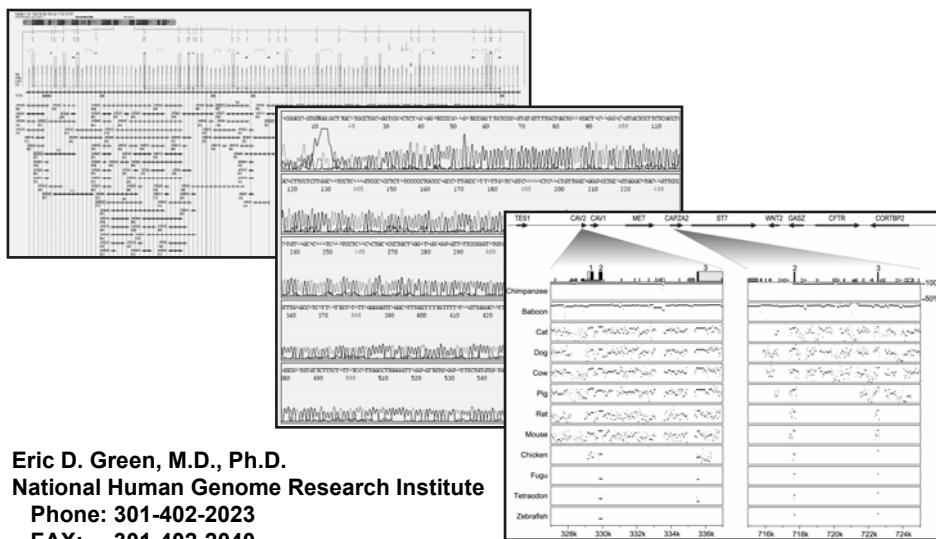
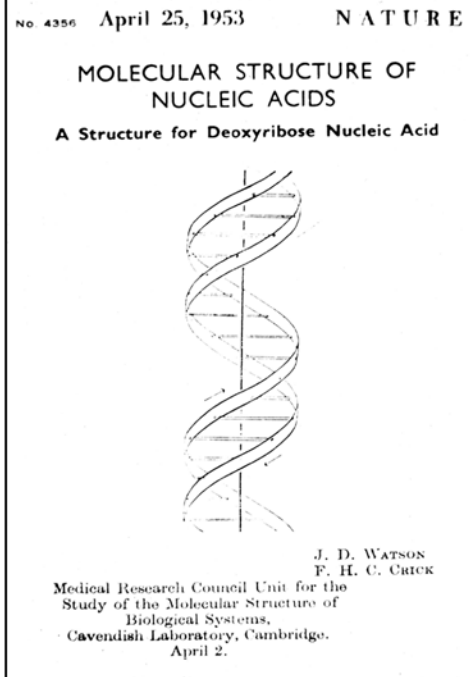
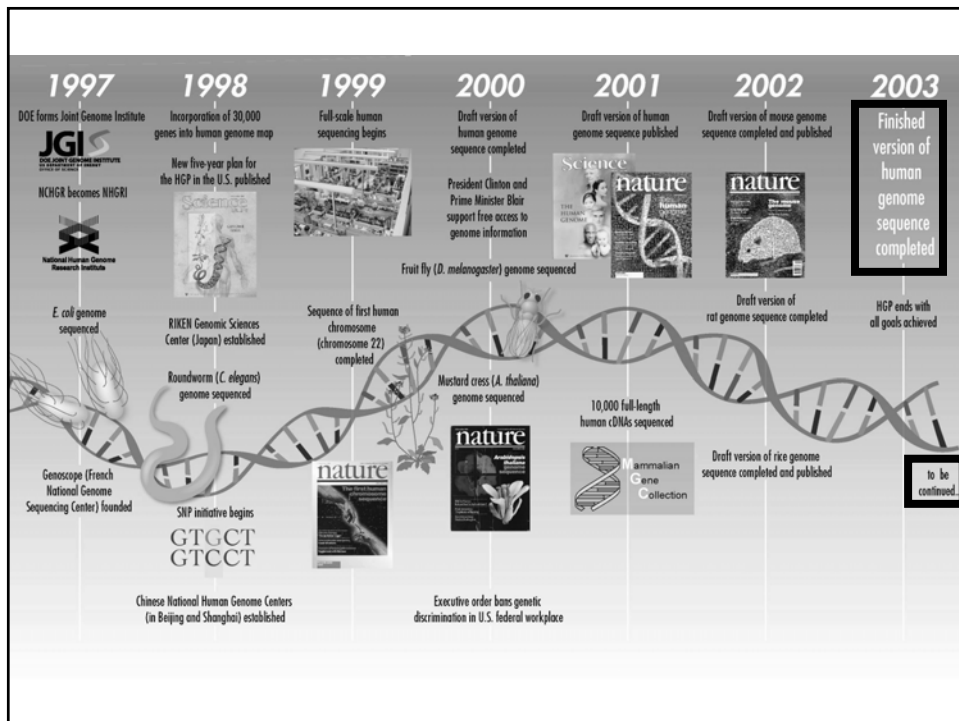
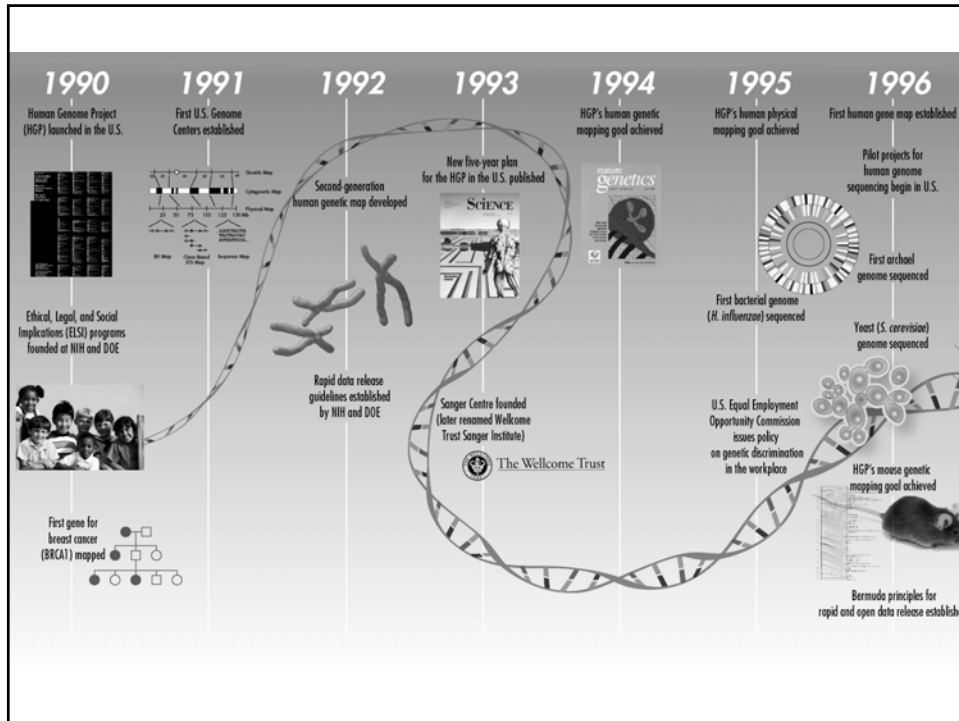


Techniques for Genome Mapping & Sequencing



Eric D. Green, M.D., Ph.D.
 National Human Genome Research Institute
 Phone: 301-402-2023
 FAX: 301-402-2040
 E-Mail: egreen@nhgri.nih.gov

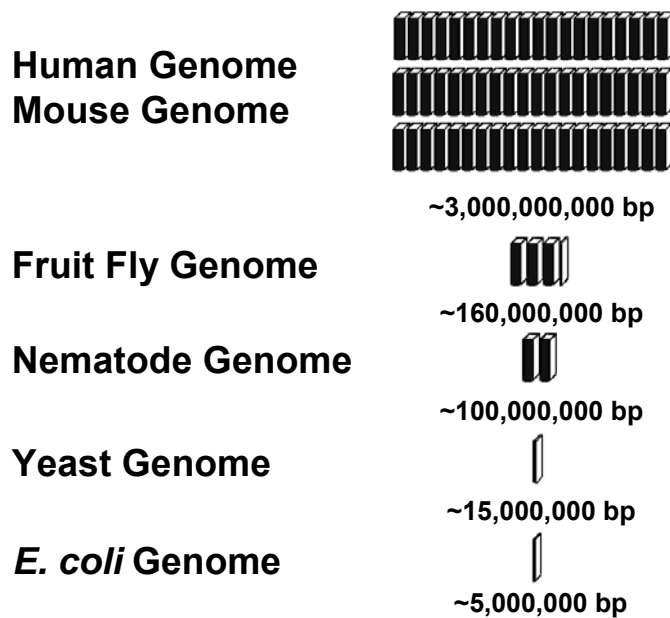


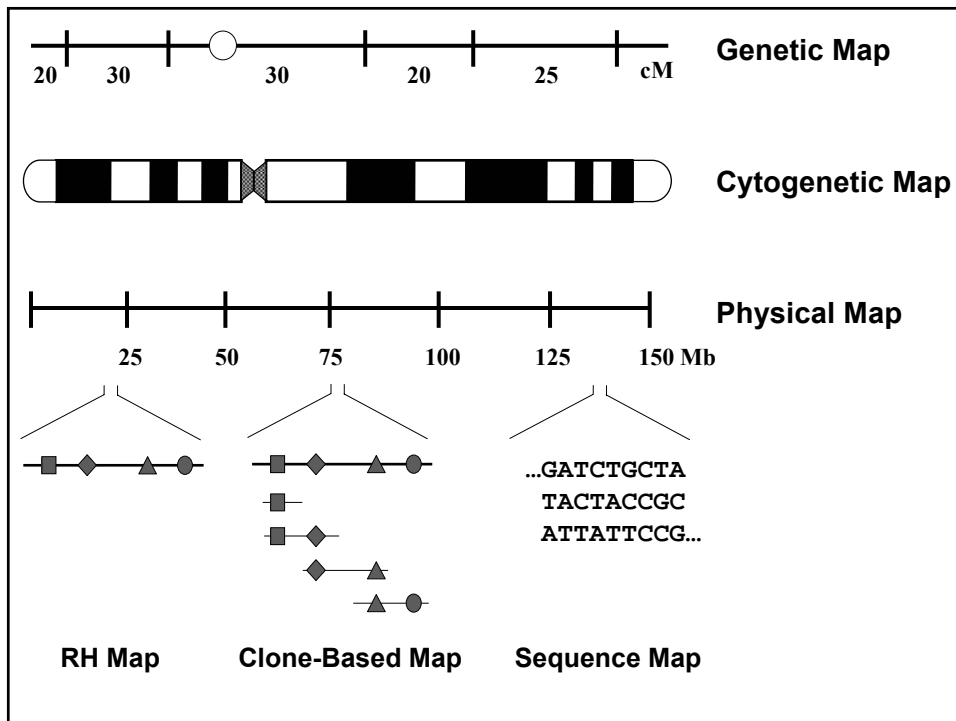
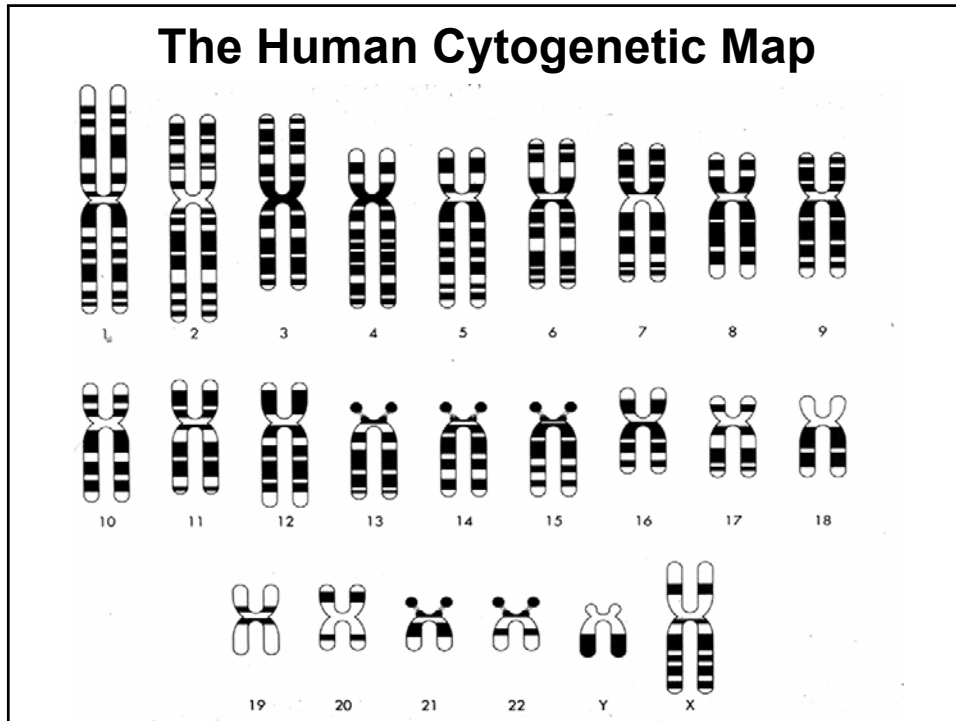


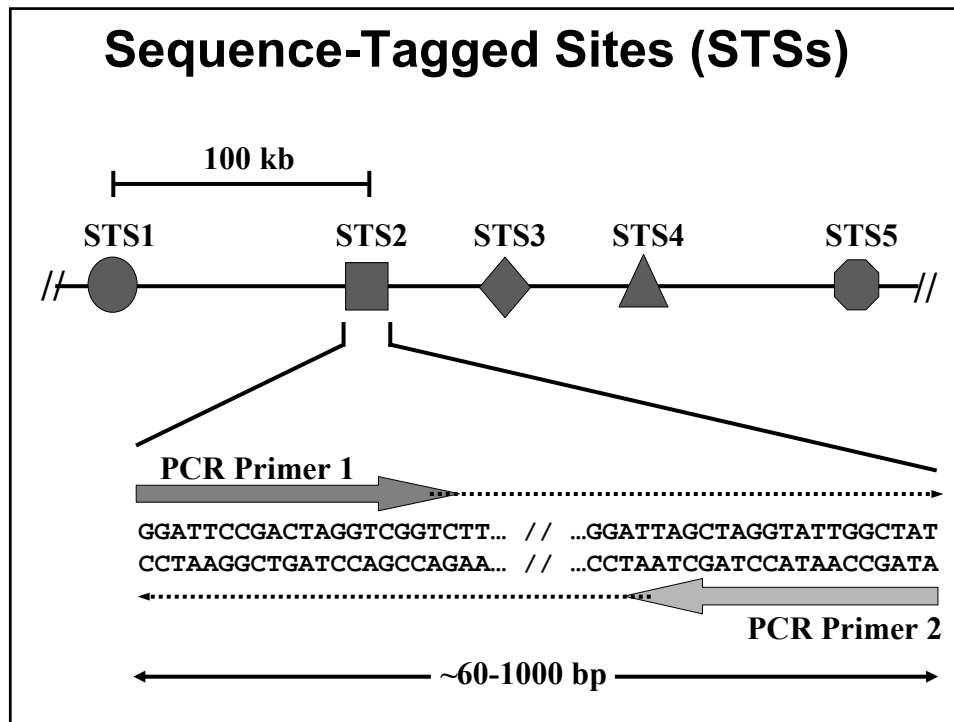
Outline

- I. Fundamentals of Genome Mapping
- II. Fundamentals of Genome Sequencing
- III. Mapping & Sequencing in the Human Genome Project... and Beyond
- IV. Comparative Sequencing

Genome Sizes

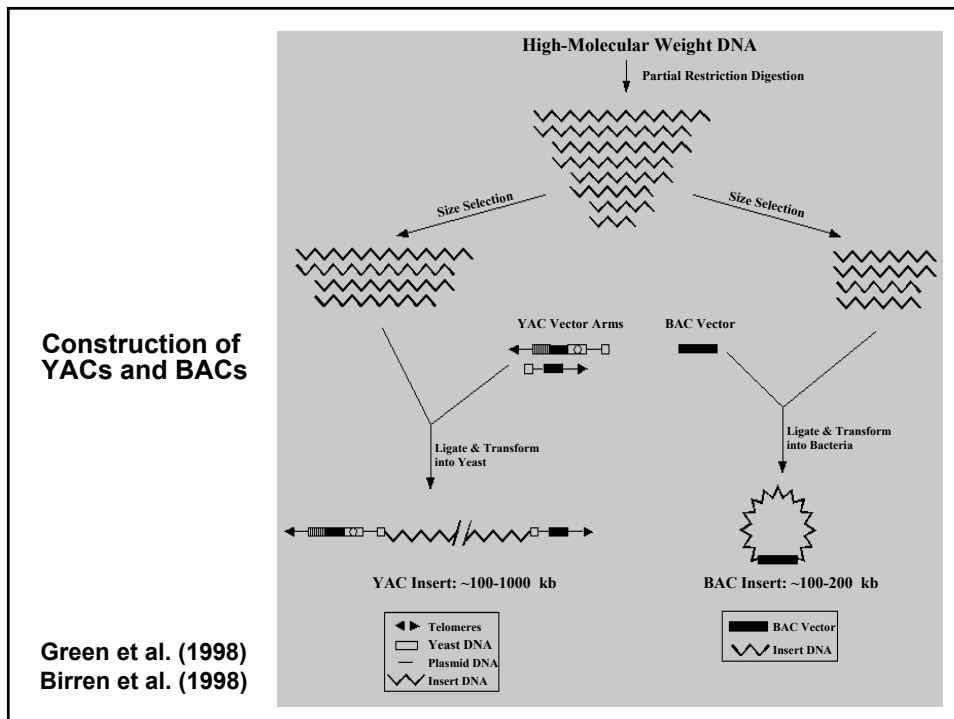
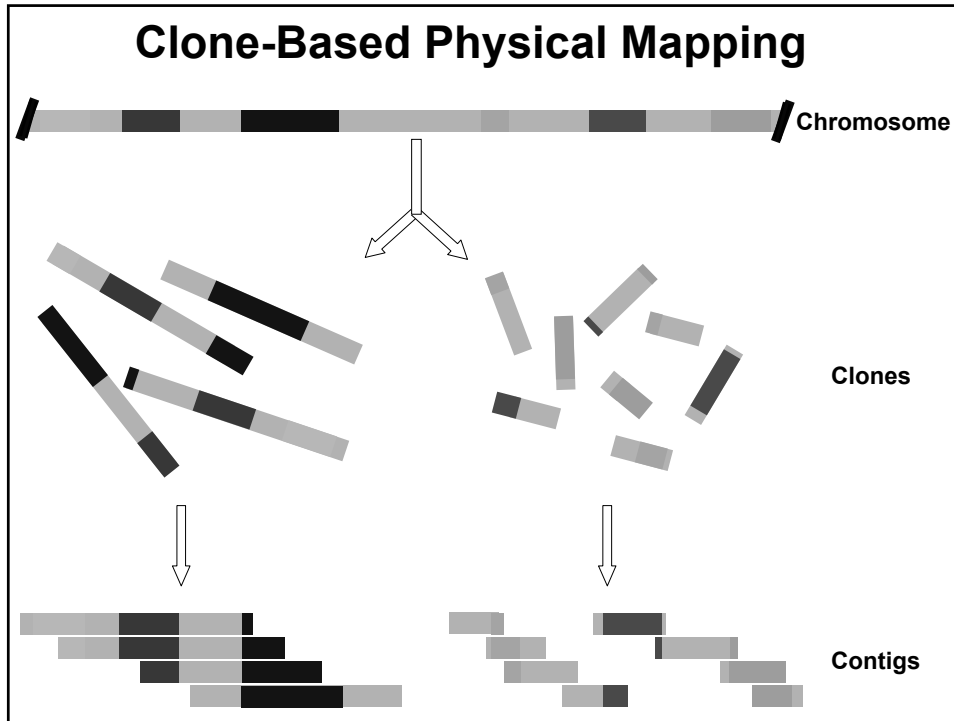


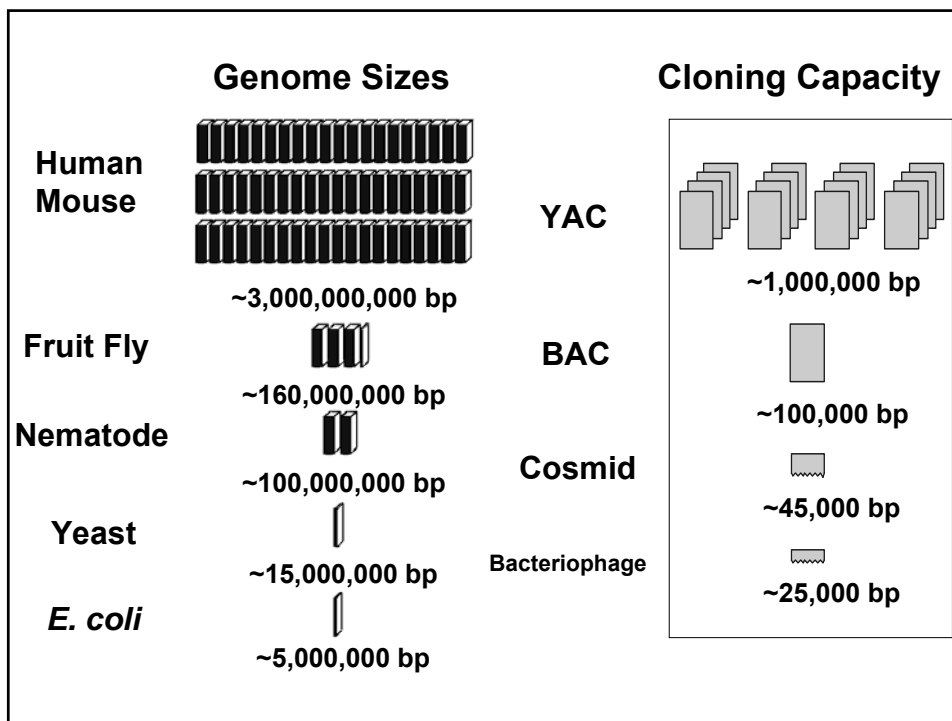




Physical Mapping: General Principles

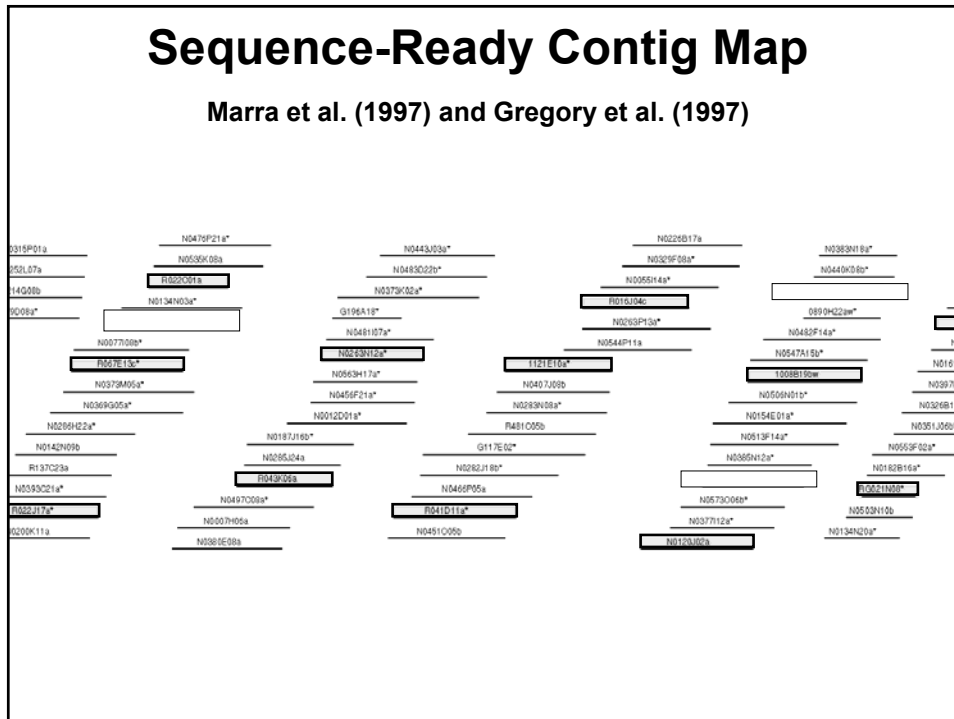
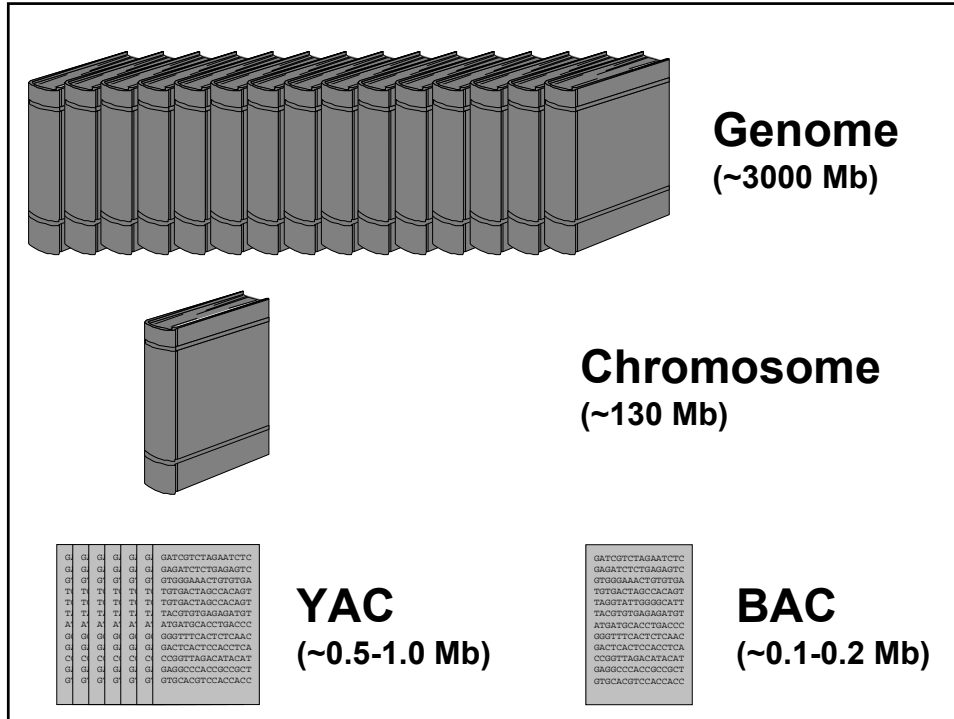
- **Importance of Physical Maps:**
 - Localization and Isolation of Genes (e.g., Positional Cloning)
 - Study of Genome Organization and Evolution
 - Framework for Genome Sequencing
- **Physical Mapping Involves Ordering Clones and/or Landmarks**
- **General Types of Physical Maps:**
 - Landmark Only (e.g., Radiation Hybrid Maps)
 - Clone-Based
 - Sequence





Bacterial Artificial Chromosomes (BACs)

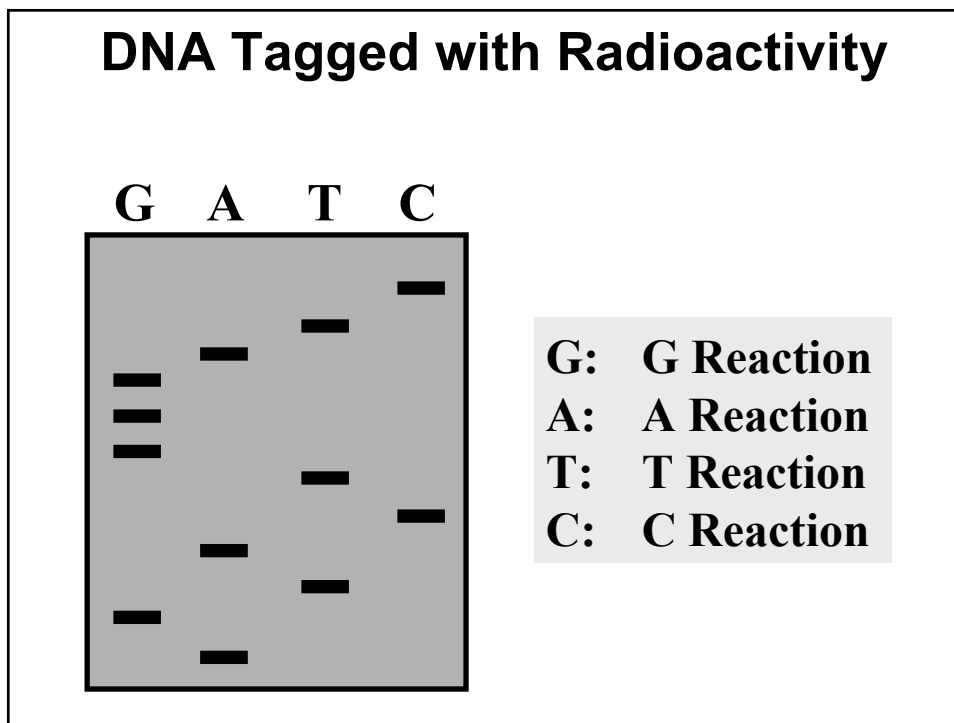
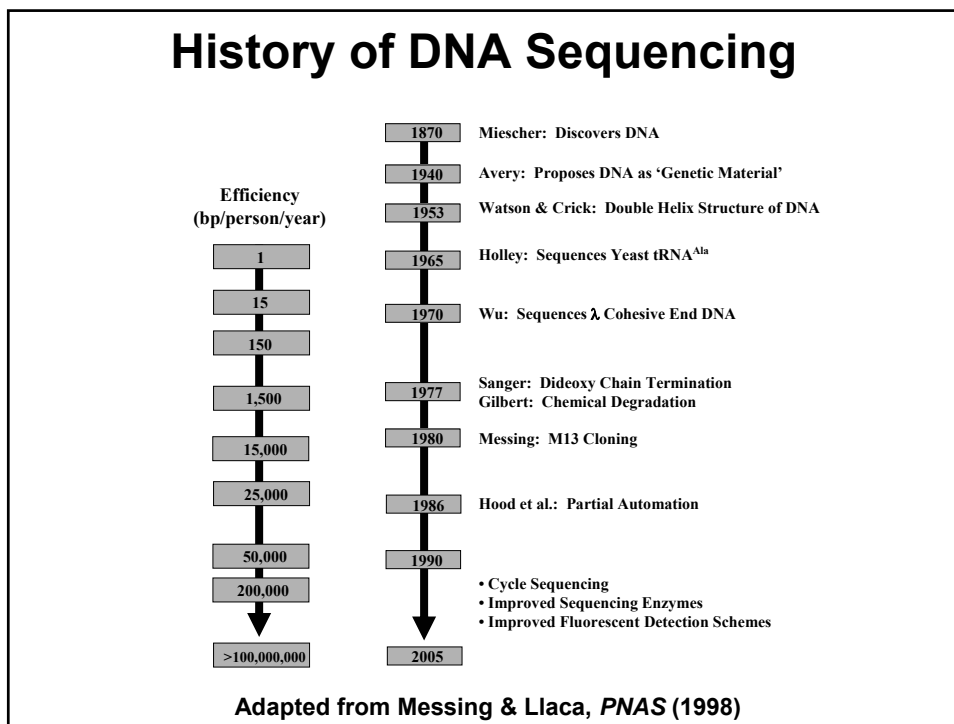
- Bacterial-Based Cloning System Developed by Shizuya et al. (1992)
- Based on the *E. coli* F Factor (Fertility Plasmid): Replication Control
- Cloned Inserts: 100-200 kb, Circular DNA
- Low Copy Number
 - Low Yields of DNA by Standard Methods
 - Reasonably Stable
- Relatively Non-Chimeric
- BAC Libraries from Many Different Species now Available (e.g., www.chori.org/bacpac)
- See Birren et al. (1998)



Physical Mapping: Future Prospects

- **Strategies for Physical Mapping are Radically Changing in the Sequence-Based Era**
- **Will Now See a Closer Interplay of Mapping and Sequencing in the Exploration of New Genomes**
- **Construction of New BAC Libraries will Allow Physical Mapping Studies of More Species' Genomes**

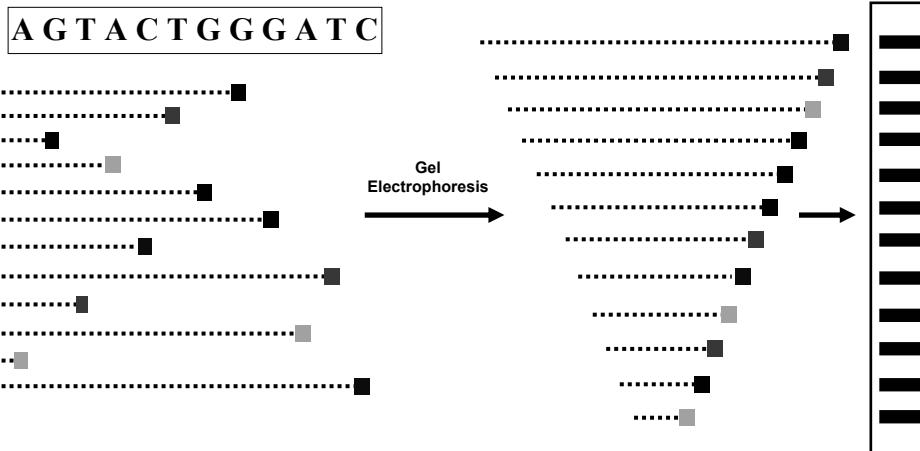
DNA Sequencing



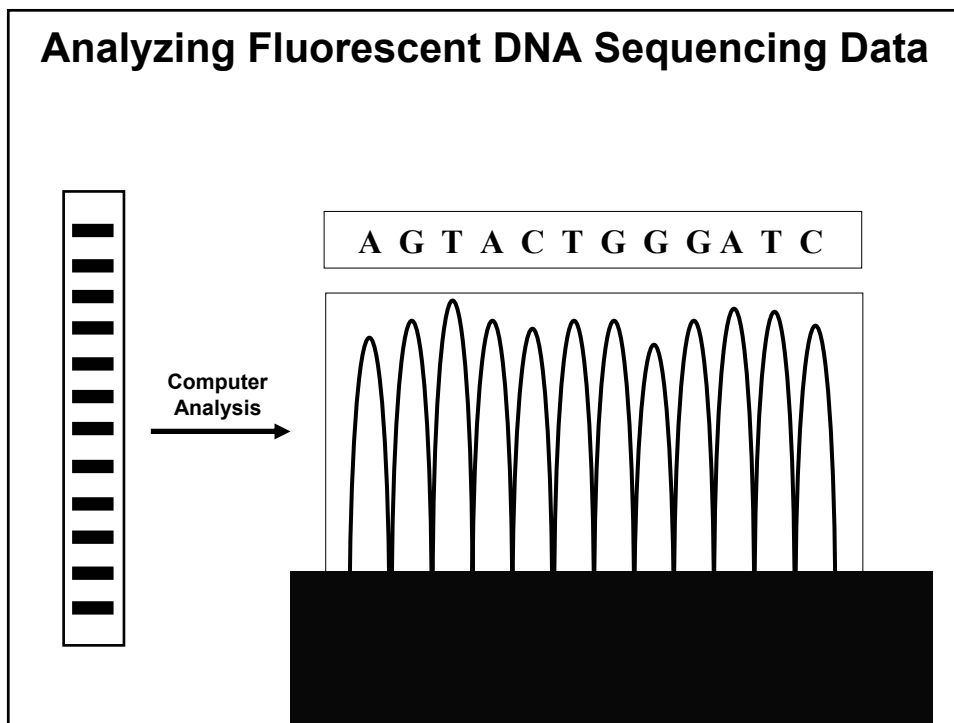
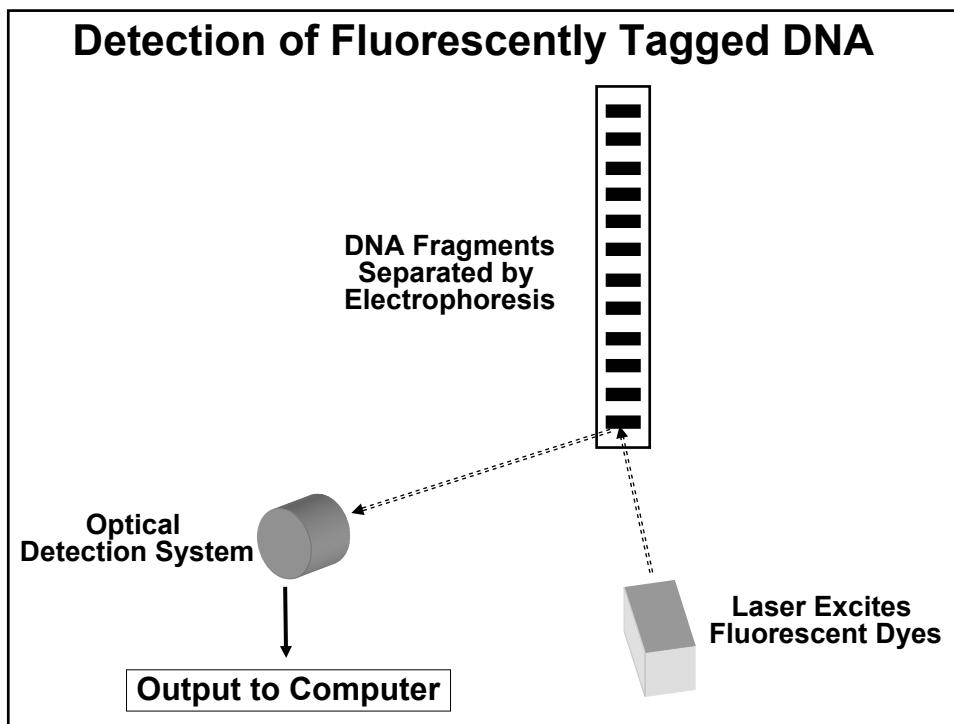
Radioactive Sequencing



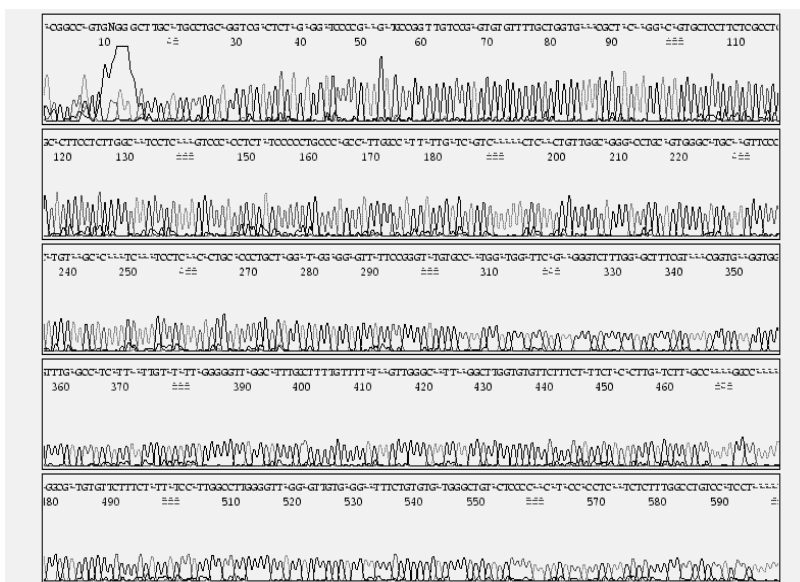
Fluorescent DNA Sequencing



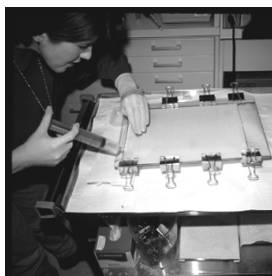
Wilson & Mardis (1997)



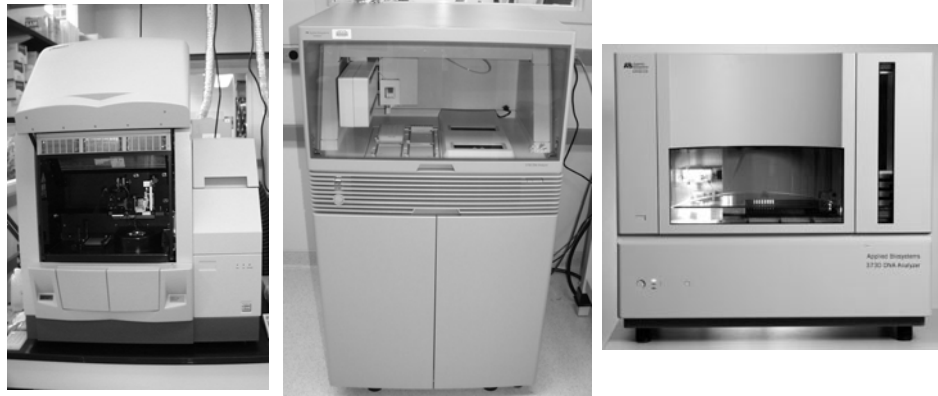
Fluorescent DNA Sequencing Results



Slab Gel-Based DNA Sequencing Instruments

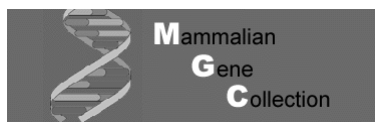


Capillary-Based DNA Sequencing Instruments



Large-Scale cDNA Sequencing

- **ESTs: Expressed-Sequence Tags**
- **SAGE: Serial Analysis of Gene Expression**
- **Full-Insert (Full-Length) cDNA Sequencing**



mgc.nci.nih.gov

Large-Scale Genomic Sequencing



Shotgun Sequencing

Wilson & Mardis (1997)
Green (2001)

Subclone Construction

```
GATGCTTAGAATTC
GAGCTCTGAGATC
CTGGAAACTGTGA
TGTGACTGACAGT
TACTGTGAGATTC
ATGATGCACTTACC
GGATTTCCTGAG
GACTCACTGACTCA
GAGGCTCACCCGCT
TTCGACTGACAGC
GATTATTACATTTA
ATCTTAGATTGACA
```

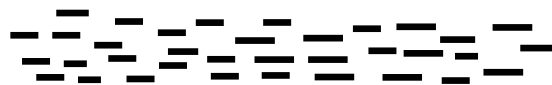
————— BAC DNA



Prepare Multiple Copies



Randomly Fragment

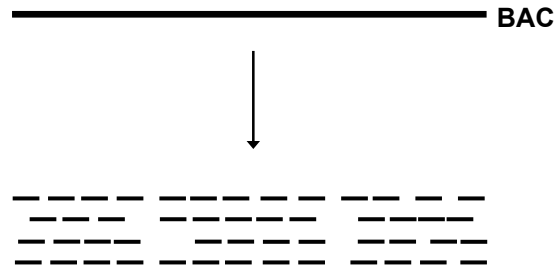


Subclone Fragments



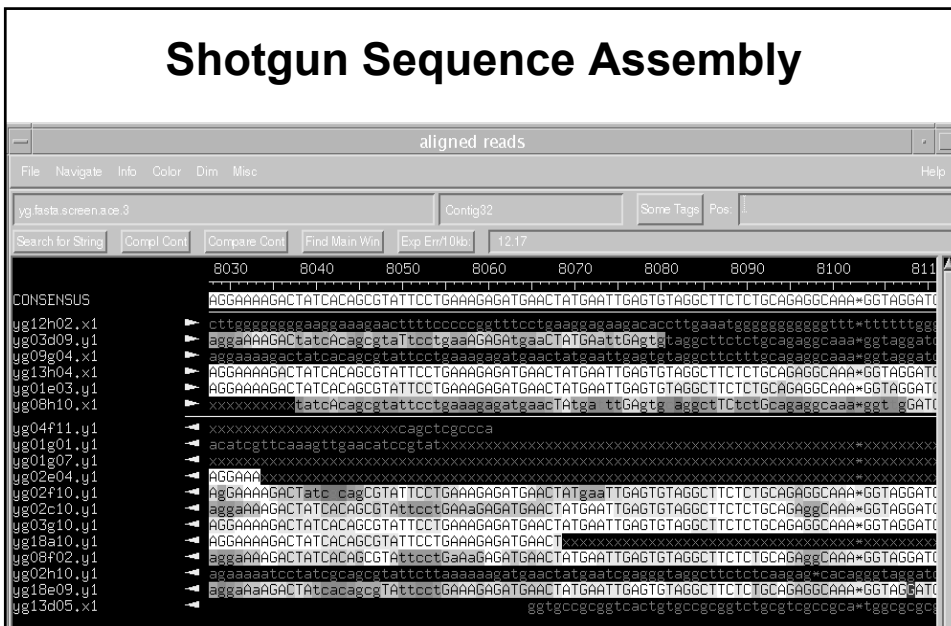
```
GA GA GA GATGCTTAGAATTC
GA GA GA GAGCTCTGAGATC
GT GT GT CTGGAAACTGTGA
TG TG TG TGTGACTGACAGT
AT AT AT TACTGTGAGATTC
GA GA GA ATGATGCACTTACC
GA GA GA GGATTTCCTGAG
GA GA GA GACTCACTGACTCA
GA GA GA GAGGCTCACCCGCT
GA GA GA TTCGACTGACAGC
GA GA GA GATTATTACATTTA
AT AT AT ATCTTAGATTGACA
```

Shotgun Sequencing Strategy

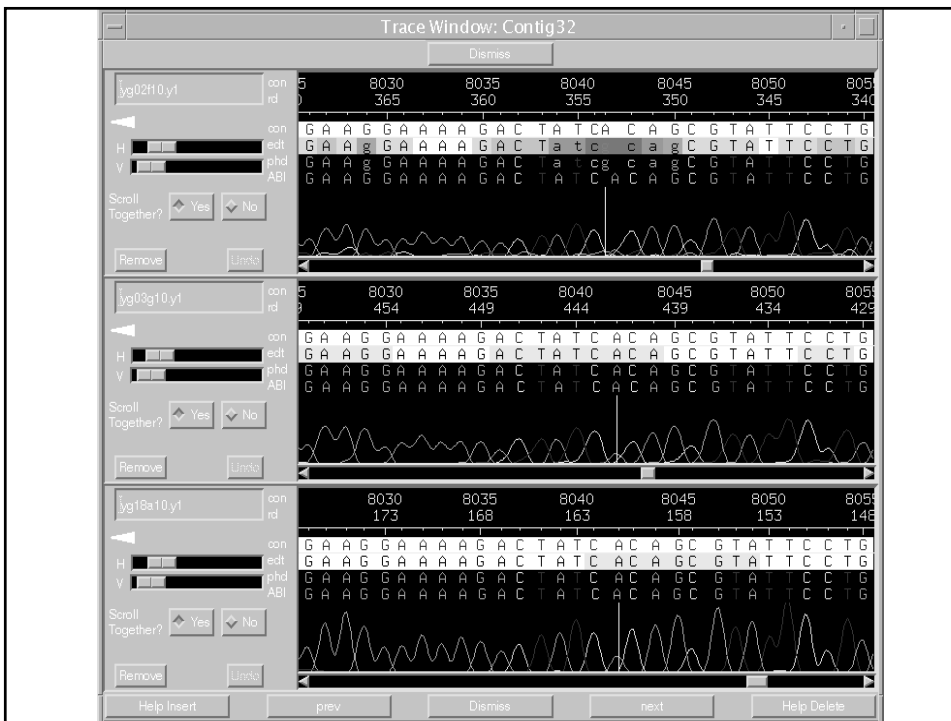


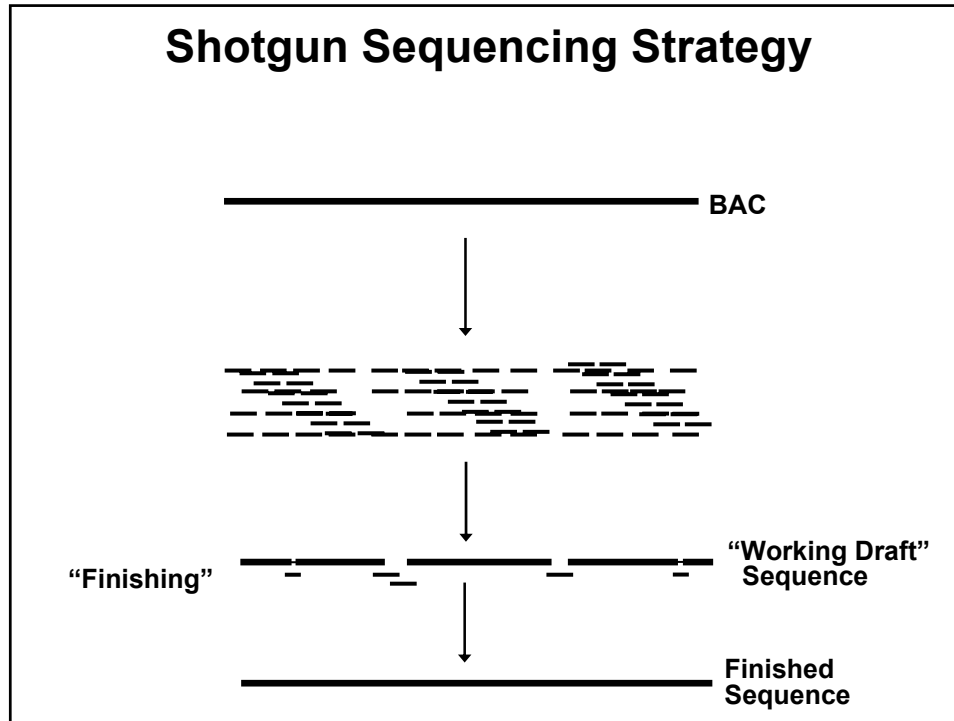
Poisson Calculations

Shotgun Sequence Assembly



“Consed” (Gordon et al., 1998)





Sequence Finishing: Resolving Ambiguities

The screenshot shows a software interface for sequence finishing. It displays two contigs, **b103h09.f1** and **b122i05.y1**, with their respective sequence data and chromatograms. The sequence data is shown in a grid format, with columns representing sequence positions and rows representing different sequencing reads. The chromatograms show the signal intensity for each base (A, T, C, G) at each position. The interface includes various controls such as **Dismiss**, **Remove**, **Insert**, **Delete**, and **Help** buttons, as well as a **Scroll Together?** option.

***** Sequence Finishing: Remains Relatively Expensive *****

Historically Significant Genome Sequencing Projects

Bacterial Genome Sequences



Welcome to the Comprehensive Microbial Resource (CMR) Home Page

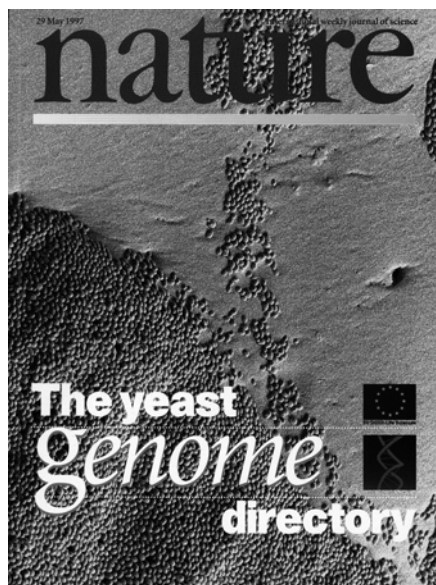
Funded by: The U.S. Department of Energy and The National Science Foundation

The Comprehensive Microbial Resource (CMR) is a tool that allows the researcher to access all of the bacterial genome sequences completed to date. For each genome not sequenced at TIGR two kinds of annotation are displayed: the Primary annotation taken from the genome sequencing center and the TIGR annotation generated by an automated annotation process at TIGR. Use the CMR to access information on all of the bacterial genomes or any subset of them.

The CMR is fully described in: J.D. Peterson, L.A. Umayam, T.M. Dickinson, E.K. Hickey and O. White. The Comprehensive Microbial Resource. Nucleic Acids Research, 29:1 (2001), 123-125.

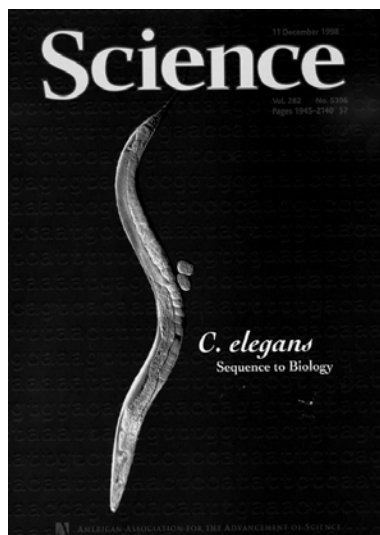
www.tigr.org

First Eukaryotic Genome Sequence



Nature 387:1-105, 1997

First Animal Genome Sequence

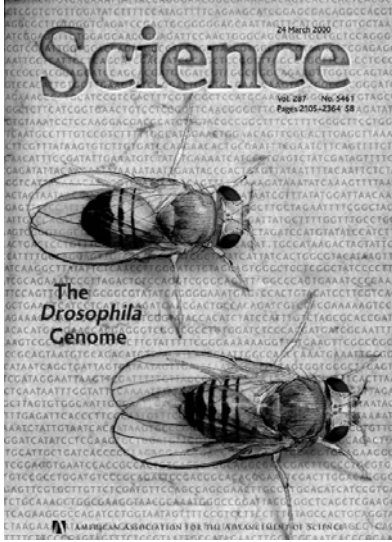


Genome Sequence of the Nematode *C. elegans*:
A Platform for Investigating Biology

The *C. elegans* Sequencing Consortium*

Science 282:1012-2018, 1998

Second Animal Genome Sequence



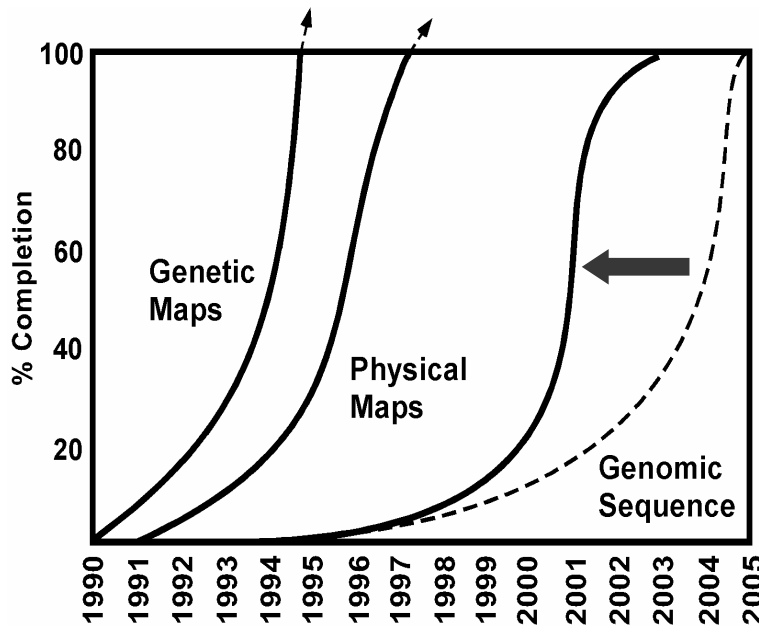
THE DROSOPHILA GENOME
REVIEW

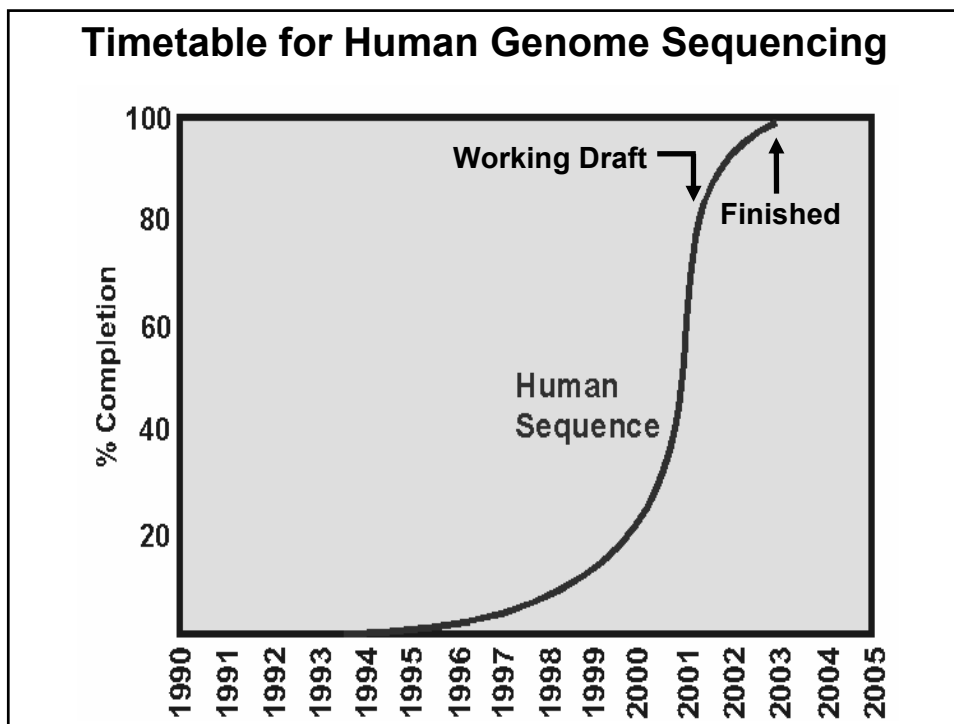
The Genome Sequence of *Drosophila melanogaster*

Mark D. Adams,^{1*} Susan E. Celniker,² Robert A. Holt,¹ Cheryl A. Evans,¹ Jeannine D. Gocayne,¹ Peter G. Amanatides,¹ Steven E. Scherer,¹ Peter W. Li,¹ Roger A. Hoskins,¹ Richard F. Gallie,¹ René A. George,² Suzanna E. Lewis,¹ Stephen Richards,¹ Michael Ashburner,¹ Scott N. Henderson,¹ Granger G. Sutton,¹ Jennifer R. Wortman,¹ Mark D. Vandell,¹ Qing Zhang,¹ Lin X. Chen,¹ Rhonda C. Brandon,¹ Yu-Hui C. Rogers,¹ Robert G. Blazaj,¹ Mark Champe,¹ Barret D. Pfeiffer,¹ Kenneth H. Wan,¹ Clara Doyle,¹ Evan G. Baxter,¹ Gregg Helt,¹ Catherine R. Nelson,¹ George L. Gabor Miklos,¹ Joseph F. Abril,¹ Anna Aghayani,¹ Hai-Jun An,¹ Cynthia Andrews-Fennko,¹ Danita Baldwin,¹ Richard M. Balow,¹ Anand Basu,¹ James Baxterdale,¹ Leyla Bayraktaroglu,¹ Ellen M. Beasley,¹ Karen Y. Reeson,¹ P. V. Benos,¹ Benjamin P. Berman,¹ Deepali Bhandari,¹ Slave Bolshakov,¹ Dana Borokov,¹ Michael R. Botchan,¹ John Bock,¹ Peter Brokstein,¹ Philippe Brotnier,¹ Kenneth C. Burtis,¹ Dana A. Busam,¹ Heather Butler,¹ Edouard Cadieu,¹ Angela Center,¹ Ishwar Chandra,¹ J. Michael Cherry,¹ Simon Cawley,¹ Carl Dahlke,¹ Lionel B. Davenport,¹ Peter Davies,¹ Beatriz de Pablos,¹ Arthur Delcher,¹ Zuoming Deng,¹ Anne Deslattes Hays,¹ Ian Dew,¹ Suzanne H. Dietz,¹ Kristina Dodson,¹ Lisa E. Doup,¹ Michael Drenth,¹ Shannon Dugan-Rocha,¹ Boris C. Dunkov,¹ Patrick Dunn,¹ Kenneth J. Durbin,¹ Carlos C. Evangelista,¹ Concepcion Ferraz,¹ Steven Ferreira,¹ Wolfgang Fleischmann,¹ Carl Fodier,¹ Andrei E. Gabrielian,¹ Nehe S. Garg,¹ William M. Gelbert,¹ Ken Glasser,¹ Anna Glöckl,¹ Fangcheng Gong,¹ J. Harley Gorrell,¹ Zhiping Gu,¹ Ping Guan,¹ Michael Harris,¹ Nomi L. Harris,¹ Damon Harvey,¹ Thomas J. Heiman,¹ Judith R. Hernandez,¹ Jarrett Houck,¹ Damon Hostin,¹ Kathryn A. Houston,¹ Timothy J. Howland,¹ Ming-Hid Wei,¹ Chinyere Ibegwam,¹ Mena Jalali,¹ Francis Kalush,¹ Gary H. Karpen,¹ Zhaodl Ke,¹ James A. Kennison,¹ Karen A. Ketchum,¹ Bruce E. Kimmel,¹ Chinnappa D. Kodira,¹ Cheryl Kraft,¹ Seul Kravitz,¹ David Kulp,¹ Zhongwei Lai,¹ Paul Lasko,¹ Yiding Lei,¹ Alexander A. Levitsky,¹ Jianlin Li,¹ Zhanyu Li,¹ Yong Liang,¹ Xiaoying Lin,¹ Xiangjun Liu,¹ Bettina Mattel,¹ Tina C. McIntosh,¹ Michael P. McLeod,¹ Duncan McPherson,¹ Genady Mykhalov,¹ Natalia V. Milhina,¹ Clark Moberly,¹ Joe Morris,¹ Ali Hoshrevi,¹ Stephen H. Mount,¹ Mei Moy,¹ Brian Murphy,¹ Lee Murphy,¹ Donna M. Murray,¹ David L. Nelson,¹ David R. Nelson,¹ Keith A. Nelson,¹ Katherine Nixon,¹ Deborah R. Nusser,¹ Joanne M. Paclab,¹ Michael Palazzolo,¹ Gjang S. Pittman,¹ Sue Pan,¹ John Pollard,¹ Vinita Puri,¹ Martin G. Reese,¹ Knut Reinert,¹ Karin Remington,¹ Robert D. C. Saunders,¹ Frederick Scheeler,¹ Hua Shen,¹ Biliang Christopher Shou,¹ Inga Sidén-Kiamos,¹ Michael Simpson,¹ Marian P. Skupski,¹ Tom Smith,¹ Eugene Spler,¹ Allan C. Spreading,¹ Mark Stapleton,¹ Renee Strong,¹ Eric Sun,¹ Robert Svirskii,¹ Cyndee Tector,¹ Russell Turner,¹ Eli Venter,¹ Alhui H. Wang,¹ Xin Wang,¹ Zhen-Yuan Wang,¹ David A. Wassarman,¹ George H. Weinstock,¹ Jean Weissenbach,¹ Sherita M. Williams,¹ Trevor Woodage,¹ Kim C. Worley,¹ David Wu,¹ Song Yang,¹ Q. Alison Yao,¹ Jian Ye,¹ Ru-Fang Yeh,¹ Jaytree S. Zaveri,¹ Ming Zhu,¹ Guangren Zhang,¹ Qi Zhao,¹ Lianheng Zheng,¹ Xiangqun H. Zheng,¹ Fei N. Zhong,¹ Wenyan Zhong,¹ Xiaojun Zhou,¹ Shaoping Zhu,¹ Xiaohong Zhu,¹ Hamilton O. Smith,¹ Richard A. Gibbs,¹ Eugene W. Myers,¹ Gerald M. Rubin,¹ J. Craig Venter¹

Science 287:2185-2195, 2000



Revised Timetable for Human Genome Sequencing



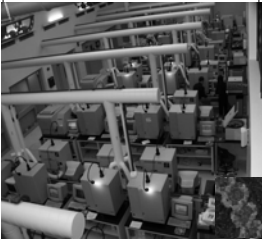


Human Genome Sequencing Centers

Genome Sequencing Center
Washington University
School of Medicine
St. Louis, MO USA






**Whitehead Institute/MIT
Genome Sequencing Center**




Baylor College of Medicine



HGSC
HUMAN GENOME SEQUENCING CENTER

JGI
JOINT GENOME INSTITUTE



The Sanger Centre

Human Genome Sequencing Centers



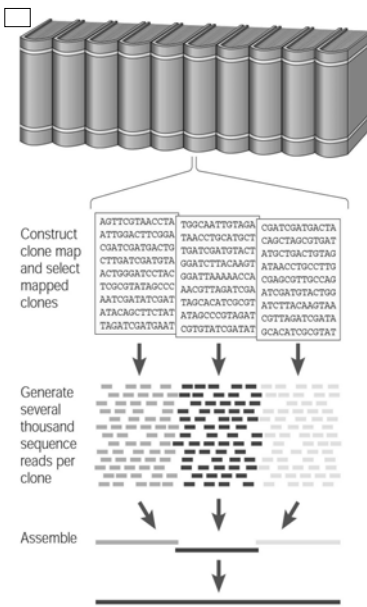
June, 2000 Announcement



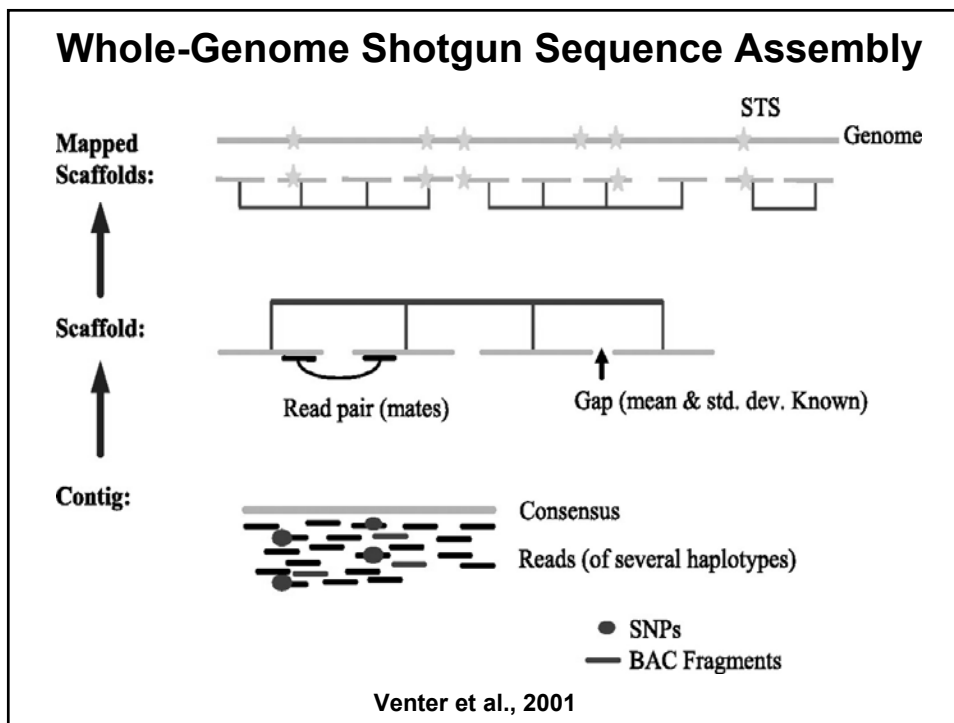
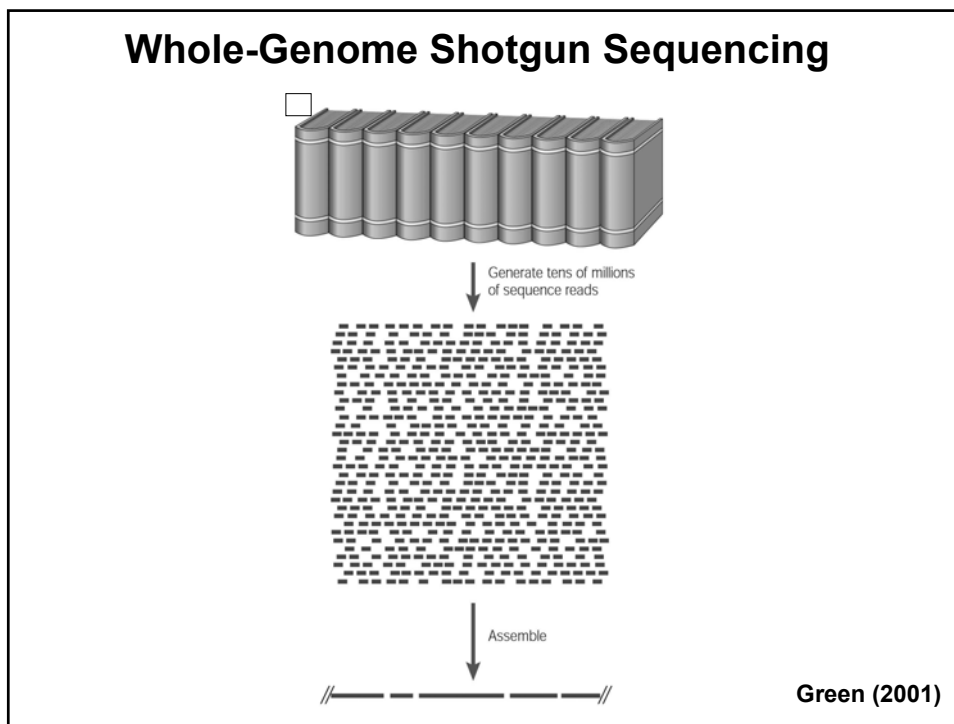
February, 2001 Publications



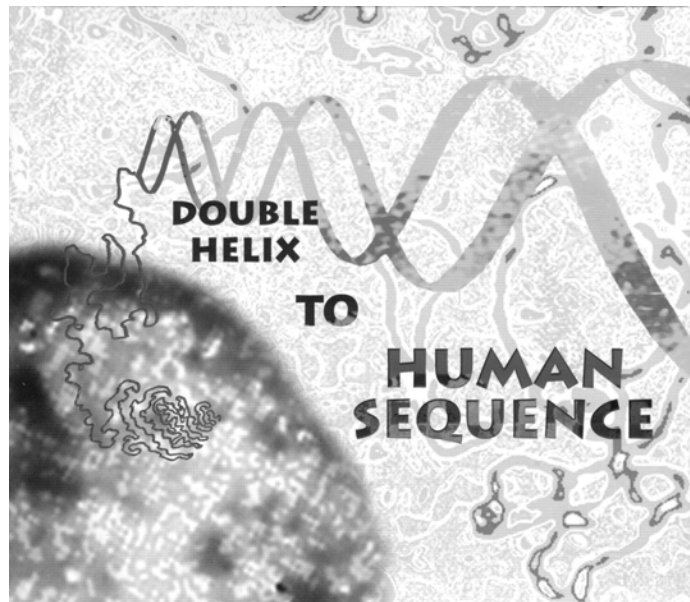
BAC-by-BAC Shotgun Sequencing



Green (2001)



April, 2003 Completion



International Human Genome Sequencing Consortium



- 6 Countries
- 20 Sequencing Centers
- 1000's of Individuals
- ~1,000 bases per second, 24 hours per day, 7 days per week



108TH CONGRESS
1ST SESSION

S. CON. RES. 10

Designating April 2003 as "Human Genome Month" and April 25 as "DNA Day".

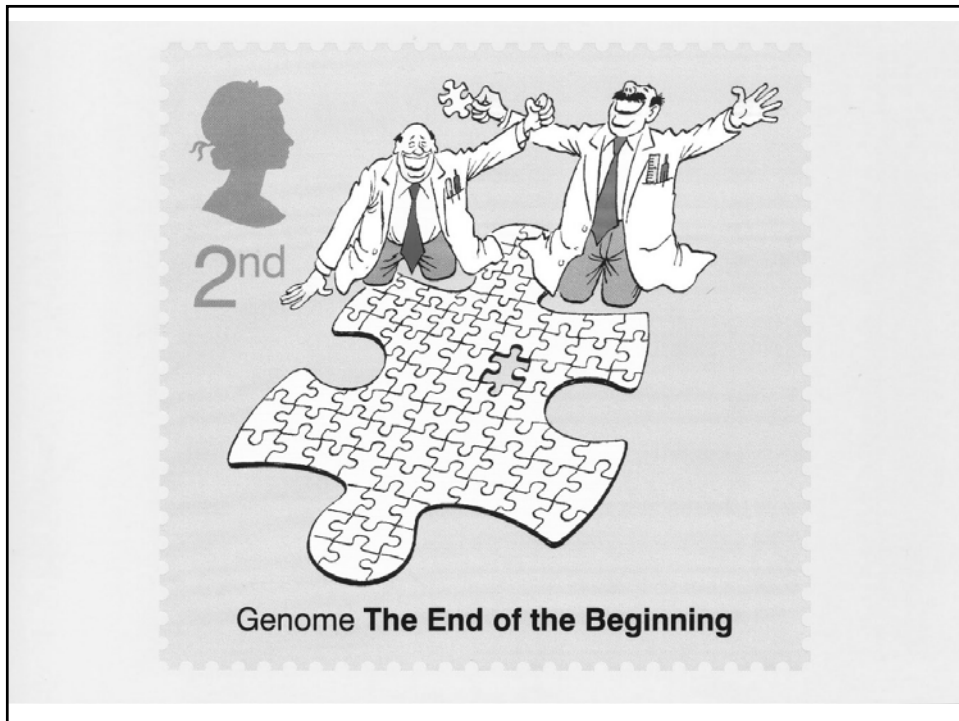
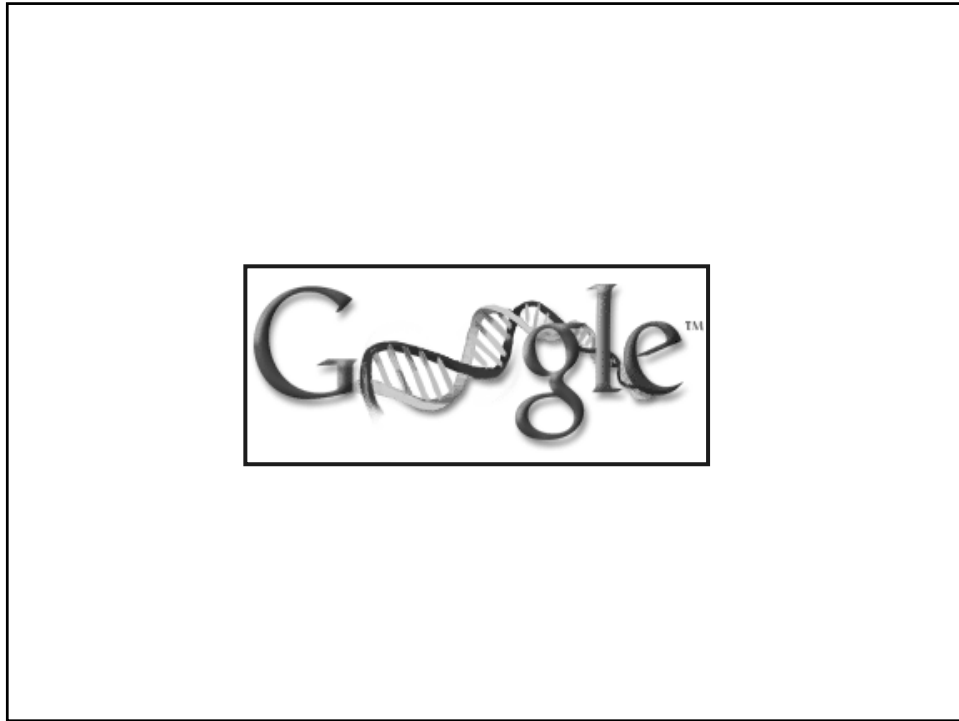
IN THE SENATE OF THE UNITED STATES

FEBRUARY 27, 2003

Mr. GREGG (for himself, Mr. KENNEDY, Ms. SNOWE, and Mr. DASCHLE) submitted the following concurrent resolution; which was considered and agreed to

CONCURRENT RESOLUTION

Designating April 2003 as "Human Genome Month" and April 25 as "DNA Day".



October, 2004 Publication

21 October 2004

nature

Tetraodon to human
Evolutionary history in genome sequences

General relativity
Did the orbit move for you?

The human genome
Going the last mile

Antibiotics crisis
Market forces fail to deliver

Medical ethics
Choosing deafness

naturejobs think Finland

articles

Finishing the euchromatic sequence of the human genome

International Human Genome Sequencing Consortium*

* A list of authors and their affiliations appears in the Supplementary Information.

The sequence of the human genome encodes the genetic instructions for human physiology, as well as rich information about human evolution. In 2001, the International Human Genome Sequencing Consortium reported a draft sequence of the euchromatic portion of the human genome. Since then, the international collaboration has worked to convert this draft into a genome sequence with high accuracy and nearly complete coverage. Here, we report the result of this finishing process. The current genome sequence (Build 25) contains 2.9 billion nucleotides interspersed by only 141 gaps. Coverage – 99% of the euchromatic genome and is accurate to an error rate of ~1 error per 100,000 bases. Many of the remaining euchromatic gaps are associated with segmental duplications and will require focused work with new methods. The near-complete sequence, the first for a vertebrate, greatly improves the precision of biological analyses of the human genome including studies of gene number, birth and death. Notably, the human genome encodes only 28,000–31,000 protein-coding genes. The genome sequence reported here should serve as a firm foundation for biomedical research in the decades ahead.

The Human Genome Project (HGP) was launched in 1990 with the goal of obtaining a high-quality sequence of the full majority of the euchromatic portion of the human genome. The initial work followed a two-pronged approach: (1) the mapping of the human and mouse genomes^{1,2} to allow the study of inherited disease and provide a scaffold for genome assembly; and (2) the sequencing of organisms with smaller, single genomes^{3,4} to serve as models for method development and as test cases for sequencing the human genome. With success along both paths, the sequencing of the human genome itself eventually became feasible. The International Human Genome Sequencing Consortium (IHGSC), an open collaboration involving twenty centres in six countries, was formed to carry out this component of the HGP.

In February 2001, the IHGSC⁵ and Celera Genomics⁶ each reported draft sequences providing a first overall view of the human genome. These sequences allowed systematic study of the human genome itself, including identification of genes, structural annotation of proteins, regional differences in genome composition, distribution and history of transposable elements, distribution of polymorphisms, and relationships between genomic coordinates and physical distance. Moreover, systematic knowledge of the human genome has enabled new tools and approaches that have markedly accelerated biological discovery.

Both draft sequences, however, had important shortcomings. The IHGSC sequence, for example, omitted ~3% of the euchromatic genome; it was interrupted by ~150,000 gaps and the order and orientation of some sequences could not be determined but had not been established. The Celera⁶ sequence, that aimed at the challenge of completing the euchromatic genome. Operationally, a finished sequence was defined as having an error rate of, at most, one error per 10⁵ bases, and the goal for the project was coverage of finished sequence of at least 95% of the euchromatic genome, with the only gaps being those explained by all available techniques (see <http://www.genome.gov/100923>). The goal was challenging because the human genome has complex and variable duplications⁷ and large segmental duplications, which greatly complicate the determination of genome structure and sequence. In fact, near-complete sequences have been obtained so far only for three multi-chloroplast sequences^{8,9}, mammalian¹⁰ and bird¹¹ genomes. These genomes are all roughly 30-fold smaller than the human genome and have much simpler structure.

We describe here the results of a mid-tier effort by the IHGSC towards the goal of a complete human sequence. The number of gaps that have reduced 40-fold to only 141, most of which are associated with segmental duplications and will require new methods for resolution. The assembled near-complete sequence has an error rate of only ~1 error per 100,000 bases. It contains 245 Mbases nucleotide and covers ~99% of the euchromatic genome. This paper describes the current genome sequence and the process used to produce it, examines the accuracy and completeness of the sequence, and illustrates biological studies made possible by the sequence. We do not attempt here a complete analysis of the contents of the human genome. An initial analysis was previously reported¹² and a series of papers is being written describing the individual chromosomes^{13–16}, including association of genes and other features.

Current genome sequence

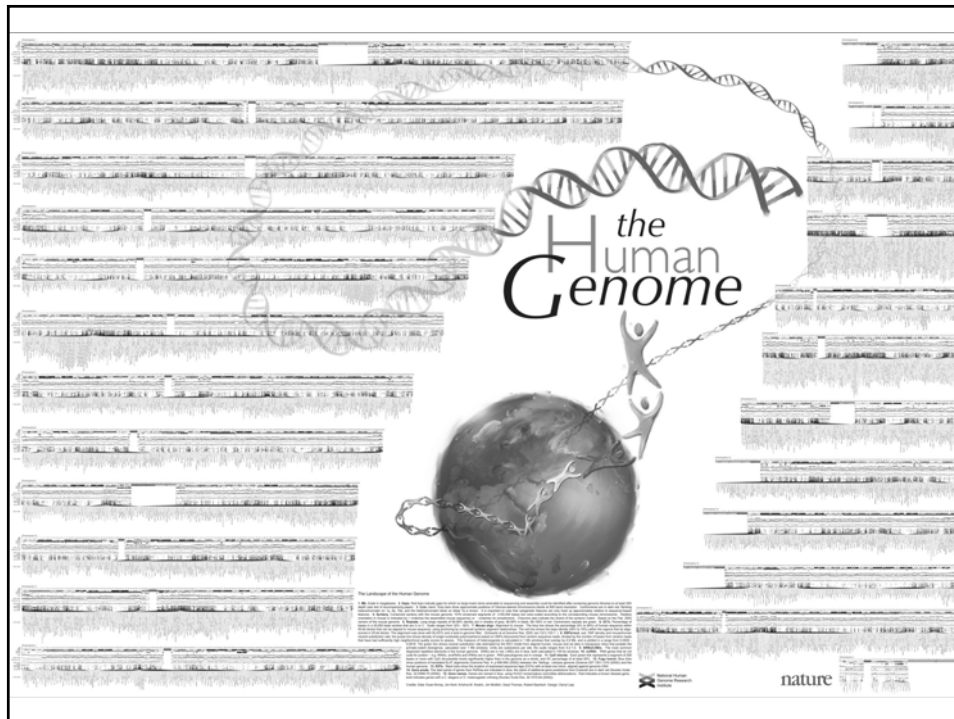
Finishing process

The process of converting the initial draft sequence into a near-complete sequence is referred to as 'finishing'. It is a complex iterative process that proceeds simultaneously at multiple scales, ranging from single nucleotides to the integrity of whole chromosomes. The fundamental challenge is that genomic regions that are not well represented in draft sequences through random shotgun sequencing tend to be highly enriched for polymorphic sequences. Finishing such regions requires the development of special approaches, which evolved substantially over time and varied among centres.

Initially, the finishing process involved two distinct components: (1) producing finished maps, consisting of continuous and accurate paths of overlapping large-insert clones spanning the euchromatic region of the genome; and (2) producing finished clones, consisting of continuous and accurate nucleotide sequence across each large insert clone. In practice, these two components were tightly interrelated in that progress in each often depended on results from the other. Further information about the finishing process and finishing standards can be found in the Supplementary Information (Site 1), and at <http://www.genome.gov/100923>.

In total, we generated a shotgun sequence from 98,268 large-insert clones (total length ~3.16 gigabases (Gb)) and finished the sequence from 45,742 of these clones (total length ~1.67 Gb). The clones consisted primarily of bacterial artificial chromosomes

Nature 431:931-945, 2004



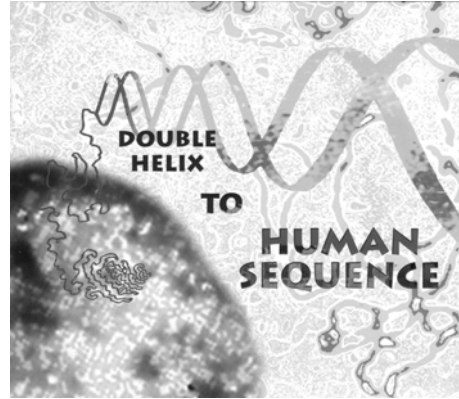
April, 1953 → **April, 2003**

NO. 4356 April 25, 1953 NATURE

MOLECULAR STRUCTURE OF
NUCLEIC ACIDS
A Structure for Deoxyribose Nucleic Acid

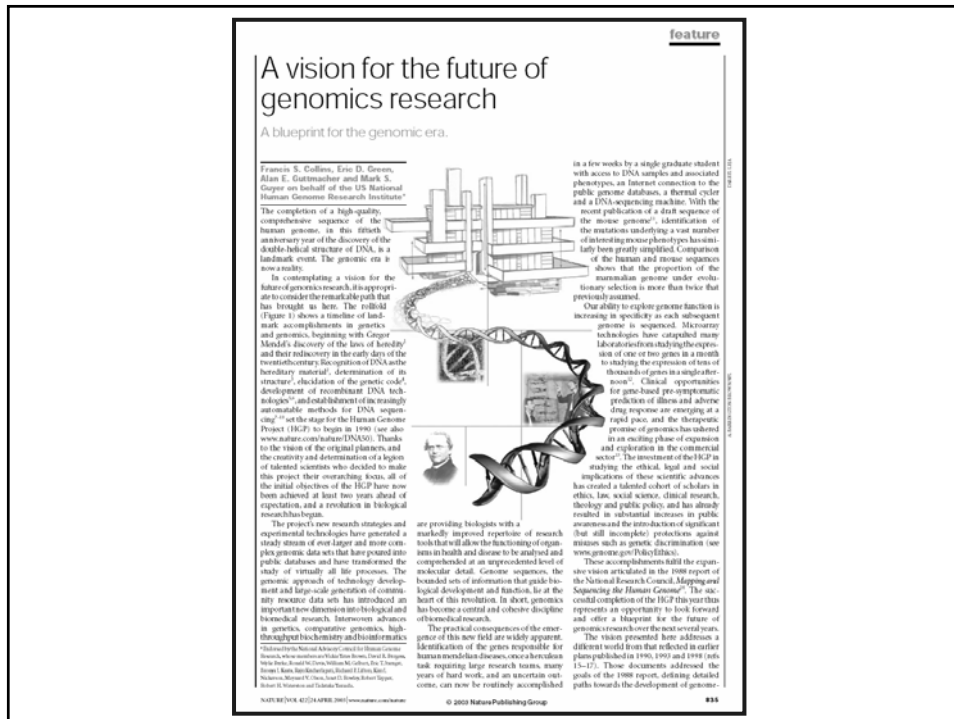


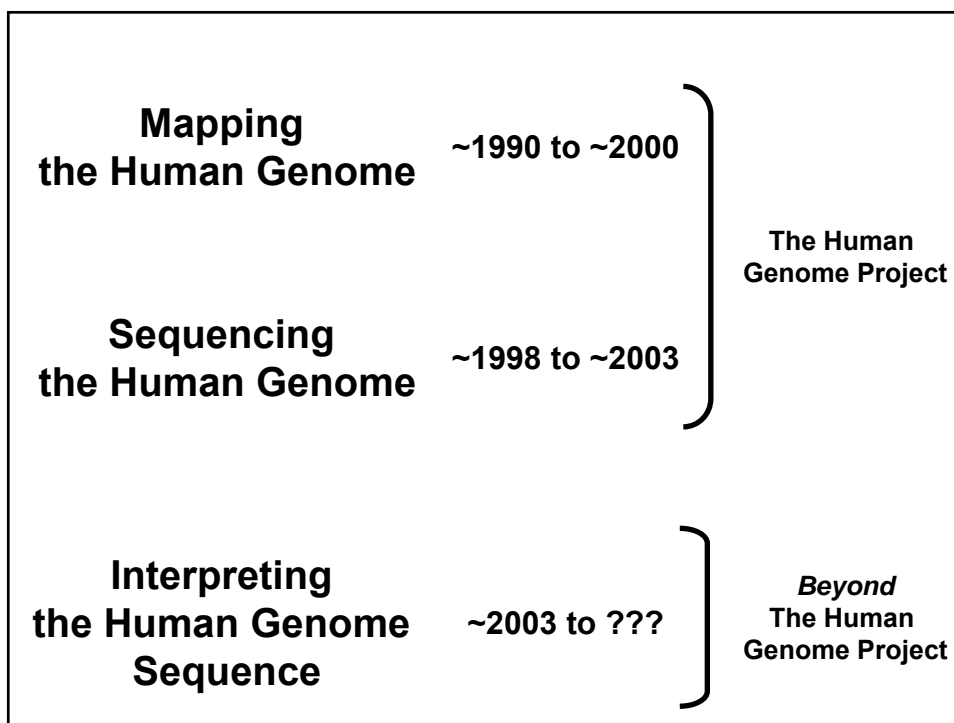
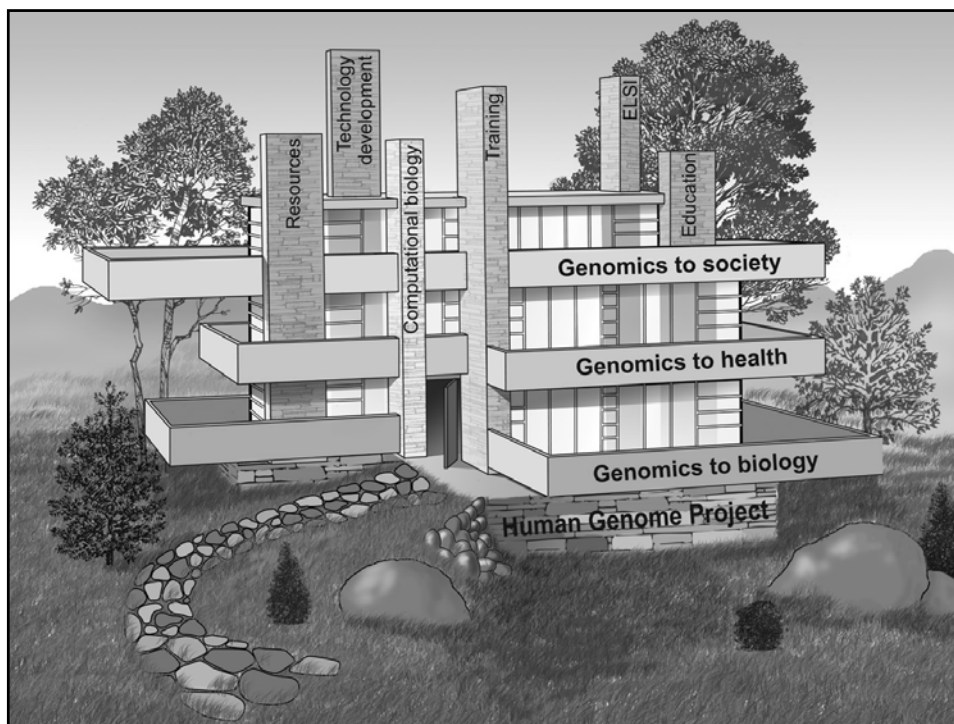
J. D. WATSON
F. H. C. CRICK
Medical Research Council Unit for the
Study of the Molecular Structure of
Biological Systems,
Cavendish Laboratory, Cambridge.
April 2.

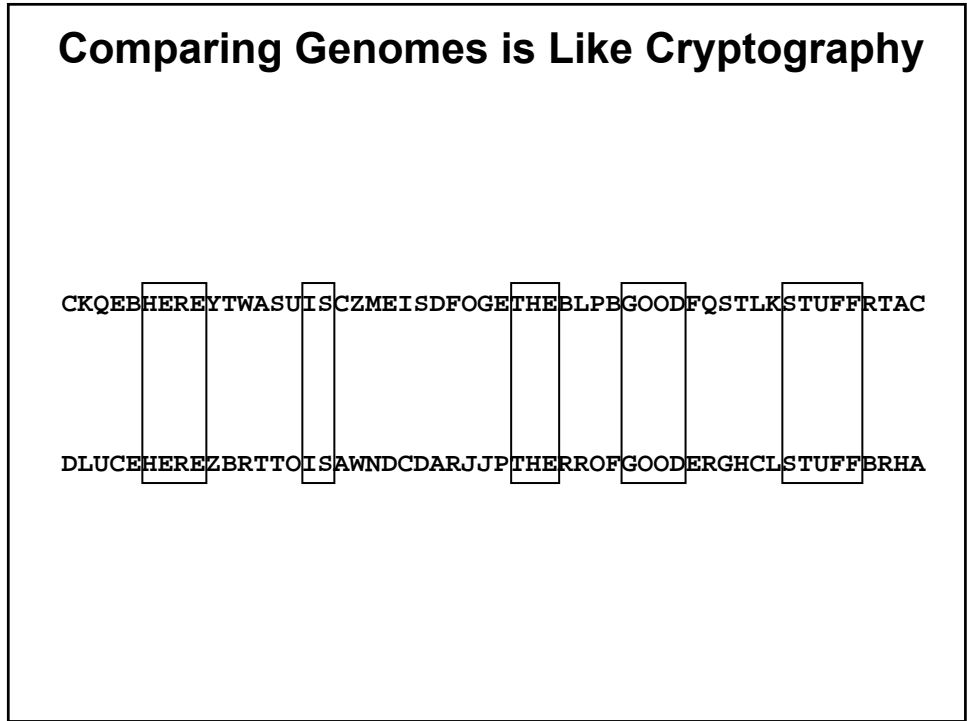
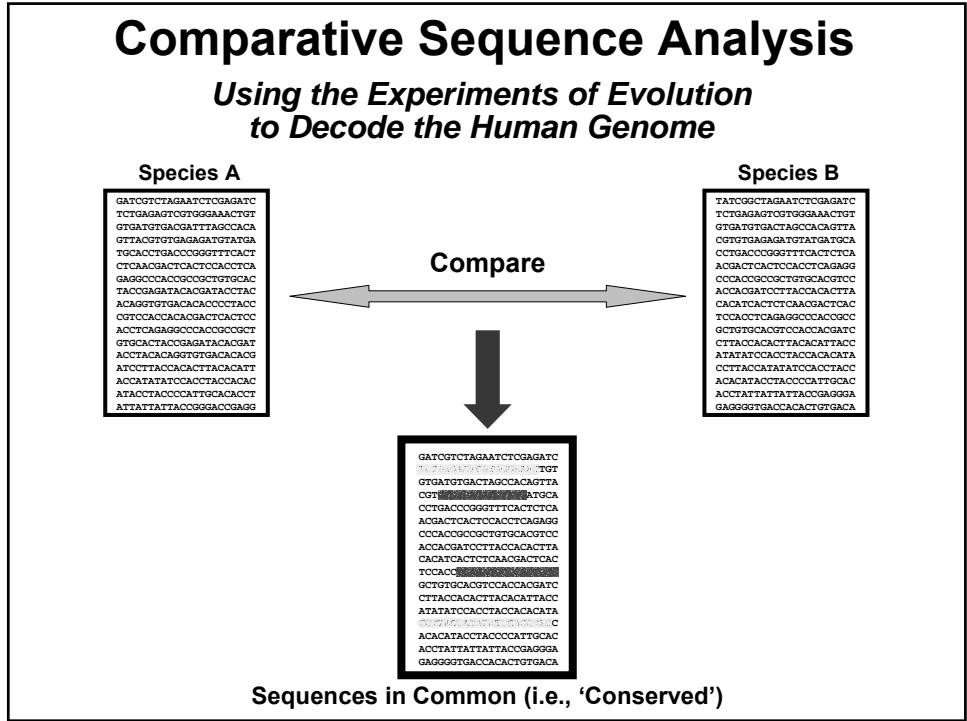


**All of the original goals of the
Human Genome Project have
been accomplished!**

What's Next?





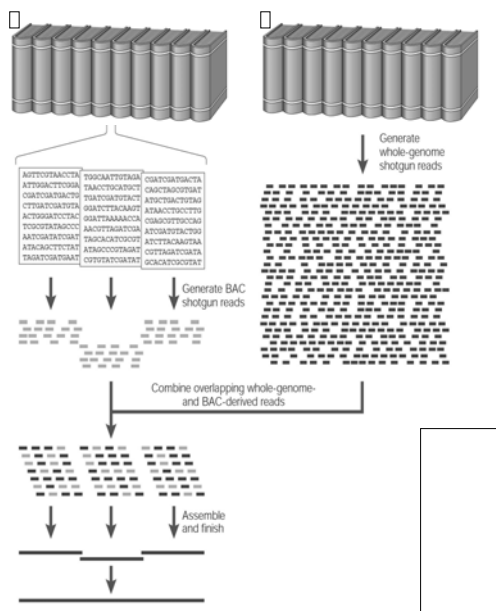


Functional Elements: Coding vs. Non-Coding

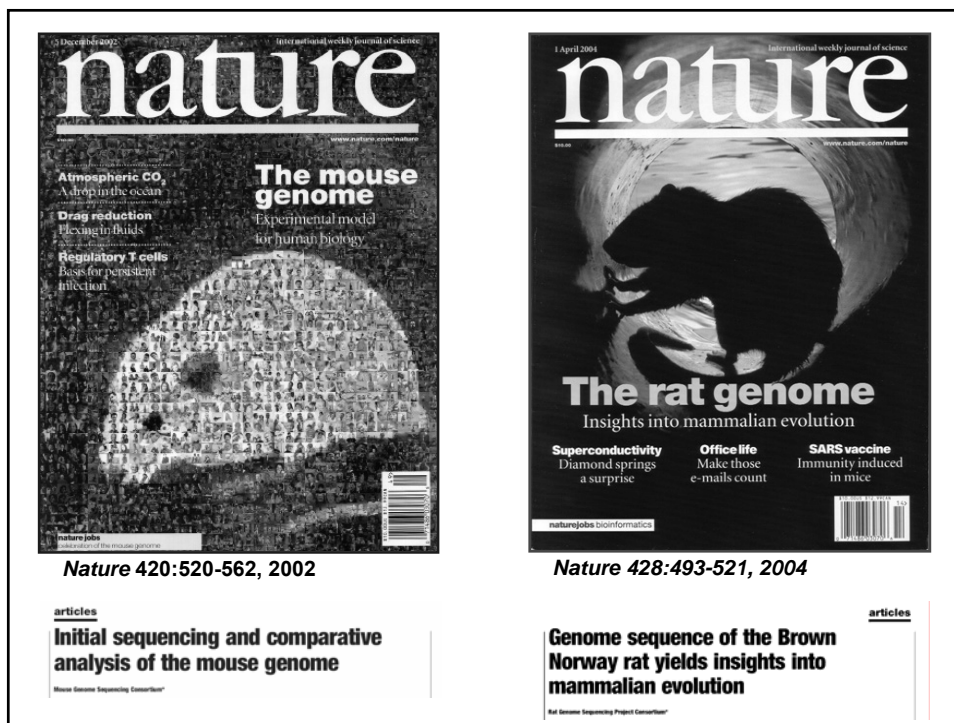
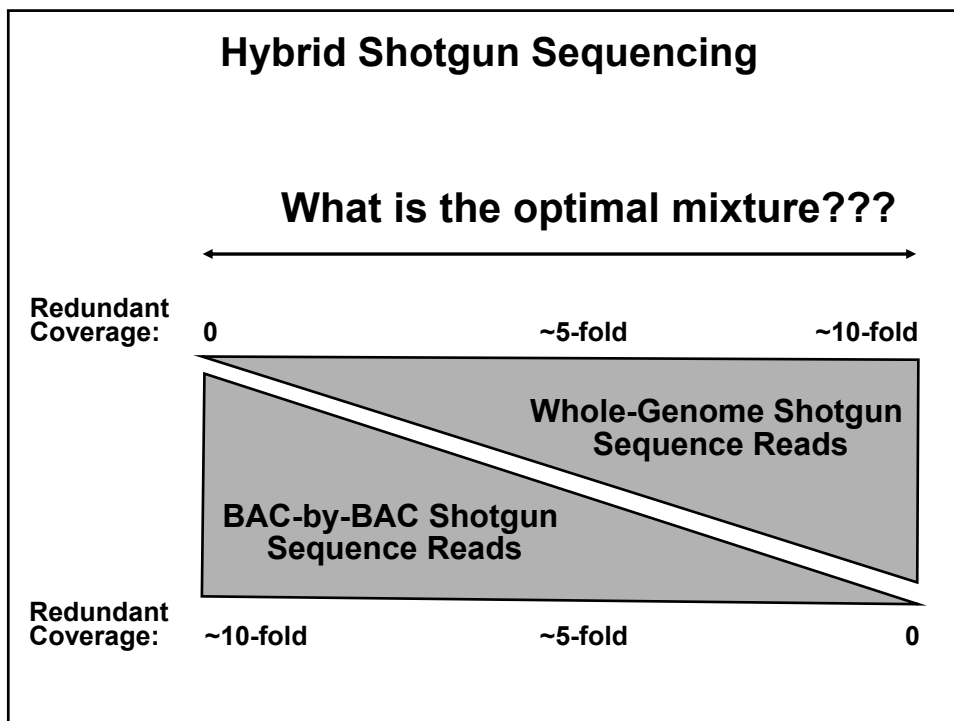
- **Coding Sequences (i.e., Genes)**
 - Relatively EASY to Identify
 - Mostly Know What to Look For
 - Complementary Data Sets Available (ESTs, cDNAs)
 - Ever-Improving Computational Gene Predictions
- **Non-Coding Functional Sequences**
 - HARD to Identify
 - Very Little Known About What to Look For
 - Virtually No Complementary Data Sets Available
 - Poor Computational Predictions

Major role for comparative sequence analysis will be the identification of functionally important, non-coding sequences

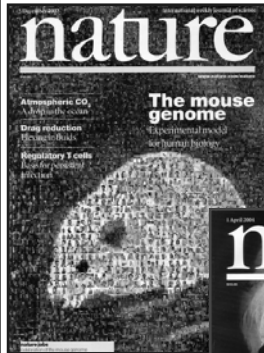
Hybrid Shotgun Sequencing



Green (2001)



Human-Rodent Sequence Comparisons



Nature
420:520-562, 2002



Nature
428:493-521, 2004

- ~40% in Alignments
- ~5% Under Selection
- ~1.5% Protein Coding
- ~3.5% Non-Coding

Multi-Species Sequence Comparisons

GATGTCCTAGAACTCTCG AGATCTCTGAGAGTCGT GGGAAACTGTGTGATGT GACTAGCCACGATTACG TGTGAGAGATGTATGAT GCACCTGACCCGGGTTT GACTCTCAACGACTCAC TCCACCTGAGAGGCCCA CCGCCCTGTGCAGTTC CACCAGATCTCTTACCA CACTTACGATTAACAT ATACTACCCCATTTGCA GACCTATATATATACC	GATGTCCTAGAACTCTCG AGATCTCTGAGAGTCGT GGGAAACTGTGTGATGT GACTAGCCACGATTACG TGTGAGAGATGTATGAT GCACCTGACCCGGGTTT GACTCTCAACGACTCAC TCCACCTGAGAGGCCCA CCGCCCTGTGCAGTTC CACCAGATCTCTTACCA CACTTACGATTAACAT ATACTACCCCATTTGCA GACCTATATATATACC	GATGTCCTAGAACTCTCG AGATCTCTGAGAGTCGT GGGAAACTGTGTGATGT GACTAGCCACGATTACG TGTGAGAGATGTATGAT GCACCTGACCCGGGTTT GACTCTCAACGACTCAC TCCACCTGAGAGGCCCA CCGCCCTGTGCAGTTC CACCAGATCTCTTACCA CACTTACGATTAACAT ATACTACCCCATTTGCA GACCTATATATATACC	GATGTCCTAGAACTCTCG AGATCTCTGAGAGTCGT GGGAAACTGTGTGATGT GACTAGCCACGATTACG TGTGAGAGATGTATGAT GCACCTGACCCGGGTTT GACTCTCAACGACTCAC TCCACCTGAGAGGCCCA CCGCCCTGTGCAGTTC CACCAGATCTCTTACCA CACTTACGATTAACAT ATACTACCCCATTTGCA GACCTATATATATACC	GATGTCCTAGAACTCTCG AGATCTCTGAGAGTCGT GGGAAACTGTGTGATGT GACTAGCCACGATTACG TGTGAGAGATGTATGAT GCACCTGACCCGGGTTT GACTCTCAACGACTCAC TCCACCTGAGAGGCCCA CCGCCCTGTGCAGTTC CACCAGATCTCTTACCA CACTTACGATTAACAT ATACTACCCCATTTGCA GACCTATATATATACC	GATGTCCTAGAACTCTCG AGATCTCTGAGAGTCGT GGGAAACTGTGTGATGT GACTAGCCACGATTACG TGTGAGAGATGTATGAT GCACCTGACCCGGGTTT GACTCTCAACGACTCAC TCCACCTGAGAGGCCCA CCGCCCTGTGCAGTTC CACCAGATCTCTTACCA CACTTACGATTAACAT ATACTACCCCATTTGCA GACCTATATATATACC	GATGTCCTAGAACTCTCG AGATCTCTGAGAGTCGT GGGAAACTGTGTGATGT GACTAGCCACGATTACG TGTGAGAGATGTATGAT GCACCTGACCCGGGTTT GACTCTCAACGACTCAC TCCACCTGAGAGGCCCA CCGCCCTGTGCAGTTC CACCAGATCTCTTACCA CACTTACGATTAACAT ATACTACCCCATTTGCA GACCTATATATATACC
---	---	---	---	---	---	---

HUMAN

Multi-Species Comparative Sequence Analysis

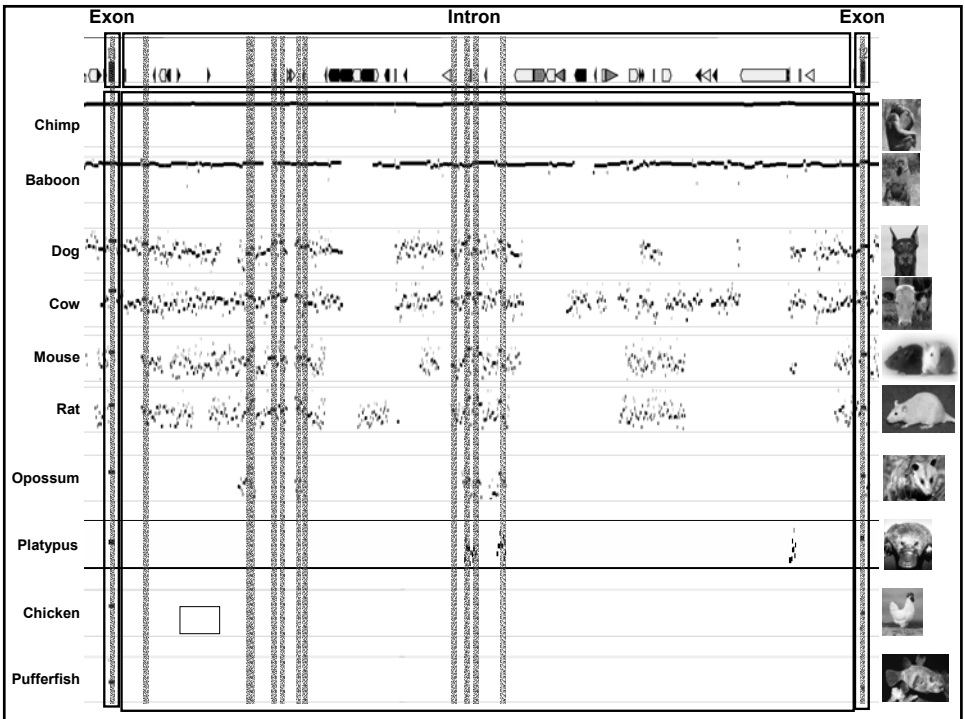
Comparative analyses of multi-species sequences from targeted genomic regions

J. M. Thomas¹, J. W. Tomblin², A. W. Rickaby³, A. G. Burt⁴, S. M. Rockowitz⁵, E. S. Bergler⁶, M. Bracken⁷, S. C. Cooper⁸, P. A. Thompson⁹, G. Hochstadt¹⁰, R. Mendenhall¹¹, M. S. Schmitt¹², S. J. Weller¹³, W. J. Kauf¹⁴, D. Kordish¹⁵, T. G. Gross¹⁶, R. Sauer¹⁷, D. E. Galter¹⁸, S. Schmitt¹⁹, L. Hahn²⁰, J. S. Mui²¹, A. H. Parnell²², S.-H. Lee²³, V. S. R. Muth²⁴, T. J. Summers²⁵, M. S. Parvaz²⁶, S. L. Schmitt²⁷, M. Miller²⁸, A. Agari²⁹, S. Sengupta³⁰, K. Corbett³¹, C. P. Brinkley³², S. T. Smith³³, S. Oishi³⁴, L. Oishi³⁵, J. Gupta³⁶, P. Hengstenberg³⁷, J. C. Lee³⁸, M. C. Hwang³⁹, J. Kasper⁴⁰, L. L. Lerner⁴¹, R. L. Lerner⁴², M. J. Lee⁴³, G. L. Muth⁴⁴, C. A. Maudsley⁴⁵, S. S. Maudsley⁴⁶, J. C. McInerney⁴⁷, R. Pearson⁴⁸, J. S. Shultz⁴⁹, E. S. Thompson⁵⁰, J. T. Tsai⁵¹, C. Thompson⁵², J. L. Vogel⁵³, M. A. Walker⁵⁴, K. S. Winkler⁵⁵, J. S. Winters⁵⁶, A. C. Wang⁵⁷, L.-H. Zhang⁵⁸, K. Chong⁵⁹, K. Zhu⁶⁰, R. Zhu⁶¹, C. L. Zhu⁶², P. J. De Jong⁶³, G. S. Lawrence⁶⁴, A. P. Dear⁶⁵, A. Chakravarti⁶⁶, D. Haussler⁶⁷, P. Green⁶⁸, M. Miller⁶⁹ & E. S. Green⁷⁰

¹Genome Technology Branch, National Human Genome Research Institute, and ²NIH Intramural Sequencing Center, National Institutes of Health, Bethesda, Maryland 20892, USA
³Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA
⁴Department of Genetics, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA
⁵Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA
⁶Children's Hospital Oakland Research Institute, Oakland, California 94612, USA
⁷The Weizmann Institute for Genome Sciences and Technology, New York State Department of Health, Albany, New York 12242, USA
⁸The Institute for Genome Sciences and Policy, Johns Hopkins University, Baltimore, Maryland 21205, USA
⁹Department of Biology, University of California, Santa Cruz, California 95064, USA
¹⁰Department of Biology, University of California, Santa Cruz, California 95064, USA
¹¹Department of Biology, University of California, Santa Cruz, California 95064, USA
¹²Department of Biology, University of California, Santa Cruz, California 95064, USA
¹³Department of Biology, University of California, Santa Cruz, California 95064, USA
¹⁴Department of Biology, University of California, Santa Cruz, California 95064, USA
¹⁵Department of Biology, University of California, Santa Cruz, California 95064, USA
¹⁶Department of Biology, University of California, Santa Cruz, California 95064, USA
¹⁷Department of Biology, University of California, Santa Cruz, California 95064, USA
¹⁸Department of Biology, University of California, Santa Cruz, California 95064, USA
¹⁹Department of Biology, University of California, Santa Cruz, California 95064, USA
²⁰Department of Biology, University of California, Santa Cruz, California 95064, USA
²¹Department of Biology, University of California, Santa Cruz, California 95064, USA
²²Department of Biology, University of California, Santa Cruz, California 95064, USA
²³Department of Biology, University of California, Santa Cruz, California 95064, USA
²⁴Department of Biology, University of California, Santa Cruz, California 95064, USA
²⁵Department of Biology, University of California, Santa Cruz, California 95064, USA
²⁶Department of Biology, University of California, Santa Cruz, California 95064, USA
²⁷Department of Biology, University of California, Santa Cruz, California 95064, USA
²⁸Department of Biology, University of California, Santa Cruz, California 95064, USA
²⁹Department of Biology, University of California, Santa Cruz, California 95064, USA
³⁰Department of Biology, University of California, Santa Cruz, California 95064, USA
³¹Department of Biology, University of California, Santa Cruz, California 95064, USA
³²Department of Biology, University of California, Santa Cruz, California 95064, USA
³³Department of Biology, University of California, Santa Cruz, California 95064, USA
³⁴Department of Biology, University of California, Santa Cruz, California 95064, USA
³⁵Department of Biology, University of California, Santa Cruz, California 95064, USA
³⁶Department of Biology, University of California, Santa Cruz, California 95064, USA
³⁷Department of Biology, University of California, Santa Cruz, California 95064, USA
³⁸Department of Biology, University of California, Santa Cruz, California 95064, USA
³⁹Department of Biology, University of California, Santa Cruz, California 95064, USA
⁴⁰Department of Biology, University of California, Santa Cruz, California 95064, USA
⁴¹Department of Biology, University of California, Santa Cruz, California 95064, USA
⁴²Department of Biology, University of California, Santa Cruz, California 95064, USA
⁴³Department of Biology, University of California, Santa Cruz, California 95064, USA
⁴⁴Department of Biology, University of California, Santa Cruz, California 95064, USA
⁴⁵Department of Biology, University of California, Santa Cruz, California 95064, USA
⁴⁶Department of Biology, University of California, Santa Cruz, California 95064, USA
⁴⁷Department of Biology, University of California, Santa Cruz, California 95064, USA
⁴⁸Department of Biology, University of California, Santa Cruz, California 95064, USA
⁴⁹Department of Biology, University of California, Santa Cruz, California 95064, USA
⁵⁰Department of Biology, University of California, Santa Cruz, California 95064, USA
⁵¹Department of Biology, University of California, Santa Cruz, California 95064, USA
⁵²Department of Biology, University of California, Santa Cruz, California 95064, USA
⁵³Department of Biology, University of California, Santa Cruz, California 95064, USA
⁵⁴Department of Biology, University of California, Santa Cruz, California 95064, USA
⁵⁵Department of Biology, University of California, Santa Cruz, California 95064, USA
⁵⁶Department of Biology, University of California, Santa Cruz, California 95064, USA
⁵⁷Department of Biology, University of California, Santa Cruz, California 95064, USA
⁵⁸Department of Biology, University of California, Santa Cruz, California 95064, USA
⁵⁹Department of Biology, University of California, Santa Cruz, California 95064, USA
⁶⁰Department of Biology, University of California, Santa Cruz, California 95064, USA
⁶¹Department of Biology, University of California, Santa Cruz, California 95064, USA
⁶²Department of Biology, University of California, Santa Cruz, California 95064, USA
⁶³Department of Biology, University of California, Santa Cruz, California 95064, USA
⁶⁴Department of Biology, University of California, Santa Cruz, California 95064, USA
⁶⁵Department of Biology, University of California, Santa Cruz, California 95064, USA
⁶⁶Department of Biology, University of California, Santa Cruz, California 95064, USA
⁶⁷Department of Biology, University of California, Santa Cruz, California 95064, USA
⁶⁸Department of Biology, University of California, Santa Cruz, California 95064, USA
⁶⁹Department of Biology, University of California, Santa Cruz, California 95064, USA
⁷⁰Department of Biology, University of California, Santa Cruz, California 95064, USA

- Targeted Genomic Regions
- BAC-Based Sequencing in Multiple Vertebrates
- Identify Highly Conserved Non-Coding Sequences
- Conserved Sequences Correlate with Functional Elements

Thomas et al. (2003)



Additional Vertebrate Genome Sequencing Efforts



Chimpanzee



Macaque



Dog



Cow



Monodelphis



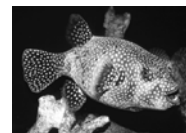
Chicken



Xenopus

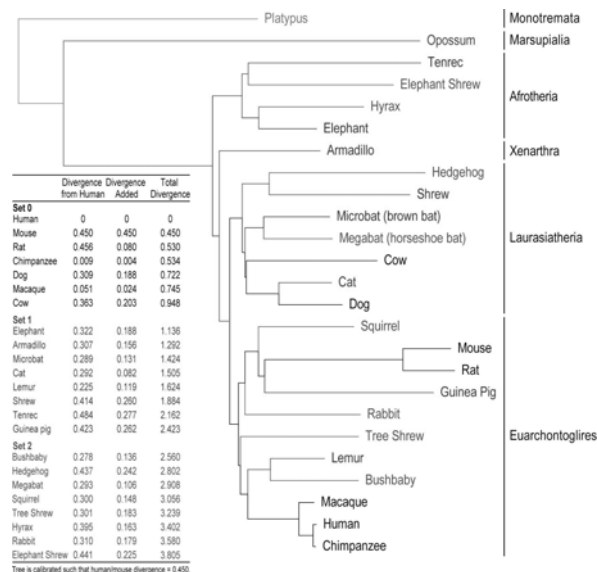


Zebrafish



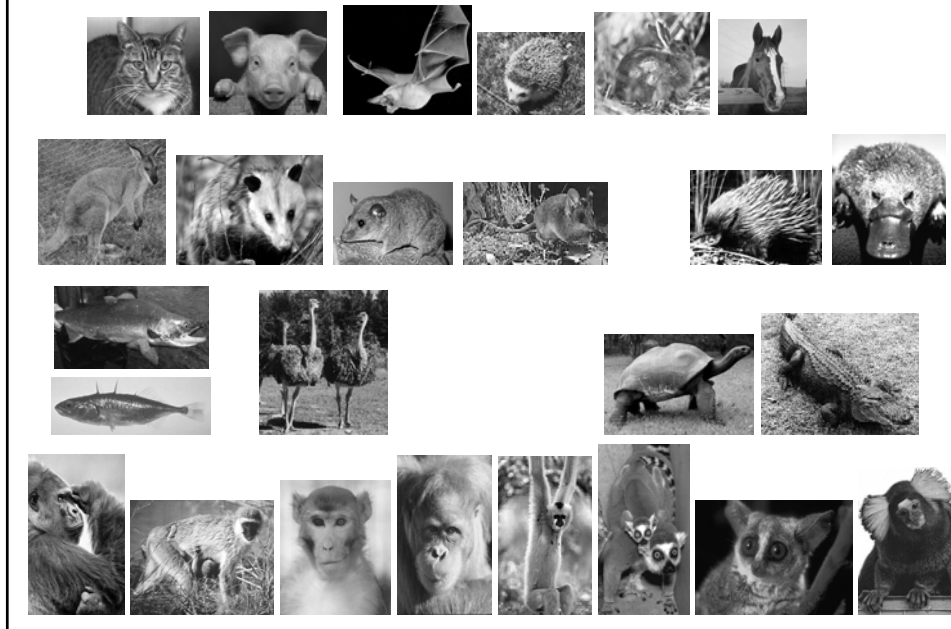
Pufferfish

Low-Redundancy Sequencing of Multiple Vertebrate Genomes



Margulies et al., *PNAS*, in press, 2005

Future Genomes to Sequence???



ENCODE Project

- **ENCODE: ENCyclopedia Of DNA Elements**
- **Goal: Compile a *comprehensive encyclopedia* of all functional elements in the human genome**
- **Initial pilot project: 1% of human genome**
- **Apply multiple approaches to study and analyze that 1% in a consortium fashion**

ENCODE Project: Web Sites

The left screenshot displays the 'ENCODE Target Regions (January 2004)' page. It includes a table with columns for 'Region', 'Description', and 'Chr'. The table lists various regions such as EN001, EN002, EN003, etc., with their corresponding descriptions and chromosome locations. Below the table, there are sections for 'Non-Exonic Conservation 90%, Gene Density 90%' and 'Non-Exonic Conservation 80%, Gene Density 80% - 100%', each with a table of genomic coordinates and region names.

The right screenshot shows the 'UCSC Genome Browser on Human July 2003 Freeze' interface. It features a search bar, a list of tracks, and a main display area showing a genomic track with various annotations and data points. The interface includes navigation controls and a search bar at the top.

genome.gov/ENCODE

genome.ucsc.edu/ENCODE

Current Big Challenges...

- Defining “Saturation Points” in Terms of Information Gained by Comparative Sequence Analyses
- The “\$1000 Genome”
- Medical Sequencing (aka, Human Re-Sequencing)