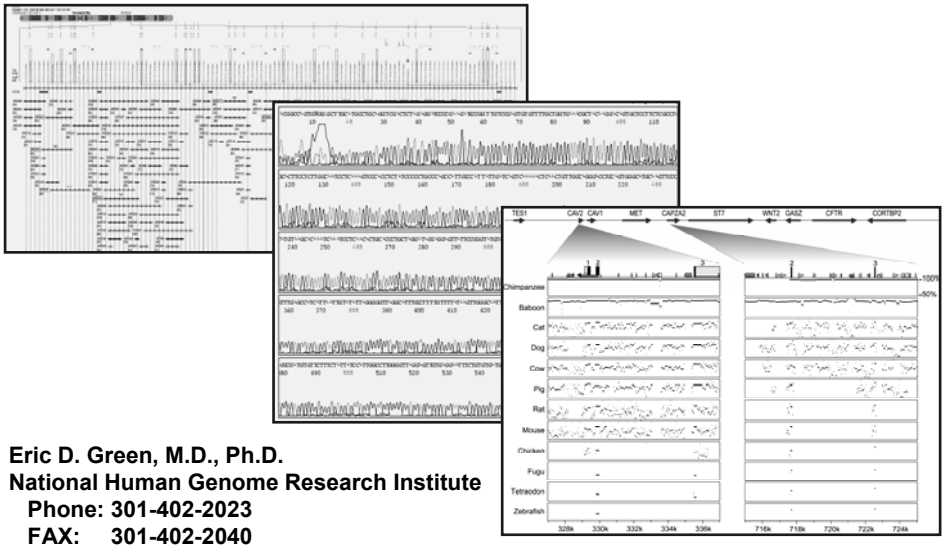
Techniques for Genome Mapping & Sequencing

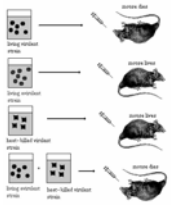


Eric D. Green, M.D., Ph.D.
National Human Genome Research Institute
Phone: 301-402-2023
FAX: 301-402-2040
E-Mail: egreen@nhgri.nih.gov

Foundational Milestones in Genetics & Genomics

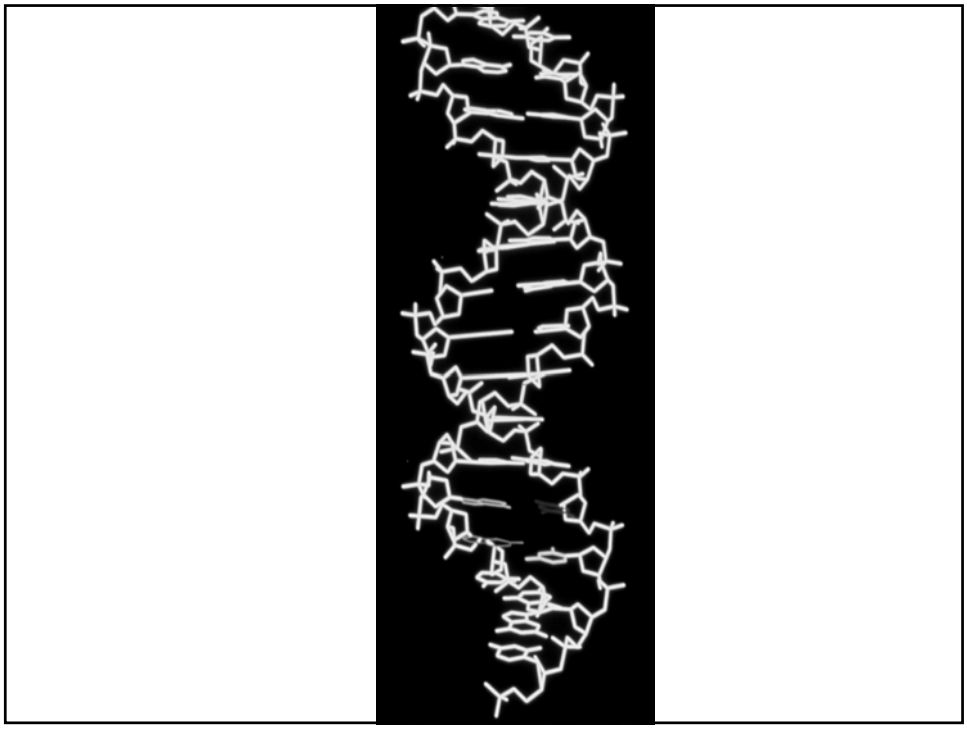
Mendel ***1865***

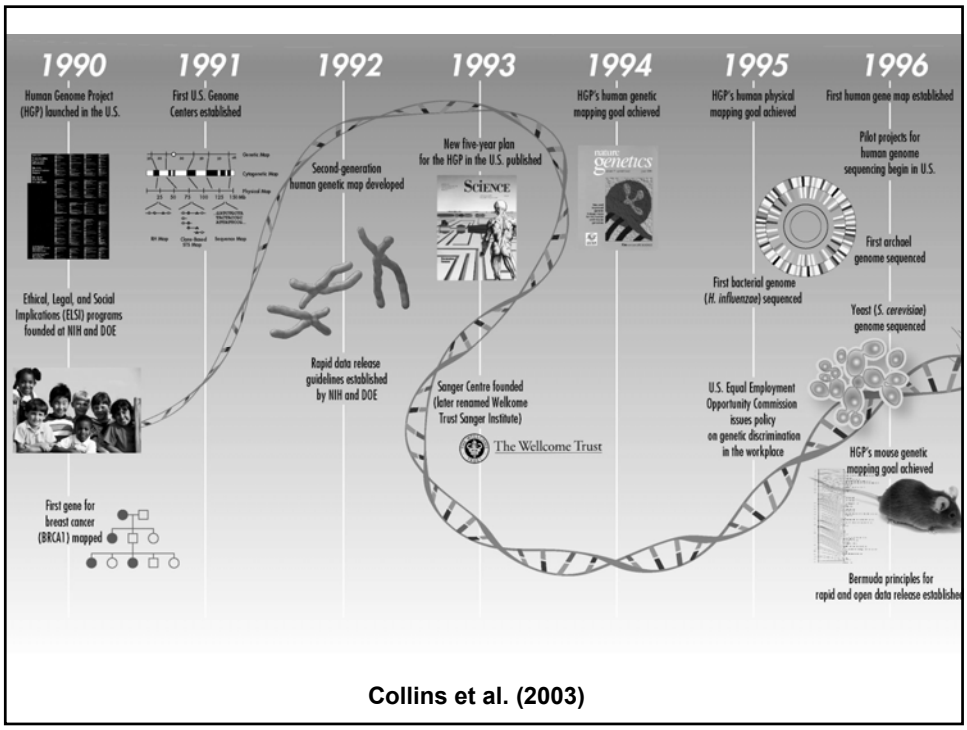
Miescher ***1871***



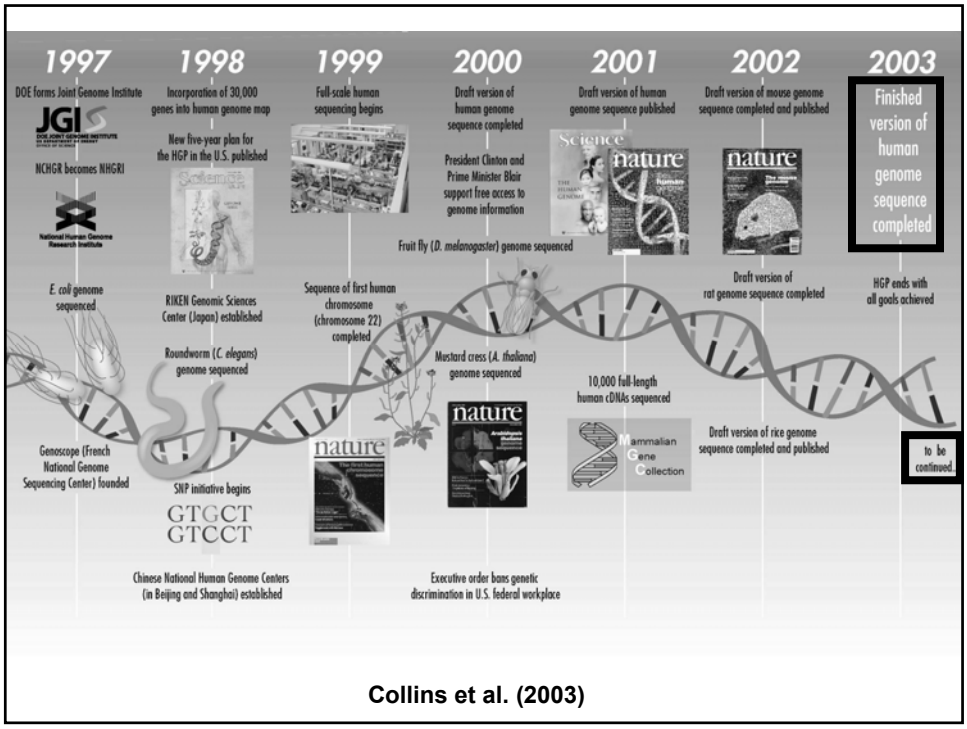
Avery ***1944***

Watson & Crick ***1953***





Collins et al. (2003)



Collins et al. (2003)

Outline

I. Fundamentals of Genome Mapping

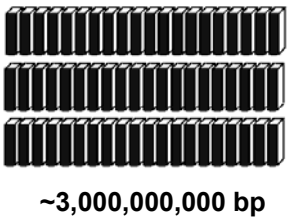
II. Fundamentals of Genome Sequencing

III. Mapping & Sequencing in the Human Genome Project

IV. Comparative Sequencing

V. New DNA Sequencing Technologies

Genome Sizes



Human Genome
Mouse Genome

~3,000,000,000 bp

Fruit Fly Genome

~160,000,000 bp

Nematode Genome

~100,000,000 bp

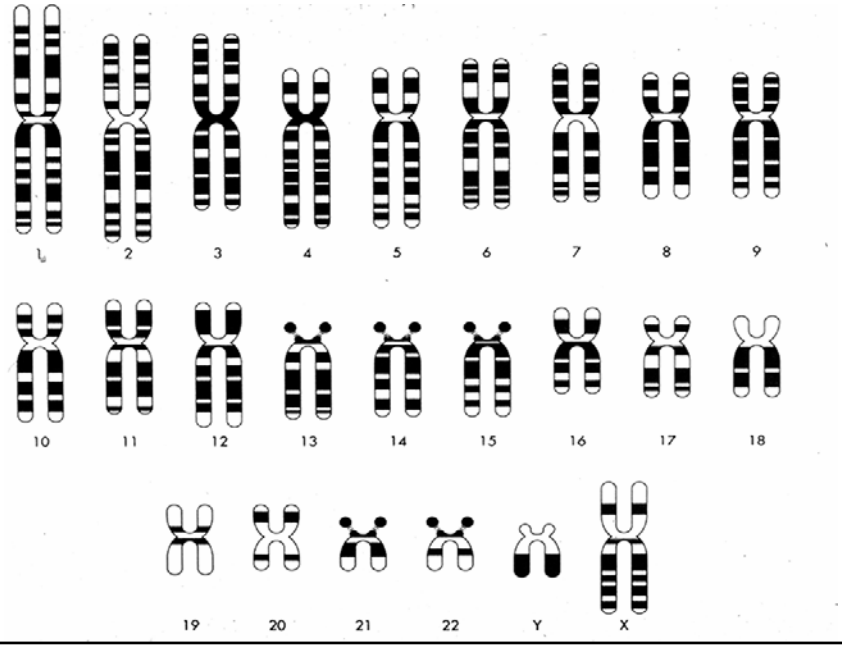
Yeast Genome

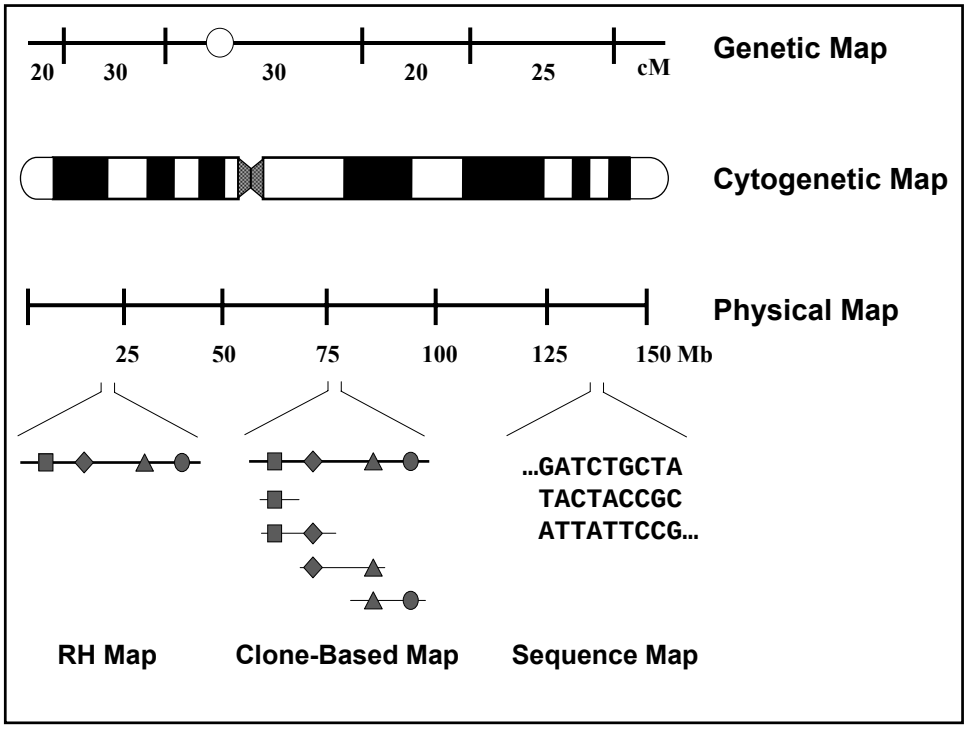
~15,000,000 bp

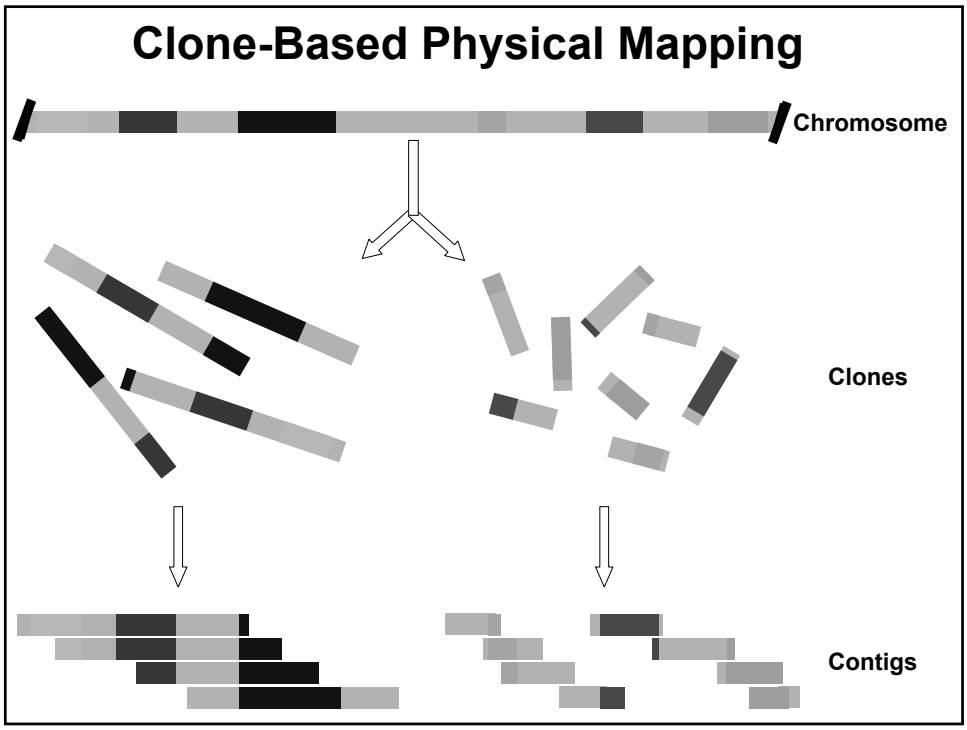
E. coli Genome

~5,000,000 bp

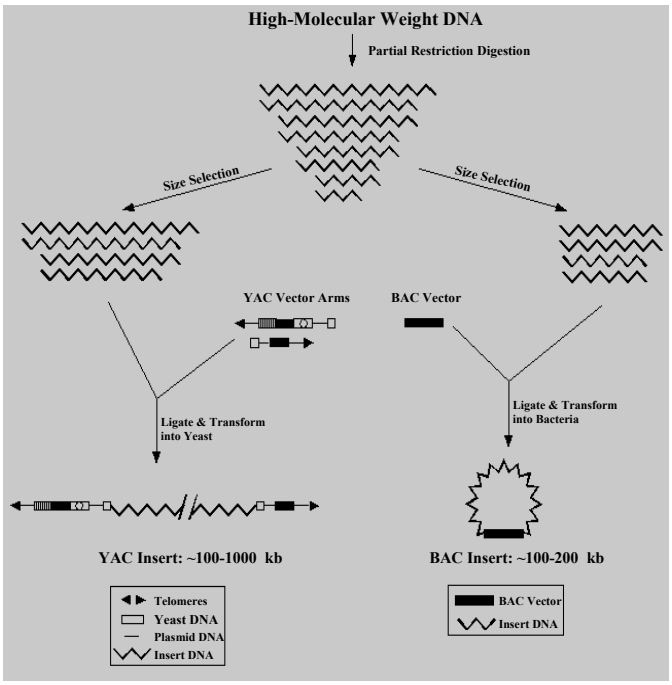
The Human Cytogenetic Map



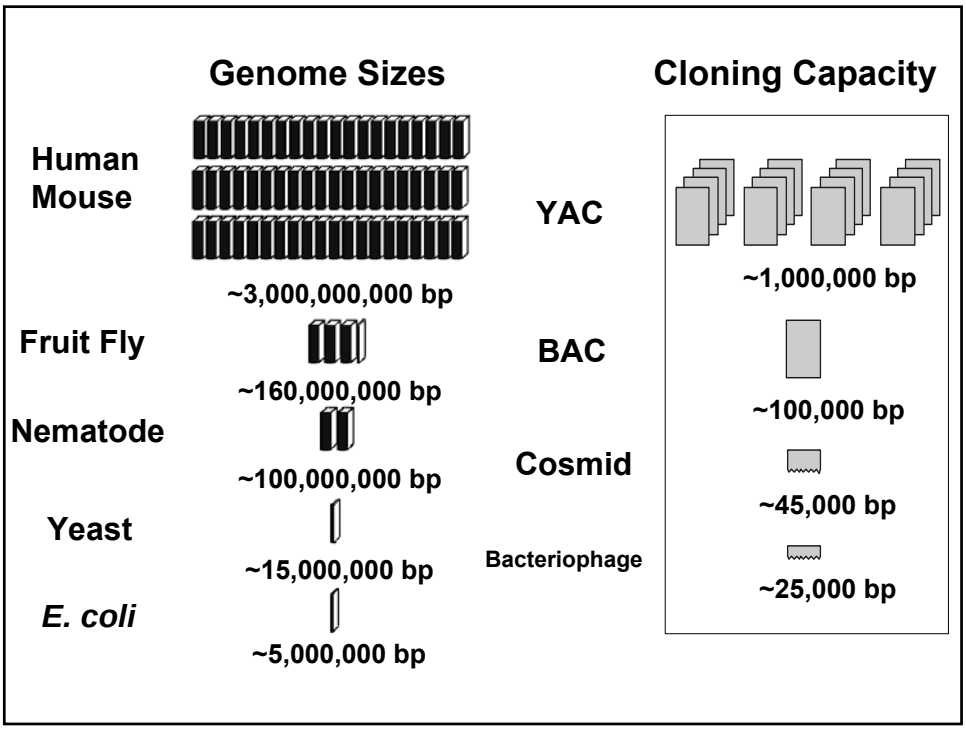

Genetic Map
20
30
30
20
25
cM
Cytogenetic Map
Physical Map
25
50
75
100
125
150 Mb
...GATCTGCTA
TACTACCGC
ATTATTCCG...
RH Map
Clone-Based Map
Sequence Map


Clone-Based Physical Mapping
Chromosome
Clones
Contigs

Construction of YACs and BACs



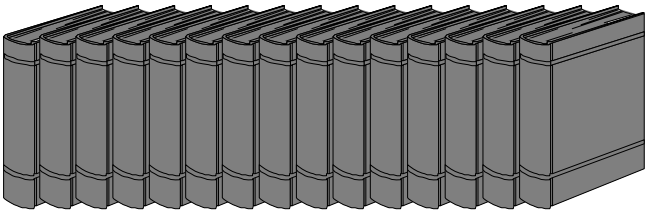
Green et al. (1998)
Birren et al. (1998)



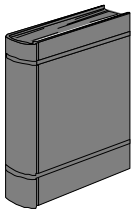
Bacterial Artificial Chromosomes (BACs)

- 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

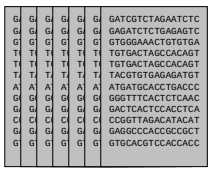
  Low Yields of DNA by Standard Methods
  Reasonably Stable
- BAC Libraries from Many Different Species Available
  (e.g., www.chori.org/bacpac)
- See Birren et al. (1998)



Genome
(~3000 Mb)



Chromosome
(~130 Mb)

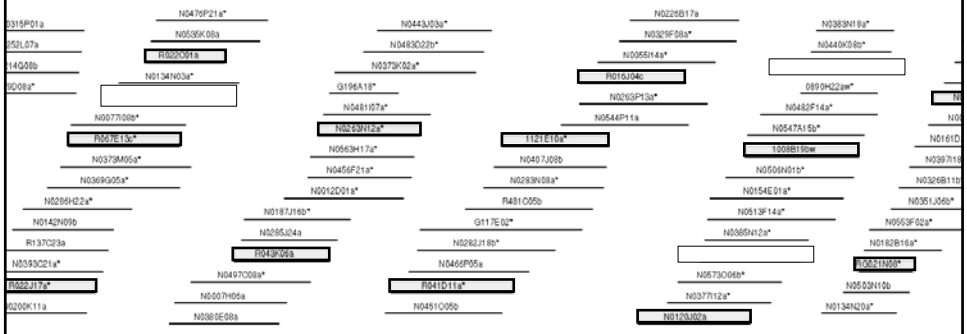


YAC
(~0.5-1.0 Mb)

BAC
(~0.1-0.2 Mb)

Sequence-Ready Contig Map

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

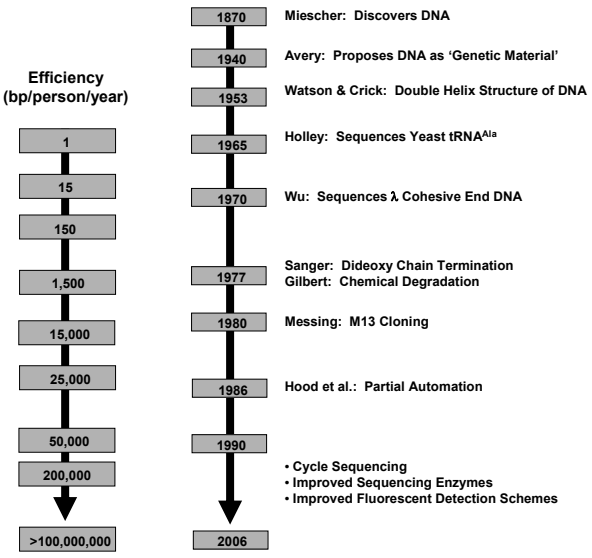


Physical Mapping: Future Prospects

- **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**

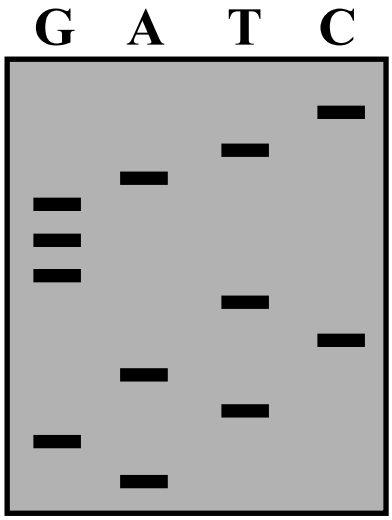
DNA Sequencing

History of DNA Sequencing



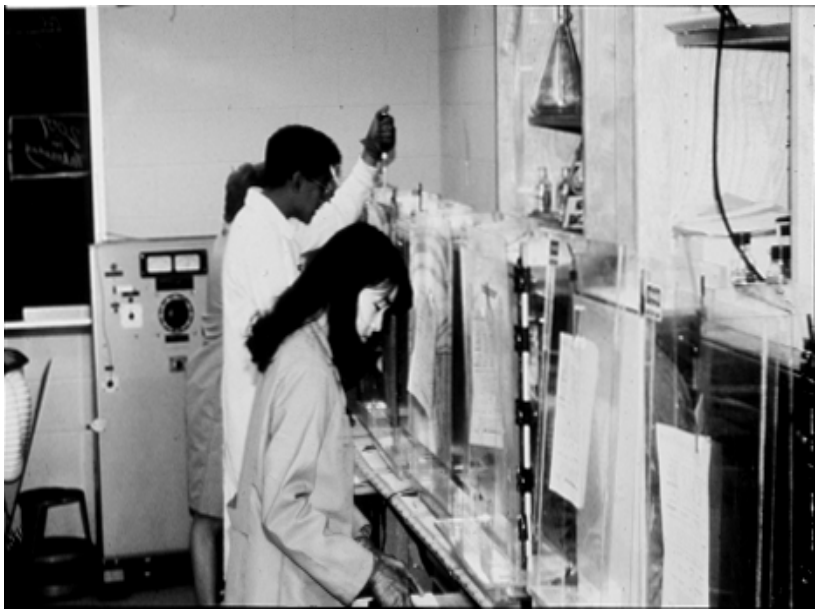
Adapted from Messing & Llaca, *PNAS* (1998)

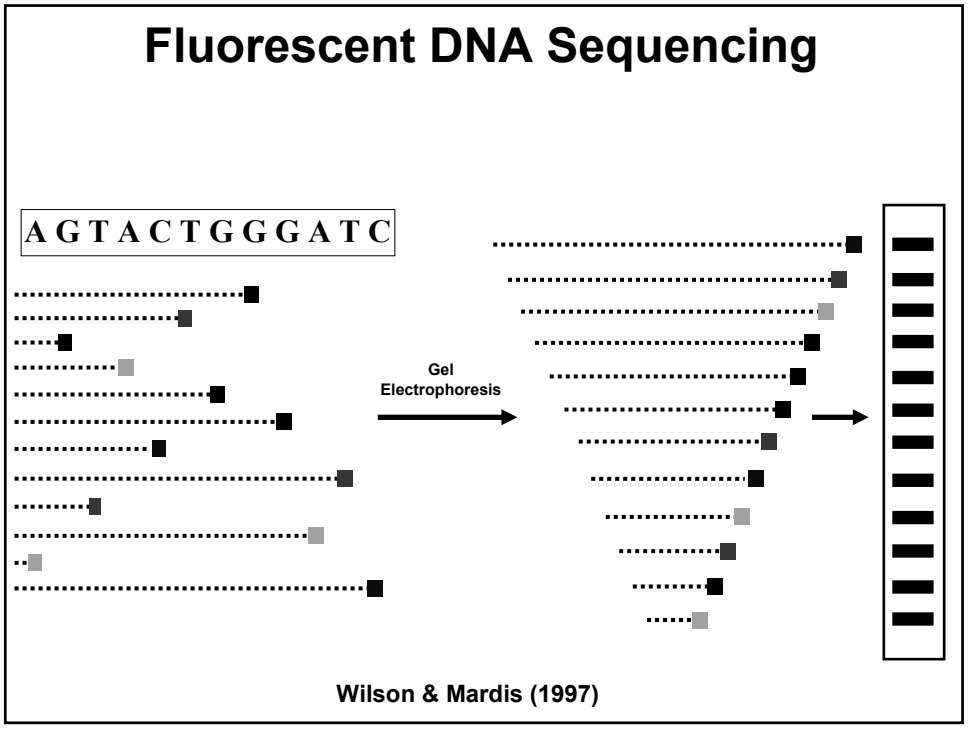
DNA Tagged with Radioactivity

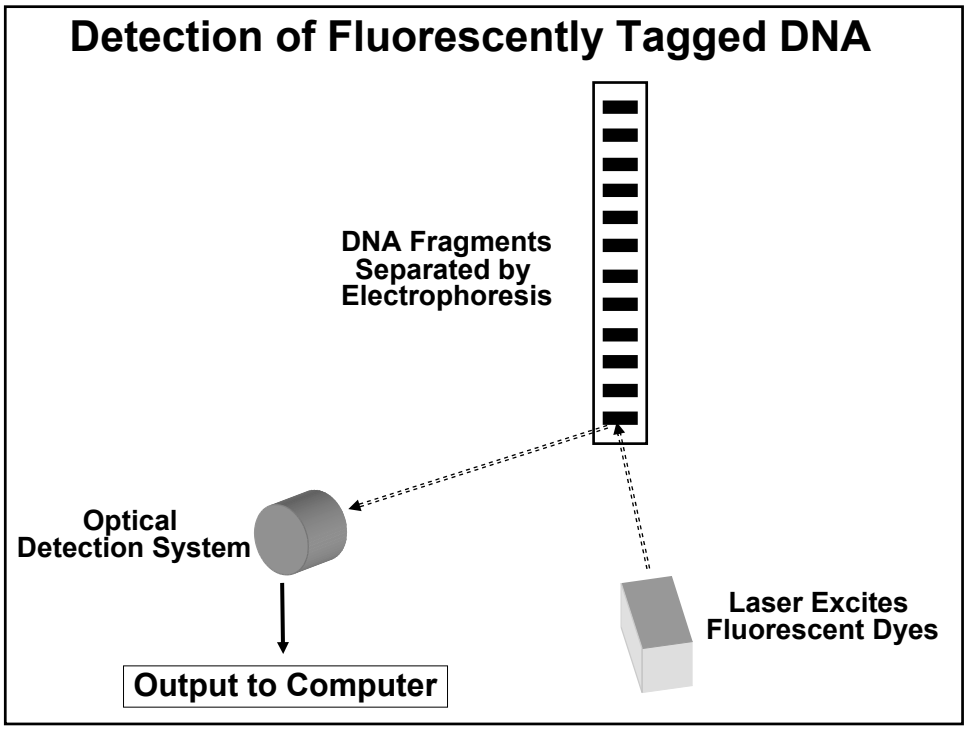


G: G Reaction
A: A Reaction
T: T Reaction
C: C Reaction

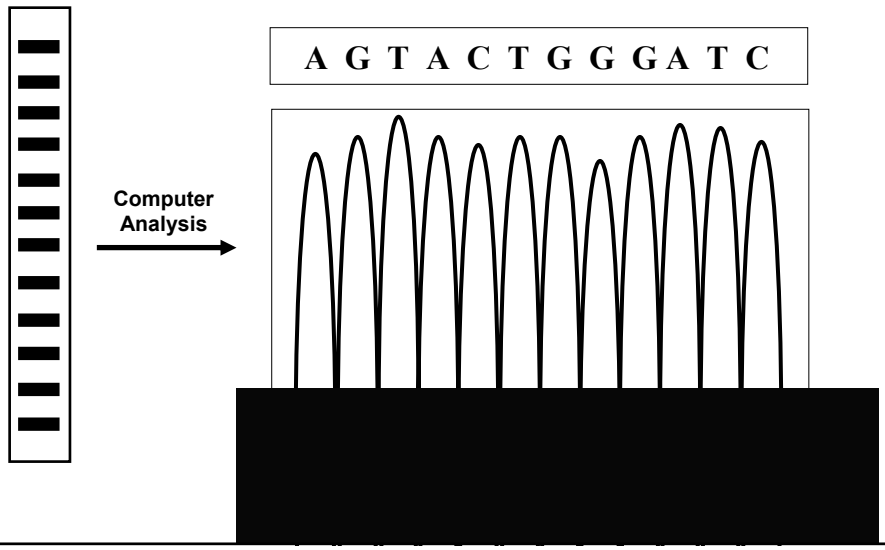
Radioactive Sequencing



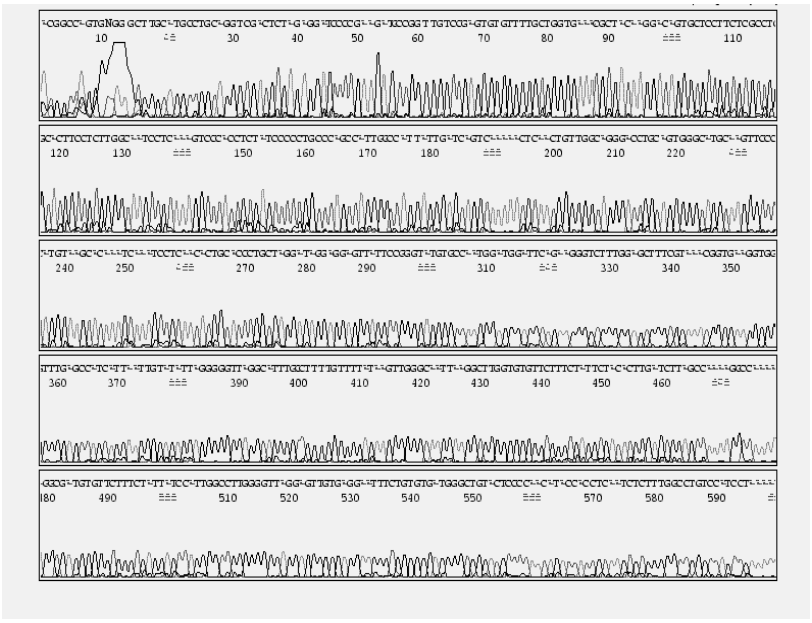

Fluorescent DNA Sequencing
A G T A C T G G G A T C
Gel
Electrophoresis
Wilson & Mardis (1997)


Detection of Fluorescently Tagged DNA
DNA Fragments
Separated by
Electrophoresis
Optical
Detection System
Laser Excites
Fluorescent Dyes
Output to Computer

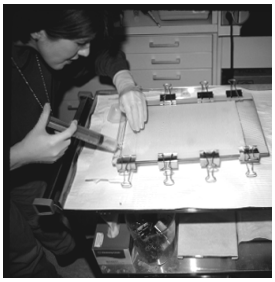
Analyzing Fluorescent DNA Sequencing Data



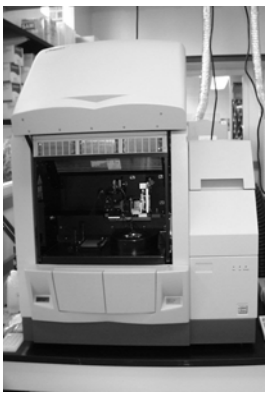
Fluorescent DNA Sequencing Results



Slab Gel-Based DNA Sequencing Instruments

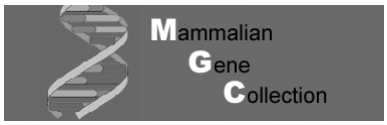


Capillary-Based DNA Sequencing Instruments



Large-Scale cDNA Sequencing

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



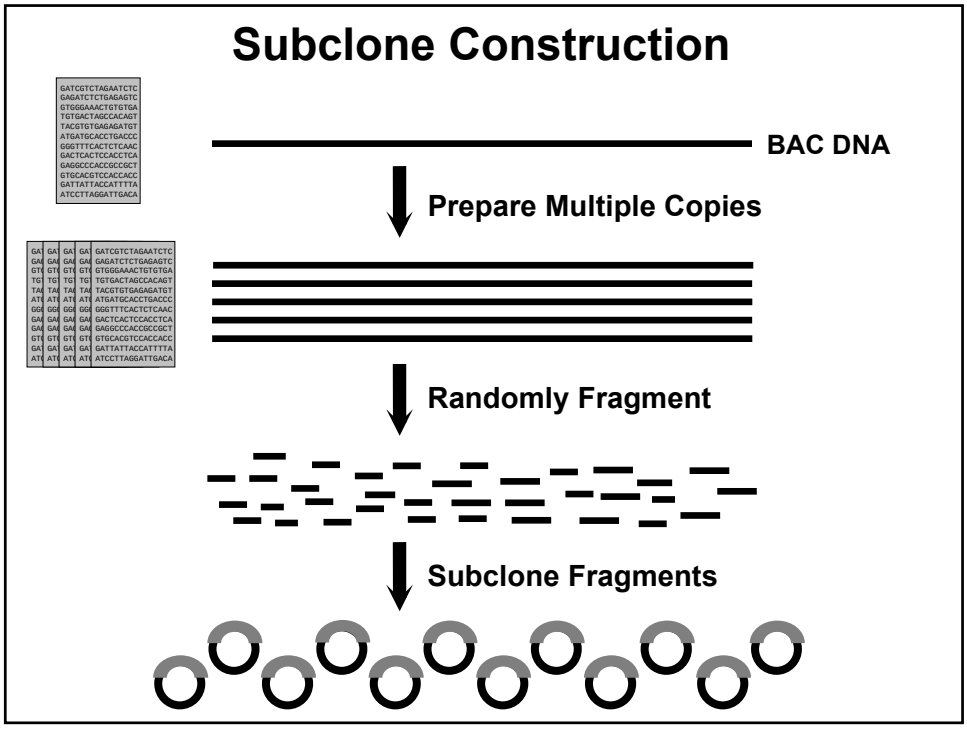
mgc.nci.nih.gov

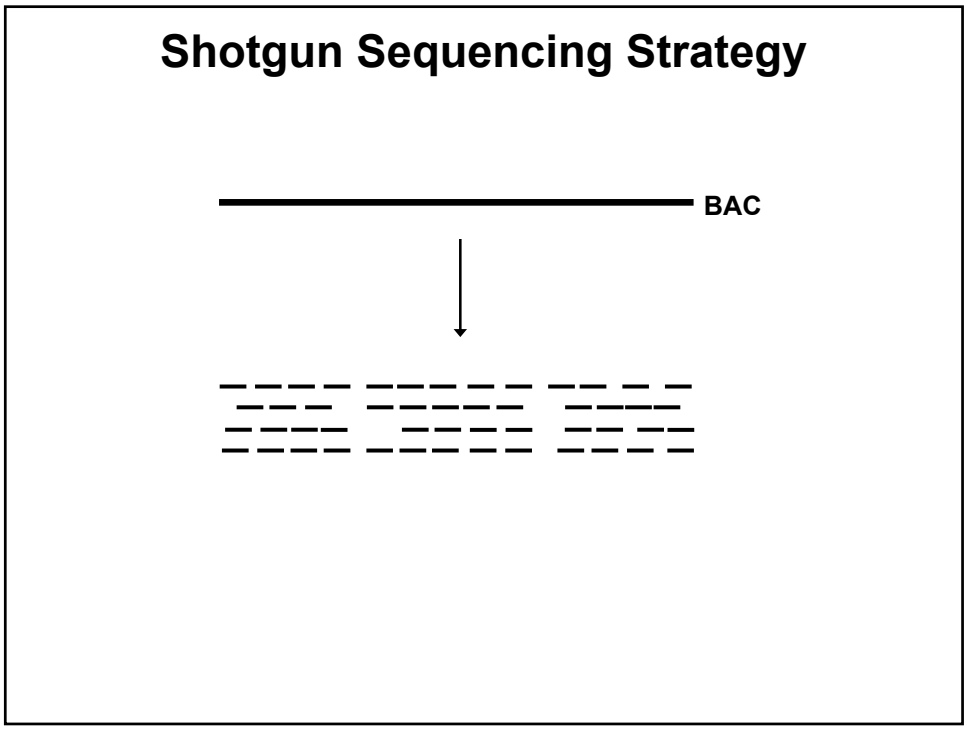
Gerhard et al. (2004)

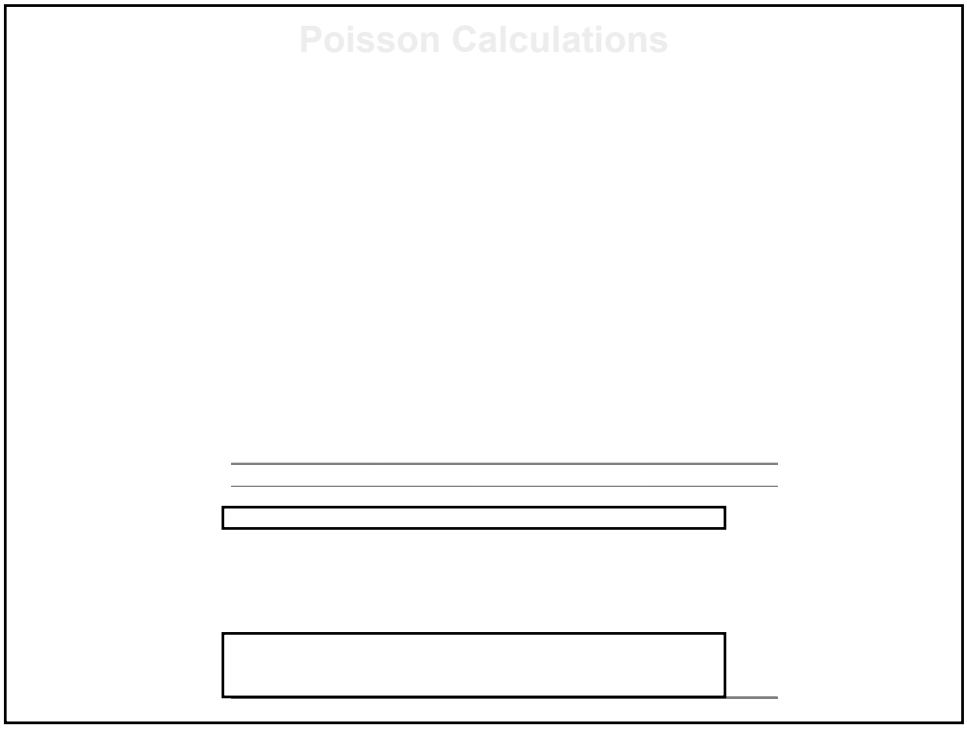
Large-Scale Genome Sequencing

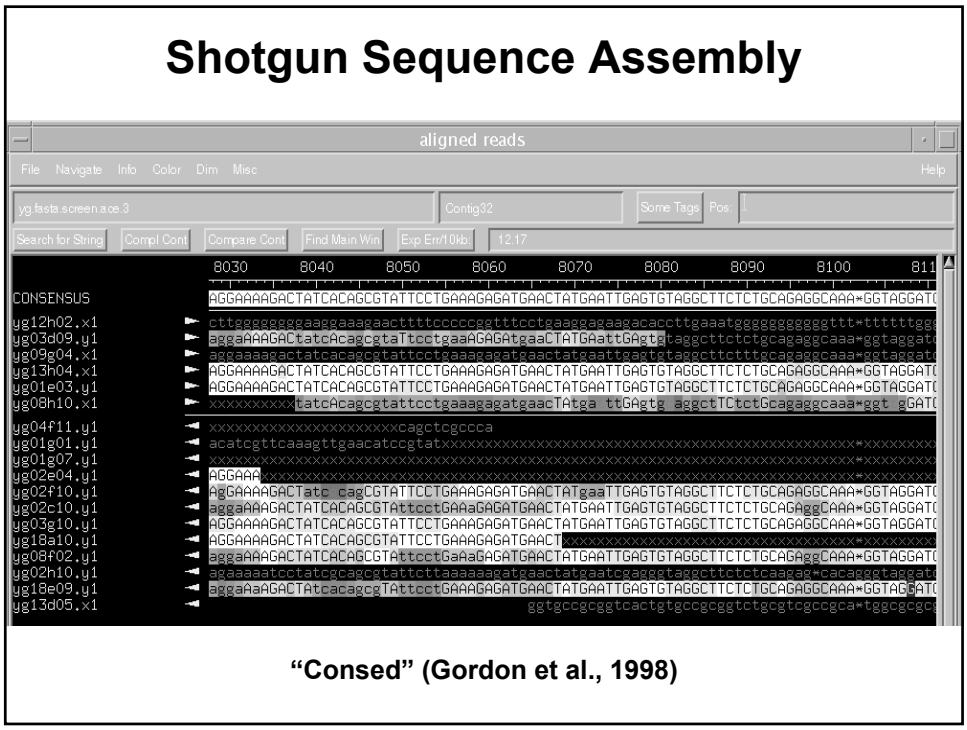
Shotgun Sequencing

Wilson & Mardis (1997)
Green (2001)

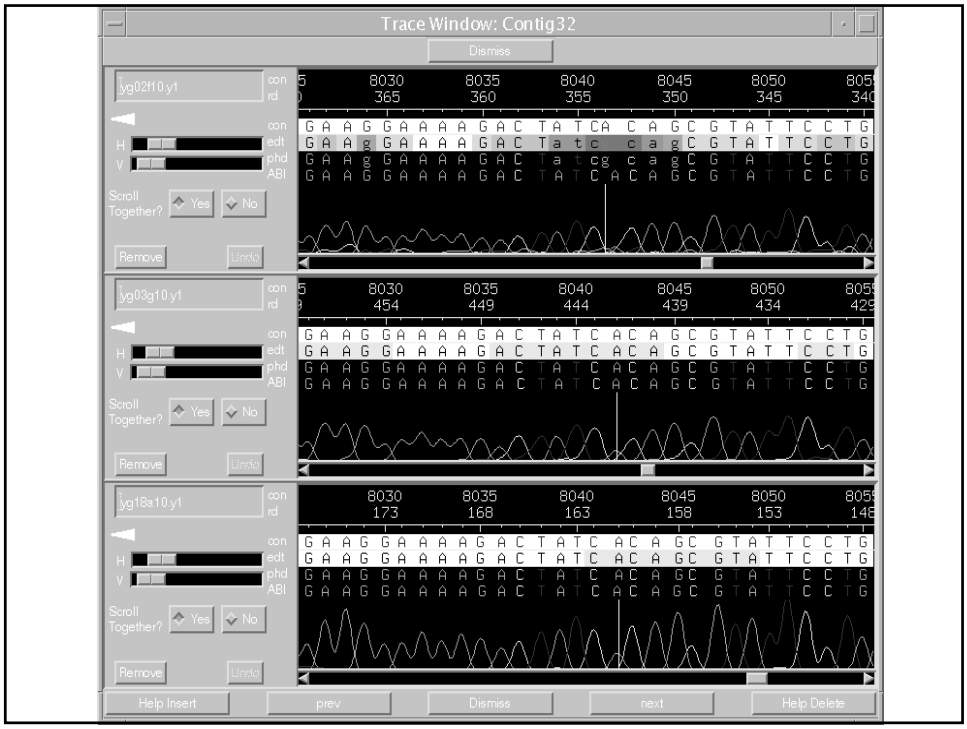

Subclone Construction
BAC DNA
Prepare Multiple Copies
Randomly Fragment
Subclone Fragments

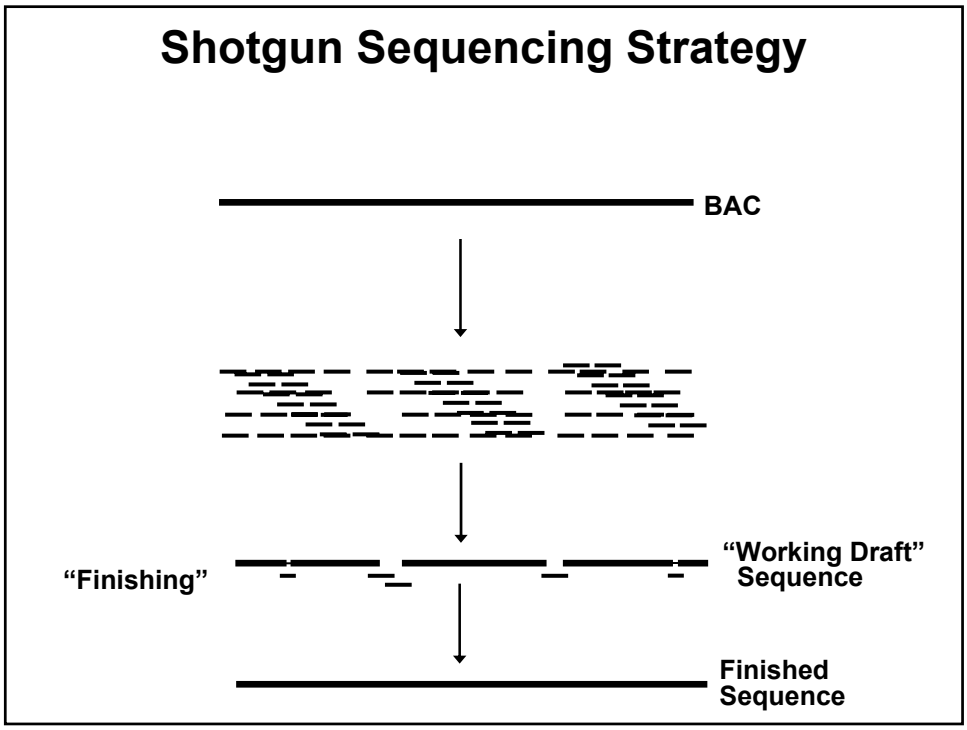

Shotgun Sequencing Strategy
BAC



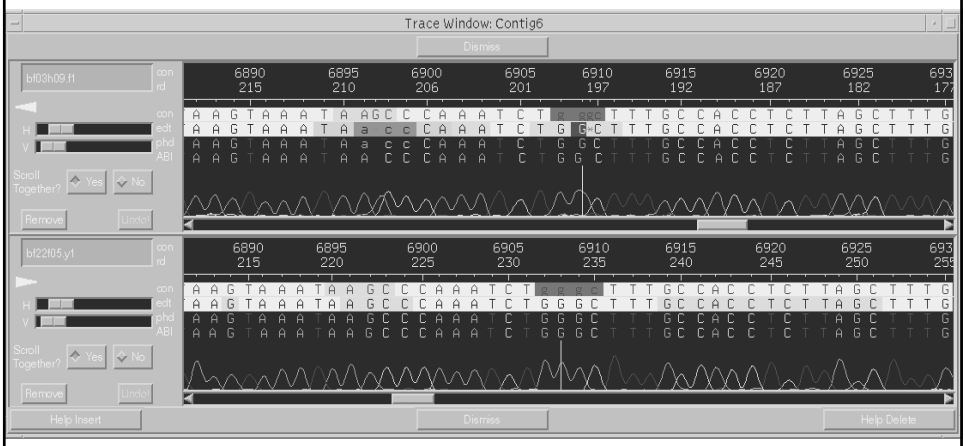


"Consed" (Gordon et al., 1998)


Trace Window: Contig32
Dismiss
Scroll Together?
Yes
No
Remove
Help Insert
prev
Dismiss
next
Help Delete


Shotgun Sequencing Strategy
BAC
"Finishing"
"Working Draft" Sequence
Finished Sequence

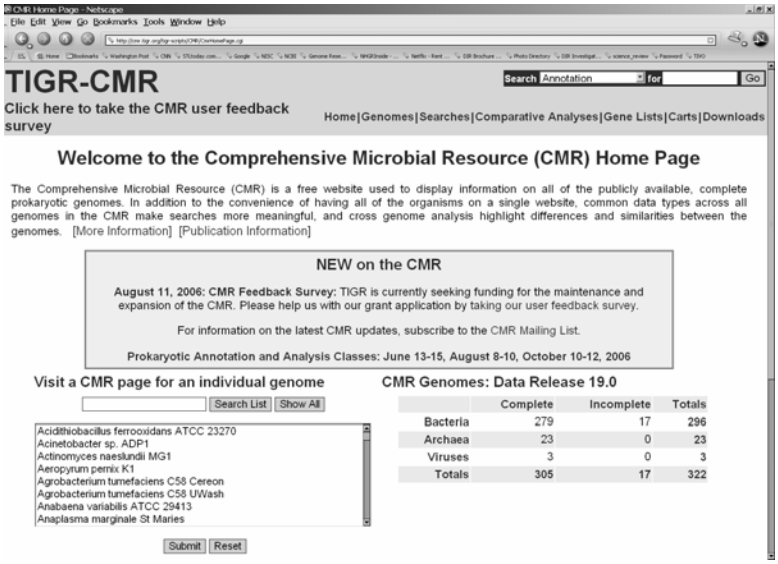
Sequence Finishing: Resolving Ambiguities



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

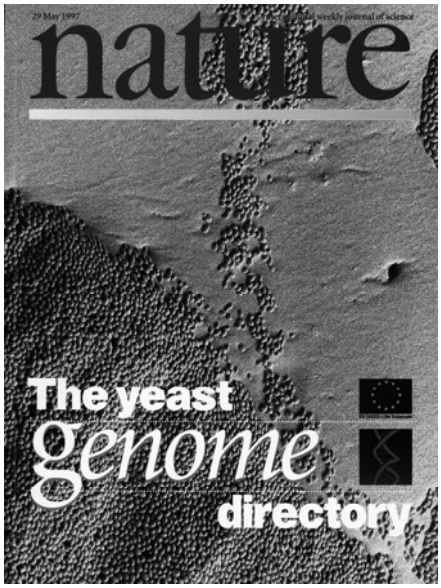
Historically Significant Genome Sequencing Projects

Microbial Genome Sequences



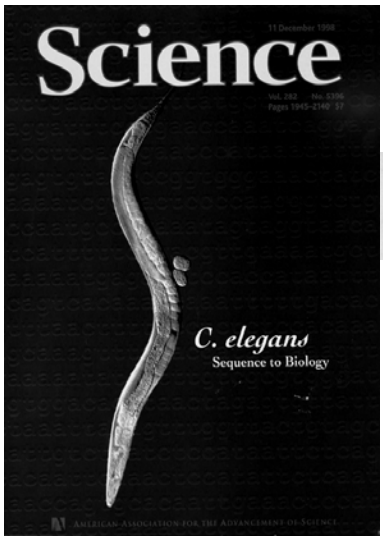
www.tigr.org

First Eukaryotic Genome Sequence



Goffeau et al. (1997)

First Animal Genome Sequence

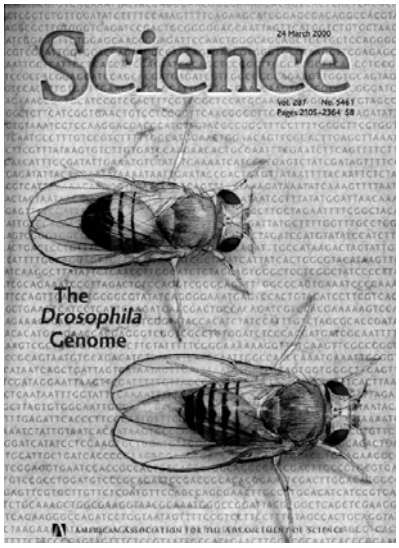


Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology

The *C. elegans* Sequencing Consortium*

***C. elegans* Sequencing Consortium (1998)**

Second Animal Genome Sequence



THE *DROSOPHILA* GENOME

REVIEW

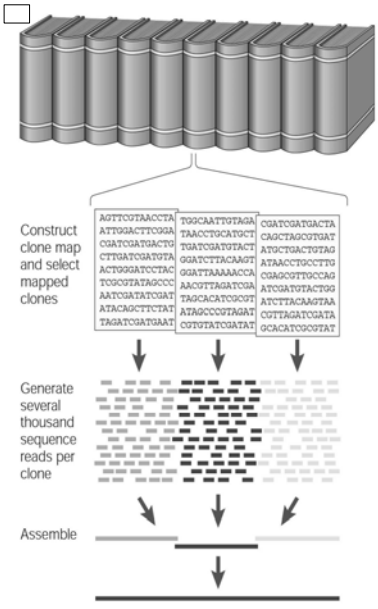
The Genome Sequence of *Drosophila melanogaster*

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,[3] Peter W. Li,[1] Roger A. Hoskins,[2] Richard F. Galle,[4] Reed A. George,[2] Suzanna E. Lewis,[4] Stephen Richards,[2] Michael Ashburner,[5] Scott N. Henderson,[1] Granger G. Sutton,[1] Jennifer R. Wortman,[1] Mark D. Yandell,[1] Qing Zhang,[1] Lin X. Chen,[1] Rhonda C. Brandon,[1] Yu-Hui C. Rogers,[1] Robert G. Blazej,[3] Mark Champe,[2] Barret D. Pfeiffer,[2] Kenneth H. Wan,[2] Clare Doyle,[2] Evan G. Baxter,[2] Gregg Helt,[4] Catherine R. Nelson,[4] George L. Gabor Miklos,[7] Josep F. Abril,[8] Anna Agbayani,[2] Hui-Jin An,[1] Cynthia Andrews-Pfannkoch,[1] Danita Baldwin,[1] Richard M. Ballew,[1] Anand Basu,[1] James Baxendale,[1] Leyla Bayraktaroglu,[9] Ellen M. Beasley,[1] Karen Y. Beeson,[1] P. V. Benos,[10] Benjamin P. Berman,[4] Deepali Bhandari,[1] Slava Bolshakov,[11] Dana Borkova,[12] Michael R. Botchan,[13] John Bouck,[3] Peter Brokstein,[4] Phillipe Brottier,[14] Kenneth C. Burtis,[15] Dana A. Busam,[1] Heather Butler,[16] Edouard Cadieu,[17] Angela Center,[1] Ishwar Chandra,[1] J. Michael Cherry,[18] Simon Cawley,[19] Carl Dahlke,[1] Lionel B. Davenport,[1] Peter Davies,[1] Beatriz de Pablos,[20] Arthur Delcher,[1] Zuoming Deng,[1] Anne Deslattes Mays,[1] Ian Dew,[1] Suzanne M. Dietz,[1] Kristina Dodson,[1] Lisa E. Doup,[1] Michael Downes,[21] Shannon Dugan-Rocha,[2] Boris C. Dunkov,[22] Patrick Dunn,[1] Kenneth J. Durbin,[3] Carlos C. Evangelista,[1] Concepcion Ferraz,[23] Steven Ferriera,[1] Wolfgang Fleischmann,[5] Carl Fosler,[1] Andrei E. Gabrielian,[1] Neha S. Garg,[1] William M. Gelbart,[3] Kan Glasser,[1] Anna Glodek,[1] Fangcheng Gong,[1] J. Harley Gorrell,[3] Zhiping Gu,[1] Ping Guan,[1] Michael Harris,[1] Nomi L. Harris,[2] Damon Harvey,[4] Thomas J. Heiman,[1] Judith R. Hernandez,[3] Jarrett Houck,[1] Damon Hostin,[1] Kathryn A. Houston,[2] Timothy J. Howland,[1] Ming-Hui Wei,[1] Chinyere Ibegwam,[1] Mena Jalali,[1] Francis Kalush,[1] Gary H. Karpen,[21] Zhaoxi Ke,[1] James A. Kennison,[24] Karen A. Ketchum,[3] Bruce E. Kimmel,[2] Chinnappa D. Kodira,[1] Cheryl Kraft,[1] Saul Kravitz,[1] David Kulp,[4] Zhongwu Lai,[1] Paul Lasko,[25] Yiding Lei,[1] Alexander A. Levitsky,[1] Jiayin Li,[1] Zhenya Li,[1] Yong Liang,[1] Xiaoying Lin,[26] Xiangjun Liu,[1] Bettina Mattei,[1] Tina C. McIntosh,[1] Michael P. McLeod,[3] Duncan McPherson,[1] Gennady Merkulov,[1] Natalia V. Milshina,[1] Clark Mobarry,[1] Joe Morris,[4] Ali Moshrefi,[2] Stephen M. Mount,[27] Mee Moy,[1] Brian Murphy,[1] Lee Murphy,[28] Donna M. Muzny,[2] David L. Nelson,[2] David R. Nelson,[29] Keith A. Nelson,[1] Katherine Nixon,[2] Deborah R. Nusskern,[1] Joanne M. Pacleb,[2] Michael Palazzolo,[2] Gjange S. Pittman,[1] Sue Pan,[1] John Pollard,[1] Vinita Puri,[1] Martin G. Reese,[4] Knut Reinert,[1] Karin Remington,[1] Robert D. C. Saunders,[30] Frederick Scheeler,[1] Hua Shen,[3] Bixiang Christopher Shue,[1] Inga Silén-Kiamos,[11] Michael Simpson,[1] Marian P. Skupski,[1] Tom Smith,[1] Eugene Spier,[1] Allan C. Spradling,[31] Mark Stapleton,[2] Renee Strong,[1] Eric Sun,[1] Robert Svirskas,[32] Cyndee Tector,[1] Russell Turner,[1] Eli Venter,[1] Alhui H. Wang,[1] Xin Wang,[1] Zhen-Yuan Wang,[1] David A. Wassarman,[33] George M. Weinstock,[2] Jean Weissenbach,[14] Sherita M. Williams,[1] Trevor Woodage,[1] Kim C. Worley,[3] David Wu,[1] Song Yang,[2] Q. Alison Yao,[1] Jane Ye,[1] Ru-Fang Yeh,[10] Jayshree S. Zaveri,[1] Ming Zhan,[1] Guangren Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Xiangqun H. Zheng,[1] Fei N. Zhong,[1] Wenyan Zhong,[1] Xiaojun Zhou,[2] Shiaoping Zhu,[1] Xiaohong Zhu,[1] Hamilton O. Smith,[1] Richard A. Gibbs,[3] Eugene W. Myers,[1] Gerald M. Rubin,[34] J. Craig Venter[1]

Adams et al. (2000)

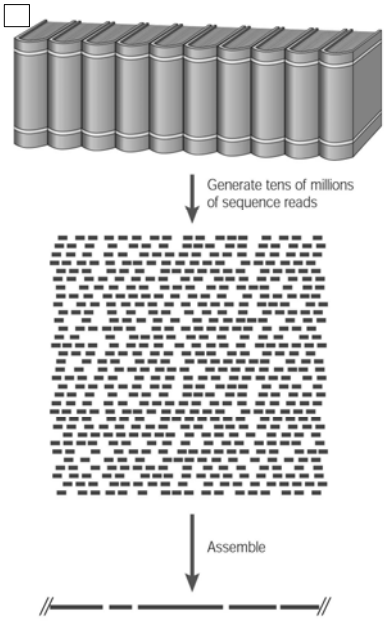
Human Genome Sequencing Centers

Clone-Based Shotgun Sequencing



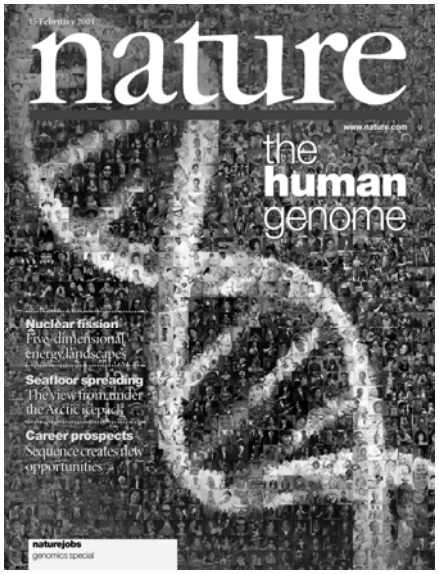
Green (2001)

Whole-Genome Shotgun Sequencing

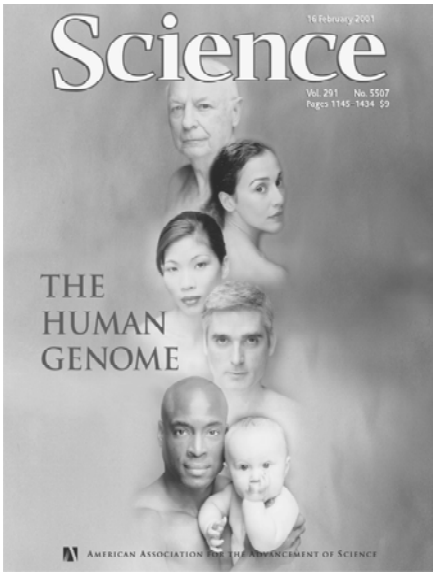


Green (2001)

February, 2001 Draft Sequence



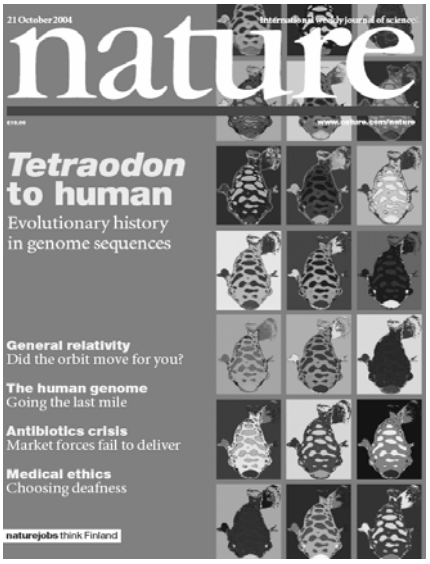
International Human Genome Sequencing Consortium (2001)



Venter et al. (2001)

April, 2003 Completion

October, 2004 Publication



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 35) contains 2.85 billion nucleotides interrupted by only 341 gaps. It covers ~99% of the euchromatic genome and is accurate to an error rate of ~1 event 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 exome seems to encode only 20,000–25,000 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 highly accurate sequence of the vast 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[1–6] to allow the study of inherited disease and provide a crucial scaffold for genome assembly; and (2) the sequencing of organisms with smaller, simpler genomes[7–9] to serve as a testbed for method 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 centres in six countries, was formed to carry out this component of the HGP.

In February 2001, the IHGSC[10] and Celera Genomics[11] each reported draft sequences providing a first overall view of the human genome. Those sequences allowed systematic study of the human genome itself, including identification of genes, combinatorial architecture of proteins, regional differences in genome composition, distribution and history of transposable elements, distribution of polymorphism and relationship between genetic recombination and physical distance. Moreover, systematic knowledge of the human genome has enabled new tools and approaches that have markedly accelerated biomedical research.

Both draft sequences, however, had important shortcomings. The IHGSC sequence, for example, omitted ~10% of the euchromatic genome; it was interrupted by ~150,000 gaps and the order and orientation of many segments within local regions had not been established. The IHGSC thus turned to the challenge of completing the sequence of the euchromatic genome. Operationally, a finished sequence was defined as having an error rate of, at most, one event per 10^4 bases, and the goal for completion was coverage in finished sequence of at least 95% of the euchromatic genome, with the only gaps being those refractory to all available techniques[12] (see http://www.genome.gov/10000923). The goal was challenging because the human genome is replete with such features as dispersed repeats 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 multicellular organisms: the nematode[13], mustard weed[14] and the fruitfly[15]. Those genomes are all roughly 30-fold smaller than the human genome and have much simpler structure.

We describe here the results of a multiyear effort by the IHGSC towards the goal of a complete human sequence. The number of gaps has been reduced 400-fold to only 341, most of which are associated with segmental duplications and will require new methods for resolution. The assembled near-complete genome sequence has an error rate of only ~1 event per 100,000 bases; it contains 2.85 billion nucleotides 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 analyses made possible by the sequence. We do not attempt here a comprehensive analysis of the contents of the human genome. An initial analysis was previously reported[16] and a series of papers is being written describing the individual chromosomes[17,18], including annotation 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 or readily resolved through random shotgun sequencing tend to be highly enriched in problematic sequences. Resolving such regions required the development of special approaches, which evolved substantially over time and varied among centres.

Broadly, 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 each chromosome arm; and (2) producing finished clones, consisting of continuous and accurate nucleotide sequence across each large-insert clone. In practice, these two components were tightly intertwined in that progress in each often depended on results from the other. The components are described in Boxes 1 and 2. Further information about the finishing process and finishing standards can be found in the Supplementary Information (Note 1) and at http://www.genome.gov/10000923.

In total, we generated a shotgun sequence from 59,208 large-insert clones (total length ~5.84 gigabases (Gb)) and finished the sequence from 45,742 of those clones (total length ~3.67 Gb). The clones consisted primarily of bacterial artificial chromosomes

International Human Genome Sequencing Consortium (2004)

CNN's #1 Medical Story of Past 25 Years

CNN.com

PRINT THIS
Powered by Clickability

Click to Print

SAVE THIS | EMAIL THIS | Close

Top 25: Medical stories

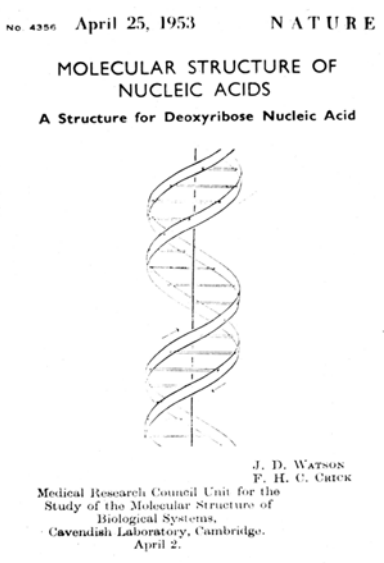
Human genome mapping ranks No. 1 in health news

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

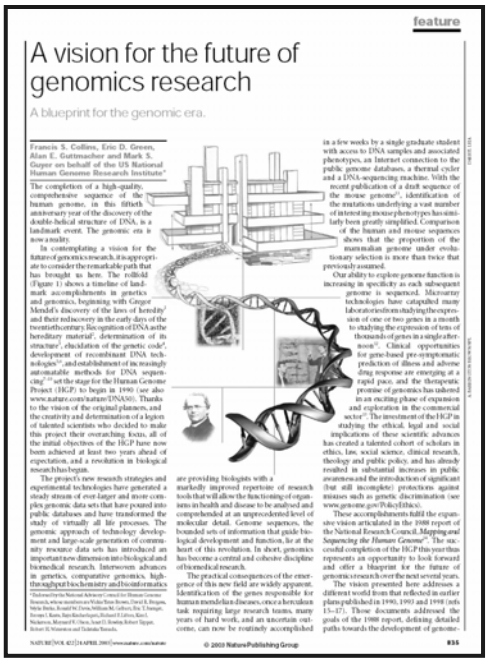
A Structure for Deoxyribose Nucleic Acid

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.

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. Guyer on behalf of the US National Human Genome Research Institute*

The completion of a high-quality, comprehensive sequence of the human genome, in this fiftieth anniversary year of the discovery of the double-helical structure of DNA, is a landmark event. The genomic era is now a reality.

In contemplating a vision for the future of genomics research, it is appropriate to consider the remarkable path that has brought us here. The rollfold (Figure 1) shows a timeline of landmark accomplishments in genetics and genomics, beginning with Gregor Mendel's discovery of the laws of heredity[1] and their rediscovery in the early days of the twentieth century. Recognition of DNA as the hereditary material[2], determination of its structure[3], elucidation of the genetic code[4], development of recombinant DNA technologies[5,6], and establishment of increasingly automatable methods for DNA sequencing[7,8] set the stage for the Human Genome Project (HGP) to begin in 1990 (see also www.nature.com/nature/DNA50). 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 overarching focus, all of the initial objectives of the HGP have now been achieved at least two years ahead of expectation, and a revolution in biological research has begun.

The project's new research strategies and experimental technologies have generated a steady stream of ever-larger and more complex genomic data sets that have poured 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 biomedical research. Interwoven advances in genetics, comparative genomics, high-throughput biochemistry and bioinformatics are providing biologists with a markedly improved repertoire of research tools that will allow the functioning of organisms in health and disease to be analysed and comprehended at an unprecedented level of molecular detail. Genome sequences, the bounded 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 biomedical research.

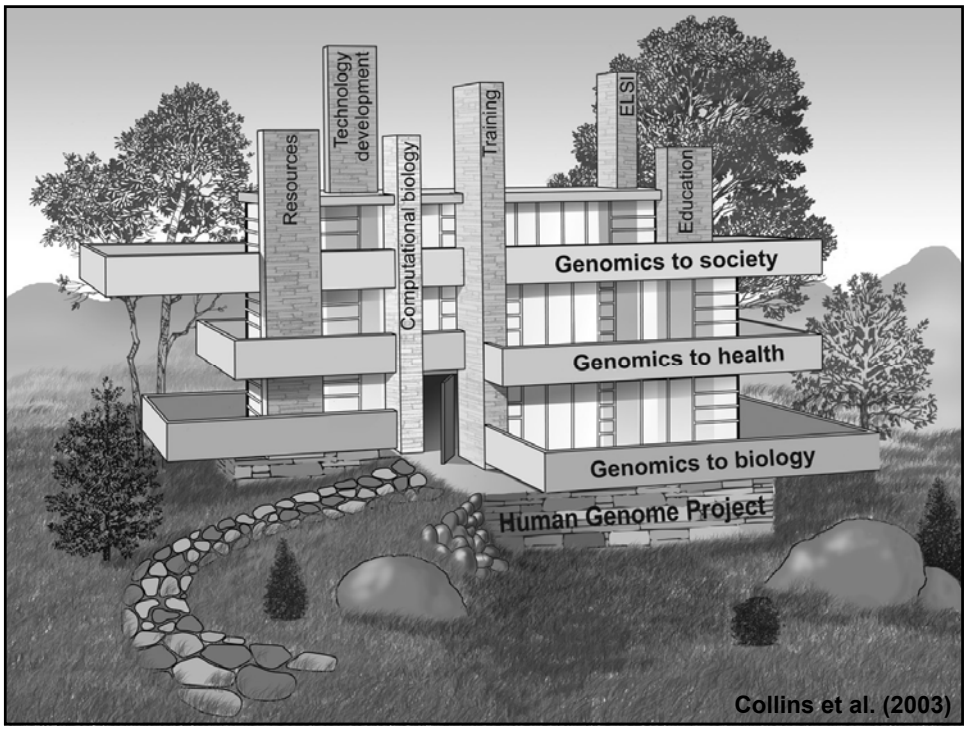
The practical consequences of the emergence of this new field are widely apparent. Identification of the genes responsible for human mendelian diseases, once a herculean task requiring large research teams, many years of hard work, and an uncertain outcome, can now be routinely accomplished in a few 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 genome[9–11], identification of the mutations underlying a vast number of interesting mouse phenotypes has similarly been greatly simplified. 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 catapulted many laboratories from studying the expression of one or two genes in a month to studying the expression of tens of thousands of genes in a single afternoon[12]. Clinical opportunities for gene-based pre-symptomatic prediction of illness and adverse drug response are emerging at a rapid pace, and the therapeutic promise of genomics has ushered in an exciting phase of expansion and exploration in the commercial sector[13]. 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 (but still incomplete) protections against misuses such as genetic discrimination (see www.genome.gov/Policy/Ethics).

These accomplishments fulfil the expansive vision articulated in the 1988 report of the National Research Council, *Mapping and Sequencing the Human Genome*[14]. The successful completion of the HGP thus very much 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 addresses a different world from that reflected in earlier plans published in 1990, 1993 and 1998 (refs 15–17). Those documents addressed the goals of the 1988 report, defining detailed paths towards the development of genome-

Collins et al. (2003)



Collins et al. (2003)

Mapping the Human Genome	**~1990 to ~2000**	**The Human Genome Project**
Sequencing the Human Genome	**~1998 to ~2003**	
Interpreting the Human Genome Sequence	**~2003 to ???**	***Beyond*** **The Human Genome Project**

~3,000 bp (0.0001%) of Human Genome Sequence

TGCCGCGGAACTTTTCGGCTCTCTAAGGCTGTATTTTGATATACGAAAGGCACATTTTCCTTCCCTTTTCAAAATGCACCTTGCAAACGTAACAG
GAACCCGACTAGGATCATCGGGAAAAGGAGGAGGAGGAGGAAGGCAGGCTCCGGGGAAGCTGGTGGCAGCGGGTCCTGGGTCTGGCGGACCCTGA
CGCGAAGGAGGGTCTAGGAAGCTCTCCGGGGAGCCGGTTCTCCCGCCGGTGGCTTCTTCTGTCCTCCAGCGTTGCCAACTGGACCTAAAGAGAGG
CCGCGACTGTCGCCCACCTGCGGGATGGGCCTGGTGCTGGGCGGTAAGGACACGGACCTGGAAGGAGCGCGCGCGAGGGAGGGAGGCTGGGAGTC
AGAATCGGGAAAGGGAGGTGCGGGGCGGCGAGGGAGCGAAGGAGGAGAGGAGGAAGGAGCGGGAGGGGTGCTGGCGGGGGTGCGTAGTGGGTGGA
GAAAGCCGCTAGAGCAAATTTGGGGCCGGACCAGGCAGCACTCGGCTTTTAACCTGGGCAGTGAAGGCGGGGGAAAAGAGCAAAAGGAAGGGGTGG
TGTGCGGAGTAGGGGTGGGTGGGGGGAATTGGAAGCAAATGACATCACAGCAGGTCAGAGAAAAAGGGTTGAGCGGCAGGCACCCAGAGTAGTAG
GTCTTTGGCATTAGGAGCTTGAGCCCAGACGGCCCTAGCAGGGACCCCAGCGCCCGAGAGACCATGCAGAGGTCGCCTCTGGAAAAGGCCAGCGT
TGTCTCCAAACTTTTTTTCAGGTGAGAAGGTGGCCAACCGAGCTTCGGAAAGACACGTGCCCACGAAAGAGGAGGGCGTGTGTATGGGTTGGGTT
TGGGGTAAAGGAATAAGCAGTTTTTAAAAAGATGCGCTATCATTCATTGTTTTGAAAGAAAATGTGGGTATTGTAGAATAAAACAGAAAGCATTA
AGAAGAGATGGAAGAATGAACTGAAGCTGATTGAATAGAGAGCCACATCTACTTGCAACTGAAAAGTTAGAATCTCAAGACTCAAGTACGCTACT
ATGCACTTGTTTTATTTCATTTTTCTAAGAAACTAAAAATACTTGTTAATAAGTACCTAAGTATGGTTTATTGGTTTTCCCCCTTCATGCCTTGG
ACACTTGATTGTCTTCTTGGCACATACAGGTGCCATGCCTGCATATAGTAAGTGCTCAGAAAACATTTCTTGACTGAATTCAGCCAACAAAAATT
TTGGGGTAGGTAGAAAATATATGCTTAAAGTATTTATTGTTATGAGACTGGATATATCTAGTATTTGTCACAGGTAAATGATTCTTCAAAAATTG
AAAGCAAATTTGTTGAAATATTTATTTTGAAAAAAGTTACTTCACAAGCTATAAATTTTAAAAGCCATAGGAATAGATACCGAAGTTATATCCAA
CTGACATTTAATAAATTGTATTCATAGCCTAATGTGATGAGCCACAGAAGCTTGCAAACTTTAATGAGATTTTTTAAAATAGCATCTAAGTTCGG
AATCTTAGGCAAAGTGTTGTTAGATGTAGCACTTCATATTTGAAGTGTTCTTTGGATATTGCATCTACTTTGTTCCTGTTATTATACTGGTGTGA
ATGAATGAATAGGTACTGCTCTCTCTTGGGACATTACTTGACACATAATTACCCAATGAATAAGCATACTGAGGTATCAAAAAAGTCAAATATGT
TATAAATAGCTCATATATGTGTGTAGGGGGGAAGGAATTTAGCTTTCACATCTCTCTTATGTTTAGTTCTCTGCATGTGCAGTTAATCCTGGAAC
TCCGGTGCTAAGGAGAGACTGTTGGCCCTTGAAGGAGAGGTCCTCCCTGTGGATGAGAGAGAAGGACTTTACTCTTTGGAATTATCTTTTTGTGT
TGATGTTATCCACCTTTTGTTACTCCACCTATAAAATCGGCTTATCTATTGATCTGTTTTCCTAGTCCTTATAAAGTCAAAATGTTAATTGGCAT
AAATTATAGACTTTTTTTAGCAGAGAACTTTGAGGAACCTAAATGCCAACCAGTCTAAAAATGCAGTTTTCAGAAGAATGAATATTTCATGGATA
GTTCTAAATACTAATGAACTTTAAAATAGCTTACTATTGATCTGTCAAAGTGGGTTTTTATATAATTTTCTTTTTACAAATCACCTGACACATTT
AATATAGGTTAAAAAATGCTATCAGGCTGGTTTGCAAAGAAAATGTATTACAAAGGCTGCTAAGTGTGTTAAGAGCATACTCATTTCTGTTCTCC
AAAATATTTCATAAGGTGCTTTAAGAATAGGTATGTTTTTAAAAGTTAAGTTCCTACTATTTATAGGAACTGACAATCACCTAAAATACCAATGA
TTACAAACTTCCTTCTGGCCTTCTGGACTGCAATTCTAAAAGTGTAAAAAACATATTTTCTGCATTAAGTTAGGCAGTATTGCTTAGTTTTCAAA
GTGGTAGGCTTTGGAGTCAGATTATTTTGATTCAGATCCTACATCTACTGTTTAGTAGCTCTGTTGCCTGAGGCAGGTCCCTTAACATCTCTGTG
TGTGACTTGACCTTTAAAATTTGGAGACTGTCATAGGGGTTAATCCCTTGAGAAAATGAATGTGAAAAGTTAGCCTAATGTTAACTGCTATTATT
ATGGATTACCATATTTTCACATTCATCACAGTACATGCACCTTGTTAATATAAGATGCTCAATTCATCTTTGAGTATAATTTTGTGACTCTCAAT
CTGGATATGCAATGAGTGGGCCTGTATGAGAATTTAATTTATGAAAAATTGTGTTTCACATGGCCTTACCAGATATACAGGAAACACGTCACATG
TTTCTATTGTATGTTGTTAAATGCCTTAGAATTTAACTTTCTGAATAGGATCCCTTCAGTTTGAGAGTCATAAAAGAGTAAAATTATTATGGTAT

The Human Genome… by the Numbers

~5% of Human Genome is Functionally Important

- 5% of 3B Bases = ~150M Bases
- Do NOT Yet Know the Position of these ~150M Functional Bases

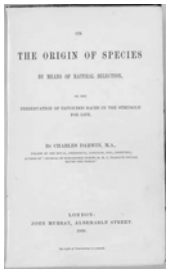
~1.5% Encodes for Protein (Genes)

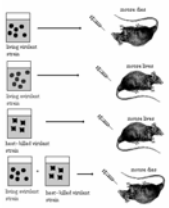
- Corresponds to ~18-22K Genes
- Many More than ~22K Different Proteins
- <u>Good</u> Inventory at Present

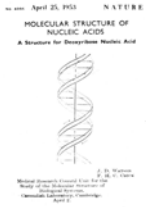
~3.5% Functional But Non-Coding

- Gene Regulatory Elements
- Chromosomal Functional Elements
- Undiscovered Functional Elements (NOT Yet in Textbooks!)
- <u>Poor</u> Inventory at Present

Foundational Milestones in Genetics & Genomics







Darwin	Mendel	Miescher	Avery	Watson & Crick
1859	*1865*	*1871*	*1944*	*1953*

Comparing Genomes is Like Cryptography

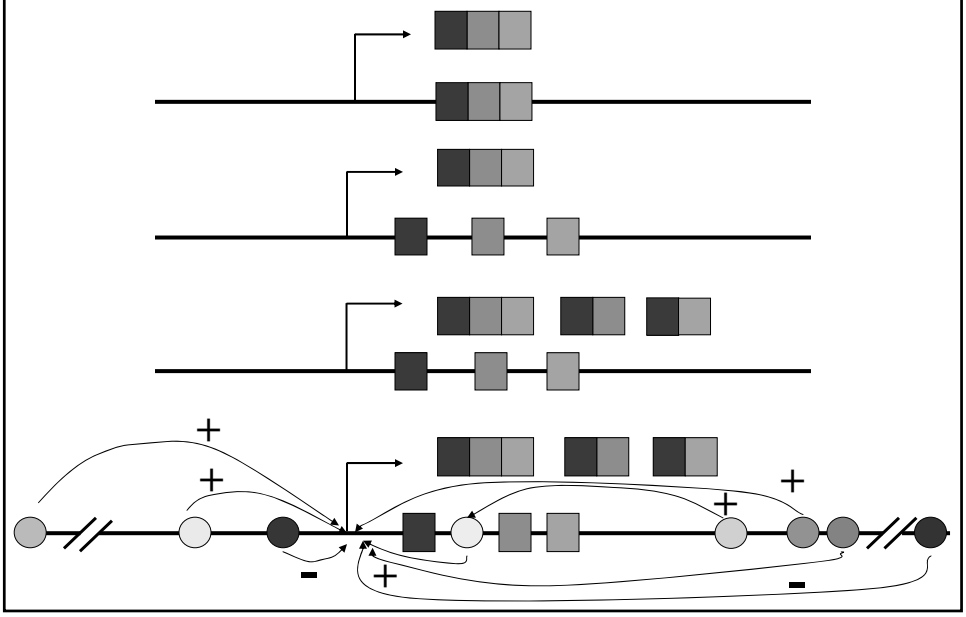
CKQEBHEREYTWASUISCZMEISDFOGETHEBLPBGOODFQSTLKSTUFFRTAC

DLUCEHEREZBRTTOISAWNDCDARJJPTHERROFGOODERGHCLSTUFFBRHA

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



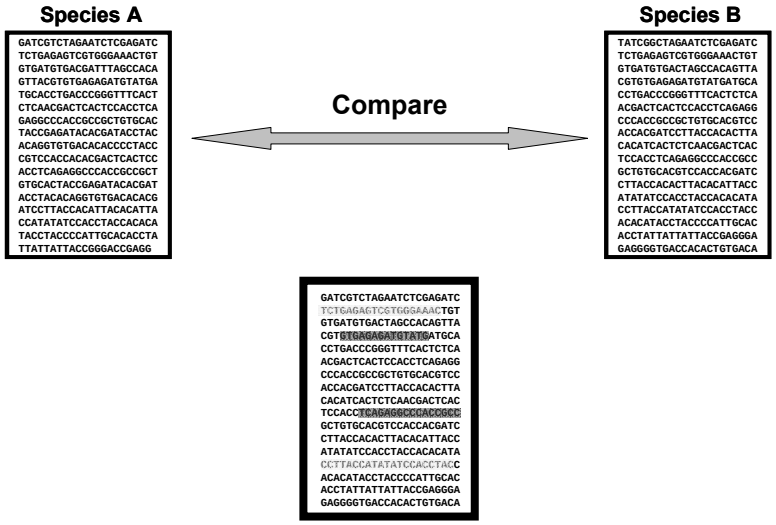
Functional Elements: Coding vs. Non-Coding

- **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

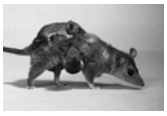
Dog


Macaque


Orangutan


Marmoset

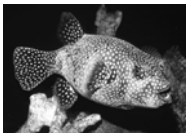

Cow


Monodelphis

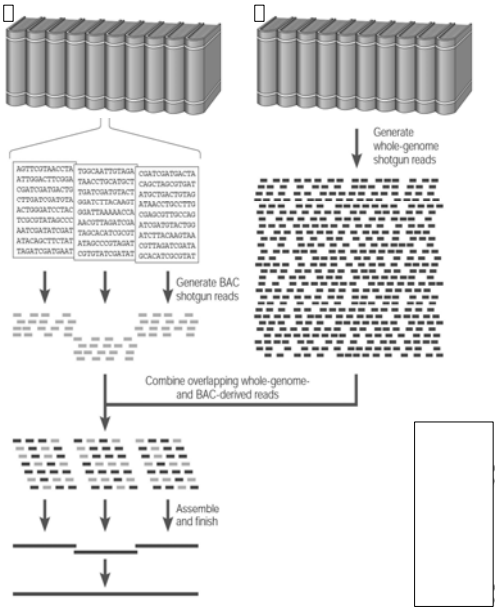

Platypus


Xenopus


Zebrafish


Pufferfish

Hybrid Shotgun Sequencing



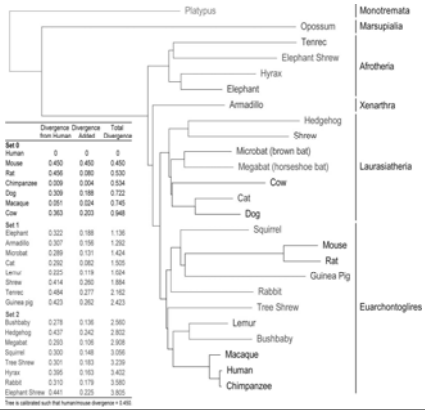
Green (2001)

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*†, Jade Vinson†‡, NISC Comparative Sequencing Program*§¶, Webb Miller‖, David B. Jaffe#, Kerstin Lindblad-Toh#, Jean Chang#, Eric D. Green*§, Eric S. Lander#, James C. Mullikin*§**, and Michele Clamp†**

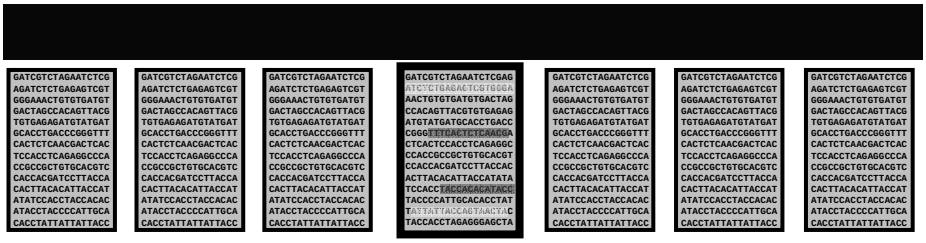
Margulies EM et al. (2005)



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…)

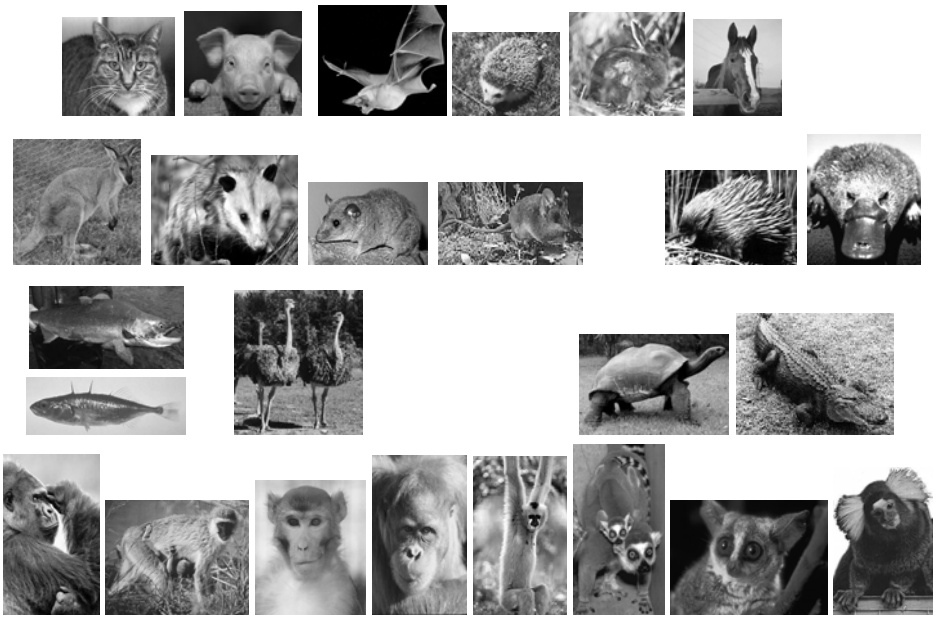
Multi-Species Sequence Comparisons



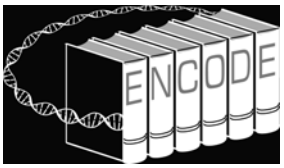
Multi-Species Conserved Sequences (MCSs)

Margulies et al. (2003)

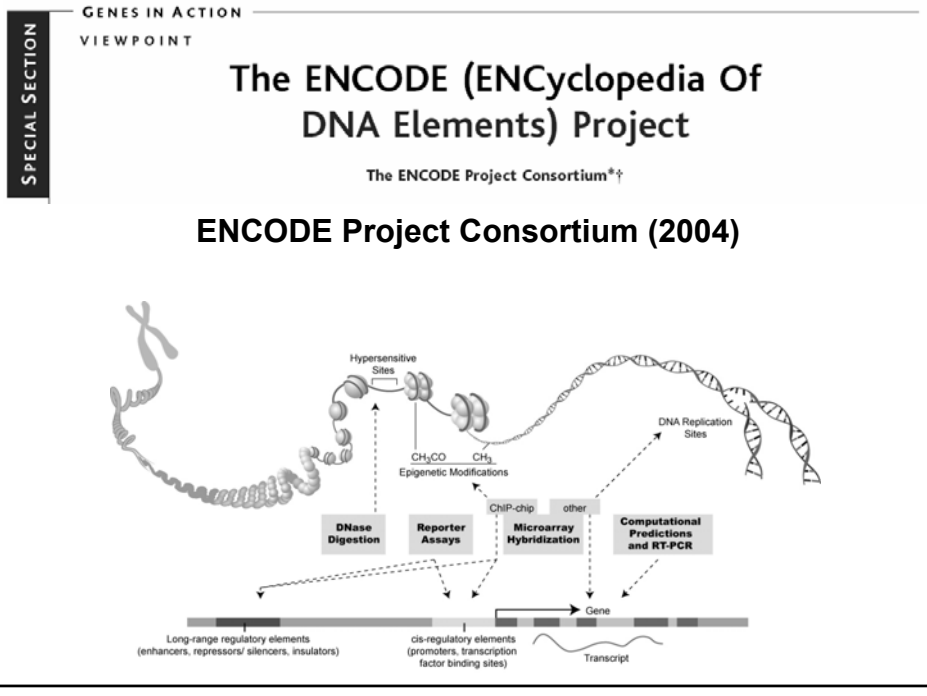
Future Genomes to Sequence???



ENCODE Project

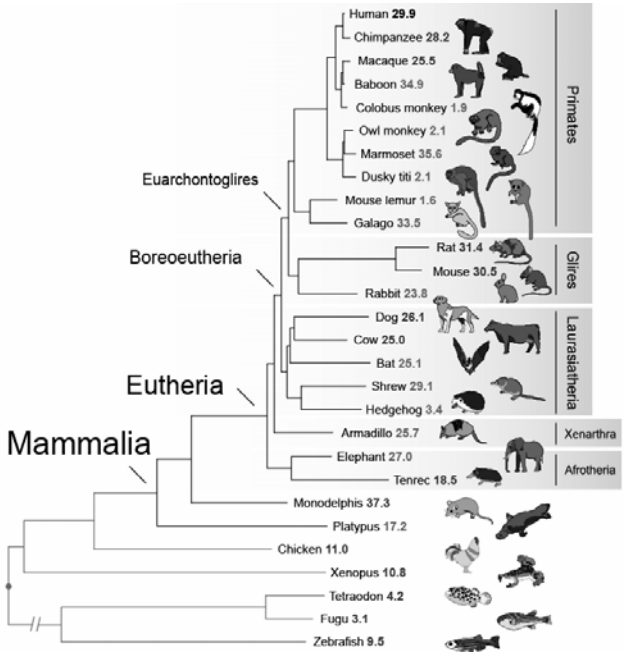


- ENCODE: ENCyclopedia Of DNA Elements
- Goal: Compile a *Comprehensive Encyclopedia* of All Functional Elements in the Human Genome
- Initial Pilot Project: 1% of Human Genome
- Apply Multiple, Diverse Approaches to Study and Analyze that 1% in a Consortium Fashion

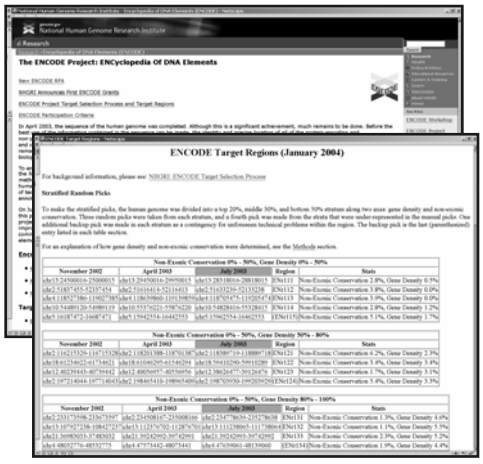


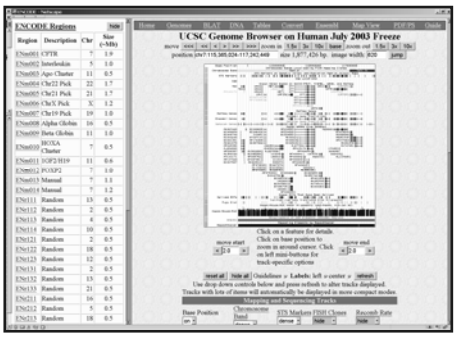
ENCODE Project Consortium (2004)

ENCODE Project: Comparative Sequencing



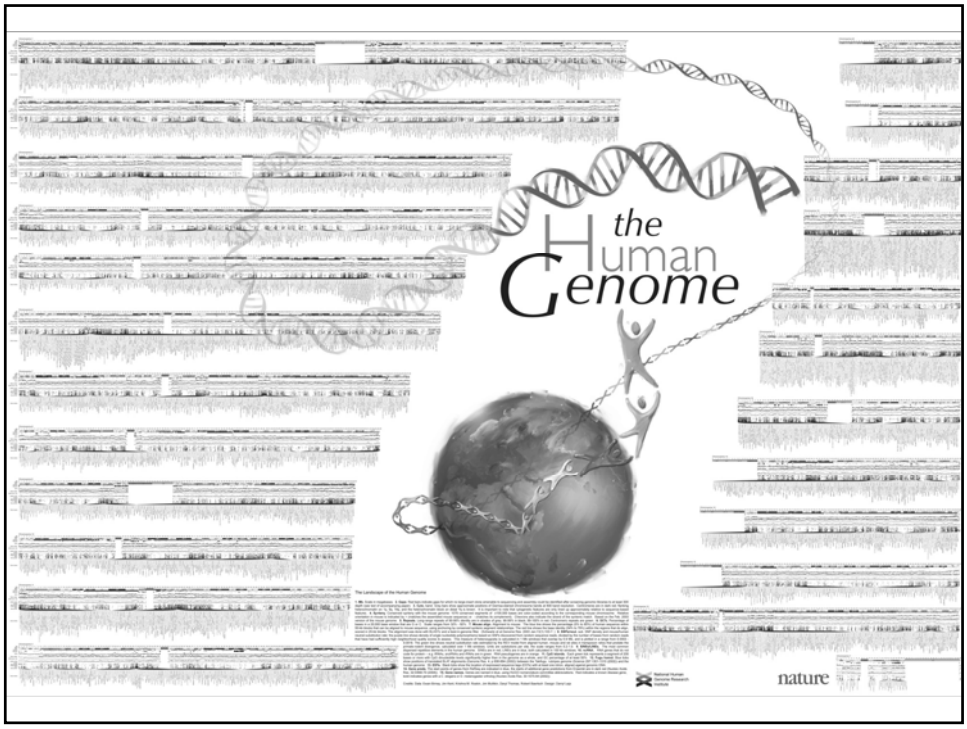
ENCODE Project: Web Sites





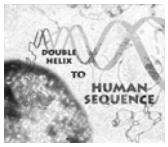
genome.gov/ENCODE

genome.ucsc.edu/ENCODE



Human Genome Sequence

>$1,000,000,000

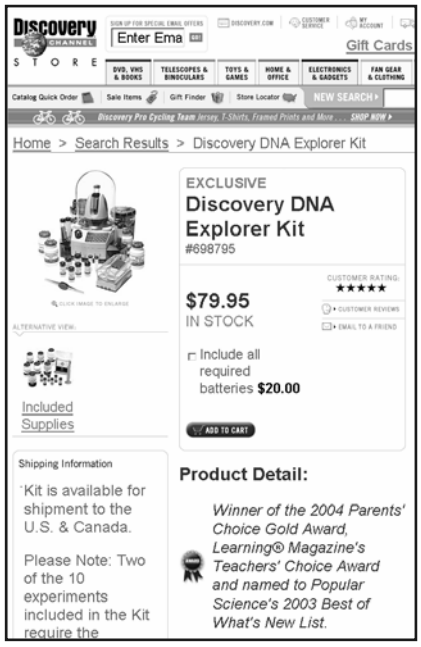


↓

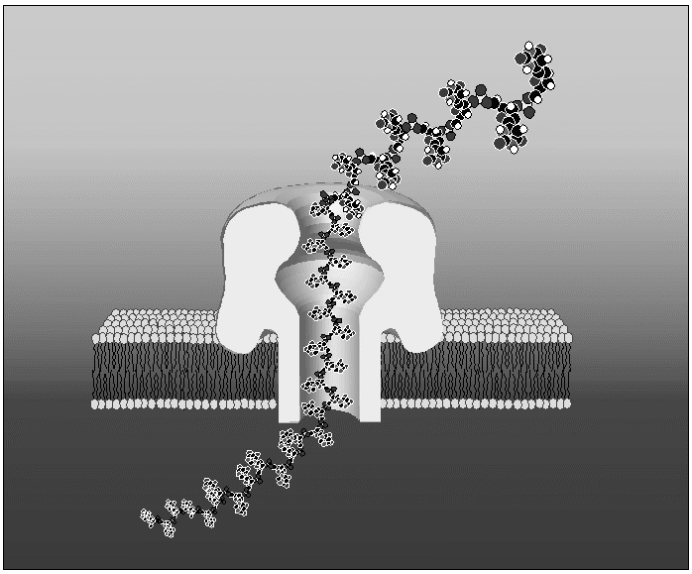
~$100,000

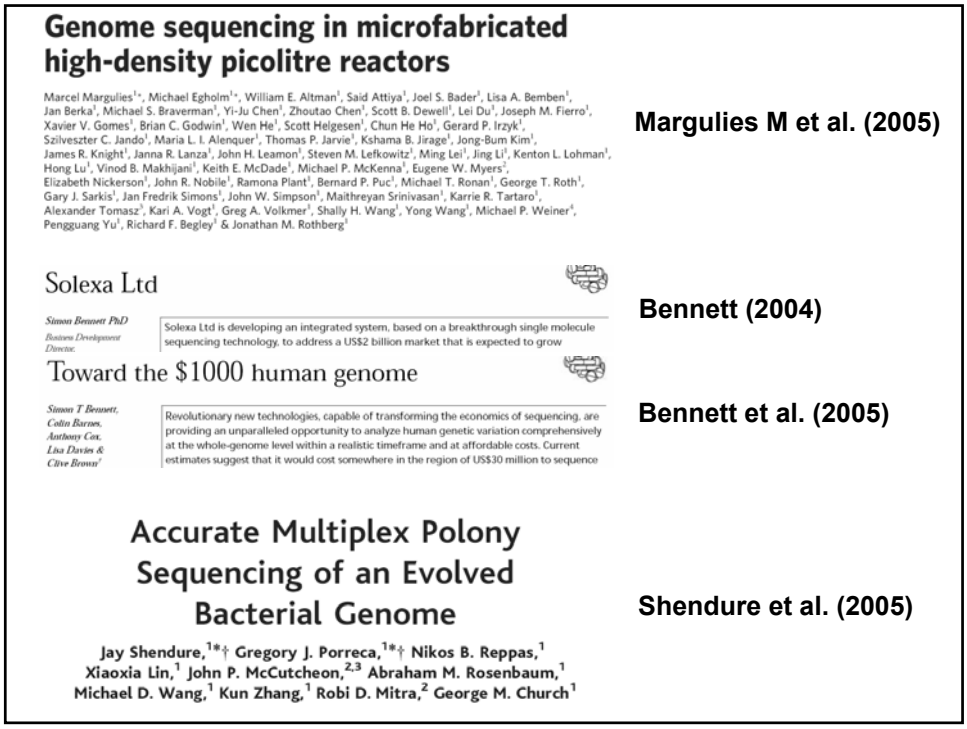
↓

~$1,000



En Route to the $1000 Genome





Genome sequencing in microfabricated high-density picolitre reactors

Marcel Margulies[1*], Michael Egholm[1*], William E. Altman[2], Said Attiya[1], Joel S. Bader[3], Lisa A. Bemben[1], Jan Berka[1], Michael S. Braverman[1], Yi-Ju Chen[1], Zhoutao Chen[1], Scott B. Dewell[1], Lei Du[1], Joseph M. Fierro[1], Xavier V. Gomes[1], Brian C. Godwin[1], Wen He[1], Scott Helgesen[1], Chun He Ho[1], Gerard P. Irzyk[1], Szilveszter C. Jando[1], Maria L. I. Alenquer[1], Thomas P. Jarvie[1], Kshama B. Jirage[1], Jong-Bum Kim[1], James R. Knight[1], Janna R. Lanza[1], John H. Leamon[1], Steven M. Lefkowitz[1], Ming Lei[1], Jing Li[1], Kenton L. Lohman[1], Hong Lu[1], Vinod B. Makhijani[1], Keith E. McDade[1], Michael P. McKenna[1], Eugene W. Myers[2], Elizabeth Nickerson[1], John R. Nobile[1], Ramona Plant[1], Bernard P. Puc[1], Michael T. Ronan[1], George T. Roth[1], Gary J. Sarkis[1], Jan Fredrik Simons[1], John W. Simpson[1], Maithreyan Srinivasan[1], Karrie R. Tartaro[1], Alexander Tomasz[2], Kari A. Vogt[1], Greg A. Volkmer[1], Shally H. Wang[1], Yong Wang[1], Michael P. Weiner[4], Pengguang Yu[1], Richard F. Begley[1] & Jonathan M. Rothberg[1]

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, Colin Barnes, Anthony Cox, Lisa Davies & Clive Brown

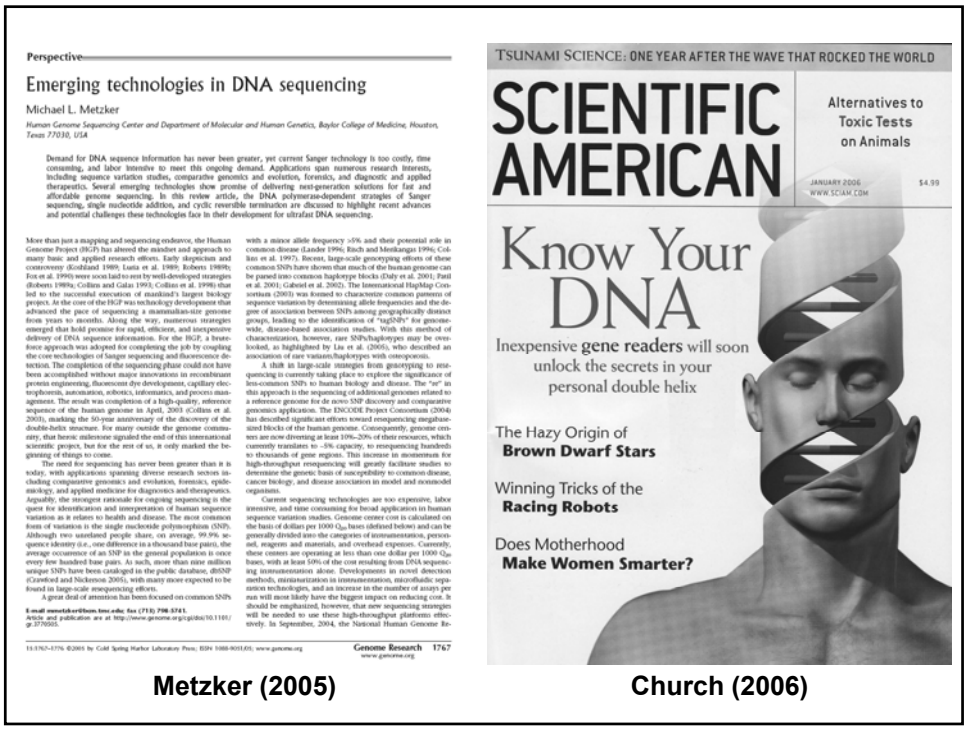
Revolutionary new technologies, capable of transforming the economics of sequencing, are providing an unparalleled opportunity to analyze human genetic variation comprehensively at the whole-genome level within a realistic timeframe and at affordable costs. Current estimates suggest that it would cost 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. Reppas,[1] Xiaoxia Lin,[1] John P. McCutcheon,[2,3] Abraham M. Rosenbaum,[1] Michael D. Wang,[1] Kun Zhang,[1] Robi D. Mitra,[2] George M. Church[1]

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 sequence variation 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 ultrafast 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 skepticism and controversy (Kruhlund 1989; Lutia et al. 1989; Roberts 1989b; Fox et al. 1990) were soon laid to rest by well-developed strategies (Roberts 1989a; Collins and Galas 1993; Collins et al. 1998) that led to the successful execution of mankind's largest biology project. At the core of the HGP was technology development that advanced the pace of sequencing a mammalian-size genome from years to months. Along the way, numerous strategies emerged that hold promise for rapid, efficient, and inexpensive delivery of DNA sequence information. For the HGP, a brute-force approach was adopted for completing the job by coupling the core technologies of Sanger sequencing and fluorescence detection. The completion of the sequencing phase could not have been accomplished without major innovations in recombinant protein engineering, fluorescent dye development, capillary electrophoresis, automation, 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 heroic 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 ongoing sequencing is the quest for identification and interpretation of human sequence variation as it relates 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 few hundred base pairs. As such, more than nine million unique SNPs have been cataloged in the public database, dbSNP (Crawford and Nickerson 2005), with many more expected to be found in large-scale resequencing 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; Risch and Merikangas 1996; Collins et al. 1997). Recent, 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; Patil et al. 2001; Gabriel et al. 2002). The International HapMap Consortium (2003) was formed to characterize common patterns of sequence variation by determining allele frequencies and the degree of association between SNPs among geographically distinct groups, leading to the identification of "tag SNPs" for genome-wide, disease-based association studies. With this method of characterization, however, rare SNPs/haplotypes may be overlooked, as highlighted by Liu et al. (2005), who described an association of rare variants/haplotypes with osteoporosis.

A shift in large-scale strategies from genotyping to resequencing is currently taking place to explore the significance of less-common SNPs to human biology and disease. The "re" in this approach is the sequencing of additional genomes related to a reference genome for de novo SNP discovery and comparative genomics application. The ENCODE Project Consortium (2004) has described significant efforts toward resequencing megabase-sized blocks of the human genome. Consequently, genome centers are now clustering at least 10%–20% of their resources, which currently translates to ~5% capacity, to resequencing hundreds to thousands of gene regions. This increase in momentum for high-throughput resequencing 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 too expensive, labor intensive, and time consuming for broad application in human sequence variation studies. Genome center cost is calculated on the basis of dollars per 1000 Q_{20} bases (defined below) and can be generally divided into the categories of instrumentation, personnel, reagents and materials, and overhead expenses. Currently, these centers are operating at less than one dollar per 1000 Q_{20} bases, with at least 50% of the cost resulting from DNA sequencing instrumentation alone. Developments in novel detection methods, miniaturization in instrumentation, microfluidic separation technologies, and an increase in the number of assays per run will most likely have the biggest impact on reducing cost. It should be emphasized, however, that new sequencing strategies will be needed to use these high-throughput platforms effectively. In September, 2004, the National Human Genome Re-

E-mail mmetzker@bcm.tmc.edu; fax (713) 798-5741.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3706505.

11:1767–1776 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

Genome Research 1767
www.genome.org

Metzker (2005)

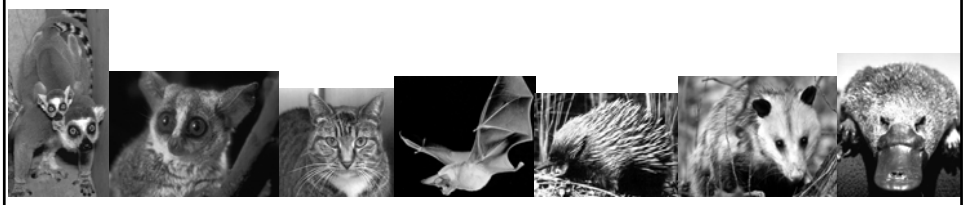
Church (2006)

DNA Sequencing Technologies

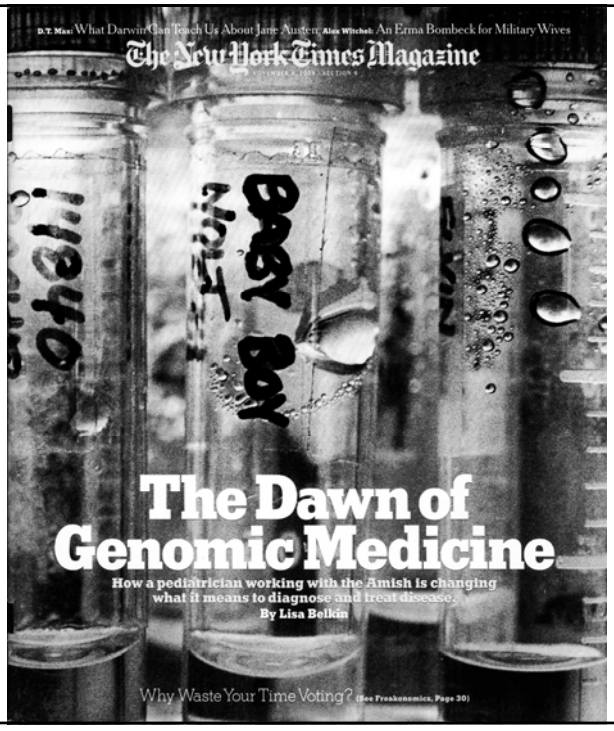
Method	Feasibility	Read Length	Data Quality	Raw Data Production
Sanger Sequencing	Well Established	Long (800-1200 bases)	+++	+
Stepwise Synthesis	Becoming Established	Short (25-100 bases)	+	+++++
Single Molecule	Far from Established	Long (>1000 bases ?)	???	+++++ ?

Realities of New DNA Sequencing Technologies…

"Inter-Species" Comparisons



"Intra-Species" Comparisons



The Pathway to Genomic Medicine

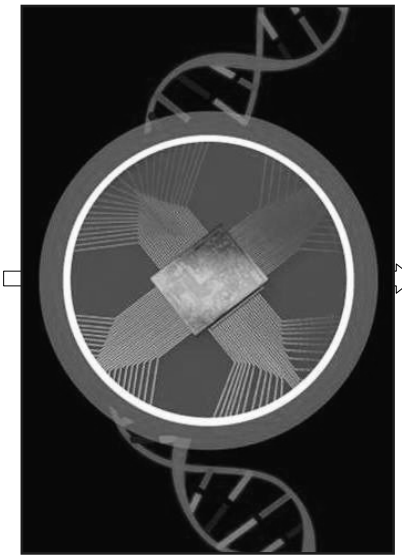
HGP

?

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.