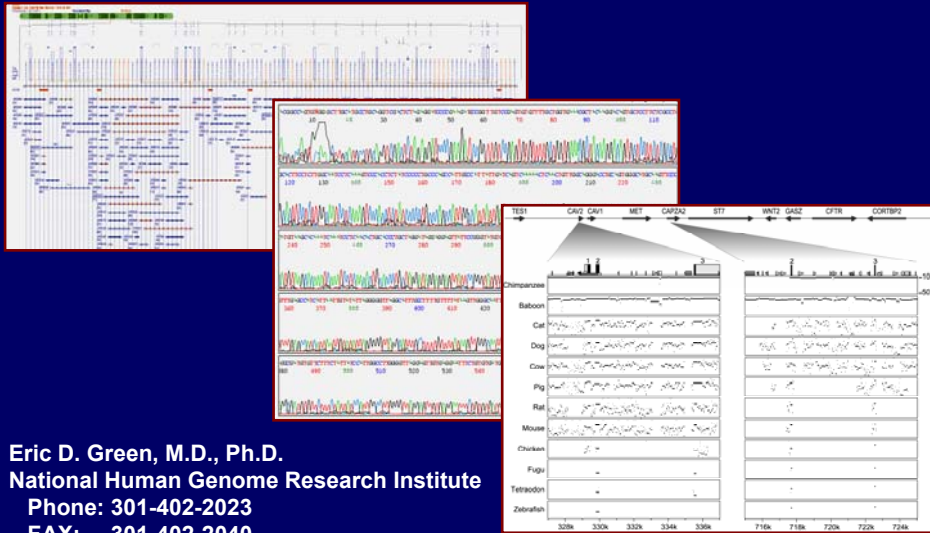


Techniques for Genome Mapping & Sequencing



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

Foundational Milestones in Genetics & Genomics



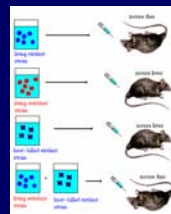
Mendel

1865



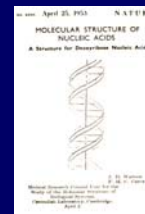
Miescher

1871



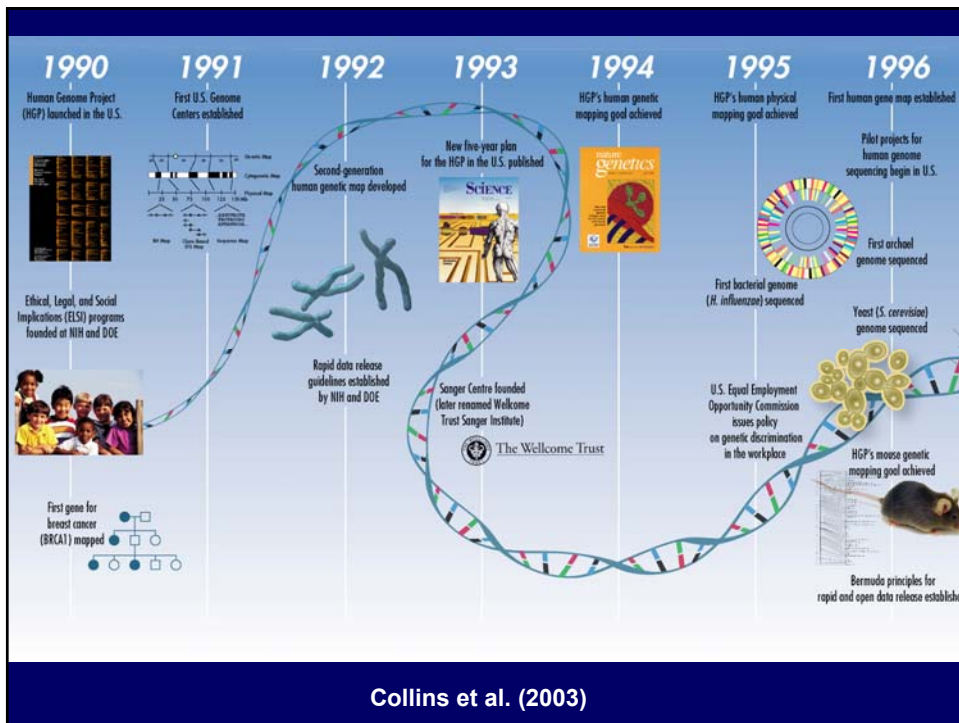
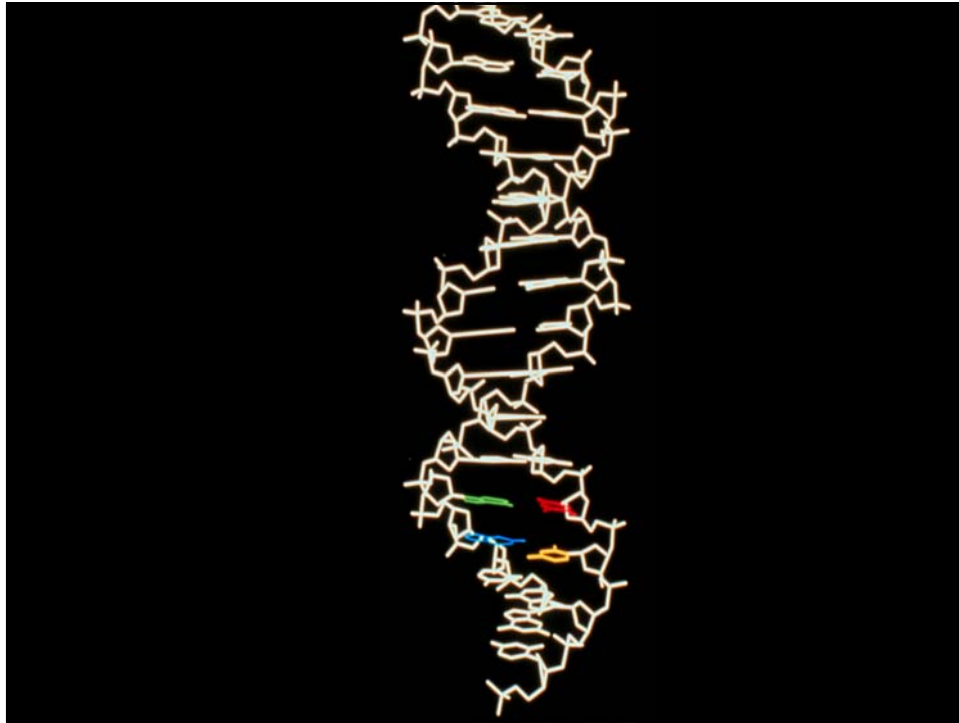
Avery

1944

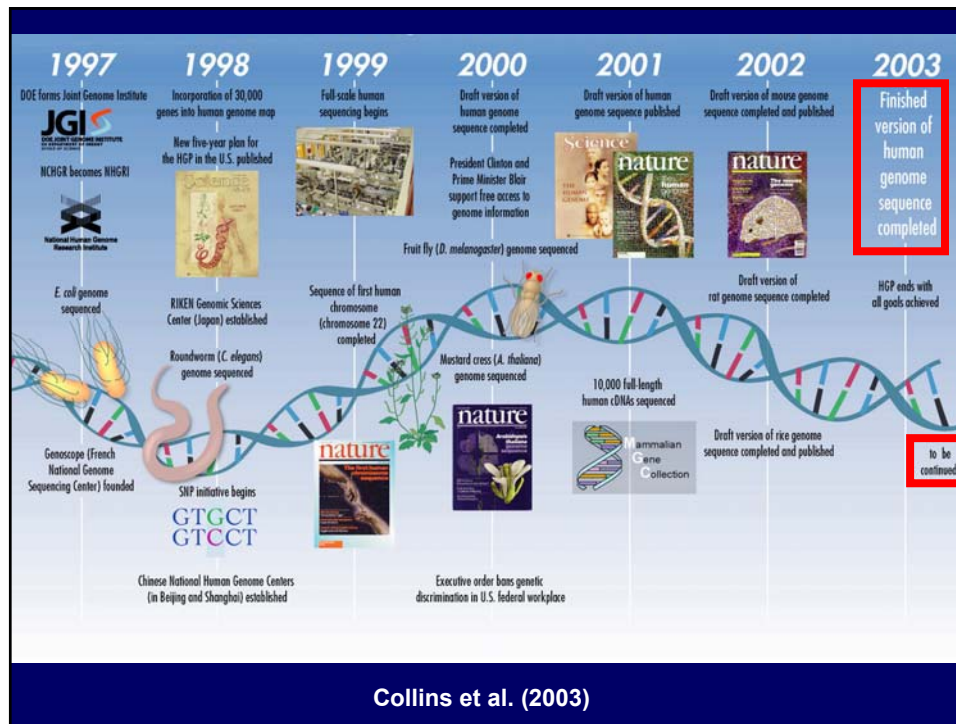


Watson & Crick

1953

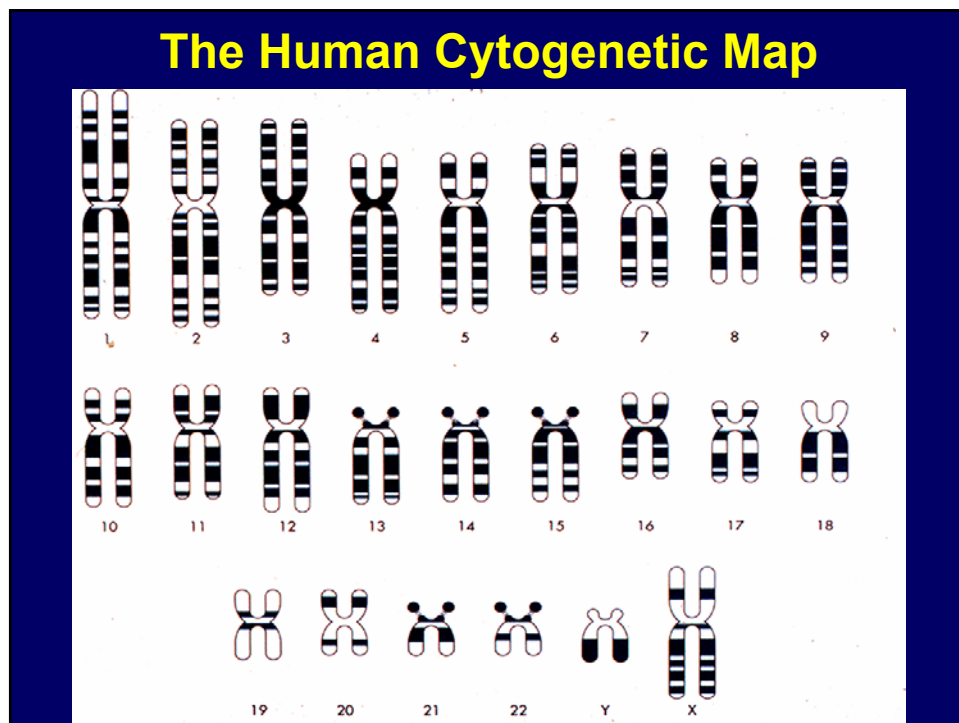
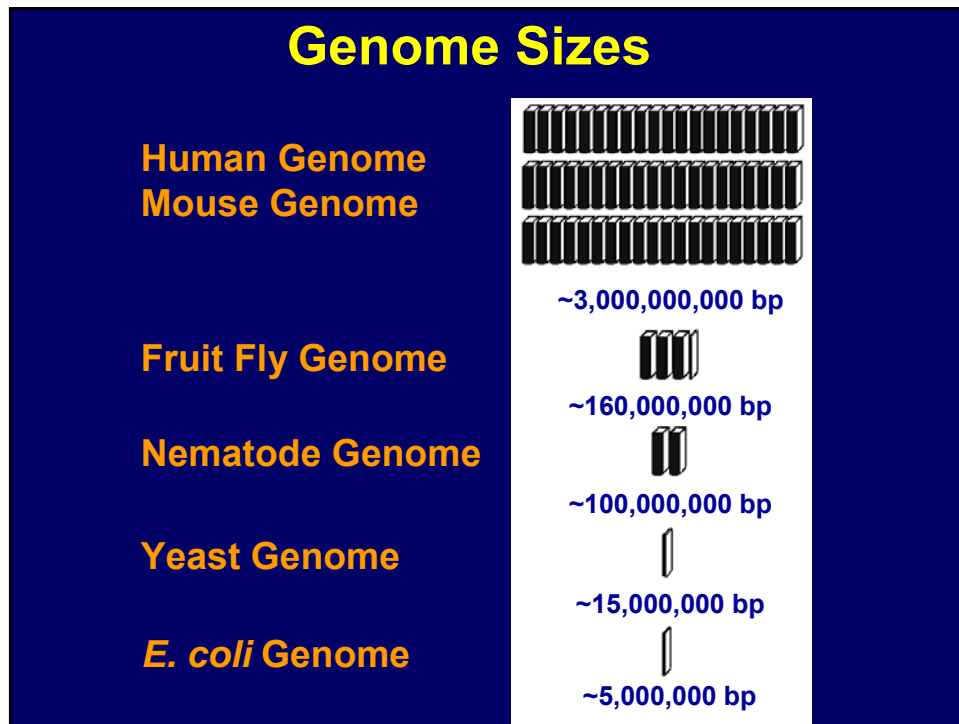


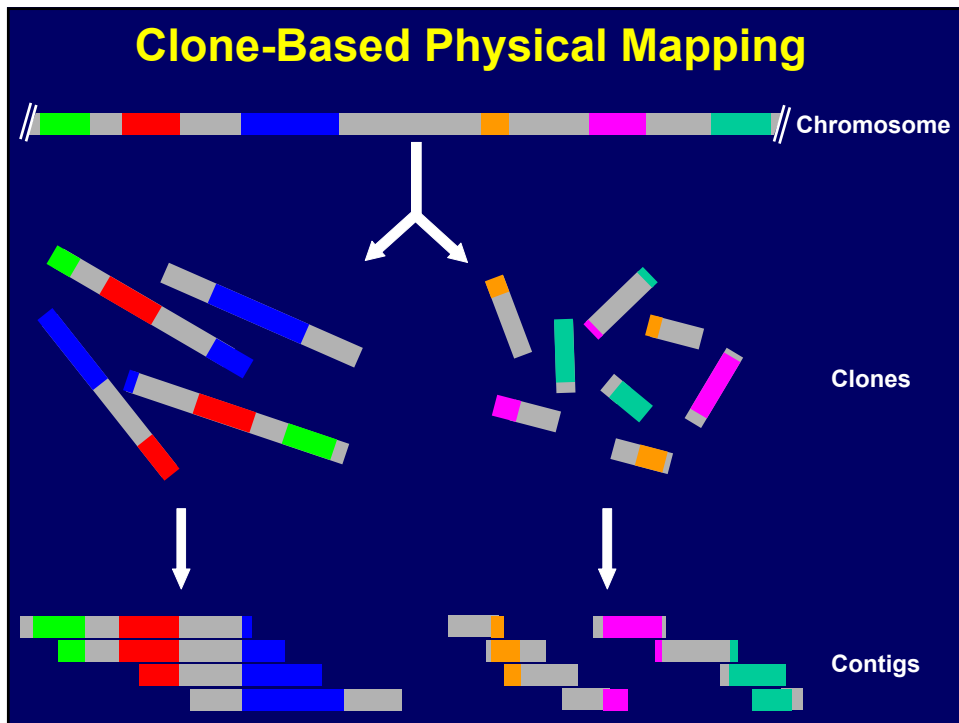
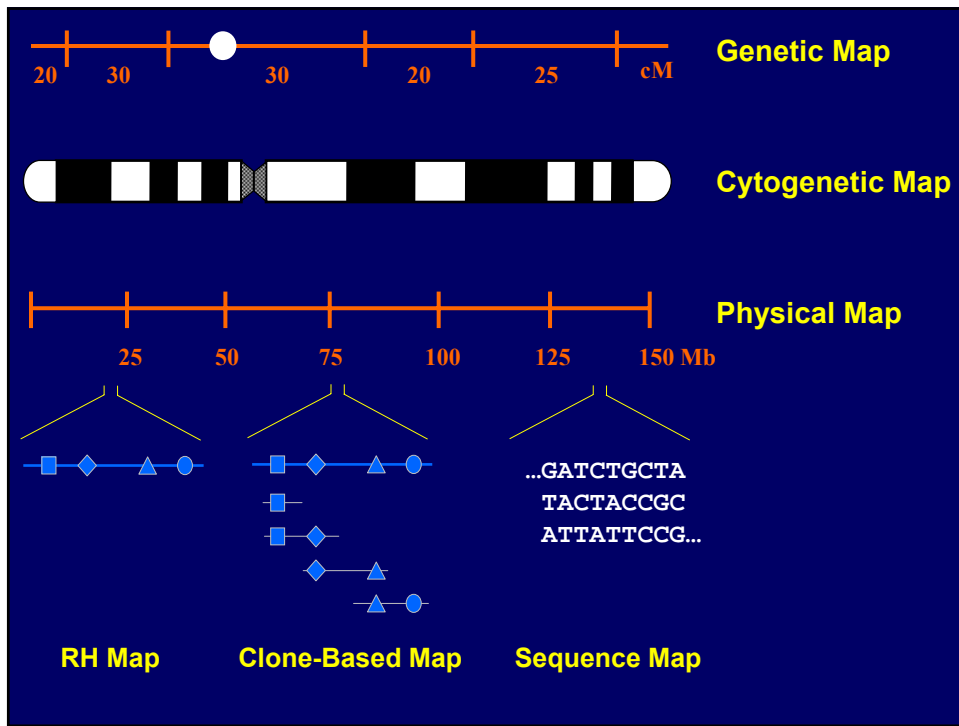
Collins et al. (2003)

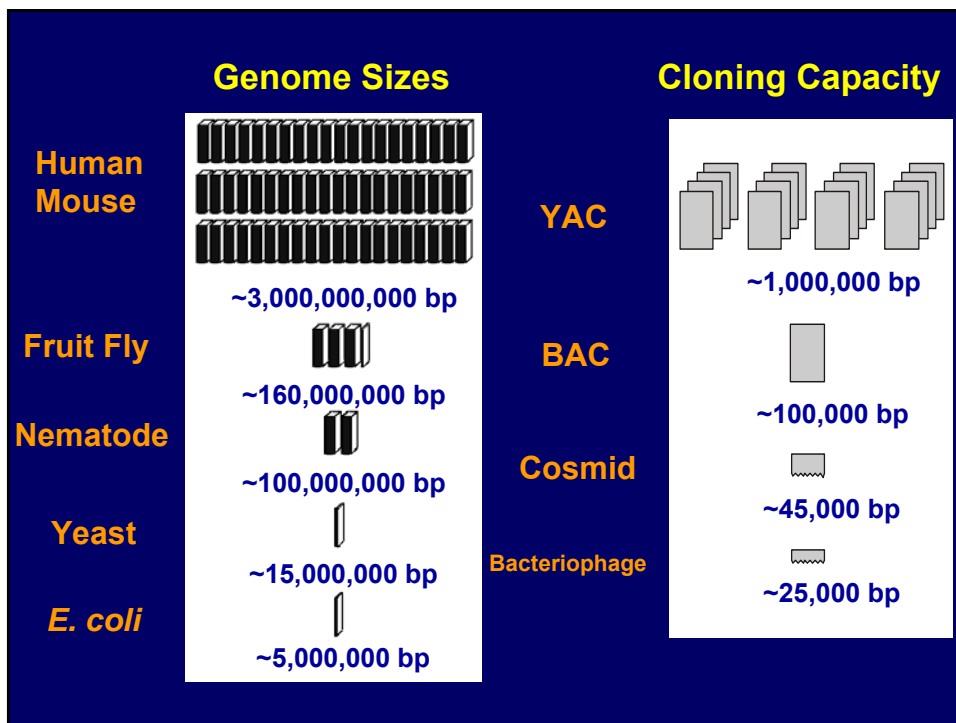
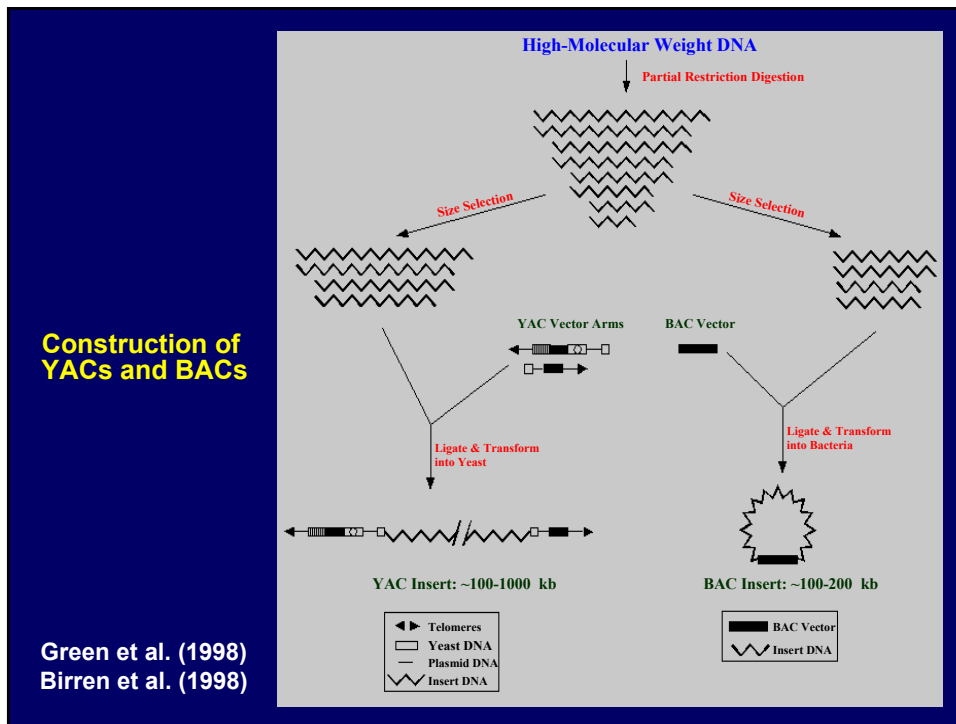


Outline

- I. Fundamentals of Genome Mapping
- II. Fundamentals of Genome Sequencing
- III. Mapping & Sequencing in the Human Genome Project
- IV. Comparative Sequencing
- V. New DNA Sequencing Technologies

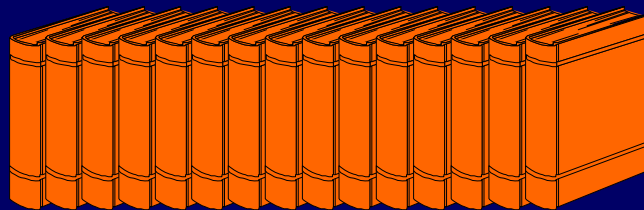




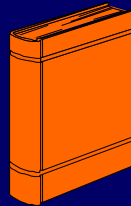


Bacterial Artificial Chromosomes (BACs)

- Bacterial-Based Cloning System
- 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
- BAC Libraries from Many Different Species Available (e.g., www.chori.org/bacpac)
- See Birren et al. (1998)



Genome
(~3000 Mb)



Chromosome
(~130 Mb)

G	A	T	C	T	C	T	A	G	A	A	T	C	T	C
G	A	G	A	T	C	T	C	T	A	G	A	G	T	C
G	T	G	G	A	C	T	G	T	T	T	A			
T	T	T	T	T	T	T	T	T	T	T	T			
T	T	T	T	T	T	T	T	T	T	T	T			
A	A	A	A	A	A	A	A	A	A	A	A			
G	G	G	G	G	G	G	G	G	G	G	G			
G	G	G	G	G	G	G	G	G	G	G	G			
C	C	C	C	C	C	C	C	C	C	C	C			
G	T	G	G	A	C	T	G	T	T	T	A			
G	T	G	G	A	C	T	G	T	T	T	A			

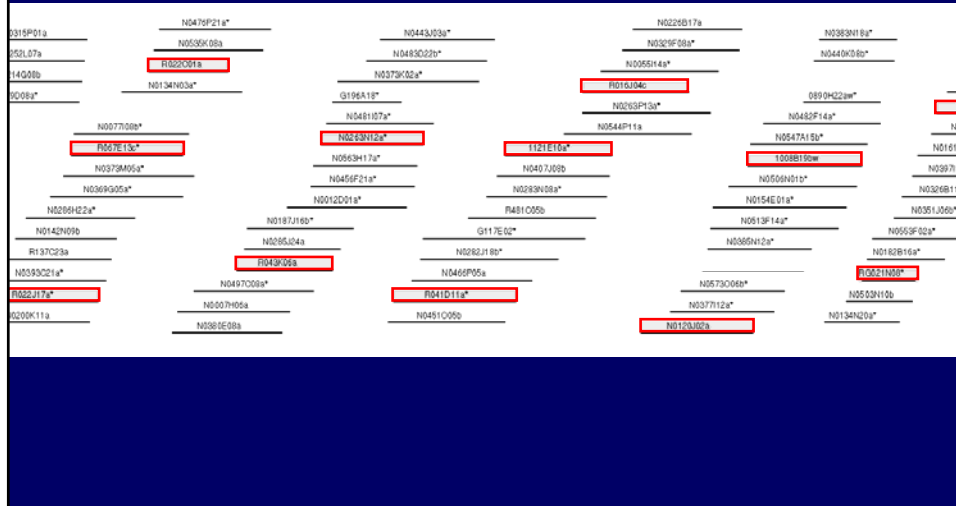
YAC
(~0.5-1.0 Mb)

G	A	T	C	T	C	T	A	G	A	A	T	C	T	C
G	A	G	A	T	C	T	C	T	A	G	A	G	T	C
G	T	G	G	A	C	T	G	T	T	T	A			
T	T	T	T	T	T	T	T	T	T	T	T			
T	T	T	T	T	T	T	T	T	T	T	T			
A	A	A	A	A	A	A	A	A	A	A	A			
G	G	G	G	G	G	G	G	G	G	G	G			
G	G	G	G	G	G	G	G	G	G	G	G			
C	C	C	C	C	C	C	C	C	C	C	C			
G	T	G	G	A	C	T	G	T	T	T	A			
G	T	G	G	A	C	T	G	T	T	T	A			

BAC
(~0.1-0.2 Mb)

Sequence-Ready Contig Map

Marra et al. (1997) and Gregory et al. (1997)

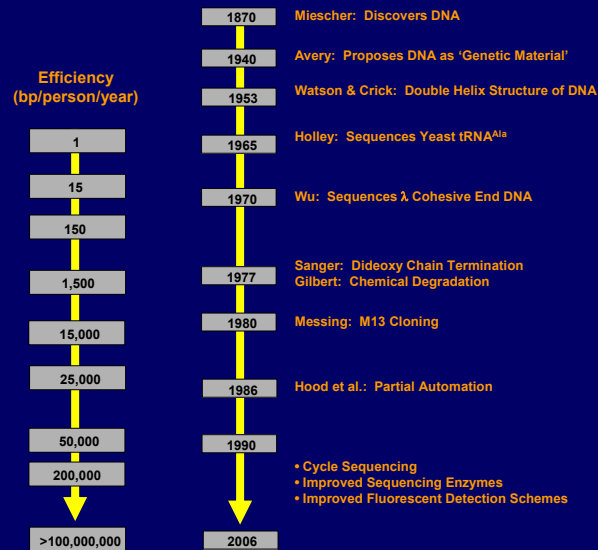


Physical Mapping: Future Prospects

- Strategies for Physical Mapping have Advanced Greatly in the Sequence-Based Era
- Close Interplay of Mapping and Sequencing in the Exploration of Genomes
- Availability of Many BAC Libraries is Allowing Physical Mapping of More Species' Genomes

DNA Sequencing

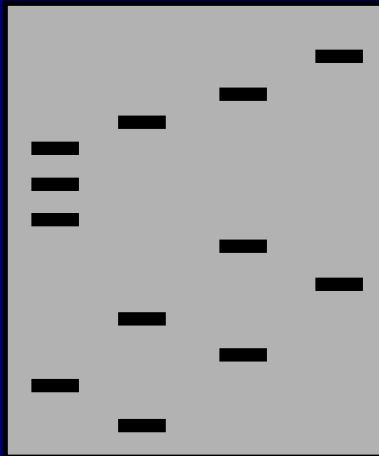
History of DNA Sequencing



Adapted from Messing & Liaca, *PNAS* (1998)

DNA Tagged with Radioactivity

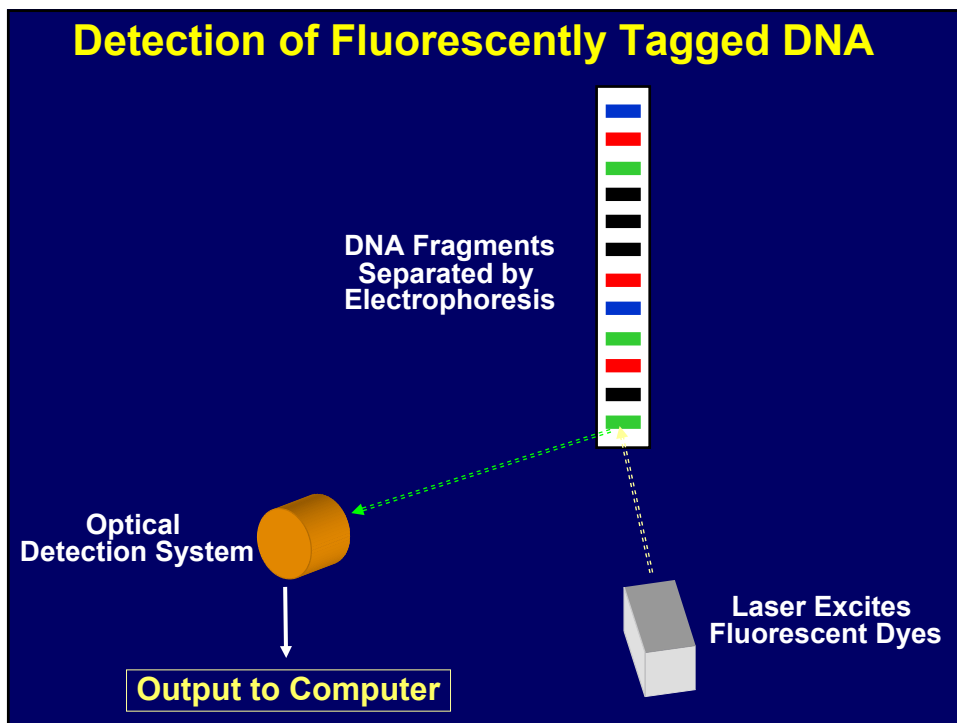
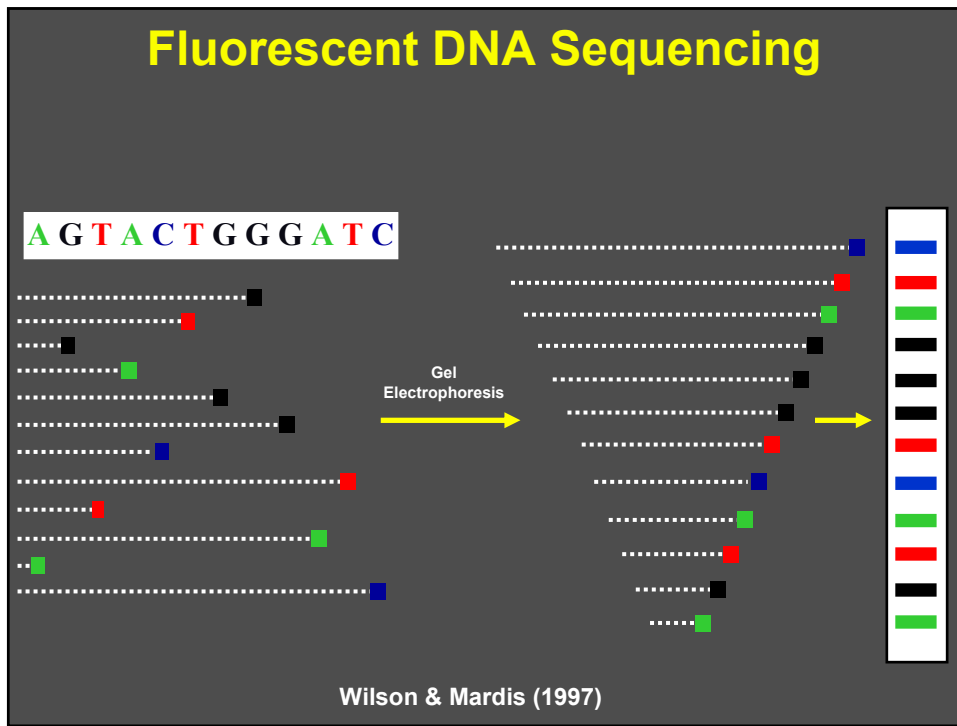
G A T C

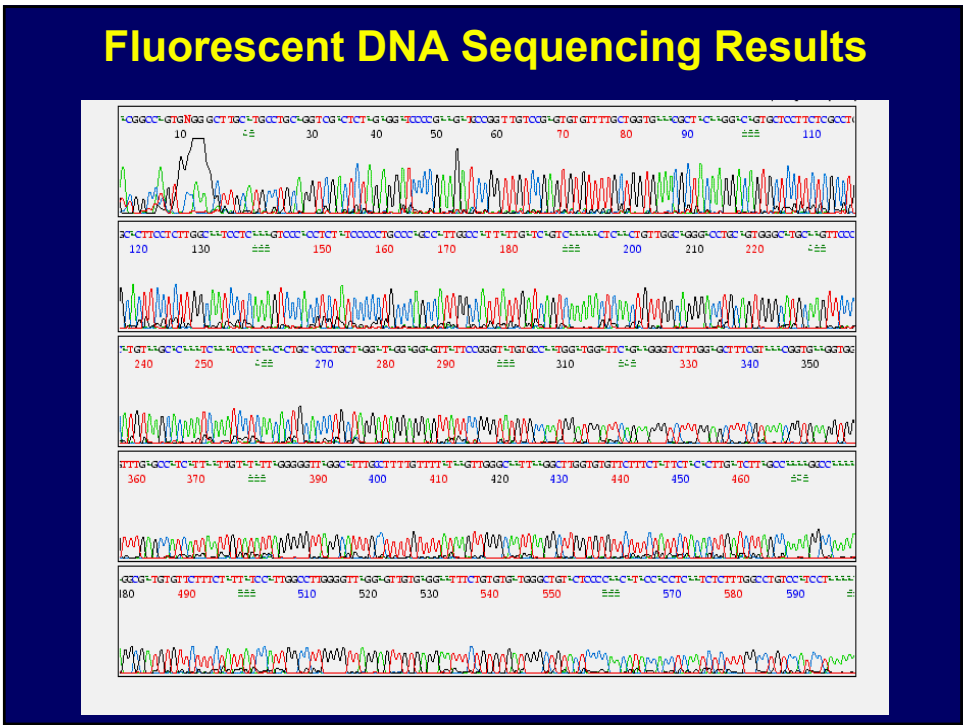
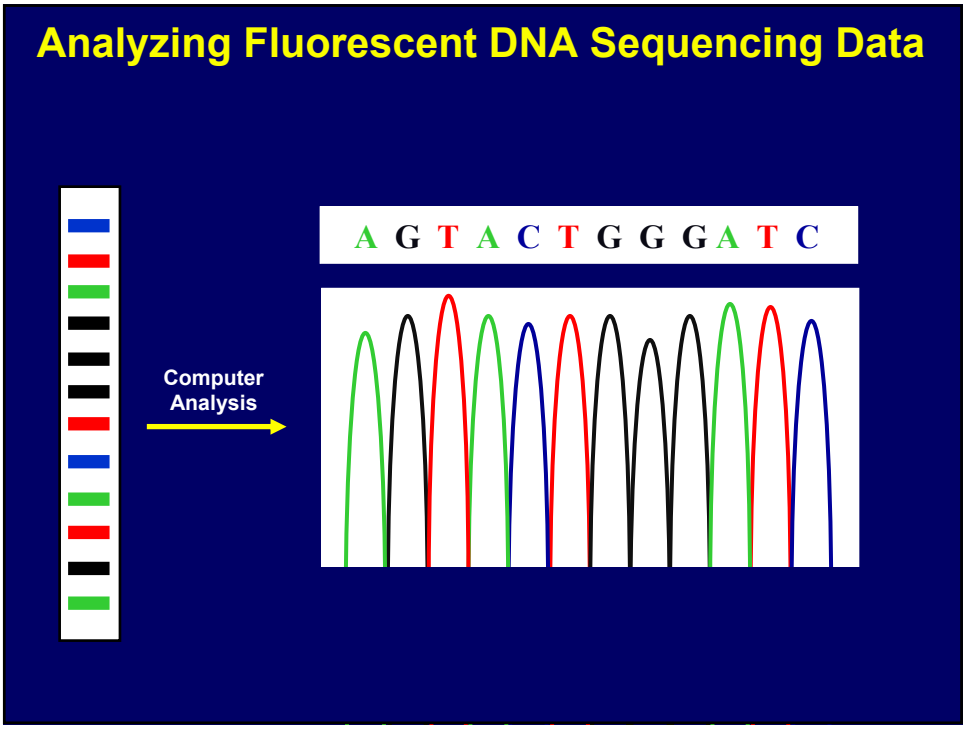


G: G Reaction
A: A Reaction
T: T Reaction
C: C Reaction

Radioactive Sequencing







Slab Gel-Based DNA Sequencing Instruments



Capillary-Based DNA Sequencing Instruments



Large-Scale cDNA Sequencing

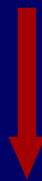
- ESTs: Expressed-Sequence Tags
- SAGE: Serial Analysis of Gene Expression
- Full-Insert (Full-Length) cDNA Sequencing



mgc.nci.nih.gov

Gerhard et al. (2004)

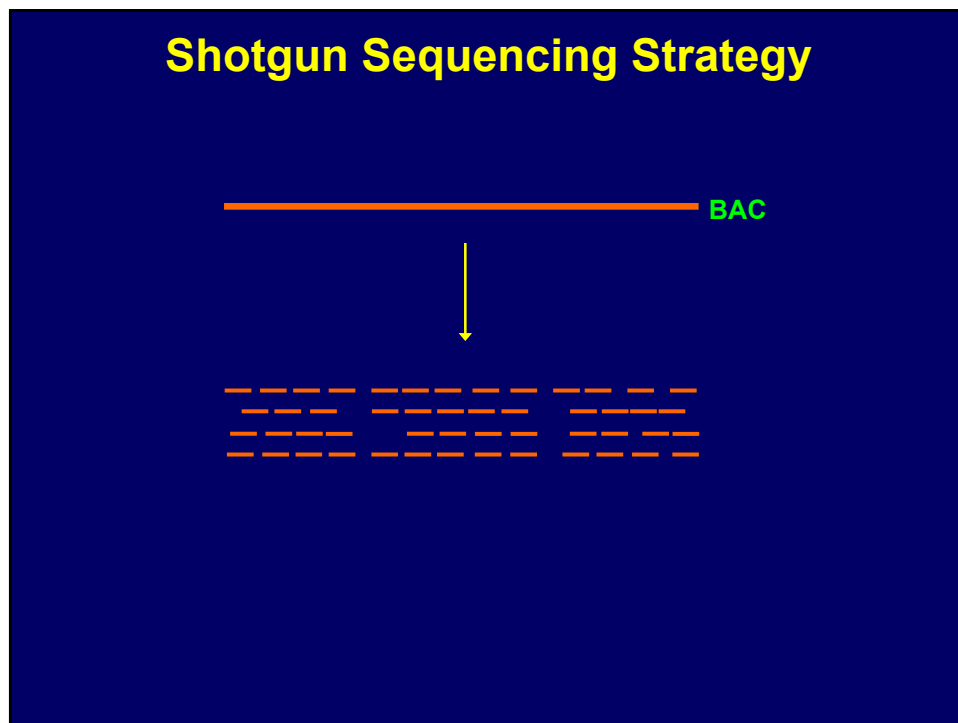
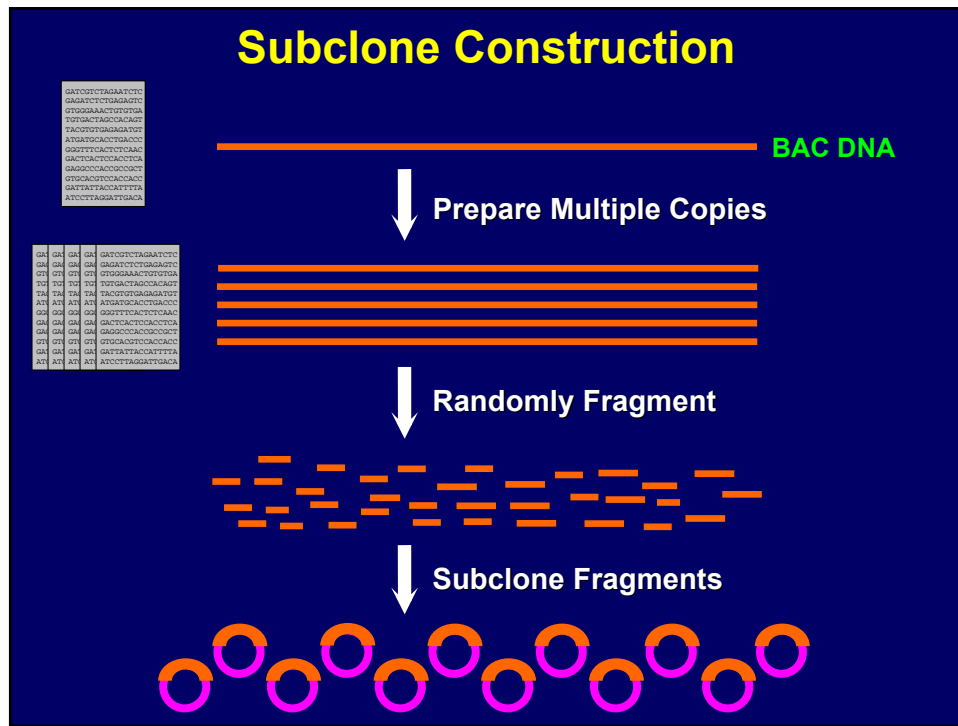
Large-Scale Genome Sequencing



Shotgun Sequencing

Wilson & Mardis (1997)

Green (2001)



Poisson Calculations

The sequencing strategy for the shotgun approach follows the Lander and Waterman application of the Poisson distribution

The probability a base is not sequenced is given by:

$$P_0 = e^{-c}$$

Where:

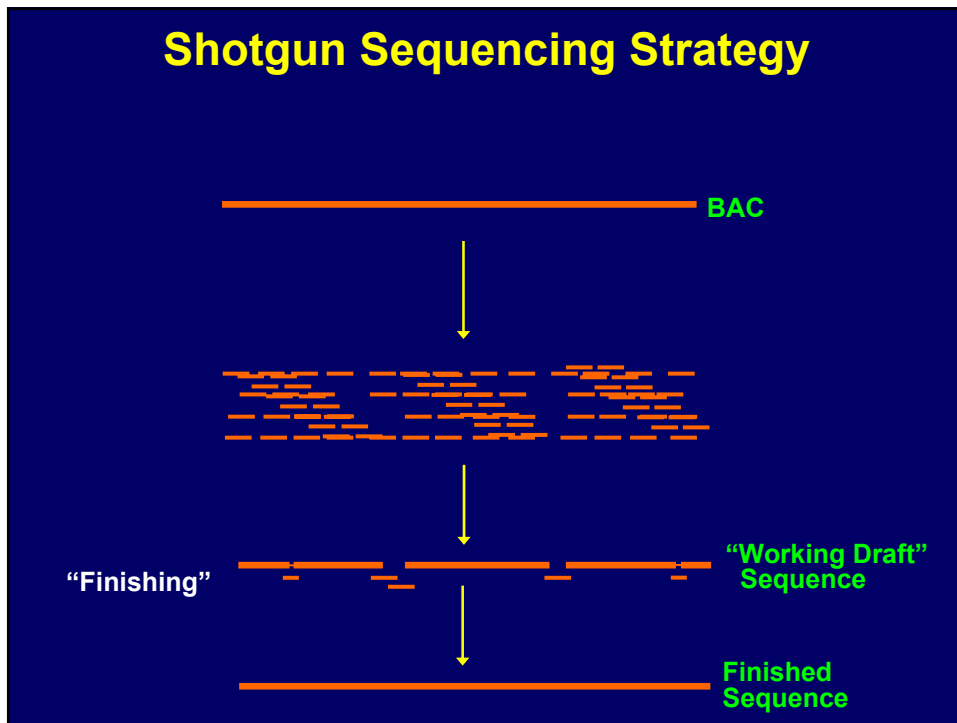
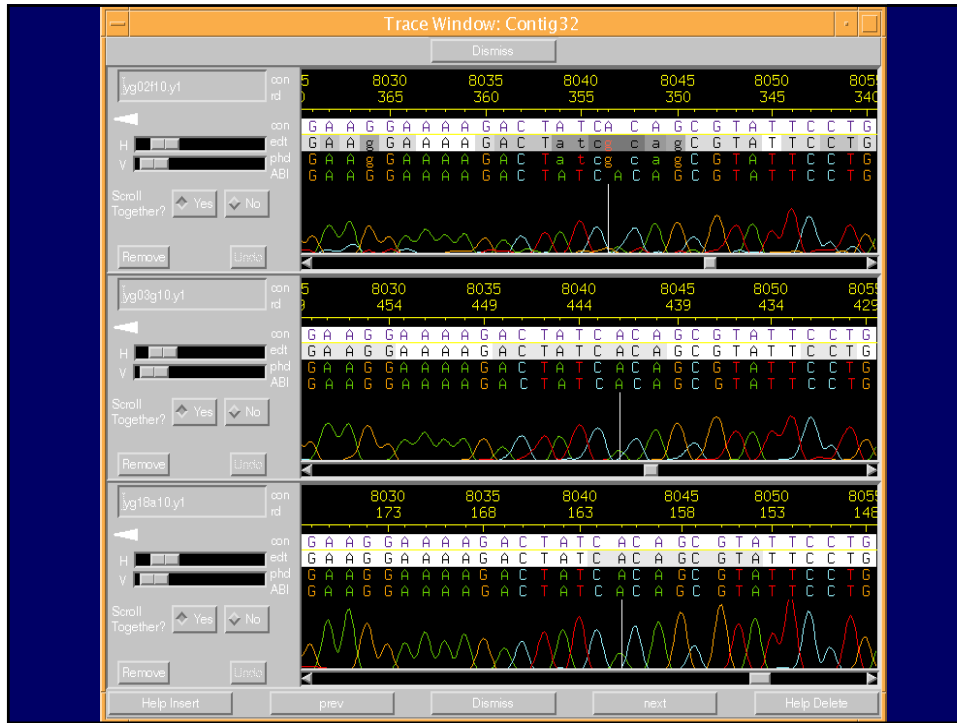
- c = fold sequence coverage ($c=LN/G$),
- LN = # bases sequenced, i.e. L = average sequencing read length and N = # reads
- G = target sequence length
- $e = 2.718$ ($e=2.718281828459$)

Fold Coverage	$P_0=e^{-c}$	% not sequenced	% sequenced
1	0.37	37%	63%
2	0.135	13.5%	87.5%
3	0.05	5%	95%
4	0.018	1.8%	98.2%
5	0.0067	0.6%	99.4%
6	0.0025	0.25%	99.75%
7	0.0009	0.09%	99.91%
8	0.0003	0.03%	99.97%
9	0.0001	0.01%	99.99%
10	0.000045	0.005%	99.995%

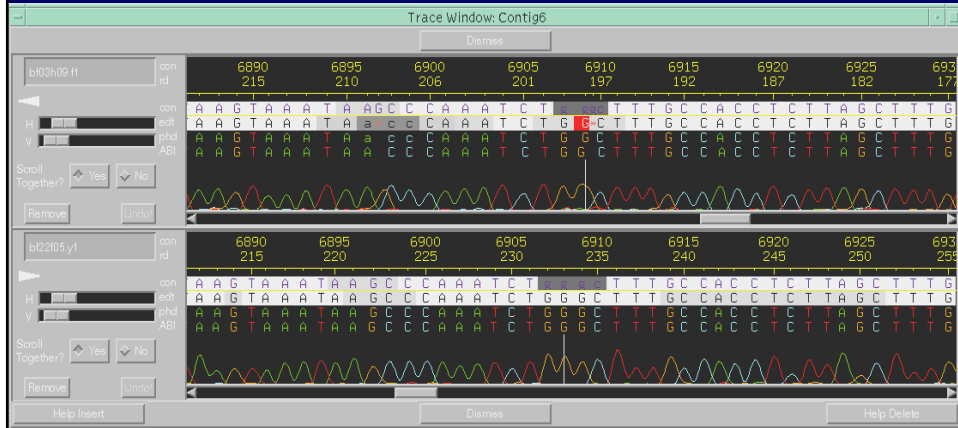
Shotgun Sequence Assembly

The screenshot shows a sequence alignment viewer window titled "aligned reads". The main display area shows a consensus sequence at the top, followed by several individual sequencing reads. The reads are aligned to the consensus sequence, with some mismatches indicated by arrows. The consensus sequence is: `AGGAAAAGACTATCACAGCGTATTCTGAAAGAGATGAACTATGAATTGAGTGTAGGCTTCTCTGCAGAGGCCAAA*GGTAGGAT`. The reads include: `yg12h02.x1`, `yg03d09.y1`, `yg09g04.x1`, `yg13h04.x1`, `yg01e03.y1`, `yg08h10.x1`, `yg04f11.y1`, `yg01g01.y1`, `yg01g07.y1`, `yg02e04.y1`, `yg02f10.y1`, `yg02c10.y1`, `yg03g10.y1`, `yg18a10.y1`, `yg08f02.y1`, `yg02h10.y1`, `yg18e09.y1`, and `yg13d05.y1`. The viewer also includes a search bar and various navigation options.

“Consed” (Gordon et al., 1998)



Sequence Finishing: Resolving Ambiguities



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

Historically Significant Genome Sequencing Projects

Microbial Genome Sequences

TIGR-CMR

Click here to take the CMR user feedback survey

Home|Genomes|Searches|Comparative Analyses|Gene Lists|Carts|Downloads

Welcome to the Comprehensive Microbial Resource (CMR) Home Page

The Comprehensive Microbial Resource (CMR) is a free website used to display information on all of the publicly available, complete prokaryotic genomes. In addition to the convenience of having all of the organisms on a single website, common data types across all genomes in the CMR make searches more meaningful, and cross genome analysis highlight differences and similarities between the genomes. [More Information] [Publication Information]

NEW on the CMR

August 11, 2006: CMR Feedback Survey: TIGR is currently seeking funding for the maintenance and expansion of the CMR. Please help us with our grant application by taking our user feedback survey.

For information on the latest CMR updates, subscribe to the CMR Mailing List.

Prokaryotic Annotation and Analysis Classes: June 13-15, August 8-10, October 10-12, 2006

Visit a CMR page for an individual genome

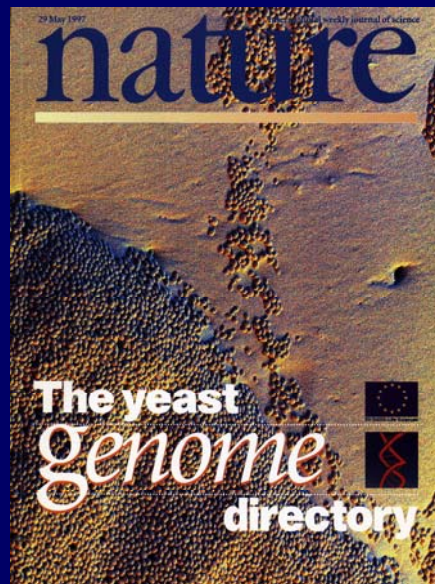
- Acidithiobacillus ferrooxidans ATCC 23270
- Acinetobacter sp. ADP1
- Actinomyces naeslundii MG1
- Aeropyrum pernix K1
- Agrobacterium tumefaciens C58 Cereon
- Agrobacterium tumefaciens C58 UWash
- Anabaena variabilis ATCC 29413
- Anaplasma marginale St Maries

CMR Genomes: Data Release 19.0

	Complete	Incomplete	Totals
Bacteria	279	17	296
Archaea	23	0	23
Viruses	3	0	3
Totals	305	17	322

www.tigr.org

First Eukaryotic Genome Sequence



Goffeau et al. (1997)

First Animal Genome Sequence

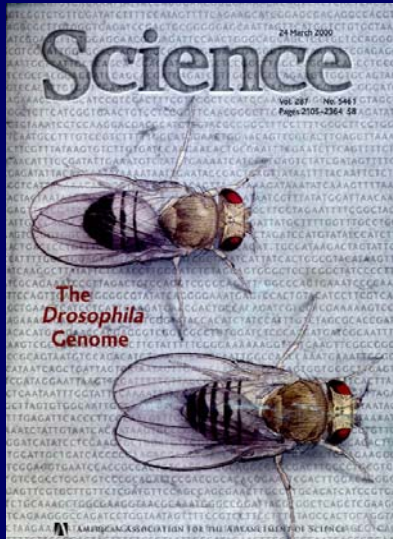


Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology

The *C. elegans* Sequencing Consortium*

C. elegans Sequencing Consortium (1998)

Second Animal Genome Sequence



The Genome Sequence of *Drosophila melanogaster*

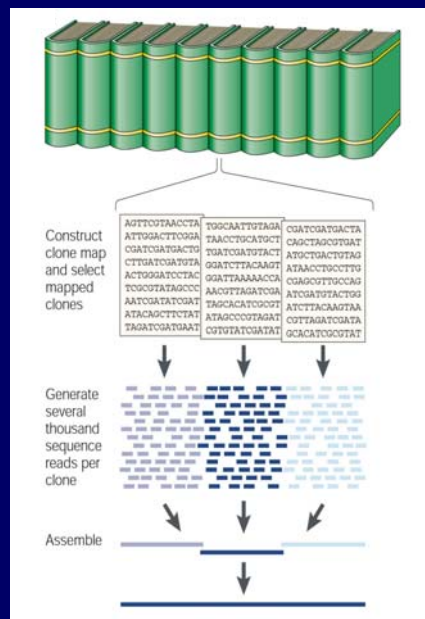
Mark D. Adams,^{1*} Susan E. Celniker,² Robert A. Holt,¹ Cheryl A. Evans,¹ Jeannine D. Gocayne,¹ Peter G. Amanatides,¹ Steven E. Scherer,¹ Peter W. Li,¹ Roger A. Hoskins,¹ Richard F. Gallie,¹ Reed A. George,¹ Suzana E. Lewis,¹ Stephen Richards,¹ Michael Ashburner,¹ Scott W. Henderson,¹ Granger G. Sutton,¹ 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 Champs,¹ Barrett D. Pfeiffer,¹ Kenneth H. Wan,¹ Clare Doyle,¹ Evan G. Baxter,¹ Gregg Heist,¹ Catherine R. Nelson,¹ George L. Gaber Miklos,¹ Joseph F. Alari,¹ Anna Aghayani,¹ Hai Jin An,¹ Cynthia Andrews-Pfannkoch,¹ Danita Baldwin,¹ Richard M. Ballwey,¹ Anand Ban,¹ James Bonkendale,¹ Leyla Bayraktaroglu,¹ Ellen M. Beasley,¹ Karen Y. Beeson,¹ P. V. Benos,¹ Benjamin P. Berman,¹ Deepali Bhandari,¹ Slava Bolshakov,¹ Dana Borokova,¹ Michael R. Botchan,¹ John Bouch,¹ Peter Brokstein,¹ Philippe Brottier,¹ Kenneth C. Burtis,¹ Dana A. Busam,¹ Heather Butler,¹ Edward Cadieu,¹ Angela Center,¹ Ishwer 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 Mays,¹ Ian Dew,¹ Suzanne M. Dietsi,¹ Kristina Dodson,¹ Lisa E. Doup,¹ Michael Dovey,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 Fiolet,¹ Andrei E. Gabrielian,¹ Neha S. Garg,¹ William M. Galbraith,¹ Ken Glasser,¹ Anna Glodok,¹ Fangcheng Gong,¹ J. Harley Gorrell,¹ Zhiping Guo,¹ Ping Guo,¹ Michael Harris,¹ Naomi L. Harris,¹ Damon Harvey,¹ Thomas J. Heiman,¹ Judith H. Hernandez,¹ Jarrett Houck,¹ Damon Houston,¹ Kathryn A. Houston,¹ Timothy J. Howland,¹ Ming-Hui Wei,¹ Chiyere Ibegwam,¹ Mena Jalali,¹ Francis Kalush,¹ Gary H. Karpen,¹ Zhaod Ke,¹ James A. Kennison,¹ Karen A. Ketchum,¹ Bruce E. Kimmel,¹ Chinnappa D. Kodira,¹ Cheryl Kraft,¹ Saul Kravitz,¹ David Kulp,¹ Zhongwei Lai,¹ Paul Lasko,¹ Yiding Lai,¹ Alexander A. Lavinsky,¹ Jinyin Li,¹ Zhanya Li,¹ Yong Liang,¹ Xiaoying Lin,¹ Xiangjun Liu,¹ Bettina Mattat,¹ Tina C. Mcintosh,¹ Michael P. McLeod,¹ Duncan McPherson,¹ Gennady Merkulov,¹ Natalia V. Milshina,¹ Clark Moberly,¹ Joe Morris,¹ All Moshrefi,¹ Stephen M. 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 Nicox,¹ Deborah R. Nuskern,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 Bixiang Christopher Shuo,¹ Inga Siöden-Klamos,¹ Michael Simpson,¹ Marian P. Skupski,¹ Tom Smith,¹ Eugene Spier,¹ Allan C. Spradling,¹ Mark Stapleton,¹ Renee Strong,¹ Eric Sun,¹ Robert Swkrak,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 M. Weinstock,¹ Jean Weissenbach,¹ Sherita H. Williams,¹ Trevor Woodage,¹ Kim C. Wortley,¹ David Wu,¹ Song Yang,¹ Q. Allison Yao,¹ Jian Ye,¹ Bo-Yang Yeh,¹ Jayshree S. Zaveri,¹ Ming Zhao,¹ Guoqiang Zhang,¹ Qi Zhao,¹ Liansheng Zheng,¹ Xiangqun H. Zheng,¹ Fei N. Zhong,¹ Wenyan Zhong,¹ Xiaojun Zhou,¹ Shaoqing Zhu,¹ Xiaohong Zhu,¹ Hamilton O. Smith,¹ Richard A. Gibbs,¹ Eugene W. Myers,¹ Gerald H. Rubin,¹ J. Craig Venter¹

Adams et al. (2000)

Human Genome Sequencing Centers

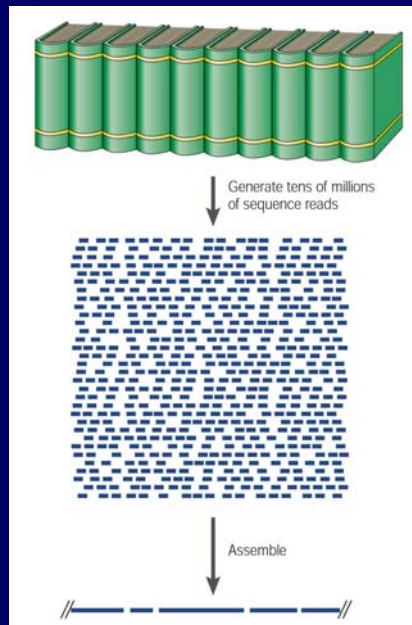


Clone-Based Shotgun Sequencing



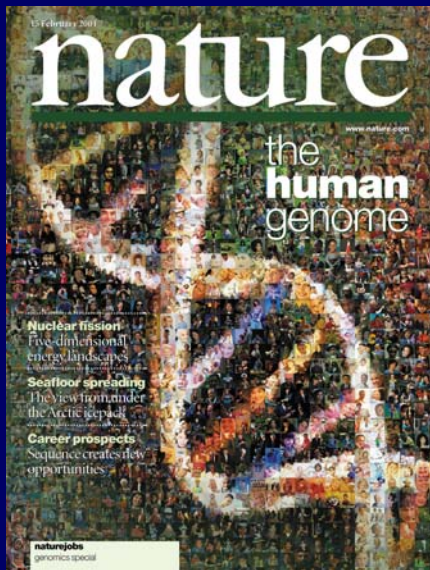
Green (2001)

Whole-Genome Shotgun Sequencing

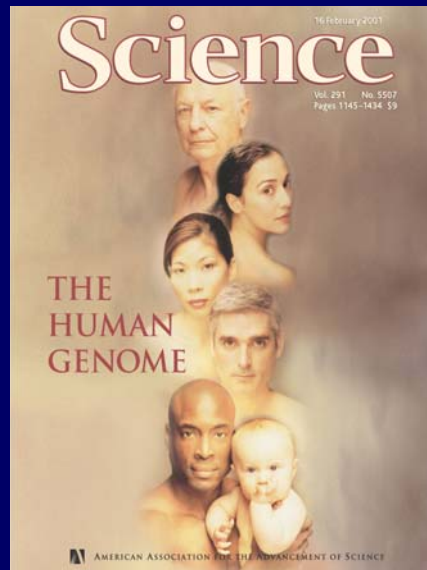


Green (2001)

February, 2001 Draft Sequence



International Human Genome Sequencing Consortium (2001)



Venter et al. (2001)

April, 2003 Completion



October, 2004 Publication

21 October 2004

International journal of science

nature

www.nature.com/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 end of this finishing process. The current genome sequence (build 30) contains 2.95 billion nucleotides interrogated by only 241 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 error per 100,000 bases. Half of the remaining euchromatic gaps are associated with repetitive elements 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 seems to encode only 23,000–25,000 protein-coding genes. The genome sequence reported here should serve as a fine foundation for biomedical research in the decades ahead.

The International Genome Project (IGP) was launched in 1986 with the goal of obtaining a high-quality sequence of the nucleotide order 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 to the study of inherited disease and provide a useful scaffold for genome assembly; and (2) the sequencing of sequences with simple, repetitive elements to serve as a method for scaffold development and assist in interpreting 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 centers in six continents, was formed to carry out the completion of the IGP.

In February 2001, the IHGSC[†] and Celera Genomics[‡] 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 gene, candidate and structural variation, regional differences in genome composition, distribution and history of transposable elements, distribution of polymorphisms and relationships between genetic, morphological and physical distance. However, systematic knowledge of the human genome has enabled new work and approaches that have accelerated scientific biomedical research.

Both draft sequences, however, had important shortcomings. The IHGSC sequence, for example, covered ~90% of the euchromatic genome. It was annotated by ~130,000 genes and the number and structure of many sequences and features had not been established. The Celera sequence, while containing a complete sequence of the euchromatic genome, operationally a finished genome of 1.27 billion bases, and the goal of complete coverage in finished gaps, was not achieved. The draft sequences were not well annotated, and the quality of the sequences was not uniform. The IHGSC sequence is higher with such features as repeat regions and large repetitive elements, which greatly complicate the determination of gene structure and content. In fact, some complete sequences have been obtained so far only for three nucleotide sequences: the mouse[§], the rat^{||} and the fly[¶]. These genomes are all roughly 10-fold smaller than the human genome and have much simpler structure.

We describe here the results of a worldwide effort by the IHGSC to finish the sequence of the human genome. The number of gaps has been reduced 40-fold to only 241, most of which are associated with repetitive elements and will require new methods for resolution. The assembled near-complete genome sequence has an error rate of only ~1 error per 100,000 bases. It contains 2.95 billion nucleotides and covers ~99% of the euchromatic genome. This paper describes the current genome sequence and the process used to produce it, maintain the accuracy and completeness of the sequence, and illustrate biological evidence made possible by the sequence. We do not attempt here a comprehensive review of the content of the human genome, an initial analysis was previously reported[†] and a series of papers is being written describing the individual chromosomes^{††}, including association of genes and other features.

Current genome sequence

Human genome

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 generally simultaneously at multiple scales, ranging from regional resolution to the mapping of whole chromosomes. The fundamental challenge in this process is regions that are not well represented or easily resolved through random shotgun sequencing tend to be highly enriched in problematic sequences. Finishing such regions required the development of special approaches, which varied substantially over time and varied among centers.

Initially, the finishing process involved one-to-one comparisons: (1) produce finished gaps, covering of contigs and overlaps; (2) produce large-scale clones, covering of contigs and overlaps; (3) produce large-scale clones, covering of contigs and overlaps; (4) produce finished gaps, covering of contigs and overlaps. In each large-scale clone, the process, these two comparisons were highly automated in that projects. In each clone, the results from the other. The comparisons are described in Boxes 1 and 2. Further information about the finishing process and finishing standards can be found in the Supplementary Information (Boxes 1) and 2. <http://www.genome.gov/25520001>

In total, we generated a draft sequence from 28,208 large-scale clones (total length ~1.24 gigabases (Gb)) and finished the sequence from 43,742 of these clones (total length ~1.47 Gb). The clones contained primarily all bacterial artificial chromosome

© 2004 Nature Publishing Group

International Human Genome Sequencing Consortium (2004)

CNN's #1 Medical Story of Past 25 Years

CNN.com PRINT THIS
Powered by clickability

[SAVE THIS](#) | [EMAIL THIS](#) | [Close](#)

Top 25: Medical stories

Human genome mapping ranks No. 1 in health news

Tuesday, March 29, 2005 Posted: 4:24 PM EST (2124 GMT)


(CNN) -- Much of the marvel of medicine has to do with discovery. Mapping the human genome, the complete sequence of DNA, gave scientists a blueprint for building a person, making it the No. 1 medical story, according to a distinguished panel CNN gathered to rank the top 25 medical stories of the past quarter-century.

Two men from two separate groups -- Francis Collins of the National Institutes of Health and Craig Venter of Celera Genomics Inc., a pharmaceutical-development company -- worked independently to discover the sequence of the human genome and identify the genes that it contains. This

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.



**DOUBLE
HELIX
TO
HUMAN
SEQUENCE**

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

What's Next?



feature

A vision for the future of genomics research

A blueprint for the genomic era.

Francis S. Collins, Eric D. Green, Alan E. Guttmacher and Mark S. Gerber on behalf of the US National Human Genome Research Institute

The completion of a high-quality, comprehensive sequence of the human genome – in the 100th anniversary year of the discovery of the double-helical structure of DNA, is a landmark event. The genome era is now under way.

In contemplating a vision for the future of genomics research, it is appropriate to consider the remarkable progress that has brought us here. The road to Figure 1 shows a timeline of landmark accomplishments in genetics and genomics, beginning with Gregor Mendel's discovery of the laws of heredity and their rediscovery in the early days of the twentieth century. The sequencing of DNA and the development of recombinant DNA technology, and the establishment of increasingly automatable methods for DNA sequencing – in the case for the Human Genome Project (HGP) to begin in 1990, see also www.nature.com/nature/HGP. Thanks to the vision of the original planners, and the creativity and determination of a legion of talented scientists who decided to make this project their overriding focus, all of the initial objectives of the HGP have now been achieved at least two years ahead of expectation, and a revolution in biology has begun.

The project's new research strategies and experimental techniques have generated a steady stream of ever larger and more complete genomic data that have been posted into public databases and have transformed the study of virtually all life processes. The genomic approach of technology development and large-scale generation of community resource data sets has introduced an important new dimension into biological and medical research. Intersectoral advances in genetics, computer genomics, high-throughput biochemistry and microarrays are providing biologists with a markedly improved repertoire of research tools that will allow the harnessing of organisms to health and disease to be studied and comprehended at an unprecedented level of molecular detail. Genome sequencing, the hallowed sets of information that guide biological development and function, lie at the heart of this revolution. In short, genomics has become a central and cohesive discipline of biological research.

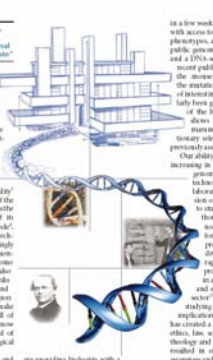
The practical consequences of the emergence of this new field are widely apparent. Identification of the genes responsible for human mendelian disorders once a herculean task, requiring large research teams, many years of hard work, and an enormous outlay of funds, can now be routinely accomplished

in a few weeks by a single graduate student with access to DNA samples and associated phenotypes, an Internet connection to the public genome databases, a thermal cycler and a DNA sequencing machine. With the recent publication of a draft sequence of the mouse genome¹, identification of the mouse orthologs underlying a vast number of interesting human phenotypes has become a greatly simplified task. Comparison of the human and mouse sequences shows that the proportion of the mammalian genome under evolutionary selection is more than twice that previously assumed.

Our ability to explore genome function is increasing in specificity as each subsequent genome is sequenced. Microarray technologies have catalogued many alternative transcripts underlying the expression of one or two genes in a month, or thousands of genes in a single afternoon². Critical opportunities for gene-based pre-emptive medicine, prediction of illness and adverse drug response are emerging at a rapid pace, and the therapeutic promise of genomics has captured the imagination of the commercial sector³. The investment of the HGP in studying the ethical, legal and social implications of these scientific advances has created a talented cohort of scholars in ethics, law, social science, clinical research, theology and public policy, and has already resulted in substantial increases in public awareness and the introduction of significant new and innovative protective social measures such as genetic discrimination law (www.genscreen.gov/ELSI).

These accomplishments fulfill the expectations stated in the 1990 report of the National Research Council, Mapping and Sequencing the Human Genome⁴. The successful completion of the HGP thus represents an opportunity to look forward and offer a blueprint for the future of genomics research over the next several years.

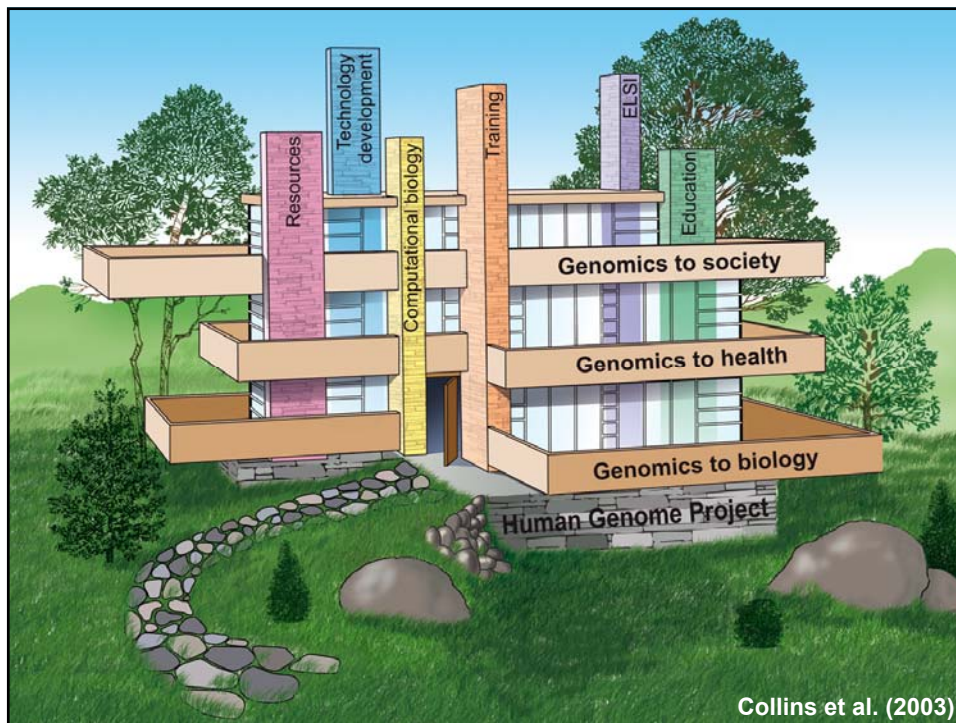
The vision presented here, although a different world from that reflected in earlier plans published in 1990, 1993 and 1995 (refs. 15–17). These documents addressed the goals of the 1990 report, defining detailed paths towards the development of genomic



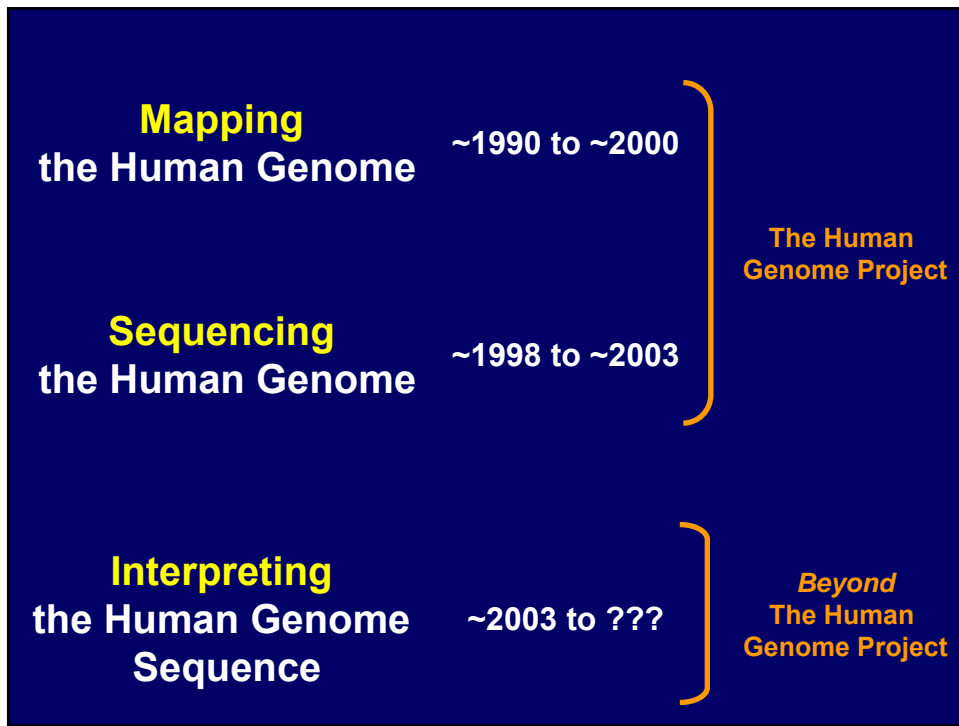
838

© 2003 Nature Publishing Group

Collins et al. (2003)



Collins et al. (2003)



Foundational Milestones in Genetics & Genomics



Darwin

1859



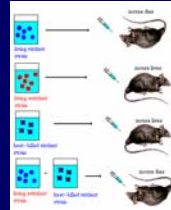
Mendel

1865



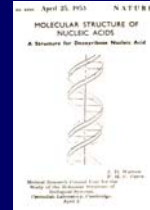
Miescher

1871



Avery

1944



Watson
& Crick

1953

Comparing Genomes is Like Cryptography

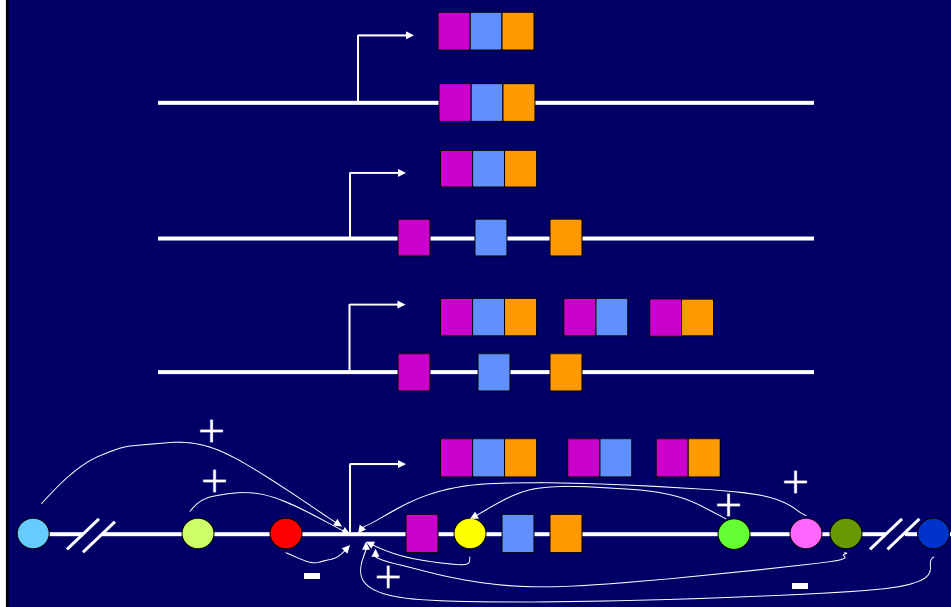
CKQEBHEREYTWASULSCZMEISDFOGETHEBLPBGODEFQSTLKS TUFFRAC

DLUCEHEREZBRTTOISAWNDCDARJJP THERROFGOODERGHCLSTUFFBRHA

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

The Language of the Genome



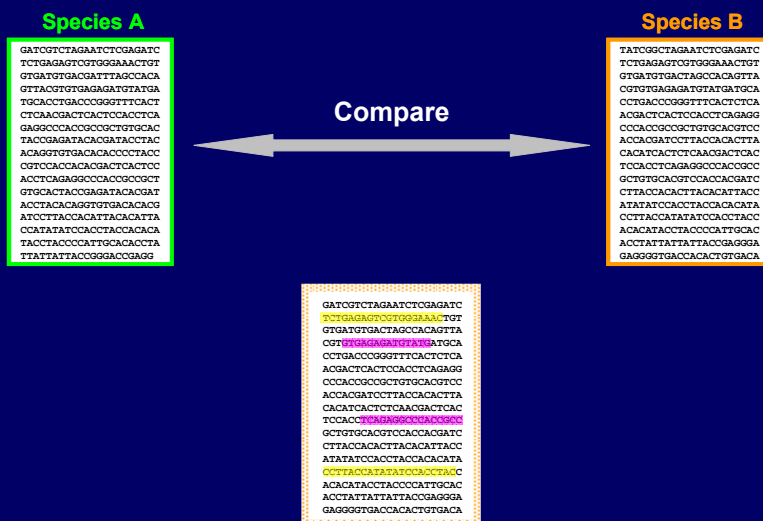
Functional Elements: Coding vs. Non-Coding

- **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

Comparative Sequence Analysis

Using the Experiments of Evolution to Decode the Human Genome



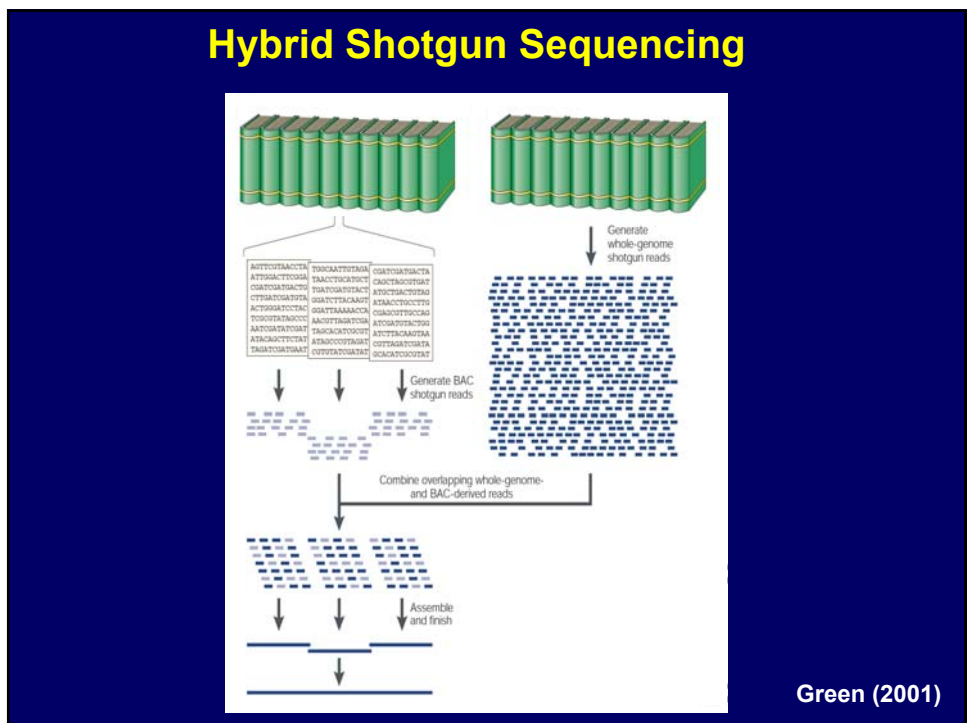
Sequences in Common (i.e., 'Conserved' or 'Constrained')

Vertebrate Genome Sequences

Mouse Rat Chicken Chimpanzee Dog

Macaque Orangutan Marmoset Cow Monodelphis Platypus

Xenopus Zebrafish Pufferfish

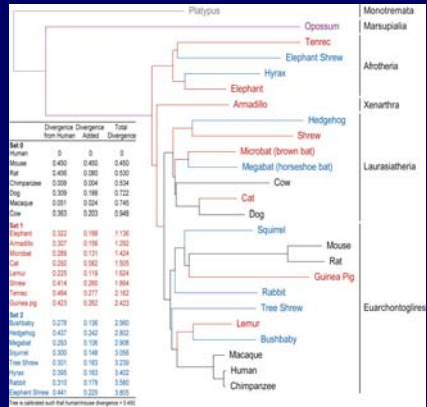


Low-Redundancy, Whole-Genome Shotgun Sequencing

An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing

Elliott H. Margulies^{1*}, Jade Vinson^{1*}, NISC Comparative Sequencing Program^{1*}, Webb Miller¹, David B. Jaffe¹, Kerstin Lindblad-Toh¹, Jean Chang¹, Eric D. Green^{1*}, Eric S. Lander¹, James C. Mullikin^{1*}, and Michele Clamp^{1**}

Margulies EM et al. (2005)



Landscape of Vertebrate Genome Sequencing

Human	=====
Mouse	=====
Rat	=====
Pufferfish	=====
Zebrafish	=====
Chicken	=====
Chimpanzee	=====
Dog	=====
Cow	=====
Xenopus	=====
Monodelphis	=====
Macaque	=====
Platypus	=====
Marmoset	=====
Orangutan	=====
Armadillo
Elephant
Tenrec
Rabbit
Cat
Shrew
Guinea Pig
Hedgehog
(and others...)	

Multi-Species Sequence Comparisons



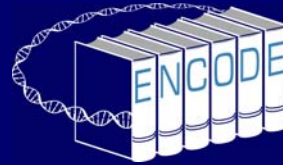
Multi-Species Conserved Sequences (MCSs)

Margulies et al. (2003)

Future Genomes to Sequence???



ENCODE Project



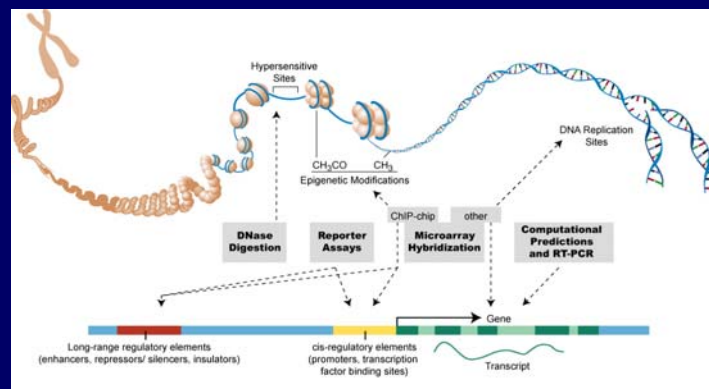
- **ENCODE: ENCyclopedia Of DNA Elements**
- **Goal: Compile a *Comprehensive Encyclopedia of All Functional Elements in the Human Genome***
- **Initial Pilot Project: 1% of Human Genome**
- **Apply Multiple, Diverse Approaches to Study and Analyze that 1% in a Consortium Fashion**

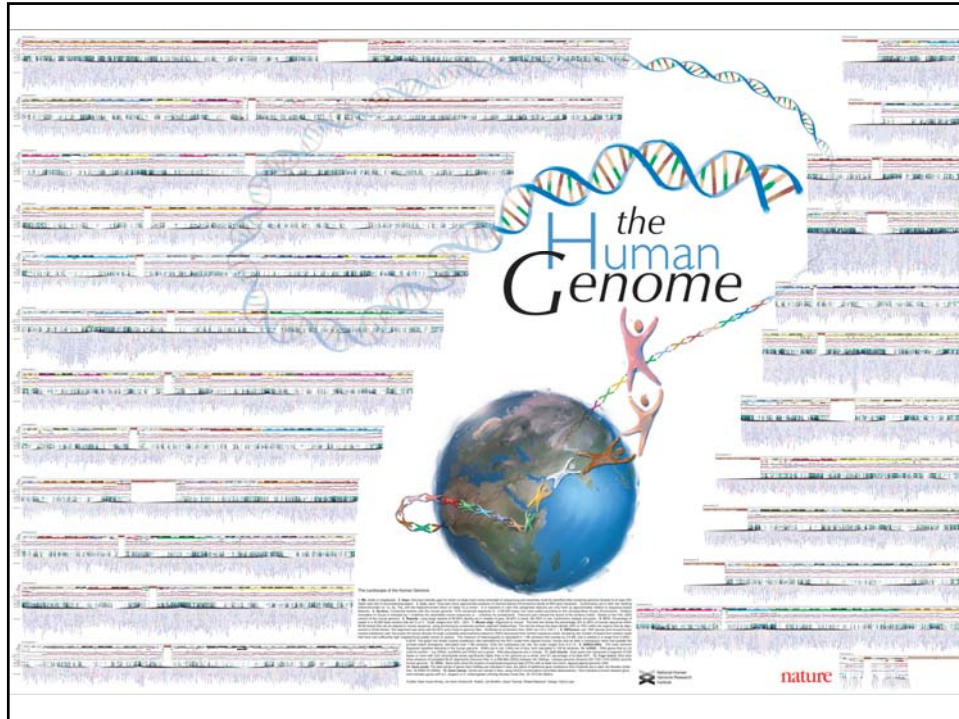
SPECIAL SECTION
GENES IN ACTION
VIEWPOINT

The ENCODE (ENCyclopedia Of DNA Elements) Project

The ENCODE Project Consortium*†

ENCODE Project Consortium (2004)






Human Genome Sequence

>\$1,000,000,000



~\$100,000

~\$1,000



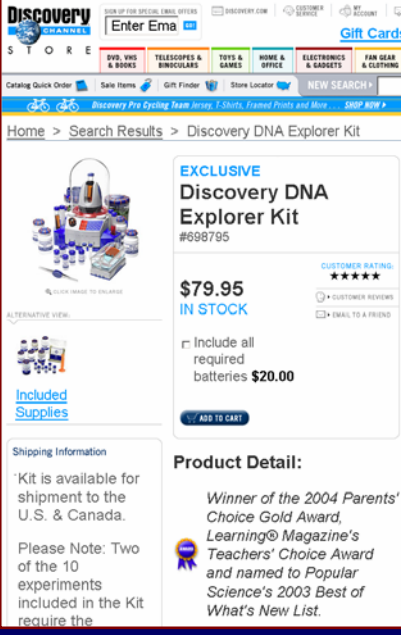
WIRED TOOLS 2K3 [TOYS] by Chris Baker

MY FIRST DNA SEQUENCE

The genetic code experiments that are containing green-up guidelines are recombining with each other. However, organisms can also give-powered running control tracks. With a programmable robot, you'll get started in the basic genome. If these aren't sophisticated enough for your genome system, the Discovery Kids DNA Explorer has just what you need. As you get started, you'll get a lot of light on the way of being able to work. Next step: cloning!

DNA Explorer Laser 10 and 100. www.discovery.com

- **Prep the Specimens**
Before extracting DNA, your specimen has to go through a few steps. The experiment works on all kinds of food: like corn, beans, or even ground chicken. The kit includes three DNA ground animal (DNA) in order to control the experiment. And, when all these are put in a beaker and mixed with distilled water, the plant, animal, or chicken starts to break down.
- **Separate the Oils**
After transferring the mixture to a test tube, Dr. Frankenstein needs to add a little soap. The tube goes inside a special green magnetic bar and centrifuge, which pulls up the oils and separates the DNA from the mixture. After 10 minutes, Dr. Frankenstein is a piece of soap, with animal and the DNA mixture. This is the surface, where they can be transferred with the "DNA stick".
- **Zip the Molecules**
To read the code, DNA, Dr. Frankenstein has to zap it with a laser. The gel goes into a battery-powered electrophoresis chamber, where it's pulled with a gel to make bands for the fragments. The molecules are transferred with a pipette. A tape of sequencing reads the molecules, which are negatively charged - moving through the gel.
- **Unravel the Mystery**
A couple of hours later, Dr. Frankenstein will see a few strips of paper that contain the code. The kit also includes a decoder, which will translate the code into a sequence. You'll get to see the code and the sequence.



Discovery CHANNEL Enter Email [Gift Cards](#)

STORE DVD, VHS & BOOKS TELESCOPES & BINOCULARS TOYS & GAMES HOME & OFFICE ELECTRONICS & GADGETS FAN GEAR & CLOTHING

Discovery Pro Cycling Team Jersey, T-Shirts, Trained Prints and More... [SHOP NOW](#)

Home > Search Results > Discovery DNA Explorer Kit

EXCLUSIVE
Discovery DNA Explorer Kit
#698795

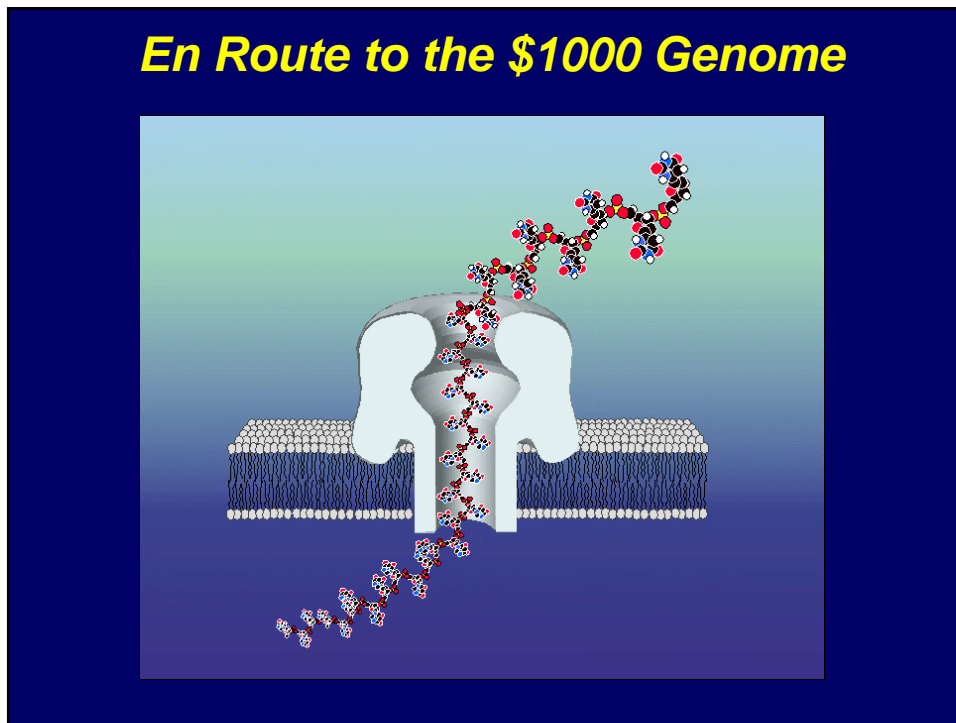
\$79.95
IN STOCK

Include all required batteries **\$20.00**

ADD TO CART

Shipping Information
Kit is available for shipment to the U.S. & Canada.

Product Detail:
Winner of the 2004 Parents' Choice Gold Award, Learning Magazine's Teachers' Choice Award and named to Popular Science's 2003 Best of What's New List.



Genome sequencing in microfabricated high-density picolitre reactors

Marcel Margulies¹, Michael Egholm¹, William E. Altman¹, Said Attiya¹, Joel S. Bader¹, Lisa A. Bemben¹, Jan Berka¹, Michael S. Braverman¹, Yi-Ju Chen¹, Zhoutao Chen¹, Scott B. Dewell¹, Lei Du¹, Joseph M. Fierro¹, Xavier V. Gomes¹, Brian C. Godwin¹, Wen He¹, Scott Helgesen¹, Chun He Ho¹, Gerard P. Irzyk¹, Szilveszter C. Jando¹, Maria L. Alenquer¹, Thomas P. Jarvie¹, Kshama B. Jirage¹, Jong-Bum Kim¹, James R. Knight¹, Janna R. Lanza¹, John H. Leamon¹, Steven M. LeKowitz¹, Ming Lei¹, Jing Li¹, Kenton L. Lohman¹, Hong Lu¹, Vinod B. Makhijani¹, Keith E. McDade¹, Michael P. McKenna¹, Eugene W. Myers¹, Elizabeth Nickerson¹, John R. Noble¹, Ramona Plant¹, Bernard P. Puc¹, Michael T. Ronan¹, George T. Roth¹, Gary J. Sarkis¹, Jan Fredrik Simons¹, John W. Simpson¹, Mithreyan Srinivasan¹, Karrie R. Tartaro¹, Alexander Tomasz², Karl A. Vogt¹, Greg A. Volkmer¹, Shailly H. Wang¹, Yong Wang¹, Michael P. Weiner¹, Pengguang Yu¹, Richard F. Beigley¹ & Jonathan M. Rothberg¹

Margulies M et al. (2005)

Solexa Ltd

*Simon Bennett PhD
Business Development
Director*
Solexa Ltd is developing an integrated system, based on a breakthrough single molecule sequencing technology, to address a US\$2 billion market that is expected to grow

Bennett (2004)

Toward the \$1000 human genome

*Simon T Bennett,
Cedric Barnes,
Anthony Cox,
Lisa Chavez &
Clive Brena**
Revolutionary new technologies, capable of transforming the economics of sequencing, are providing an unparalleled opportunity to analyze human genetic variation comprehensively at the whole-genome level within a realistic timeframe and at affordable costs. Current estimates suggest that it would cost somewhere in the region of US\$30 million to sequence

Bennett et al. (2005)

Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome

Jay Shendure^{1,†}, Gregory J. Porreca^{1,†}, Nikos B. Reppas¹, Xiaoxia Lin¹, John P. McCutcheon^{2,3}, Abraham M. Rosenbaum¹, Michael D. Wang¹, Kun Zhang¹, Robi D. Mitra², George M. Church¹

Shendure et al. (2005)

**Perspective
Emerging technologies in DNA sequencing**

Michael L. Metzker
Human Genome Sequencing Center and Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030, USA

Demand for DNA sequence information has never been greater, yet current Sanger technology is too costly, time consuming, and labor intensive to meet this ongoing demand. Applications span numerous research interests, including comparative studies, comparative genomics and evolution, forensics, and diagnostic and applied therapeutics. Several emerging technologies show promise of delivering next-generation solutions for fast and affordable genome sequencing. In this review article, the DNA polymerase-dependent strategies of Sanger sequencing, single nucleotide addition, and cyclic reversible termination are discussed to highlight recent advances and potential challenges these technologies face in their development for strand DNA sequencing.

More than just a mapping and sequencing endeavor, the Human Genome Project (HGP) has altered the mindset and approach to many basic and applied research efforts. Early disruption and controversy (Goodstadt 1990; Lusa et al. 1989; Roberts 1989); Fox et al. 1996) were soon laid to rest by well-developed strategies (Bowers 1997a; Collins and Galis 1993; Collins et al. 1998) that led to the successful completion of mankind's largest biology project. As the cost of the HGP was technology development that advanced the pace of sequencing a mammalian genome from years to months. Along the way, numerous strategies emerged that hold promise for rapid, efficient, and accurate delivery of DNA sequence information. For the HGP, a hybrid approach was adopted for completing the job by coupling the core technology of Sanger sequencing and fluorescence detection. The completion of the sequencing phase could not have been accomplished without major innovations in continuous process engineering, fluorescence dye development, capillary electrophoresis, microarrays, robotics, informatics, and process management. The result was completion of a high-quality, reference sequence of the human genome in April, 2003 (Collins et al. 2003), marking the 50-year anniversary of the discovery of the double-helix structure. For many outside the genome community, that historic milestone signaled the end of this international scientific project, but for the rest of us, it only marked the beginning of things to come.

The need for sequencing has never been greater than it is today, with applications spanning diverse research sectors including comparative genomics and evolution, forensics, epidemiology, and applied medicine for diagnostics and therapeutics. Arguably, the strongest rationale for emerging sequencing in the genes for identification and interpretation of human sequence variation is in relation to health and disease. The most common form of variation is the single nucleotide polymorphism (SNP), although two unrelated people share, on average, 99.9% sequence identity (i.e., one difference in a thousand base pairs), the average occurrence of an SNP in the general population is once every five hundred base pairs. As such, more than nine million SNPs have been cataloged in the public database, dbSNP (Zwick and Hudson 2005), with many more expected to be found in large-scale sequencing efforts.

A great deal of attention has been focused on common SNPs with a minor allele frequency >5% and their potential role in common disease (Lander 1996; Bach and Moebis 1996; Collins et al. 1997). Broad, large-scale genotyping efforts of these common SNPs have shown that much of the human genome can be parsed into common haplotype blocks (Daly et al. 2001; Feil et al. 2001; Gabriel et al. 2002). The International HapMap Consortium (2003) was formed to characterize common patterns of sequence variation by densifying allele frequencies and the degree of association between SNPs among geographically diverse groups, leading to the identification of "tagSNPs" for genome-wide, disease-based association studies. With this method of characterization, however, rare SNPs/haplotypes may be overlooked, as highlighted by Lu et al. (2005), who described an association of rare variants/haplotypes with susceptibility to disease-based association studies. With this method of characterization, however, rare SNPs/haplotypes may be overlooked, as highlighted by Lu et al. (2005), who described an association of rare variants/haplotypes with susceptibility to disease-based association studies.

A hallmark in large-scale strategies from genotyping to sequencing is currently taking place to explore the significance of less-common SNPs to human biology and disease. The "m" in this approach is the sequencing of individual genomes related to a reference genome for de novo SNP discovery and comparative genomics application. The ENCODE Project Consortium (2004) has devoted significant efforts toward sequencing megabase-sized blocks of the human genome. Consequently, genome centers are now offering at least 100x, 20x, or 5x coverage, which currently translates to ~8% capacity, or sequencing hundreds to thousands of gene regions. This increase in attention for high-throughput sequencing will greatly facilitate studies to determine the genetic basis of susceptibility to common disease, cancer biology, and disease association in model and nonmodel organisms.

Current sequencing technologies are so expensive, labor intensive, and time consuming for broad application to human sequence variation studies. Genome center cost is calculated on the basis of dollars per 1000 CpG bases (defined below) and can be generally divided into the categories of microarrays, previous, reagents and materials, and covered expenses. Currently, these costs are separating at less than one dollar per 1000 CpG bases, with at least 50% of the cost resulting from DNA sequencing instrumentation alone. Developments in novel detection methods, instrumentation in instrumentation, microfluidic separation technologies, and an increase in the number of assays per run will soon likely have the biggest impact on reducing cost. It should be emphasized, however, that new sequencing strategies will be needed to rise these high-throughput platform efficiency. In September, 2004, the National Human Genome Res-

15370-15376 ©2005 by Cold Spring Harbor Laboratory Press, ISSN 1088-0486, www.genome.org
Genome Research 17:67
www.genome.org

Metzker (2005)

TSUNAMI SCIENCE: ONE YEAR AFTER THE WAVE THAT ROCKED THE WORLD

SCIENTIFIC AMERICAN

Alternatives to Toxic Tests on Animals

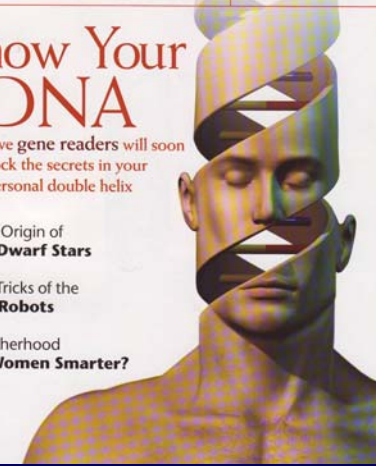
JANUARY 2006 \$4.99
WWW.SCIAM.COM

Know Your DNA
Inexpensive gene readers will soon unlock the secrets in your personal double helix

The Hazy Origin of Brown Dwarf Stars

Winning Tricks of the Racing Robots

Does Motherhood Make Women Smarter?



Church (2006)

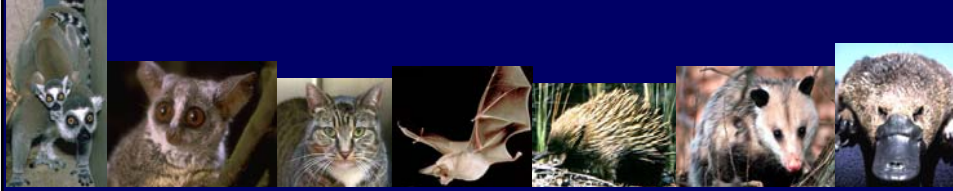
DNA Sequencing Technologies

<u>Method</u>	<u>Feasibility</u>	<u>Read Length</u>	<u>Data Quality</u>	<u>Raw Data Production</u>
Sanger Sequencing	Well Established	Long (800-1200 bases)	+++	+
Stepwise Synthesis	Becoming Established	Short (25-100 bases)	+	+++++
Single Molecule	Far from Established	Long (>1000 bases ?)	???	+++++ ?

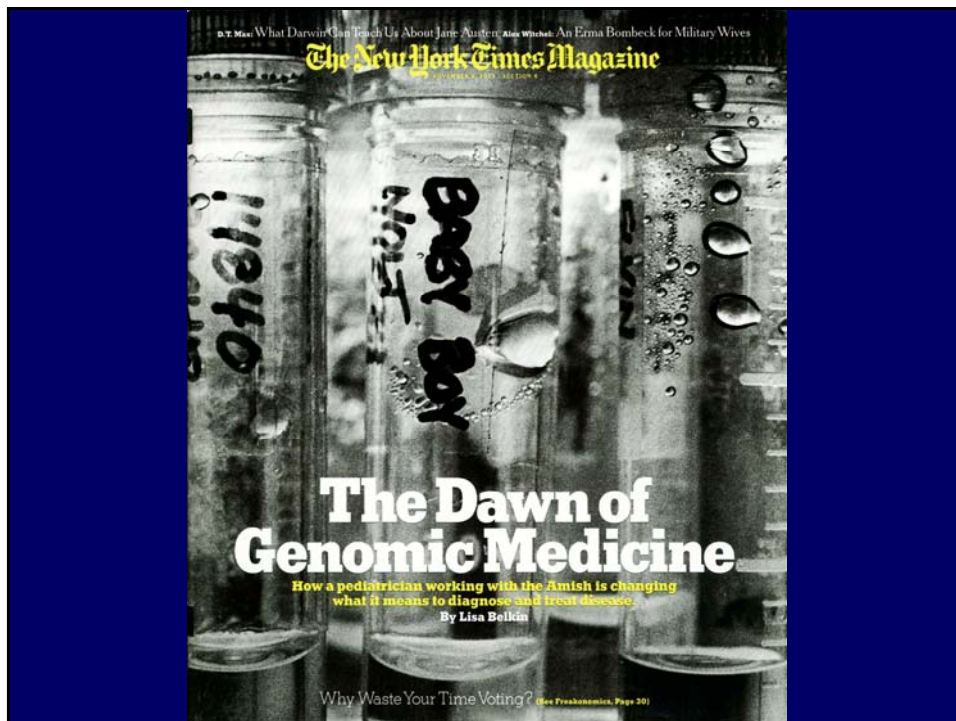
Realities of New DNA Sequencing Technologies...



“Inter-Species” Comparisons



“Intra-Species” Comparisons



The Pathway to Genomic Medicine



HGP



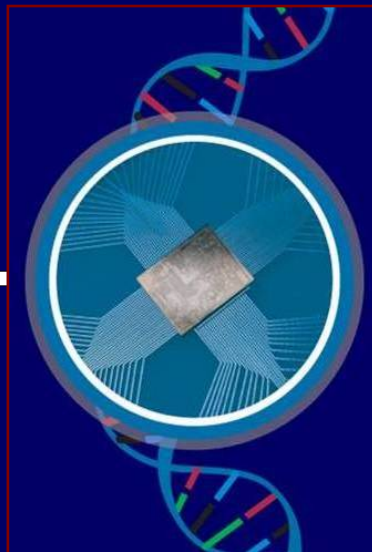
Realization of Genomic Medicine



The Pathway to Genomic Medicine



HGP



Realization of Genomic Medicine



The Current Big Challenges...

- Defining “Saturation Points” in Terms of Information Gained by Comparative Sequence Analyses
- Achieving the “\$1000 Genome”
- Large-Scale Deployment of Medical Sequencing

The Human Genome Sequence to Genomic Medicine...



...from base pairs to bedside.

Bibliography

- Adams MD et al. (2000). The genome sequence of *Drosophila melanogaster*. *Science* 287:2185-2195.
- Bennett S (2004). Solexa Ltd. *Pharmacogenomics* 5:433-438.
- Bennett ST (2005). Toward the \$1000 human genome. *Pharmacogenomics* 6:373-382.
- Birren B et al. (1998). Bacterial artificial chromosomes. In *Genome Analysis: A Laboratory Manual, Vol. 3 Cloning systems* (B Birren et al., eds.; Cold Spring Harbor Laboratory Press), pp. 241-295.
- C. elegans* Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282:2012-2018.
- Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69-87.
- Church GM (2006). Genomes for all. *Sci Am* 294:46-54.
- Collins FS et al. (2003). A vision for the future of genomics research: a blueprint for the genomic era. *Nature* 422:835-847.
- ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306:636-640.
- Gerhard DS et al. (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res* 14:2121-2127.
- Goffeau A et al. (1997). The Yeast Genome Directory. *Nature* 387S:1-105.
- Gordon D et al. (1998). Consed: a graphical tool for sequence finishing. *Genome Res* 8:195-202.
- Green ED (2001). Strategies for the systematic sequencing of complex genomes. *Nature Rev Genet* 2:573-583.
- Green ED et al. (1998). Yeast artificial chromosomes. In *Genome Analysis: A Laboratory Manual, Vol. 3 Cloning systems* (B Birren et al., eds.; Cold Spring Harbor Laboratory Press), pp. 297-565.
- Gregory SG et al. (1997). Genome mapping by fluorescent fingerprinting. *Genome Res* 7:1162-1168.
- Hillier LW et al. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695-716.
- International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409:860-921.

- International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431:931-945.
- Lindblad-Toh K et al. (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438:803-819.
- Margulies EM et al. (2003). Identification and characterization of multi-species conserved sequences. *Genome Res* 13:2507-2518.
- Margulies EM et al. (2005). An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc Natl Acad Sci* 102:4795-4800.
- Margulies M et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.
- Marra MA et al. (1997). High throughput fingerprint analysis of large-insert clones. *Genome Res* 7:1072-1084.
- Messing J and Llaca V (1998). Importance of anchor genomes for any plant genome project. *Proc Natl Acad Sci* 95:2017-2020.
- Metzker ML et al. (2005). Emerging technologies in DNA sequencing. *Genome Res* 15:1767-1776.
- Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520-562.
- Rat Genome Sequencing Project Consortium (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493-521.
- Shendure J et al. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309:1728-1732.
- Thomas JW et al. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424:788-793.
- Venter JC et al. (2001). The sequence of the human genome. *Science* 291:1304-1351.
- Wilson RK and Mardis ER (1997). Fluorescence-based DNA sequencing. In *Genome Analysis: A Laboratory Manual, Vol. 1 Analyzing DNA* (B Birren et al., eds.; Cold Spring Harbor Laboratory Press), pp. 301-395.
- Wilson RK and Mardis ER (1997). Shotgun sequencing. In *Genome Analysis: A Laboratory Manual, Vol. 1 Analyzing DNA* (B Birren et al., eds.; Cold Spring Harbor Laboratory Press), pp. 397-454.