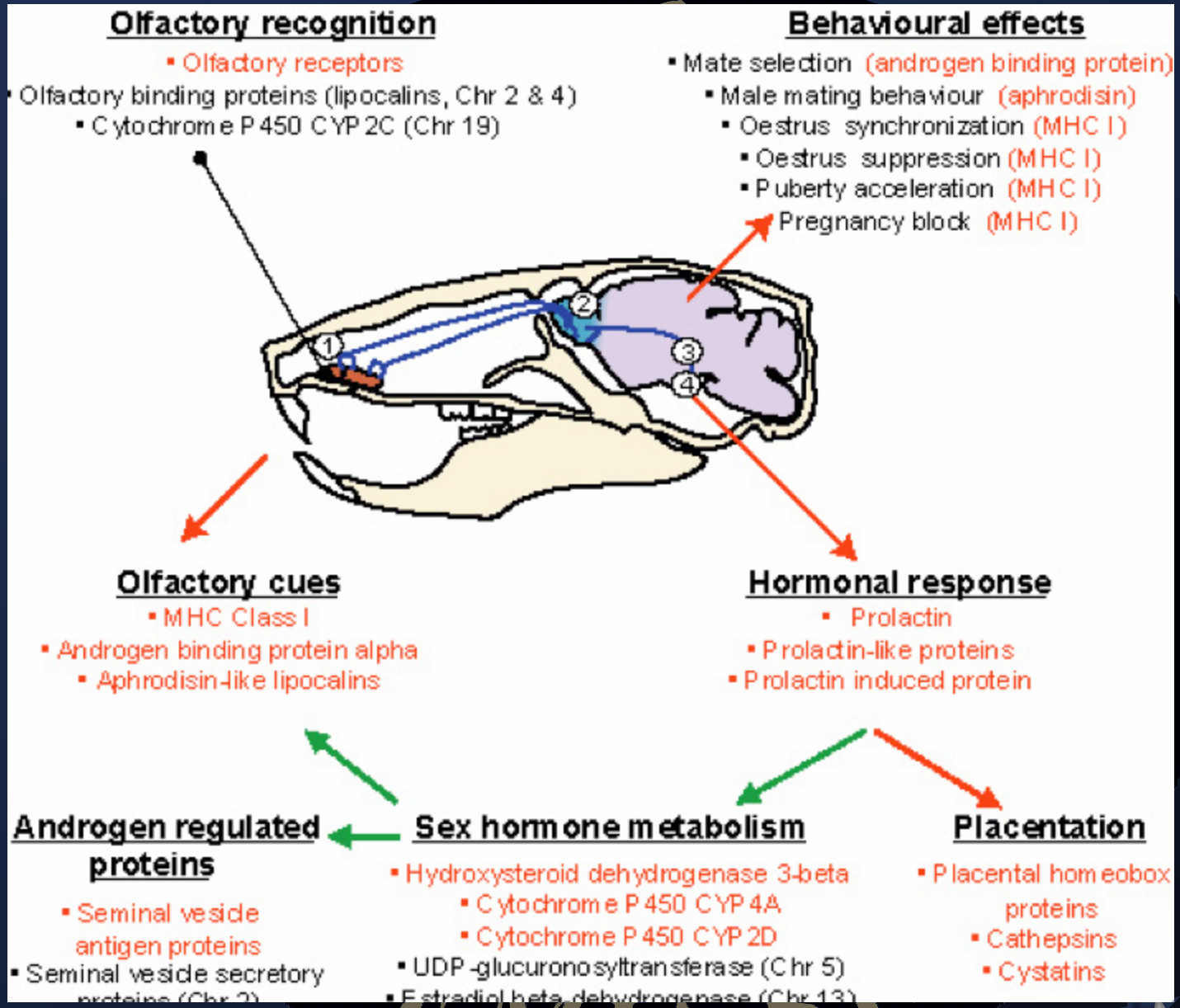
To explore systematically recent evolution of the mouse proteome, we searched for mouse-specific gene clusters.

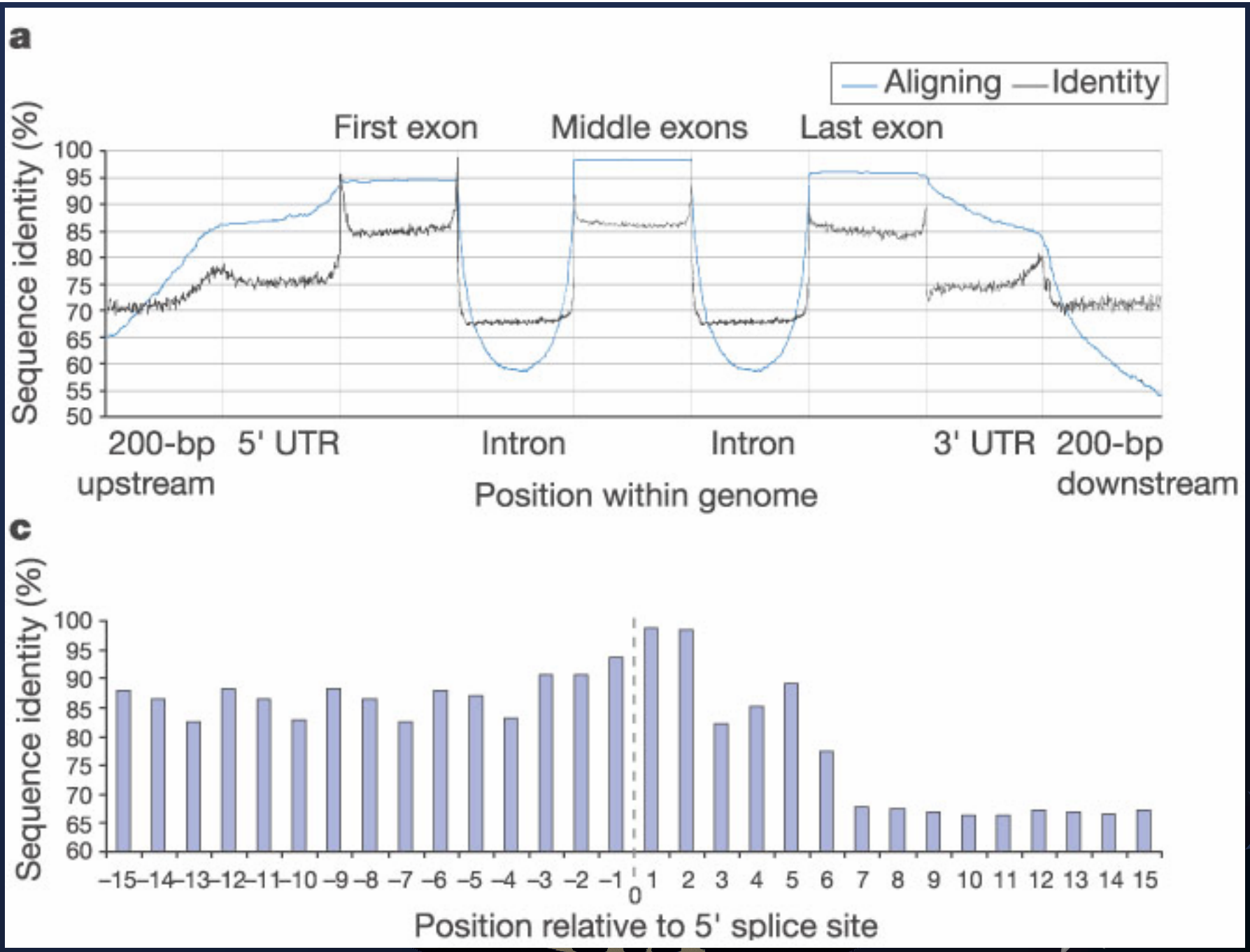
Table 15 **Mouse homologous gene clusters**

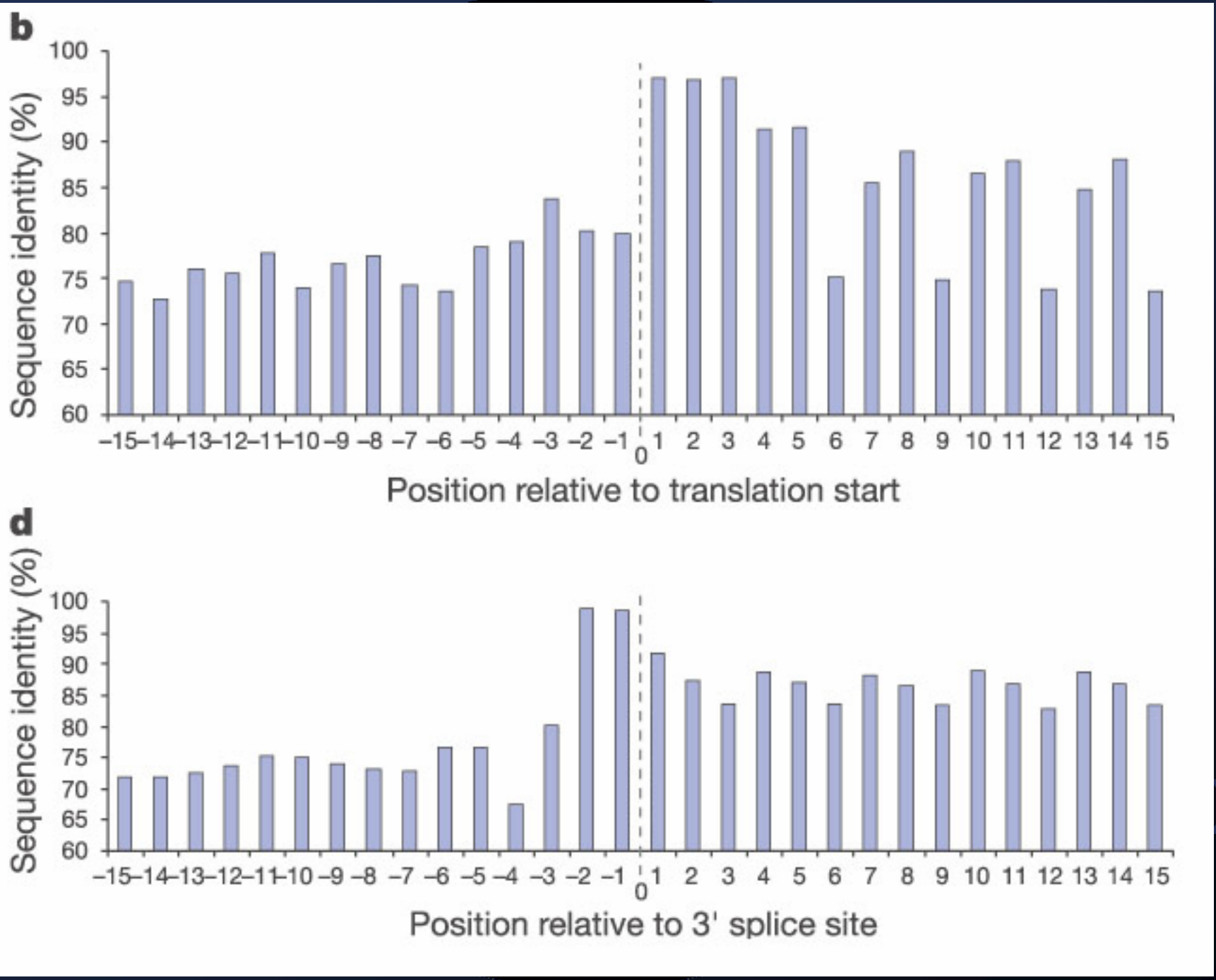
Cluster	Abbreviation*	Chr†	No. of predicted genes‡	Strand (major/minor)	Function and expression	Reference
HOX cluster including Pem, Gpbox and Psx1	Hox	X	8	5/3	Probable functions in embryogenesis, placentation and rodent oogenesis. Sequence evidence for positive selection. Expression in the placenta/embryo. Pem is under androgen control in the testis and epididymis.	204, 299
Odorant binding proteins (Obp)/aphrodisin homologues (lipocalins)	Obp	X	8	4/4	Lipocalin family members probably bind volatile odorants. Aphrodisin is an aphrodisiac hormone of hamster vaginal discharges. Obps are highly expressed in the nasal area. Prostate probasin is regulated by androgens.	300, 301
Claudins	Claudins	X	6	4/2	Claudins form an intercellular barrier in tight junctions. Mutations in claudins cause reproductive defects. Four homologues were found in a testis cDNA library.	302, 303
Elafin, eppin and antileukoproteinase 1 homologues	Elafin	2	7	4/3	Reproduction-related WAP domain proteins; antimicrobial properties. Some proteins are specific to the epididymis.	304
Hydroxysteroid dehydrogenase	Hsd	3	7	7/0	Biosynthesis of hormonal steroids. Controls binding of hormone to receptor. Gender-specific expression in olfactory bulb and olfactory tubercle.	215, 216
Class mu glutathione *S*-transferases	GST	3	7	7/0	Conjugation of glutathione to hormones or metabolites?	
Butyrophilin homologues	Butyr	4	5	3/2	Unknown functions. Contain immunoglobulin-like domain(s).	
Class Cyp4a cytochromes P450	Cyp4a	4	7	4/3	Oxidation of compounds, possibly fatty acids. Gender differences in expression (Cyp4a and Cyp4b1).	305–307
Prolactin-inducible protein; seminal vesicle-antigen (Sva)	Pip	6	4	4/0	Sva role in suppression of spermatozoa motility; 14 kDa submandibular gland protein is expressed in lacrimal, salivary and sweat	214, 308, 309

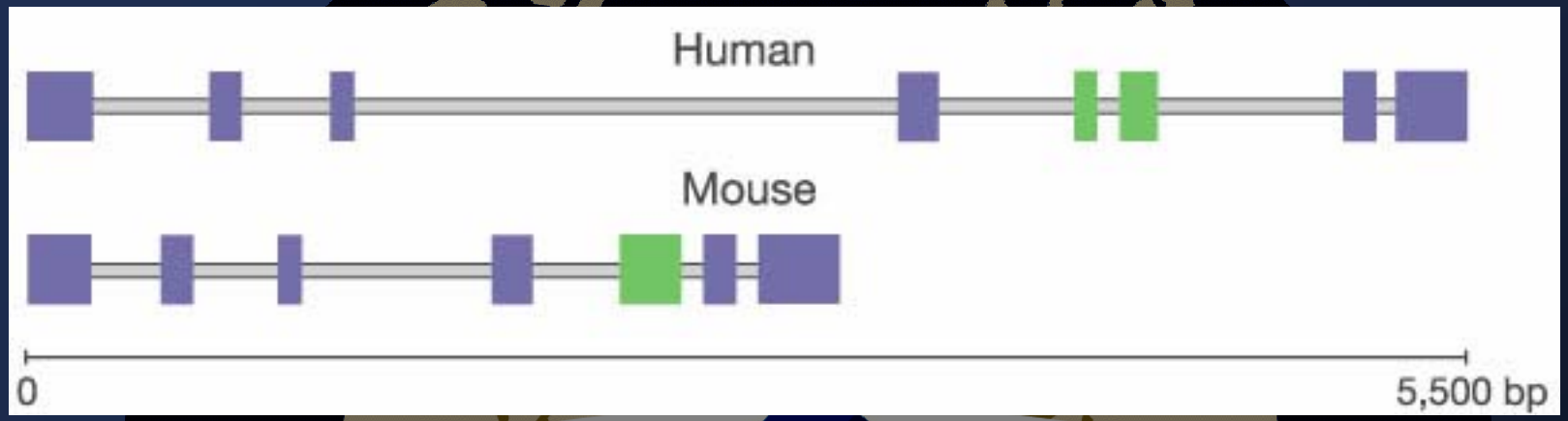
Proline-rich proteins	(–)	6	4	2/2	Salivary proteins of unknown function.	310
Submandibular gland secretory proteins	SmGSP	6	9	5/4	Salivary proteins of unknown function, related to proline-rich proteins (above). Expression is androgen-dependent.	311
Obox, family of homeobox proteins	Obox	7	6	3/3	Homeobox proteins preferentially expressed in the gonads.	312
Salivary androgen-binding protein alpha-subunit	Abpα/1 Abpα/2	7	9	7/2	Mate selection. Genes may be under positive selection due to role in subspecies recognition.	221, 222, 313, 314
Beta-defensin proteins	Bdp/1	8	5	4/1	Antimicrobial peptides. Beta-defensin 3 is expressed in salivary glands, epididymis, ovary and pancreas. Beta-defensin 5 is expressed in trachea, oesophagus and tongue.	315
Beta-defensin proteins	Bdp/2	8	5	3/2	Antimicrobial peptides. Beta-defensins 1 and 2 are expressed in kidney.	315
Carboxylesterase	CEase/1	8	6	6/0	Involved in detoxification of xenobiotics.	316
Carboxylesterase	CEase/2	8	5	4/1	Involved in detoxification of xenobiotics. Expression of egasyn is differentially regulated by androgens.	316, 317
Glioma pathogenesis-related protein homologues containing SCP domains	GPrP	10	5	3/2	Unknown.	
Prolactin-related proteins	Prolactin	13	17	11/6	Placentation; probable role in development of placental blood vessels.	212, 318, 319
Cathepsin J-like enzymes	Cath-J	13	6	6/0	Placentation; expressed in murine placenta only.	202
Eosinophil-associated ribonuclease	RNAses	14	11	7/4	Probable roles in pathogen response in eosinophils.	224, 320, 321
Class Cyp2d cytochromes P450	Cyp2d	15	5	3/2	Oxidation of compounds, possibly fatty acids. Cyp2d9 is a testosterone 16-alpha hydroxylase regulated by androgens. Cyp2d22 is expressed in mammary epithelial cells.	217, 218
Cystatins/stefins	Cy	16	7	5/2	Inhibitors of papain-like cystein proteinases.	322
Proteins of unknown function	(–)	16	6	5/1	Gly-, Cys- and Tyr-rich proteins of unknown function. ESTs (BB615096 and AV261464) are derived from a testis cDNA library.	
MHC class Ib	MHCI	17	8	4/4	Some genes involved in antigen presentation, but most are not. Very different between human and mouse and between mouse strains. Class Ia peptide-binding region shows increased non-synonymous-to-synonymous substitution.	227, 323

(–), indicates that reliable alignments could not be determined, and thus they were not included in Fig. 21.
*Where paralogues were aligned into two sequence-similar subfamilies, this is indicated by appending a slash followed by the ordinal to the abbreviation.
†Mouse chromosome number.


Olfactory recognition
▪ Olfactory receptors
▪ Olfactory binding proteins (lipocalins, Chr 2 & 4)
▪ Cytochrome P450 CYP2C (Chr 19)
Behavioural effects
▪ Mate selection (androgen binding protein)
▪ Male mating behaviour (aphrodisin)
▪ Oestrus synchronization (MHC I)
▪ Oestrus suppression (MHC I)
▪ Puberty acceleration (MHC I)
▪ Pregnancy block (MHC I)
1
2
3
4
Olfactory cues
▪ MHC Class I
▪ Androgen binding protein alpha
▪ Aphrodisin-like lipocalins
Hormonal response
▪ Prolactin
▪ Prolactin-like proteins
▪ Prolactin induced protein
Androgen regulated proteins
▪ Seminal vesicle antigen proteins
▪ Seminal vesicle secretory
Sex hormone metabolism
▪ Hydroxysteroid dehydrogenase 3-beta
▪ Cytochrome P450 CYP4A
▪ Cytochrome P450 CYP2D
▪ UDP-glucuronosyltransferase (Chr 5)
Placentation
▪ Placental homeobox proteins
▪ Cathepsins
▪ Cystatins


a
Aligning
Identity
First exon
Middle exons
Last exon
Sequence identity (%)
100
95
90
85
80
75
70
65
60
55
50
200-bp upstream
5' UTR
Intron
Intron
3' UTR
200-bp downstream
Position within genome
c
Sequence identity (%)
100
95
90
85
80
75
70
65
60
–15 –14 –13 –12 –11 –10 –9 –8 –7 –6 –5 –4 –3 –2 –1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Position relative to 5' splice site


b
Sequence identity (%)
100
95
90
85
80
75
70
65
60
–15 –14 –13 –12 –11 –10 –9 –8 –7 –6 –5 –4 –3 –2 –1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Position relative to translation start
d
Sequence identity (%)
100
95
90
85
80
75
70
65
60
–15 –14 –13 –12 –11 –10 –9 –8 –7 –6 –5 –4 –3 –2 –1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Position relative to 3' splice site



The human spermidine synthase gene (*SRM*) on chromosome 1, involved in the biosynthesis of polyamines, and its mouse orthologue (*Srm*) on chromosome 4. The fifth exon in the mouse gene (green) is interrupted by an intron in the human homologue.

The mouse and human genomes each seem to contain about 30,000 protein-coding genes. These refined estimates have been derived from both new evidence-based analyses that produce larger and more complete sets of gene predictions, and new *de novo* gene predictions that do not rely on previous evidence of transcription or homology. The proportion of mouse genes with a single identifiable orthologue in the human genome seems to be approximately 80%. The proportion of mouse genes without any homologue currently detectable in the human genome (and vice versa) seems to be less than 1%.

Dozens of local gene family expansions have occurred in the mouse lineage. Most of these seem to involve genes related to reproduction, immunity and olfaction, suggesting that these physiological systems have been the focus of extensive lineage-specific innovation in rodents.

Table 18 **SNPs generated by WGS sequencing**

Strain	Reads*	SNPs
129S1/SvImJ	67,974	47,279
C3H/HeJ	34,949	20,294
BALB/cByJ	19,686	11,696
Total	122,609	79,269

*Reads passing all filters: sequence quality, SSAHA-SNP program[324] and unique placement on the genome.

??? QuickTime?和
唯TIFF財烘縛｜貢乾?縛｜？
?螃粟??畫蚓|

9 December 2004

??? QuickTime?和
唯TIFF財烘縛｜貢乾?縛｜？
?螃粟??畫蚓|

Table 1 **Whole-genome assembly statistics**

Genome feature	>1 kb number	N50 length (kb)	N50 number	Largest (kb)
Contigs	98,612	36	7,486	442
Supercontigs	32,767	7,067	37	33,505

Statistics presented are for the whole-genome assembly before integration of physical mapping data. Contigs are contiguous sequences not interrupted by gaps, and supercontigs are ordered and oriented contigs including estimated gap sizes. The N50 statistic is defined as the largest length *L* such that 50% of all nucleotides are contained in contigs of size at least *L*. A total of 10,743,700 reads were included in the final assembly. Only 4.39% of the total sequencing reads presented to the assembler were not used in the final assembly.

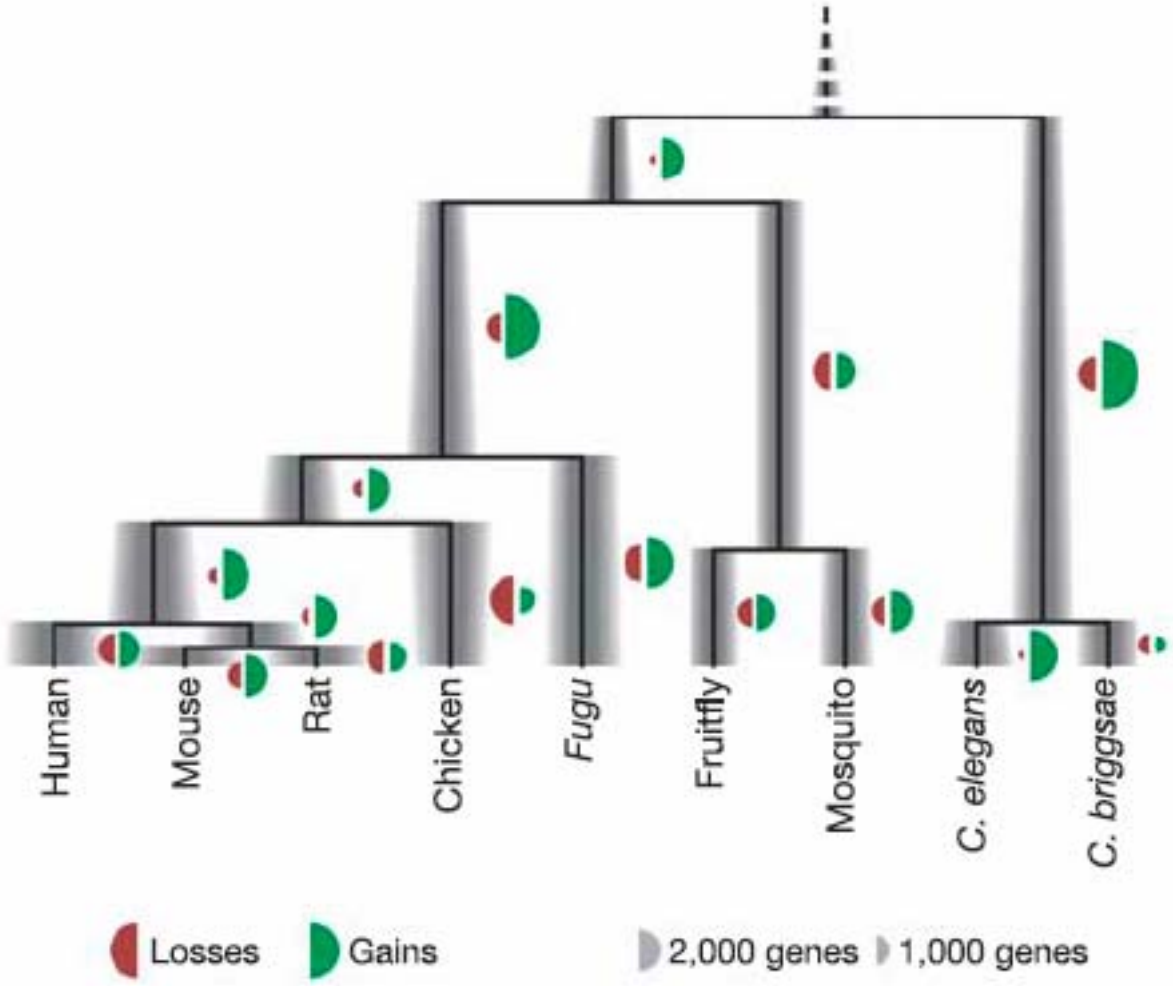


Figure 8 Approximate history of animal genomes. The gene content of ancestral

??? QuickTime?和
唯TIFF財棋縛｜ 貢乾?縛｜ ?
?螃粟??畫蚓|

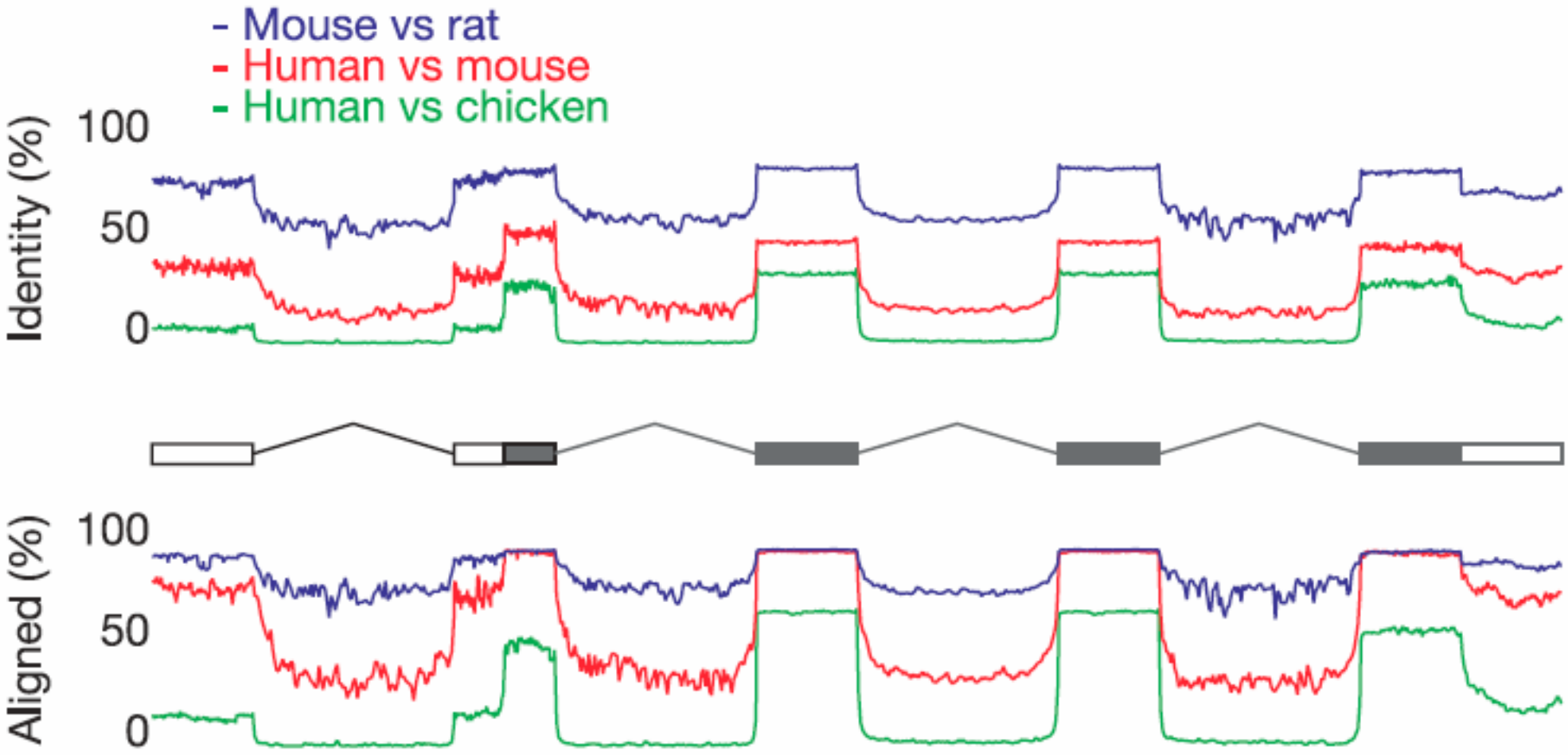


Figure 2 An idealized protein-coding gene structure showing average percentage alignment and average percentage identity (including gaps and unaligned regions) over 10,000 orthologous gene structures in either human–chicken, human–mouse or mouse–

??? QuickTime?和
唯TIFF財烘縛｜貢乾?縛｜？
?螃粟??畫蚓

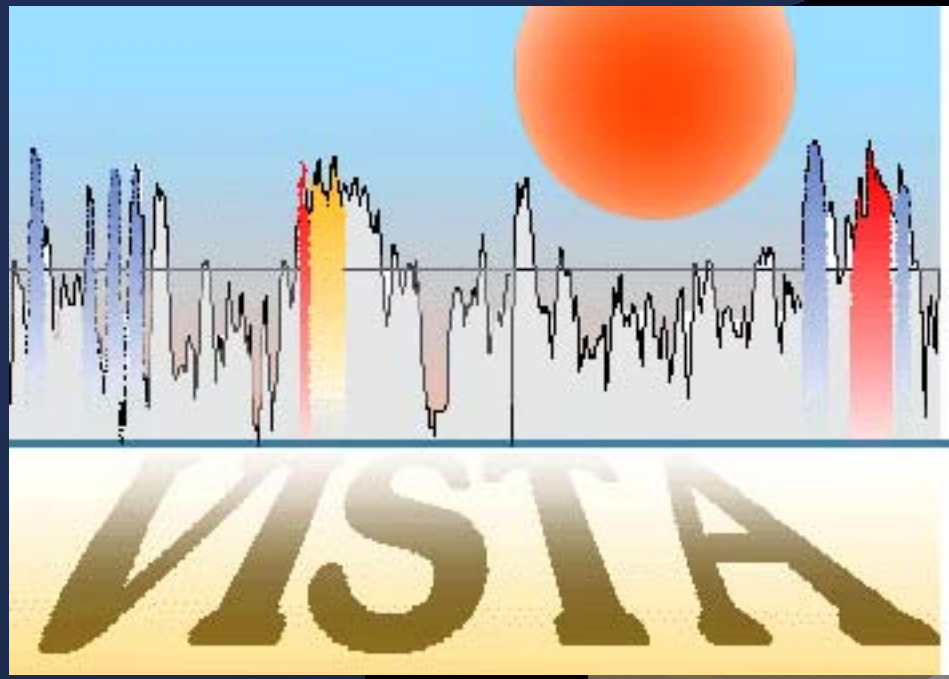
Table 7 Details of family expansions in chicken

Number of chicken genes	Number of chicken pseudogenes	Chicken:human gene ratio	Annotation
202	16	202:1	Olfactory receptors 5U1/5BF1
25	6	25:1	Immunoglobulin-like receptor CHIRs
14	3	14:1	Major histocompatibility complex, class I-like sequence
13	0	13:1	Myelin-oligodendrocyte glycoprotein
26	1	26:3	G-protein-coupled receptor 43; free fatty acid activated receptor 2
8	0	8:1	Early activation antigen CD69
6	0	6:1	T-cell surface glycoprotein CD8a-chain
5	0	5:1	Immune-associated nucleotide 6-like; guanosine triphosphatase

In-depth analysis of gene families that have been expanded at least fourfold, as revealed by automatic orthology analysis.

Table 11 **Comparison of orthologous segments**

Feature	Anchors	
	Orthologous genomic segments	Orthologous gene pairs
Number of blocks	1,068	1,009
Average size in human (Mb)	2.10	2.03
Average size in chicken (Mb)	0.78	0.76
Average human:chicken ratio	2.67	2.67
Median size in human (kb)	177	566
Median size in chicken (kb)	74	185
Median human:chicken ratio	2.39	3.06
Longest block in human (Mb)	91.5	48.5
Longest block in chicken (Mb)	38.2	19.8
Coverage per chicken chromosome (%)	71.1	66.1
Coverage per human chromosome (%)	72.7	66.8



VISUALIZATION TOOLS FOR ALIGNM

VISTA is a set of tools for comparative genomics. It was designed to visualize long sequence alignments of DNA from two or more species with annotation information. It has a clean output, allowing for easy identification of sequence similarities and differences, and is easily configurable, enabling the visualization of alignments of various lengths at different levels of resolution. Click here for instructions on how to cite VISTA.

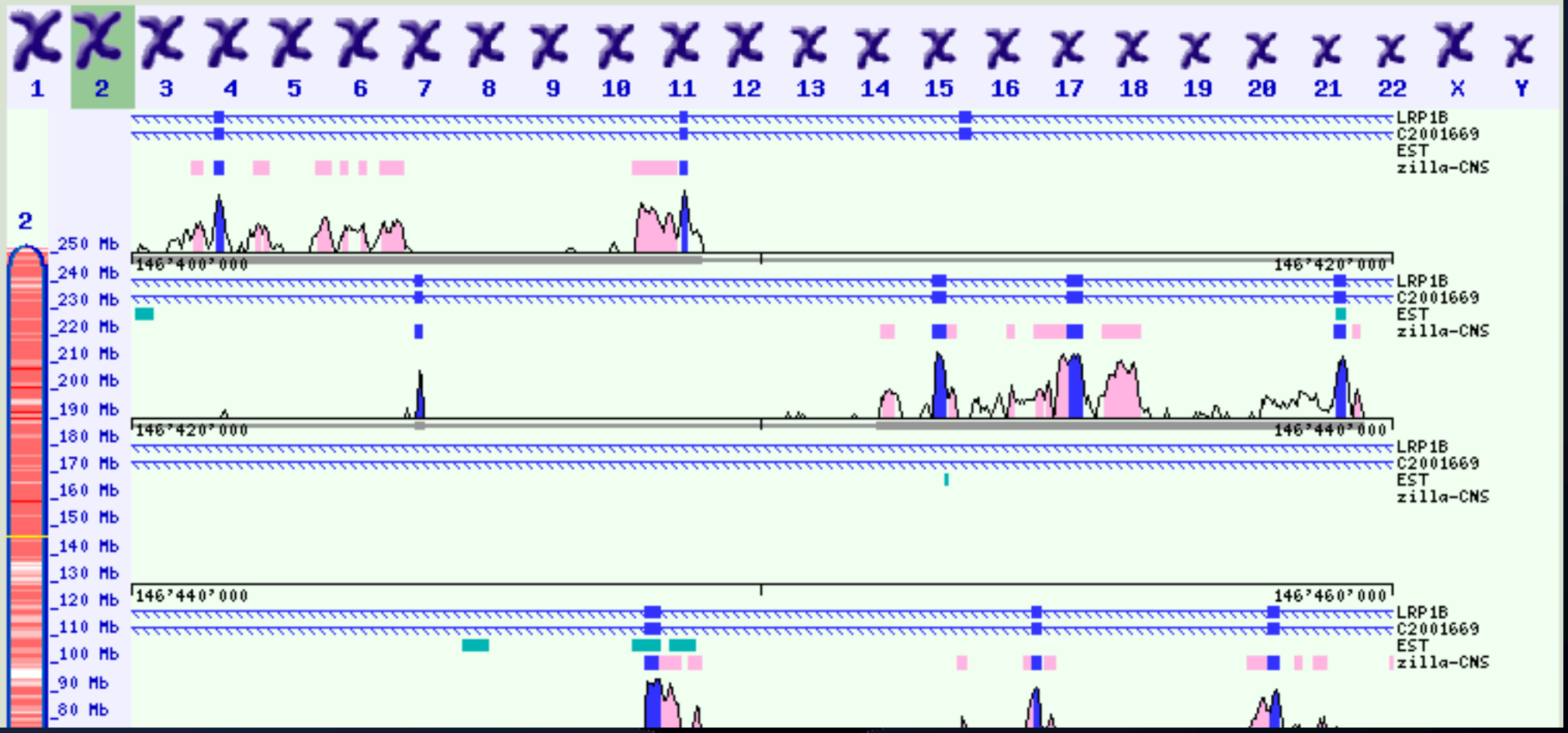
zBROWSER

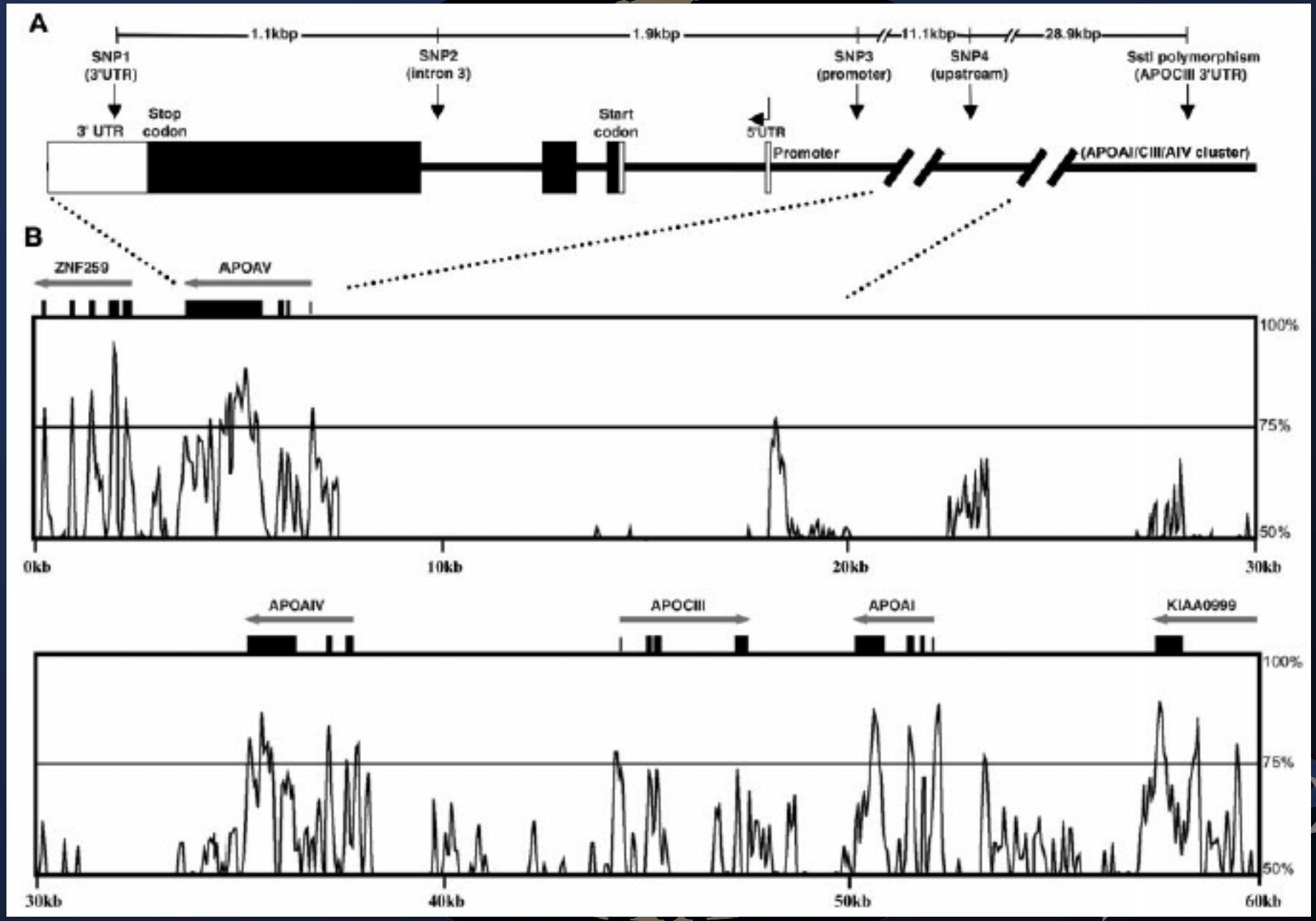
Location: chr2.146400000-146500000

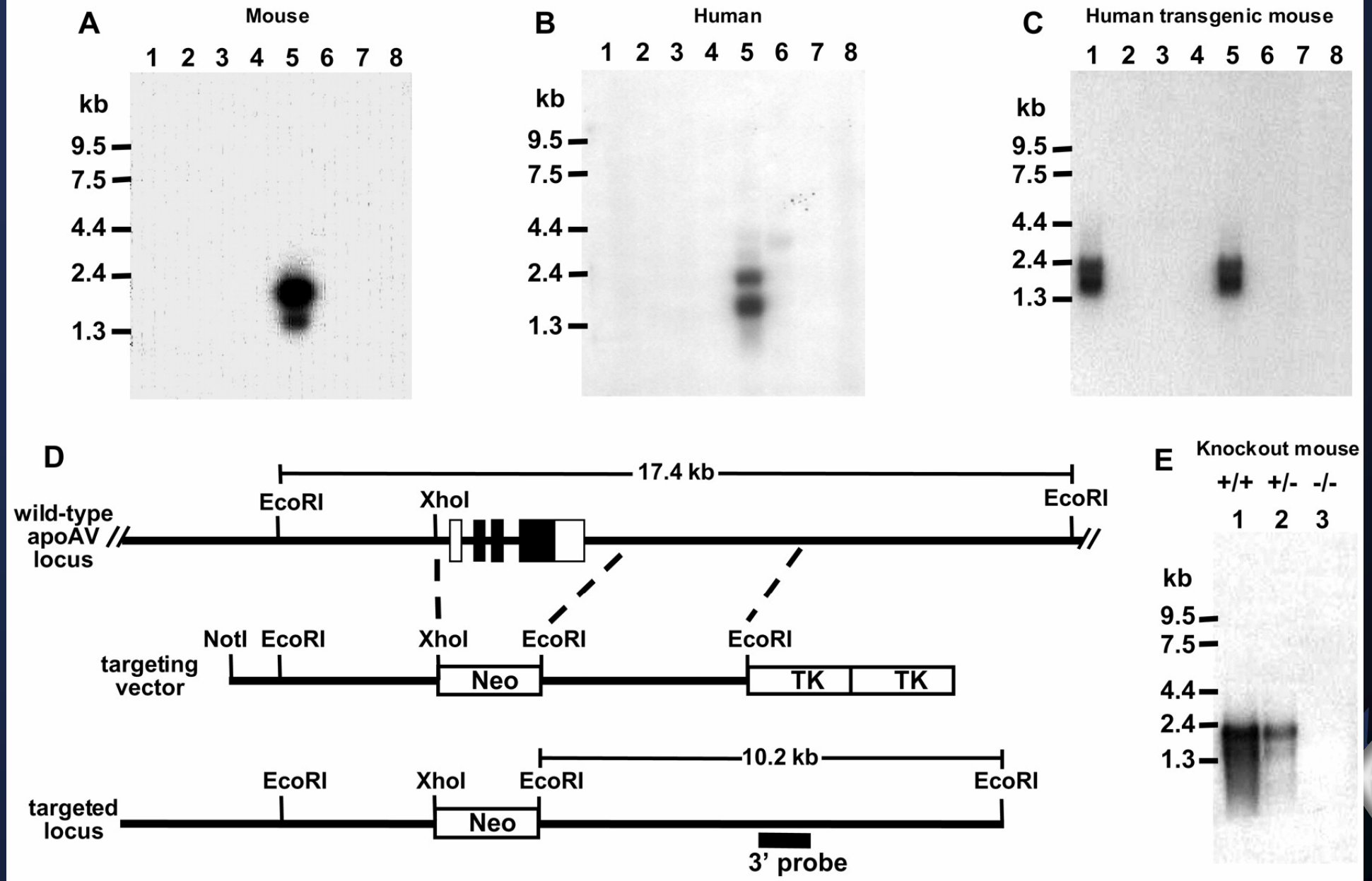
Chromosome 2. From 1464000(, to 1465000(. Go

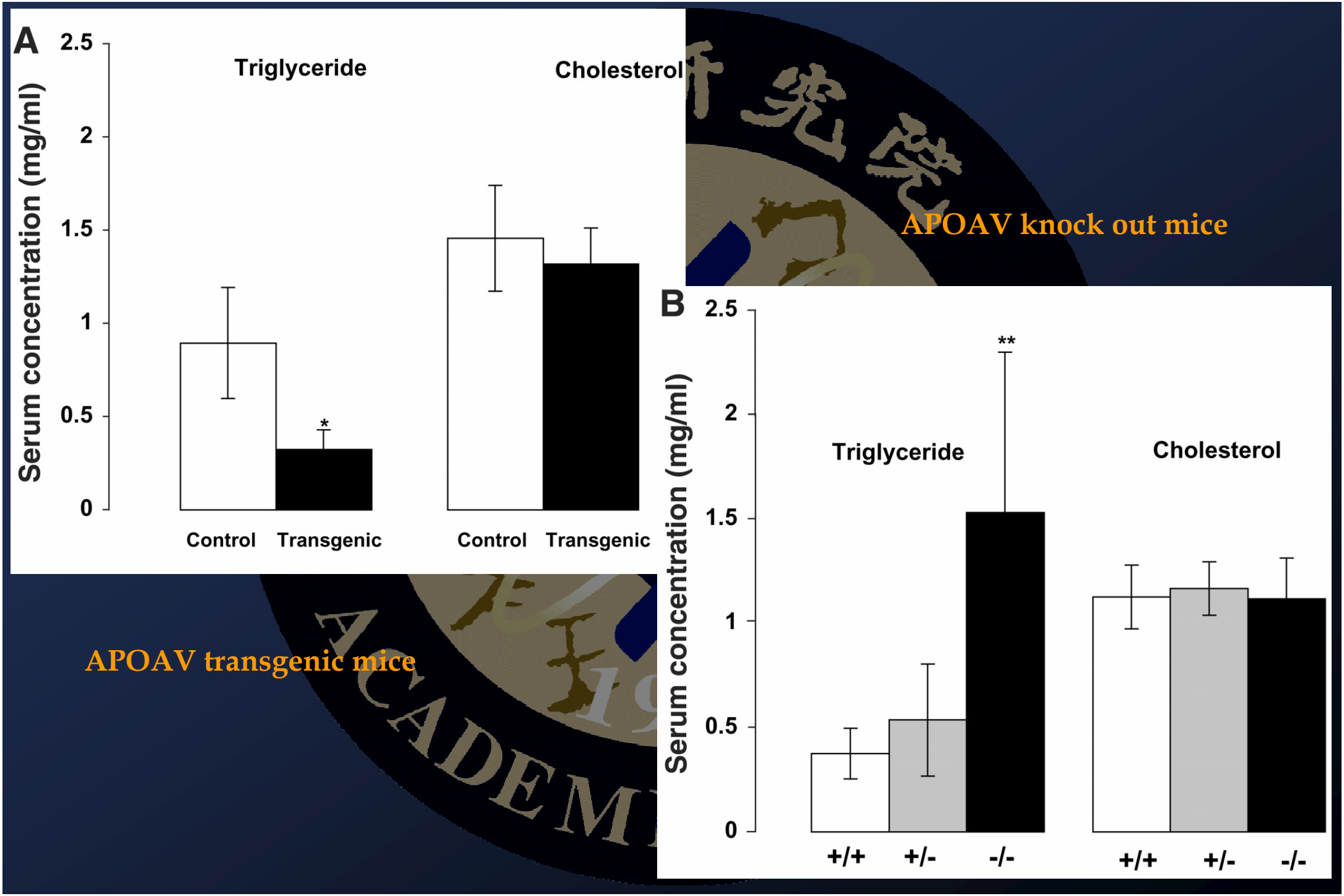
Search for RefSeq gene Gene Name Search

Note: Data fetching from central databases could take up to serveral minutes. Until that picture will not appear




A
1.1kbp
1.9kbp
11.1kbp
28.9kbp
SNP1
(3'UTR)
SNP2
(intron 3)
SNP3
(promoter)
SNP4
(upstream)
SstI polymorphism
(APOCIII 3'UTR)
3' UTR
Stop
codon
Start
codon
5'UTR
Promoter
(APOAI/CIII/AIV cluster)
B
ZNF259
APOAV
100%
75%
50%
0kb
10kb
20kb
30kb
APOAIV
APOCIII
APOAI
KIAA0999
100%
75%
50%
30kb
40kb
50kb
60kb


A
Mouse
1 2 3 4 5 6 7 8
kb
9.5
7.5
4.4
2.4
1.3
B
Human
1 2 3 4 5 6 7 8
kb
9.5
7.5
4.4
2.4
1.3
C
Human transgenic mouse
1 2 3 4 5 6 7 8
kb
9.5
7.5
4.4
2.4
1.3
D
17.4 kb
EcoRI
XhoI
EcoRI
wild-type
apoAV
locus
NotI
EcoRI
XhoI
EcoRI
EcoRI
targeting
vector
Neo
TK
TK
10.2 kb
EcoRI
XhoI
EcoRI
EcoRI
targeted
locus
Neo
3' probe
E
Knockout mouse
+/+ +/- -/-
1 2 3
kb
9.5
7.5
4.4
2.4
1.3


A
Serum concentration (mg/ml)
2.5
2
1.5
1
0.5
0
Triglyceride
Cholesterol
*
Control
Transgenic
Control
Transgenic
APOAV knock out mice
APOAV transgenic mice
B
Serum concentration (mg/ml)
2.5
2
1.5
1
0.5
0
**
Triglyceride
Cholesterol
+/+
+/-
-/-
+/+
+/-
-/-

C

Group	n	BMI	Age	TG	SNP3 1,1	SNP3 1,2	SNP3 2,2	p value
Total TG>90%	161	29.0±4.6	49.2±16.3	340.6±249.4	125 (77.6)	35 (21.7)	1 (0.6)	>0.0001
Total TG<10%	298	25.7±4.1	49.9±15.7	54.6±13.4	278 (93.3)	20 (6.7)	0	
Male TG>90%	97	28.9±3.7	50.1±16.6	336.4±155.3	67 (69.1)	29 (29.9)	1 (1.0)	>0.0001
Male TG<10%	192	26.5±3.8	54.4±15.6	57.4±12.3	184 (95.8)	8 (4.2)	0	
Female TG>90%	64	29.3±6.0	47.8±15.8	347.1±349.6	58 (90.6)	6 (9.4)	0	0.69
Female TG<10%	106	24.1±4.2	42.4±12.8	50.0±13.8	94 (88.7)	12 (11.3)	0	